Fundamentals of Cache Aided Wireless Networks

Avik Sengupta

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

T. Charles Clancy, Chair
Ravi Tandon, Co-Chair
Jeffrey H. Reed
Yaling Yang
Ing-Ray Chen

November 17, 2016
Blacksburg, Virginia

Keywords: Caching, Wireless Networks, Content Delivery

Fundamentals of Cache Aided Wireless Networks

Avik Sengupta

(ABSTRACT)

Caching at the network edge has emerged as a viable solution for alleviating the severe capacity crunch in content-centric next generation $5\mathrm{G}$ wireless networks by leveraging localized content storage and delivery. Caching generally works in two phases namely (i) *storage phase* where parts of popular content is pre-fetched and stored in caches at the network edge during time of low network load and (ii) *delivery phase* where content is distributed to users at times of high network load by leveraging the locally stored content. Cache-aided networks therefore have the potential to leverage storage at the network edge to increase bandwidth efficiency. In this dissertation we ask the following question - *What are the theoretical and practical guarantees offered by cache-aided networks for reliable content distribution while minimizing transmission rates and increasing network efficiency?*

We furnish an answer to this question by identifying fundamental Shannon-type limits for cache-aided systems. To this end, we first consider a cache-aided network where the cache storage phase is assisted by a central server and users can demand multiple files at each transmission interval. To service these demands, we consider two delivery models - (i) centralized content delivery where demands are serviced by the central server; and (ii) device-to-device-assisted distributed delivery where demands are satisfied by leveraging the collective content of user caches. For such cache-aided networks, we develop a new technique for characterizing information theoretic lower bounds on the fundamental storage-rate tradeoff. Furthermore, using the new lower bounds, we establish the optimal storage-rate tradeoff to within a constant multiplicative gap and show that, for the case of multiple demands per user, treating each set of demands independently is order-optimal. To address the concerns of privacy in multicast content delivery over such cache-aided networks, we introduce the problem of caching with secure delivery. We propose schemes which achieve information theoretic security in cache-aided networks and show that the achievable rate is within a constant multiplicative factor of the information theoretic optimal secure rate. We then extend our theoretical analysis to the wireless domain by studying a cloud and cache-aided wireless network from a perspective of low-latency content distribution. To this end, we define a new performance metric namely normalized delivery time, or NDT, which captures the worst-case delivery latency. We propose achievable schemes with an aim to minimize the NDT and derive information theoretic lower bounds which show that the proposed schemes achieve optimality to within a constant multiplicative factor of 2 for all values of problem parameters. Finally, we consider the problem of caching and content distribution in a multi-small-cell heterogeneous network from a reinforcement learning perspective for the case when the popularity of content is unknown. We propose a novel topology-aware learning-aided collaborative caching algorithm and show that collaboration among multiple small cells for cache-aided content delivery outperforms local caching in most network topologies of practical interest. The results presented in this dissertation show definitively that cache-aided systems help in appreciable increase of network efficiency and are a viable solution for the ever evolving capacity demands in the wireless communications landscape.

Fundamentals of Cache Aided Wireless Networks

Avik Sengupta

(GENERAL AUDIENCE ABSTRACT)

Caching at the network edge has emerged as a viable solution for alleviating the severe capacity crunch in content-centric next generation $5\mathrm{G}$ wireless networks by leveraging localized content storage and delivery. Caching generally works in two phases namely (i) *storage phase* where parts of popular content is pre-fetched and stored in caches at the network edge during time of low network load and (ii) *delivery phase* where content is distributed to users at times of high network load by leveraging the locally stored content. Cache-aided networks therefore have the potential to leverage storage at the network edge to increase bandwidth efficiency. In this dissertation we study cache-aided systems from an information theoretic perspective and identify fundamental Shannon-type limits for such systems. The results presented in this dissertation show definitively that cache-aided systems help in appreciable increase of network efficiency and are a viable solution for the ever evolving capacity demands in the wireless communications landscape.

Dedication

To my parents.. for their unconditional support and encouragement.

Acknowledgments

It has been quite an eventful journey to finally arrive at the point I had waited for - to pen down this section of my dissertation. The most exciting part of the journey (apart from the obvious earth-shattering impacts of my awesome research ☺) was to get to know so many people over the years, who had such a huge impact on my life in this microcosm that is Blacksburg, Virginia. The time I spent here has been some of the best years of my life and I am thankful to the town and its occupants for making it a memorable and often breathtakingly beautiful adventure.

I would like to begin this section by thanking my family. Without their unconditional love and support, this would have been impossible. This work is as much their contribution as is mine for the mere fact that they have always wholeheartedly supported every decision I have made thus far in life. Thanks for always believing in me and making things easy just so that I didn't have to worry about anything but my research. This one is for you all.

Over the course of my Ph.D., I have had the great fortune to collaborate with a number of amazing people. Firstly, I would like to acknowledge the support and constant guidance of my guru, advisor and teacher, Dr. Ravi Tandon. Starting from hours of staring a white-board together in his office, often till late in the night, to the multiple coffee breaks where we discussed almost every topic under the sun (and thanks for buying me so many coffees and giving me late-night rides back home over the years), to the thrill of small eureka moments when we solved those problems that had been bugging us for months, it was an absolute roller-coaster. I am privileged to have been one of his first doctoral students and am grateful for the time and patience that he donated towards teaching me and bringing me up to speed on a multitude of topics. The experience has better prepared me for the world outside university and I will always be indebted to him for anything I achieve in my life. I hope that we continue to collaborate well past my Ph.D.

I would also like to thank my co-advisor Dr. T. Charles Clancy for the faith he has always shown in me. He ensured that I was funded for my graduate studies throughout my time here at Virginia Tech and I am grateful to him for offering helpful feedback and for finding different projects to work on while I continued my Ph.D. I would also like to thank Dr. Reed for his support and the interest with which he has always reviewed my work and for always inviting us home for his awesome Thanksgiving parties when we were the only souls left in town. He is one of the best professors I have had the fortune to work with. I would also like to thank Dr. Yaling Yang and Dr. Ing-Ray Chen for being a part of my Ph.D. committee and reviewing my work.

Recently, I had the opportunity of collaborating with Dr. Osvaldo Simeone. I would like to express my gratitude to him for always critically reviewing my work (that too at lightning speed and from almost any part of the world) and instantly providing the most useful feedback. The work presented in Chapter 5 would not have been possible without his guidance. Additionally, I would like to thank Dr. R. Michael Buehrer for being a part of my Ph.D. qualifying exam committee and for reviewing the work presented in Chapter 6 of this dissertation. I would like to thank Dr. Joesph Ernst for his tireless support for the work I did with Hume Center and for always being patient with me in the face of absolute uncertainty owing to my lack of knowledge in the field of work. Thanks for teaching me new things and always waiting for me to catch up. I would also like to thank Jack Smith, Dr. Suman Das and Dr. Roland Rick for their guidance and teachings over the three awesome summers I spent at Blackberry, Huawei and Qualcomm respectively.

A special vote of thanks to my Masters advisor, Dr. Bala Natarajan, without whose initial guidance and encouragement, none of this would have been possible.

And now for the people who made my life in Blacksburg all the more enjoyable. I would like to thank my lab-mate (and later roommate) Aditya for the all support he has given me over years by patiently listening to all my rants and always giving awesome feedback. We began our Ph.D. journey together, weathered many a storm and hope our collaborations continue well beyond Virginia Tech. Another vote of thanks for Mahi for always being there and lending her support to everything we did. Without you, this journey would have been incomplete. I would like to thank Dr. Dhiraj for all the wisdom that he imparted over the years and for being patient enough to work with me. I would like to thank my lab-mates Munawwar, Daniel, Matt, Ali, Chris, Akshay and Tad for the fun (and often technical) times we shared. A big thank you is also in order for Hilda. Thanks for always taking care of us like a parent and making our lab a fun place to be. I would also like to thank Sonya, Leslie and Janet from the Hume Center for always looking out for me and supporting me in every way possible.

I would be amiss if I did not thank my roommates and friends Aritra, Shuvodeep, Bhela, Deven and Aditya for all the fun times we shared together from the deep all-encompassing philosophical discussions to playing guitar at the dead of the night. To my bands VT Grooves and Ande Tamatar, thanks for all the fun jams and the crazy shows we pulled off together. It was the perfect relaxing pill in the middle of the rigors of grad school. Finally, a special thanks to Sruthi for being my partner-in-crime and lending her unequivocal support over the course of this crazy ride. From our late night Math Emporium sessions to hanging out at Torg Bridge, to all the cooking, food, drives and jam sessions, this would not have been possible without you. Thanks for patiently helping me make good looking figures (actually for making many of them yourself after looking at my puny attempts) and for designing my defense slides with your usual flair. To all my friends in Blacksburg Varuni, Vishwas, Arka, Naveen, Atul, Vireshwar, thank you for all the great times we shared. Lastly, to my school friends, Partha, Urmi, Saumik, Sweta, Sohini, Bratiraj, Sayantika and Chiky, thanks for always being there for me.

Contents

List of Figures

List of Tables

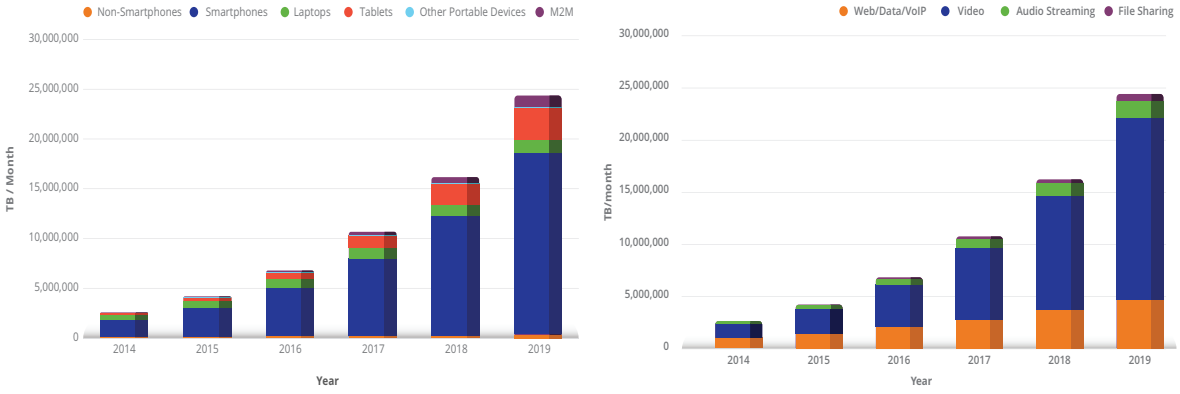
Chapter 1

Introduction

The nature of traffic over wireless cellular networks has undergone a paradigm shift in recent times to become increasingly data and content-centric. In the era of $2\mathrm{G}$ and $2.5\mathrm{G}$ systems, cellular communications was primarily circuit-switched with voice communications holding precedence over data communications. However with the advent of $3\mathrm{G}$ [1] in the form High Speed Packet Access (HSPA) and HSPA+, the nature of wireless traffic underwent a shift towards becoming increasingly data centric. Finally the advent of Long Term Evolution (LTE) [2–4] as global standard in wireless cellular communications and the proliferation of smart-phone technology, has led to a near-complete transformation of wireless networks from circuit-switched to packet-switched with data traffic holding precedence. The legacy circuit-switched networks are employed in some systems for voice communications pending the reliability of packet based Voice-over-IP (VoIP) improving to support stringent latency requirements for voice data. This paradigm shift has also changed the way consumers use wireless devices i.e., from strictly voice based usage in early 1990's to exceedingly content based usage in recent times.

Consequently, efficient content distribution over wireless networks has been an area of focus for the industry as well as academic research communities. Fig. 1.1(a) and 1.1(b) show the projected wireless data traffic (in Terabytes/month) growth over a period of 5 years from $2014-2019$[1]. It can be seen that the global usage of smart-phones is increasing and will completely dominate non-smart devices by 2019. Furthermore, devices like tablets will also become major consumers of wireless mobile data. Tablets and smart-phones are primarily used for content-centric data access over cellular networks. Based on such device usage characteristics, it is expected that for upcoming $5\mathrm{G}$ wireless systems [5–7], multimedia content like video will hold precedence over all other types of mobile data traffic. Fig. 1.1(b) validates this by showing that mobile data traffic will be dominated by video which is projected to see an increase of almost 66% in terms of generated traffic. Most of this traffic falls under the category of *elastic* traffic. Elastic traffic like buffering video and file sharing are classified under the category of Non-GBR (non guaranteed-bit-rate) traffic in the

[1]The data was obtained from Cisco's Visual Networking Index (VNI) projections which gives projected growth of mobile and web traffic over period of 5 years based on current growth rates.



(a) (b)

Figure 1.1: Projected Mobile Data Traffic Growth (a) by Device Type and (b) by Application Category (Data Source: Cisco VNI).

LTE architecture [8]. This kind of traffic is essentially more tolerant to delays and latency and has less stringent packet error requirements as compared to GBR *non-elastic* traffic like VoIP, live streaming video, interactive applications, gaming and remote control signals. However, Non-GBR traffic typically has much larger data volume and thereby the capability of overloading the network, especially at the rate of projected growth as predicted by Cisco. As a result, ensuring the reliability and scalability of present day networks to handle such large amounts of data is an area of active research.

1.1 Increasing Network Efficiency

In this section, we discuss existing approaches towards increasing network efficiency and resiliency in order to enable networks to handle extremely large volume of data traffic. The first step in this direction was taken by the introduction of Multiple-Input-Multiple-Output (MIMO) communication systems, wherein transmitters and receivers were equipped with multiple antennas to exploit diversity and multiplexing gains [9]. Current cellular systems like LTE make extensive use of MIMO and related signal processing mechanisms in the PHY layer to boost data rates. The advantages of MIMO will be further leveraged in the new 5G wireless standards where massive MIMO [7, 10] is expected to play a key role in increasing bandwidth efficiency for high frequency, high data-rate communications. However, even with the use of MIMO (or massive MIMO), state-of-the-art techniques are needed to exploit available bandwidth resources with maximal efficiency. To this end, Fig. 1.2 highlights some of the key approaches in current use. We next discuss these approaches in detail.

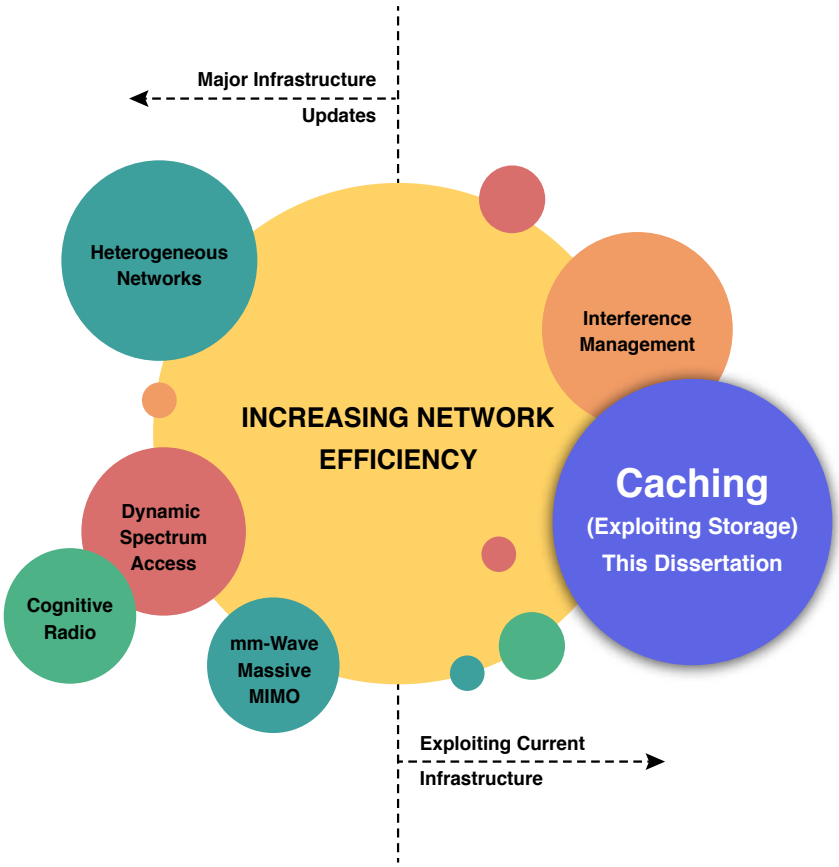


Figure 1.2: Approaches to Increasing Network Efficiency

1.1.1 Heterogeneous Networks

To increase network capacity in a cost-effective manner, a paradigm shift in cellular network infrastructure deployment has occurred away from traditional (and expensive) high-power tower-mounted base stations, towards heterogeneous network (HetNet) elements. Examples of HetNet elements include microcells, picocells, femtocells, and distributed antenna systems (such as remote radio heads), which are distinguished by their lower transmit powers and smaller coverage areas, physical size, backhaul (wired/wireless), and propagation characteristics. These elements are generally referred to as small-cell base stations (sBSs) and are deployed in large numbers within a cell leading to *network densification* [11] which is envisioned to continue into the the future through the deployment of 5G systems. The sBSs are capable of connecting to the parent BS via wired/wireless backhaul and help in increasing the coverage of cellular networks by plugging spatial spectrum holes which leads to efficient use of frequency resources. By efficiently using these sBSs for multimedia content distribution, the load on the parent BS can be considerably reduced. HetNets can also act as relays to extend coverage range and is a salient feature for LTE-Advanced

networks [8]. The gains, both economic and technological, offered by HetNets has attracted a wide body of research in modeling and design of such networks [12–14]. HetNets also entail further interference management techniques in the form of Inter-cell-interference-cancellation (ICIC), as well reintroducing the concept of Fractional Frequency Reuse (FFR) [15–17] in LTE[2] whereby different frequency bands are used for parent BS and the sBSs.

A different approach to capacity utilization in this paradigm deals with low power direct communication between devices i.e., device-to-device (D2D) communications. In LTE, this technique is classified as a part of Proximity Services [18] while Qualcomm has adopted it as LTE Direct [19]. The main idea is to cluster devices with similar shared content together and offload their communications as an underlay to the cellular networks [20–30]. This ensures efficient frequency reuse within cells since the low power direct communications causes minimal interference. Such networks are ideally suited for localized exchange and distribution of high data rate multimedia content. This makes D2D an attractive choice for public safety applications as well as aiding in elastic video distribution and sharing. Qualcomm's FlashLinQ [31] is an example of a PHY layer protocol which supports LTE Direct as an underlay to a cellular network.

1.1.2 Spectrum Sharing & Cognitive Radio

Another approach to utilizing the potential of current wireless networks is through efficient and complete use of existing spectral resources. The FCC spectrum policy task force has reported that a significant amount of the current licensed spectrum is sporadically utilized [32] and the fixed spectrum assignment policy currently in use should be modified in order to support the ever growing wireless data traffic. The key enabler in breaking the spectrum gridlock is Cognitive Radio (CR) technology [33, 34]. CR enables secondary or unlicensed users (SU) to dynamically access a licensed Primary User's (PU) band when it is not in use. This opportunistic sharing of PU bands with the SU is popularly known as Dynamic Spectrum Access (DSA) [35]. DSA exploits white spaces in the frequency spectrum usage and secondary transmissions are scheduled on an ad-hoc basis, avoiding interference to primary spectrum users.

The main constraint for the secondary devices (or cognitive radios) is to guarantee that the interference caused to the Primary Users' (PU) needs to be below a tolerable threshold [36]. To this end, spectrum sensing [37,38] is an important component of the DSA architecture. Spectrum sensing [37] is the task of obtaining awareness about the spectrum usage and existence of primary users in a geographical area. Spectrum sensing, however, is limited in its usefulness by hidden primary node problems, which leads to mis-detection of spectrum holes and causes unwanted interference. Cooperative sensing is proposed in the literature for handling hidden primary user problem [38]. Cooperative sensing increases the sensing accuracy by fusing data from multiple nodes and thus takes advantage of spatial diversity. However even cooperative sensing methods are fundamen-

[2]Legacy LTE systems are generally termed as reuse 1, meaning that each cell has the entire licensed frequency band from which to allocate resources to its users. These systems increasingly rely on efficient orthogonal resource scheduling and advanced interference management techniques at the cell edge for efficient network operation.

tally limited by factors such as SNR walls [39], presence of noise and modeling uncertainties. This motivates approaches based on Radio Environment Maps (REM) [40, 41], and hybrid techniques [42,43] which employ sensing combined with REMs. A wide range of efforts deal with the introduction of cognition into the LTE standards to facilitate spectrum sharing especially in public safety networks [17, 44–46].

Recent efforts have been focused on applications of sensing based DSA in *spectrum sharing* [47–52] for cellular networks which is expected to be one of the key enablers for $5\mathrm{G}$ [5]. If a cell of an operator is under-loaded for a certain period of time, then a part of the spectrum will be wasted, while it could have been exploited by co-located/adjacent cells of other operators experiencing high traffic [53]. In addition, exploiting diversity between end-users and different operators' base stations may result in a higher throughput for the user with the same total bandwidth utilized but from differing operators. Inter-operator spectrum sharing for 3G systems has been discussed in [54–57]. With the introduction of 4G LTE and the shift towards complete packet-services-based networks, frequency spectrum sharing among cellular operators became a more feasible architecture. Spectrum sharing policies proposed could be generally categorized into non-orthogonal and orthogonal. The former allows several base stations to use the same transmission frequency at the same time, provided that the level of interference at the intended receivers is below a desired threshold. The latter considers mutually exclusive access to the shared spectrum and hence does not tolerate any interference [53]. Non-orthogonal spectrum sharing approaches are considered as resource allocation problems under interference constraints. Solutions presented include transmit beam-forming [58], Dynamic Frequency Selection (DFS) algorithms based on interference measurements [59], and more complex game theoretical perspective [60, 61].

As a result of concerted and focused research activity in this area, recently the FCC has approved the proposal of using the $3.5\mathrm{GHz}$ band as a platform for future spectrum sharing innovation [62]. FCC calls this the "Innovation Band" and has proposed setting up the "Citizen's Broadband Radio Service". The use of advanced spectrum sharing technology will allow wireless broadband systems to share spectrum with military radars and other incumbent systems in this band, while protecting federal missions. Spectrum Sharing has received a further boost from the 3GPP which has adopted spectrum sharing in LTE Release 13 under the moniker of "LTE Licensed Assisted Access" (LAA) with industry leaders like Qualcomm and Ericsson leading the innovation in this paradigm [63,64]. Furthermore, shared spectrum access between different Radio Access Technologies (RAT) has been another area of active research in the recent past with Intel's LTE-WiFi Aggregation (LWA) providing a framework for spectrum sharing and aggregation between LTE and WiFi. To this end, the wireless industry has also made a recent thrust towards harmonious co-existence of LTE and WiFi in unlicensed bands [65] which inherently use the theoretical framework of cognition in shared spectrum.

Related Publications: The scope of inter-operator spectrum sharing as well as cross layer resource allocation for spectrum sharing in a *Cognitive-LTE* framework has been explored in our previous works presented in [17, 46].

1.1.3 Massive MIMO and mm-Wave

Another area of active research has been in the deployment of mm-Wave cellular communication systems [66–68] as a key enabler for $5\mathrm{G}$ systems. The deployment of cellular communications in the mm-Wave bands (e.g, 28 GHz) has the advantage that wide bands of unused spectrum is freely available and can potentially support Gigabit communications over the wireless links. Although the idea of mm-Wave communications has been around for almost 60 years [69], only recently has it been seriously considered as a candidate for cellular network deployment [70] in the wireless industry. The fundamental differences of mm-Wave communications with Sub-6 GHz range is the vulnerability to blocking and the need for significant directionality at receivers/transmitters [67]. To this end, massive MIMO [7, 10] has emerged as a perfect candidate for implementation of cellular transmitter and receivers in mm-Wave owing to its inherent capability of exploiting large antenna arrays for beamforming with very high directionality. Recently, the first LTE modem for mm-Wave was unveiled by Qualcomm (Snapdragon X50) which leverages the 28 GHz band using up to 800 MHz of bandwidth via 8×100 MHz carrier aggregation [70].

1.1.4 Interference Management

While the above approaches require changes (often economically prohibitive) to existing infrastructure, interference management techniques generally work within the purview of current infrastructure. The current implementations of wireless cellular networks are mostly interference limited [8, 9] with inter-cell interference being the major limiting factor. Thus, to maximize capacity utilization, efficient interference mitigation is the key. As a result, there has been an ample body of research for efficient resource allocation and interference mitigation in wireless networks [44, 71–78]. The common theme across all these works is that they aim to jointly allocate resources and while employing efficient power control mechanisms in the PHY layer to limit interference. However, joint power and resource allocation is a hard problem to solve optimally [17,44] and most solutions use approximate algorithms. Another approach that has been proposed in literature for multiple antenna (MIMO) systems is the concept of interference alignment [79–83]. The idea of interference alignment is to coordinate multiple transmitters so that their mutual interference aligns at the receivers, facilitating simple interference cancellation techniques. This potentially leads to exploitation of maximum possible Degrees-of-Freedom (DoF) of multi-user wireless systems thereby improving spectral efficiency. Interference Alignment however works on the premise that Channel State Information (CSI) is available at the transmitter which entails channel reciprocity or feedback and might not be feasible in all cases due to overhead. Some works as in [84, 85] have studied the benefits of delayed CSI at the transmitter which can still lead to exploiting higher DoFs of the system. In [83], the authors address the practical implementation issues of interference alignment in a cellular setting and design message passing feedback schemes to aid in implementation. In addition to these techniques, another promising technique adapted to LTE was coordinated multipoint transmission (CoMP) schemes [8]. In CoMP, multiple base stations coordinate with each other in order to cancel out interference at the cell-edge by efficient

scheduling of frequency resources [86]. CoMP however, has practical implementation issues with its gains saturating at practical transmit power ranges (being mostly limited to less than 30% in practice) owing to the overhead and latency in backhaul signaling [87, 88]. Interference management by itself is not capable of increasing spectrum efficiency to the extent where it can sustain the projected traffic growth, especially multimedia content distribution over the wireless networks. To this end, we next discuss a hybrid approach which leverages cheap infrastructure as well as interference management and load balancing techniques to effectively increase network efficiency.

1.1.5 Exploiting Storage

In addition to PHY layer interference mitigation, cognitive and infrastructure based approaches, higher layer (MAC) techniques like network coding and index coding [82, 89–93] have also emerged as important techniques for maximizing capacity utilization and facilitating efficient data distribution. In spite of providing theoretical guarantees, however, these approaches have yet to be deployed in practice owing to the complexity and signaling overhead. Index coding with side information [93] is a promising technique in this regime but the lack of adequate storage in devices generally is a deterrent for practical implementation. But with the advent of advanced processing power and large amounts of available storage in smart-phones and hand-held devices, these approaches are becoming increasingly realistic. An example in this case is ITLinQ [94] which applies the information theoretic notion of the optimality of treating interference as noise (TIN) [95] to improve the performance of Qualcomm's FlashLinQ architecture. An amalgamation of these techniques in conjunction with infrastructure based approaches and the availability of storage in cheaper and more compact forms have opened up the possibility of exploiting storage across the network to increase bandwidth efficiency. To this end, *caching* has as emerged as a tool to aid in efficient multimedia content distribution in wireless (or wired backhaul) networks. In this dissertation, we study the fundamental impacts of caching in next-generation $5\mathrm{G}$ wireless networks.

1.2 Caching in Wireless Networks

Caching in wireless networks has emerged as an important technique for facilitating spectrum utilization and reducing network load at times of high data (multimedia/video) traffic. Parts of popular files are pre-stored in users' devices. Once user requests are revealed, these local storages can be leveraged to deliver the complementary content thereby reducing the network load during content delivery. With the advent of smart-phones and small cells, there is no dearth of storage space in next-generation wireless networks. Thus caching presents the opportunity of *utilizing storage to save bandwidth*. Caching and complementary file delivery has been a subject of a wealth of recent research as evidenced by the works in [96–107]. The main questions that caching poses are - "When to cache?" and "How to utilize caches to reduce load?".

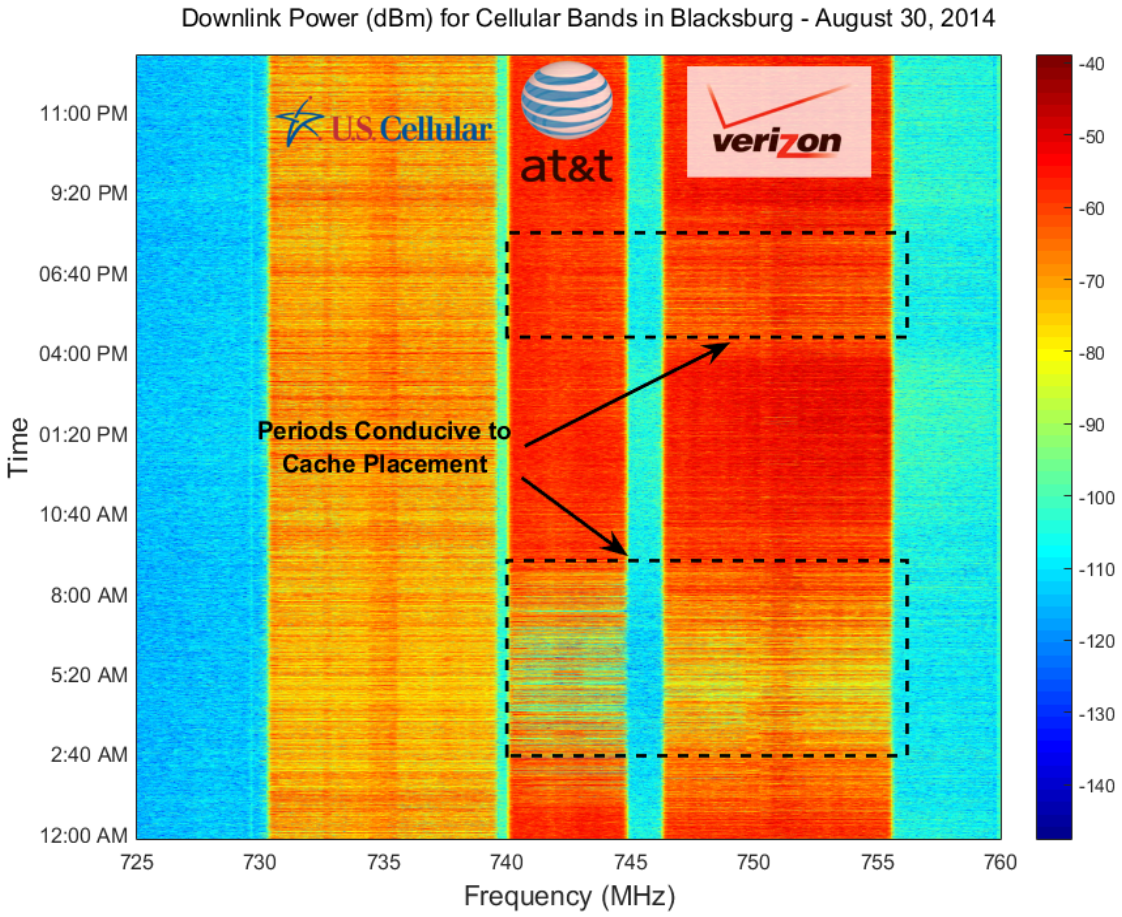


Figure 1.3: Downlink Power on Commercial Cellular Bands over a period of a day (Aug. 30, 2014) in Blacksburg, Virginia.

1.2.1 When to Cache?

Caching typically operates in two phases (1) the *storage phase* where parts of popular content is placed in users' caches and (2) the *delivery phase*, where requested content is delivered by exploiting the local cache storage of users. Based on this, we answer the question of *"When?"* by studying real data from cellular downlink bands. Cellular networks present temporal traffic variations over the course of every day. The aim of caching is to utilize the periods of low network activity to populate the local cache storage of users with parts of popular content. During periods of very high activity this can be leveraged to reduce peak data rates. This is illustrated in Fig. 1.3 with the help of downlink power data on the cellular bands[3]. The temporal variation in downlink activity is shown for the 700MHz band which is occupied by three major cellular carriers - US Cellular, At&t and Verizon. It can be clearly seen from Fig. 1.3 that there is very low network

[3]The data was collected as a part of the Global Spectrum Observatory at Wireless@VT.

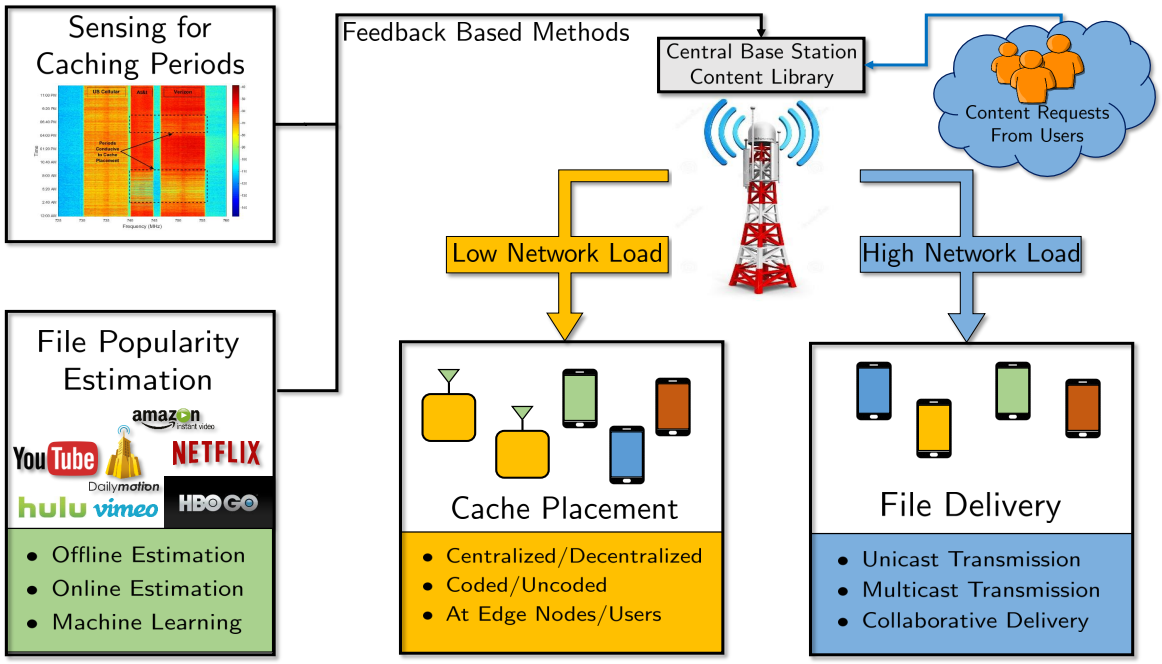


Figure 1.4: An overview of Cache-Aided Network Operation

activity during late night and early morning hours (from $2-8$ am) and again during late afternoon to early evening ($4-6$ pm). These periods of low network activity can be leveraged to operate the storage phase of caching while delivery phase can be used during the peak hours of 9 am to 4 pm and prime-time hours of $9-11$ pm. Note that the US Cellular band shows low activity throughout the day perhaps owing to minimal subscribers in the area compared to At&t and Verizon. Thus US Cellular bands can operate both storage and delivery at any time during the day.

1.2.2 How to utilize caches to reduce load?

The concept of caching has been in existence for almost 20 years with its origins traced back to the widespread adoption of the Internet and consequently the proliferation of internet congestion as a major bottleneck. The bottleneck was eased by the invention of content delivery networks (CDNs) and the exploitation of web caching [108–111]. Web caching replicates popular content in different geographical areas leading to bandwidth savings by avoiding unnecessary multihop retransmissions. This also decreases access time (latency) by decreasing the distance for accessing requested content. The notion of caching in wireless networks follows essentially the same philosophy i.e., to offer increased QoS at low data rates by essentially converting the *local storage into bandwidth* [112]. The proposition of *information-centric networking* (ICN) as a candidate for the

Internet of the future however raises the important question of *where to install network memory for maximal benefits* [113]. To this end, ICN proposed to equip routers with caches in order to facilitate network-wide content replication. Interestingly, it was shown in [114] that most of the caching gains in ICN can be obtained by caching at the edges of the network using existing CDNs. Contrary to the popular belief that caching close to users is highly inefficient, in this dissertation, we consider the problem of edge-caching in multi-tier cellular networks with an aim to understand if the gain from ICNs translate directly to the cellular paradigm.

Fig. 1.4 gives an overview of the cache-aided network operation considered for this work. The cache-aided network employs monitoring of downlink activity to dynamically identify periods of low network activity which are amenable to cache storage by pre-fetching content. It also has a popularity estimation process, wherein a set of popular content is identified based on local user behavior and content demands. Based on this feedback based architecture, the central server (BS) chooses the time and content to perform cache placement during periods of low network load. The placement phase happens over a much larger time-scale (typically of the order of days) compared to content demand by users and is hence agnostic to user requests i.e., the cache storage must not be dependent on specific user requests but should be able to accommodate any possible request of content. Once user requests are revealed, the server performs complementary delivery during times of high network activity at a potentially lower rate by exploiting the locally stored content. The complementary content delivery by the servers in response to user requests is termed as a *transmission interval* in the rest of this dissertation.

Traditionally, the delivery phase of cache-aided systems [109, 110, 115, 116] operated as a series of dedicated point-to-point unicast transmissions to individual users by transmitting fractions of requested files[4] which are not stored in their caches. However, this is not a scalable solution as the number of users in the system increases. Most of the prior works in this area tend to use a fixed delivery scheme and then optimize the storage phase to suit the delivery scheme [109, 110]. Further, their investigations are mainly based on the gains obtained from local content distribution, ignoring the scope of global (system) cache interactions across users and content sharing as a factor for extracting additional caching gain.

Recently, a series of pioneering works by Maddah-Ali and Niesen [97–101] presented a novel information theoretic formulation for caching in wireless networks. They showed that by *jointly designing* the storage and delivery phase, and using multicast transmissions to simultaneously deliver content to users, order-wise improvement in the transmission rates can be achieved as compared to traditional unicast delivery. A novel caching and multicast delivery scheme based on shared content in user caches was presented, whereby a *global caching gain* was extracted from the system in addition to the traditional local gains. The ideas presented in these works had their roots in fundamentals of network coding, specifically index coding with side information [93] and the idea of interference alignment [79, 81]. The authors modeled the cache content as side-

[4] In this dissertation, we use the concept of files as a generic block of data which is desired by users. The terms *files* and *content* are used interchangeably throughout the dissertation and can refer to any elastic multimedia content such as videos, movies etc.

information and leveraged that fact if structured side-information can be designed aprior (i.e., design of cache placement strategies), efficient index coding solutions (i.e., complementary content delivery) can then be used to send information via multicast network coding simultaneously to multiple users. The authors also proved the order optimality of their proposed schemes by showing that the achievable rate is within a constant gap to the information theoretic optimal rate of the system. The proposed algorithms are elegant due to their relative ease of implementation in case of video distribution and effectively re-kindled the interest in cache-aided networking from an information theoretic perspective.

In the initial information theoretic modeling of cache-aided systems, the authors assumed uniform file popularity i.e., that all files in the library have equal probability of being requested. However extensions to non-uniform file popularities has been extensively studied in some later works e.g., [99, 104, 107, 117–119]. In the years following this line of work by Maddah-Ali et. al., a widespread interest in evaluating the information theoretic limits of caching has seen the extension of this framework to a plethora of related frameworks e.g., hierarchical coded caching [106, 120–122], cache-aided D2D systems [105, 112], multi-server coded caching [123], multi-library coded caching [124] etc. A parallel body of recent work has aimed to better understand the fundamental limits of the system model studied in [97]. These works have been focused on improving the achievable schemes to yield better upper bounds on the optimal rate [125–131] as well as on improving known information theoretic converse arguments (one of the first contributions being the work presented in this dissertation) to tighten the lower bounds [132–140]. Alternate formulations in the form of information theoretic caching [129, 130] as well as caching and delivery via interference elimination [141] were also proposed.

The original framework of [97, 98] ignored the physical layer impairments of wireless channels. A line of recent work (including work presented in this dissertation) has considered caching in wireless networks [142–145] in the presence of noisy links and feedback. Caching as an aide to wireless interference mitigation was studied in [146–150]. It was shown in [151] that gains from caching and massive MIMO are indeed complementary.

A related line work has been the study of the impact of caching in multi-tier heterogeneous cellular networks, where small-cell base stations like femto or pico-cells are connected to a parent cellular BS. Such networks offer an important application area for local content caching by using the sBSs as cache storage units at the edge of the network. Proactive caching in small-cell networks has been recently studied in [152–161] and references therein. In [152], the authors addressed a distributed cache placement problem with an aim to reduce the delay in delivering the files to the end-users, while a single sBS cache content placement problem was addressed in [153, 162, 163] from a reinforcement learning perspective. The results show that caching offers significant improvements in terms of reduction of peak network load. A different perspective of cache-aided networks was considered in [164–170] where cache-aided networks were modeled from stochastic geometry perspective with the main aim of characterzing the outage probability in closed form and evaluating the useful of cache-aided networks from a purely physical layer standpoint. Based on benefits established by virtue of a vast and ever-increasing array of recent literature, caching has therefore been identified as one of the key enablers of $5\mathrm{G}$ wireless networks [6].

1.3 Main Contributions

In this dissertation, we study the impact of cach-aided network operation on the efficiency of modern multi-tier communication networks. To this end, we ask the following fundamental question

> *What are the theoretical and practical guarantees offered by cache-aided networks for reliable content distribution while minimizing transmission rates and increasing network efficiency?*

In order to furnish an answer to this question, this dissertation mainly focuses on characterizing Shannon-type-limits for caching through the following key contributions.

1.3.1 Information Theoretic Limits of Caching

In this contribution, the fundamental information theoretic limits of caching are investigated. In Chapter 3, the novel caching and coded multicast delivery of Maddah-Ali and Niesen are studied for a general case of multiple file demands in each transmission interval and *tighter* information theoretic lower bounds are derived which better account for the correlation between user caches, when compared to the original results presented in [97, 171]. This leads to a *better approximate characterization of the optimal transmission rate* as a function of per-user cache storage. Extensions of the technique provide strictly tighter lower bounds for caching with D2D-assisted content delivery. A common underlying assumption in the works on information theoretic modeling of cache-aided systems [96–105, 107, 125] is that each device in the network is equipped with the same cache storage. However, in practice, different types of devices in the system might possess different storage. Thus motivated, we introduce the *heterogeneous caching problem* and present a novel layered heterogeneous caching scheme which exploits maximal multicasting opportunity in cache-aided networks with storage heterogeneity. Leveraging a proposed information theoretic lower bound, we show that the proposed scheme is order-optimal for some system settings of practical interest.

1.3.2 Information Theoretic Security for Caching

In Chapter 4, we study the issue of privacy (or security) of multicast content delivery in cache-aided networks. In the information theoretic model of cache-aided systems highlighted in the previous sections, users are often treated jointly and content is shared across caches during placement phase. Furthermore, common multicast transmissions are used to deliver requested content to users. Thus *security* in the content delivery process is a cause of major concern. To this end, we introduce the *secure caching problem* with the goal of minimizing information leakage to an external wiretapper while servicing the legitimate users with the minimum possible rate. Firstly, the fundamental cache storage vs. transmission rate trade-off of the secure caching problem is characterized for the case of

uniform file popularity i.e., in the case where every file has an equal probability of being requested by users. Rather surprisingly, the results show that security can be introduced at a negligible cost, particularly for large number of files and users. It is also shown that the rate achieved by the proposed caching scheme with secure delivery is within a constant multiplicative gap from the information-theoretic optimal rate for most parameter values of practical interest. These results are then extended to the case of files with non-uniform popularity distribution.

1.3.3 Cloud and Cache-Aided Wireless Networks

In Chapter 5, we extend the analysis of cache-aided networks to the wireless domain under a multiple edge node setting. To this end, we study a *cloud and cache-aided wireless network* architecture in which edge-nodes are connected to a cloud processor via dedicated fronthaul links, while also being endowed with caches. We address the interplay between virtualization via cloud processing and localized content distribution via edge caching from an information-theoretic viewpoint by investigating the fundamental limits of a high Signal-to-Noise-Ratio (SNR) metric, termed normalized delivery time (NDT), which captures the worst-case latency for delivering any requested content to the users. The NDT is defined under the assumption of either serial or pipelined fronthaul-edge transmissions, and we study it as a function of fronthaul and cache storage constraints. We propose transmission policies that encompass the caching phase as well as the transmission phase across both fronthaul and wireless segments, with the aim of minimizing the NDT for given fronthaul and cache storage. Information-theoretic lower bounds on the NDT are also derived. Achievability arguments and lower bounds are leveraged to characterize the minimal NDT in a number of important special cases, including systems with no caching capability, as well as to prove that the proposed schemes achieve optimality within a constant multiplicative factor of at most 2 for all values of the problem parameters.

1.3.4 Learning-Aided Caching in Small Cell Networks

In Chapter 6, we study a more practical problem of learning-aided caching in a 2-tier heterogeneous networks. Since file popularity in a real network is generally unknown, we investigate the scope of caching in small-cell base stations from a reinforcement learning perspective wherein an online estimation of file popularities is studied. We present a framework for *topology-aware collaborative caching* in a multiple small cell scenario. A multi-armed bandit based framework is used to learn the popularity profile of files in a given network based on the observation of user requests over time in an online manner. However, even with complete knowledge of popularity profile, the design of optimal caching of whole files, termed as *uncoded caching*, in small cell base stations is shown to be NP-hard. In this work, we present an approximation algorithm for the uncoded caching problem by incorporating a novel graph color and local search based algorithm. We also present an alternate *coded caching* strategy which makes use of rateless coded as means to obtain a relaxation of the original NP-hard placement problem. Through simulations we show that the uncoded approximate

caching algorithm performs close to the optimal coded scheme when integrated with the learning framework in the multi-small-cell setting. We also show that for network topologies of practical interest, the collaborative caching strategies outperform local caching strategies.

The results on fundamental performance limits of cache-aided networks presented in this dissertation show that caching definitively has a positive impact as a tool for high-data-rate multimedia content distribution. The techniques studied and proposed in this work are aimed at better equipping current mobile networks to handle the expected exponential growth in data traffic especially stemming from large scale video distribution over wireless networks as the evolution towards a diverse $5\mathrm{G}$ network architecture takes shape.

1.4 List of Relevant Publications

The work presented in this dissertation is based on the following publications.

Journal Publications:

1. A. Sengupta, R. Tandon, T. C. Clancy, "Fundamental Limits of Caching with Secure Delivery," *IEEE Transactions on Information Forensics and Security*, vol. 10, issue 2, pp 355-370, 13 January 2015.

2. A. Sengupta, R. Tandon, "Improved Approximation of Storage-Rate Trade-off for Caching with Multiple Demands," in revision, *IEEE Transactions on Communications*, September 2016.

3. A. Sengupta, R. Tandon, O. Simeone, "Cloud and Cache Aided Wireless Networks: Fundamental Latency Trade-offs," submitted, *IEEE Transactions on Information Theory*, May 2016.

4. A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, T. C. Clancy, "Learning-Aided Collaborative Caching in Small Cell Networks," submitted, *IEEE/ACM Transactions on Networking*, November 2016.

5. A. Sengupta, R. Tandon, "Layered Caching for Heterogeneous Storage," to be submitted, *IEEE Transactions on Wireless Communications*, November 2016.

Conference Publications:

1. A. Sengupta, R. Tandon, O. Simeone, "Pipelined Fronthaul-Edge Content Delivery in Fog Radio Access Networks", *In Proc. IEEE Globecom Workshop on Emerging Technologies for 5G Wireless Cellular Networks*, Washington DC, USA, December 2016.

2. A. Sengupta, R. Tandon, T. C. Clancy, "Layered Caching for Heterogeneous Storage", *In Proc. Asilomar Conference on Signals and Systems*, Pacific Grove, CA USA, November 2016.

3. A. Sengupta, R. Tandon, O. Simeone, "Cloud RAN and Edge Caching: Fundamental Performance Trade-Offs," *In Proc. IEEE SPAWC*, Edinburgh, UK, July 2016.

4. A. Sengupta, R. Tandon, O. Simeone, "Cache Aided Wireless Networks: Trade-offs between Storage and Latency," *In Proc. CISS 2016*, Princeton, NJ USA, March 2016.

5. A. Sengupta, R. Tandon, T. C. Clancy, "Improved Approximation of Storage-Rate Trade-off for Caching via new Outer Bounds," *In Proc. IEEE Internation Symposium on Information Theory*, Hong Kong, June 2015.

6. A. Sengupta, R. Tandon "Beyond Cut-set Bounds - The Approximate Capacity of D2D Networks," *In Proc. Information Theory and Applications Workshop*, UCSD, San Diego, CA USA, Feb 2015.

7. A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, T. C. Clancy "Learning Distributed Caching Strategies in Small Cell Networks," *IEEE International Symposium on Wireless Communications Systems*, Barcelona, Spain, Aug 2014.

8. A. Sengupta, R. Tandon, T. C. Clancy, "Decentralized Caching with Secure Delivery," *In Proc. IEEE Internation Symposium on Information Theory*, Honolulu, Hawaii USA, July 2014.

9. A. Sengupta, R. Tandon, T. C. Clancy, "Fundamental Limits of Caching with Secure Delivery," *In Proc IEEE International Conference on Communications (Wireless Physical Layer Workshop)*, Sydney, Australia, June 2014.

10. A. Sengupta, R. Tandon, T. C. Clancy, "Secure Caching with Non-Uniform Demands", *Global Wireless Summit, aCcESS Special Session*, Aalborg, Denmark, May 2014.

1.5 Organization of Dissertation

The rest of the dissertation is organized as follows: In Chapter 2 we introduce the cache-aided network model under consideration in this work and present a quick primer on the information theoretic formulation of the caching problem. In Chapter 3, we present new lower bounds on the delivery rate of caching for centralized as well as D2D-assisted content delivery for the case when users can demand multiple files at each transmission interval. We also introduce the problem of caching with heterogeneous storage. In Chapter 4, we introduce the problem of caching with secure delivery. In Chapter 5 we extend our analysis to the wireless domain by studying a cloud and cache-aided wireless network. In Chapter 6, we study the impact of caching in small cell networks from a reinforcement learning perspective. Finally, Chapter 7 concludes the dissertation.

Chapter 2

Background

In this chapter, we provide a formal introduction to cache-aided networks and present a classification of different system models under consideration in this dissertation. We then provide a quick primer on existing preliminary results on information theoretic modeling of cache-aided systems which forms the basis of the majority of work presented in the remainder of this dissertation.

2.1 A Taxonomy of Cache-Aided Networks

In this dissertation, we study the fundamental performance limits of cache-aided networks under different system settings in order to gain a better understanding of the practical impacts of caching and complementary content delivery. The main system model under consideration throughout this work is that of a 2-tier heterogeneous cellular network [12–14], where the macro base station (BS) forms the first tier and the second tier is modeled by small cell base stations (sBS) or edge nodes (EN) which constitute the edge of the network[1] i.e., the last hop to end-users. The user devices are served by both the BS and the ENs. The ENs are connected to the central BS by means of capacity limited backhaul links. The system model is illustrated in Fig. 2.1. It is assumed that the central BS acts as the server which has a library of files (e.g., movies, videos etc.) from which users request popular content. Furthermore, the ENs as well as the users can potentially be equipped with limited cache storage. Under the general umbrella of this heterogeneous network model, we consider different sub-systems in order to effectively characterize the impact of caching as illustrated in Fig. 2.1.

Through the course of this work, we move from a theoretical analysis of single-server cache-aided networks to system models of practical interest. A major part of the work is dedicated to the

[1]The terms *Macro Base-Station, Central Base Station* and *Central Server* are used interchangeably to refer to the 1st tier of the network throughout the course of this dissertation. Similarly, the terms *Small Cell Base Station* and *Edge Nodes* are also used interchangeably to refer to the 2nd tier of the network.

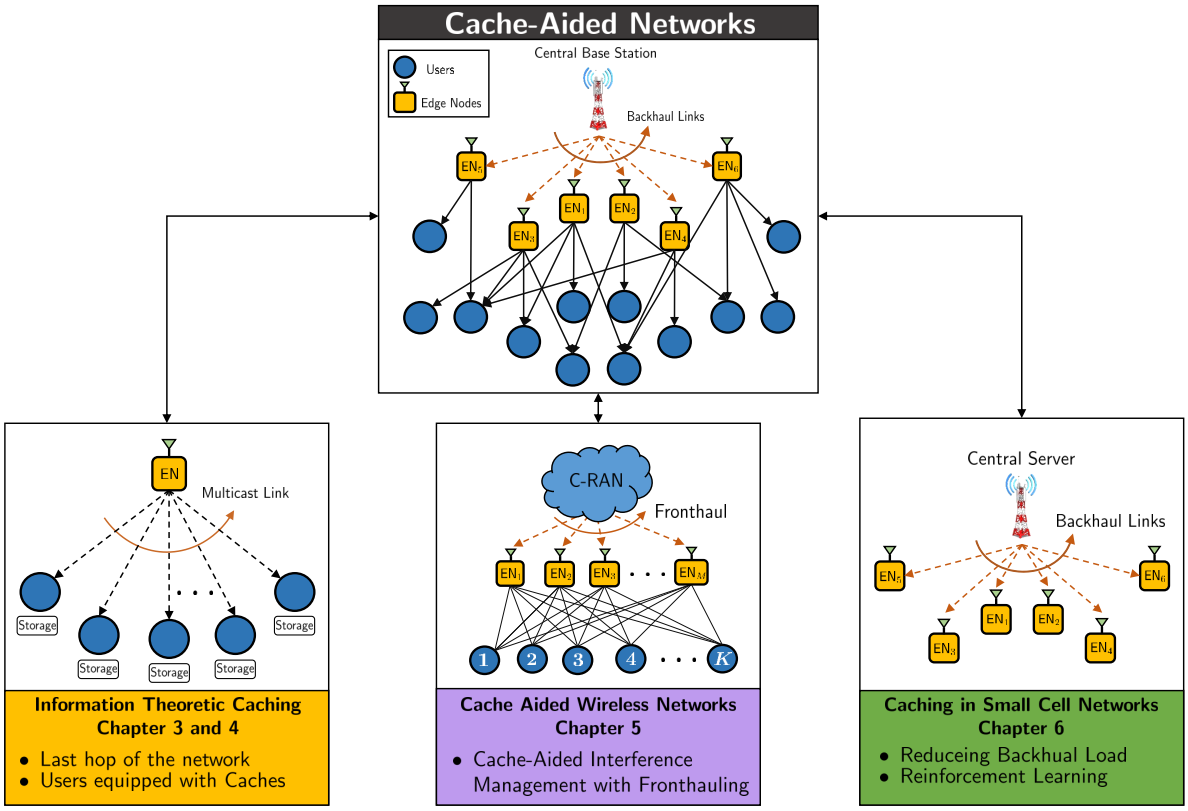


Figure 2.1: Cache-Aided Networks: A taxonomy of system models under consideration.

study of the last hop of the heterogeneous network where the EN acts as the central server and the users are equipped with cache storage. An information theoretic model of this *single-server cache-aided system* was first proposed in [97, 98], where the fundamental storage vs. transmission rate trade-off for this system was approximately characterized under the assumption of error-free bit-pipes connecting the EN to its users. This model presents a theoretical standpoint for analysis of cache-aided systems and ignores wireless physical layer impairments e.g., fading, path-loss, shadowing etc. The information theoretic model of caching is studied in detail in Chapters 3 and 4, where we provide improved approximations of the storage-rate trade-off. A latency centric analysis of cache-aided wireless networks is presented in Chapter 5. We study the information theoretic model of cache-aided interference channels first introduced in [146]. We extend this model to a cloud and cache aided wireless network[2] (which encompasses the model of [146]) where the central BS is replaced by a Cloud Radio Access Network [172] which connects to the ENs via rate-limited fronthaul links. The ENs are endowed with caches and are connected to the users via a noisy wireless interference channel [79–81, 83, 173]. For this system, we introduce a new latency-centric performance metric and provide fundamental information theoretic performance limits. Finally, Chapter 6 presents a practical analysis of cache-aided networks from a reinforcement

[2]The cloud and cache-aided wireless network was first introduced in [149] for the specific case of 2 ENs serving 2 users. In this work, we generalize this model to a M EN, K user network.

learning perspective, where the main objective is to maximally serve user requests directly from the caches at the ENs in order to reduce the backhaul load. The system is along the same lines of the "FemtoCaching" framework studied in [152] for low latency video delivery using caching helpers. Conversely, learning-aided caching was first studied in a single sBS network in [153, 163]. We extend this framework to a multi-sBS network and study the interplay of distributed learning and topology-aware caching strategies.

Next, we present a detailed overview of existing results on the information theoretic caching model of Maddah-Ali and Niesen [97, 98], which forms the basis for majority of the work presented in this dissertation.

2.2 Information Theoretic Model for Caching

We first introduce the system model considered in the single-server centralized coded caching problem in [97]. This system model will be used throughout the remainder of the discussion on the information theoretic model for caching. To this end, consider a content distribution system with K users and a central server which has a library of N files (denoted by $(F_1, F_2, \ldots, F_N)$, each of size B bits). The file have equal popularity i.e., they are modeled as uniformly distributed independent random variables. Each user $k \in \{1, \ldots, K\}$, has a cache storage Z_k of size MB bits. Caching operates in two phases:

1. *Storage phase*: where parts of popular content is placed in users' caches. The storage phase can be of two types: *centralized storage* or *decentralized storage*. In case of centralized storage, the central server stores the cache Z_k of user k with some content, which is a function of the files $(F_1, \ldots, F_N)$. In case of decentralized storage, the user k is allowed to store any random combination of bits from each file without coordination from the central server. User k (for k=$1, \ldots, K$) then requests access to one of the files, F_{d_k}, in the database.

2. *Delivery phase*: where requested content is delivered by exploiting the local cache storage of users. The central server proceeds by transmitting a *multicast* signal $X_{(d_1,\ldots,d_K)}$ of size RF bits over the shared link. Using the content Z_k (of its cache) and the received signal $X_{(d_1,\ldots,d_K)}$, the $k-$th user intends to reconstruct the requested file F_{d_k}.

A storage-rate pair (M, R) is *achievable* if for a (per-user) cache size of MF bits, and using rate RF bits, it is possible for each user to decode its requested file for *any* set of requests $(d_1, \ldots, d_K)$. Fig. 2.2 shows the system model. Let $R^*(M)$ denote the smallest rate R such that the pair (M, R) is achievable. The function $R^*(M)$ is the *fundamental storage-rate trade-off* for the caching problem.

Next we present the main results from [97, 98] which give an achievable rate for the case of centralized and decentralized caching, as well as an information theoretic lower bound on the optimal rate $R^*(M)$. The first result presents achievable rates which upper bound the optimal.

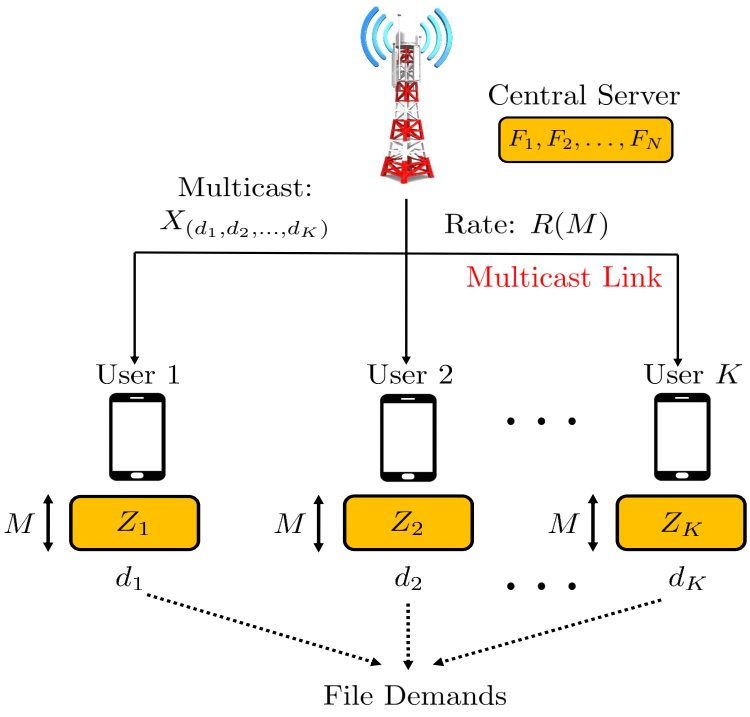


Figure 2.2: Information Theoretic Model for a Single-Server Cache-Aided Network.

Theorem 1 (*Theorem 1*, [97, 98])**.** *For any N files and K users, each with a cache size of $M \in \frac{N}{K}\{0, 1, 2, \ldots, K\}$,*

$$R^*(M) \leq R_{\mathsf{cen}}(M) \triangleq K \cdot \left(1 - \frac{M}{N}\right) \min\left\{\frac{1}{1 + \frac{KM}{N}}, \frac{N}{K}\right\} \tag{2.1}$$

is achievable for centralized caching. For any $1 \leq M \leq N$, the lower convex envelope of these points is achievable. Furthermore, for decentralized caching with a per-user cache size of MB bits with $M \in (0, N]$, the achievable rate is arbitrarily close to

$$R^*(M) \leq R_{\mathsf{dec}}(M) \triangleq K\left(1 - \frac{M}{N}\right) \cdot \min\left\{\frac{N}{K(M)} \cdot \left(1 - \left(1 - \frac{M}{N}\right)^K\right), 1\right\}. \tag{2.2}$$

The first term in the rate expression in (2.1), $K \cdot \left(1 - \frac{M}{N}\right)$, is a *local caching gain*. In traditional unicast delivery enabled systems, users are allowed to store any M/N fraction of every file in their caches and the remaining $(1 - M/N)$ was sent via point-to-point unicast transmissions once the requests were revealed. The second term, $1/\left(1 + KM/N\right)$, is a *global caching gain* which results from jointly designing the storage and delivery phase to enable the use of multicast transmissions to satisfy multiple user requests with a single transmission. The coded caching and multicast delivery scheme which achieves the rate in (2.1) is detailed in [97].

For the case of decentralized caching, each user is allowed randomly cache M/N fraction of every file. The central server then maps the contents of users' caches back to splits in the original files which reflects the parts of the files shared by the users [98] in order to exploit multicast opportunities in the system. The decentralized achievable rate in (2.2) also extracts a global caching gain from the system by virtue of multicast delivery.

The intuition behind the schemes are explained through the following examples.

Example 1 (*Centralized Caching*)**.** We consider the case for $N = K = 3$. For this case, from Theorem 1, $M \in \{0, 1, 2, 3\}$. The system and bounds for this case are illustrated in Fig. 2.3(a) and 2.3(b). First consider the two extreme points $M = 0$ and $M = 3$. In the first the users can store nothing and hence the entire files need to be sent over the air leading to an achievable rate of 3. In the second case, all users can store all files and over the air transmission are not required resulting in an achievable rate of 0. To illustrate the coded caching scheme, we first consider the case of $M = 1$ and three files A, B, C. Each file is split into 3 equal parts i.e., $A = (A_1, A_2, A_3),\ B = (B_1, B_2, B_3),\ C = (C_1, C_2, C_3)$. In this case, each sub-file is of size $B/3$ bits. Each user caches those sub-files for each file which have their index in them. For example user 1 caches sub-file A_1 of file A. The overall cache placement is as follows:

$$Z_1 = \{A_1, B_1, C_1\},\ \ Z_2 = \{A_2, B_2, C_2\},\ \ Z_3 = \{A_3, B_3, C_3\}. \tag{2.3}$$

Thus each cache has size $M = 3 \times (1/3) = 1$. Considering a worst case request where all users request different files, $(d_1, d_2, d_3) = (A, B, C)$, the server can make the transmission,

$$X_{(A,B,C)} = \{A_2 \oplus B_1, A_3 \oplus C_1, B_3 \oplus C_2\}, \tag{2.4}$$

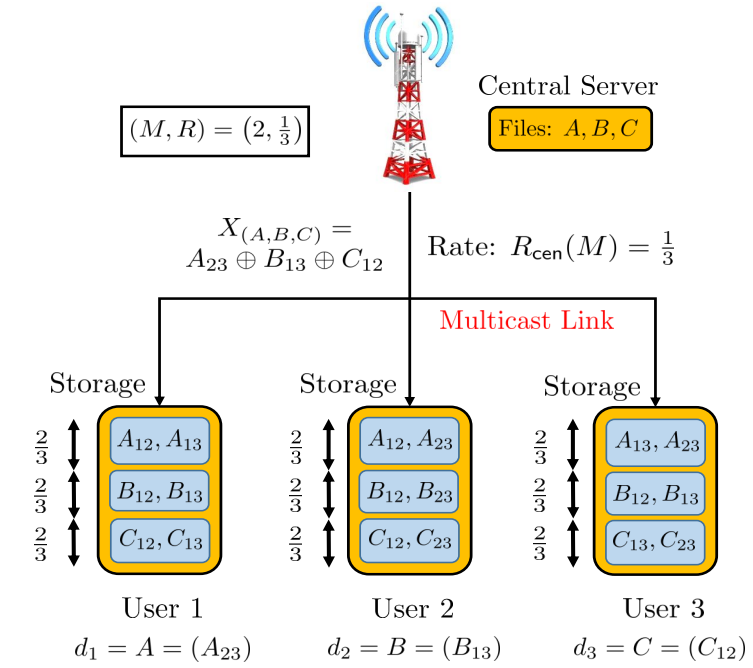
such that everyone can retrieve their requested files. For example from $A_2 \oplus B_1$, user 1 can retrieve A_2 by XOR-ing out B_1 which is already present in its cache. Similarly, user 2 can retrieve B_1 by XOR-ing out A_2. This transmission has a rate of $R = 3 \times \frac{1}{3}$. Thus $(M, R_{\mathsf{cen}}) = (1, 1)$ is achievable. We next consider the case of $M = 2$. Each file, in this case, is split into 3 equal parts i.e., $A = (A_{12}, A_{13}, A_{23}),\ B = (B_{12}, B_{13}, B_{23}),\ C = (C_{12}, C_{13}, C_{23})$ where each sub-file is of size $B/3$ bits. Again, each user caches those sub-files which have their index. For example for file A, user 1 caches the sub-files A_{12}, A_{13}. The overall cache placement is as follows:

$$\begin{aligned} Z_1 &= \{A_{12}, A_{13}, B_{12}, B_{13}, C_{12}, C_{13}\} \\ Z_2 &= \{A_{12}, A_{23}, B_{12}, B_{23}, C_{12}, C_{23}\} \\ Z_3 &= \{A_{13}, A_{23}, B_{13}, B_{23}, C_{13}, C_{23}\} \end{aligned} \tag{2.5}$$

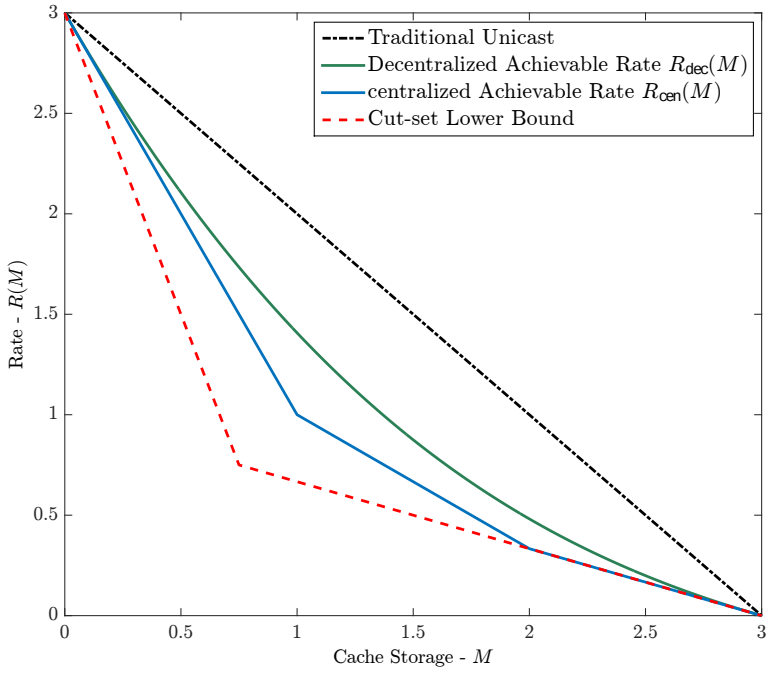
Thus each cache has size $M = 6 \times (1/3) = 2$. Now considering a worst case request where all users request different files, $(d_1, d_2, d_3) = (A, B, C)$, the server can make the transmission,

$$X_{(A,B,C)} = \{A_{23} \oplus B_{13} \oplus \mathcal{C}_{12}\}, \tag{2.6}$$

such that everyone can retrieve their requested files. This transmission is of rate $1/3$. Thus $(M, R_{\mathsf{cen}}) = (2, 1/3)$ is achievable. Thus, the achievable rates yields the blue curve in Fig. 2.3(b).



(a)



(b)

Figure 2.3: (a) Centralized caching scheme and (b) (M, R) trade-off for $N = K = 3$.

The black curve represents the unicast rate for the case when each user is allowed to store M/N fraction of files within their cache and unicast delivery is used to service all requests. It can be seen that the multicast based coded caching scheme offers improvements in the $R(M)$ trade-off. The interested reader is referred to [97] for details of the general scheme which achieves the rate of Theorem 1. □

Example 2 (*Decentralized Caching*)**.** We again consider the case for $N = 3$ files and $K = 3$ users, each with a cache of size MB bits. Let the three files be denoted as $(F_1, F_2, F_3) = (A, B, C)$. Fig. 2.3(b) shows the rate achieved by the decentralized caching scheme. In the decentralized placement phase, each of the 3 users caches a subset of $MB/3$ bits of each file independently at random. Thus, each bit of a file is cached by a specific user with probability $M/3$. Considering the file A, the server maps the storage of fragments of file A at the different users' caches into splits, $A_{\mathcal{T}}$, such that $\mathcal{T} \subseteq \{1, 2, 3\},\ |\mathcal{T}| = i$ for $i = 0, 1, 2, 3$. Thus there are $\sum_{i=0}^{3} \binom{3}{i} = 2^3 = 8$ splits of file A: $(A_\phi, A_1, A_2, A_3, A_{12}, A_{13}, A_{23}, A_{123})$, where A_ϕ consists of bits of A which are not stored in any users' cache. On the other hand, A_{123} has bits which are stored in all users cache. In general, bits in $A_{\mathcal{T}}$ are stored in user k's cache if $k \in \mathcal{T}$. By law of large numbers, we have:

$$|A_{\mathcal{T}}| \approx \left(\frac{M}{3}\right)^{|\mathcal{T}|} \left(1 - \frac{M}{3}\right)^{3-|\mathcal{T}|} B \text{ bits} \tag{2.7}$$

with probability approaching one for large enough file size B. The same analysis holds for files B, C. Next we look at the cache contents from the central server's perspective after the centralized key placement phase and before the delivery procedure begins. The cache placement for $N = K = 3$ is:

$$Z_1 = \left\{ \begin{array}{c} A_1, A_{12}, A_{13}, A_{123} \\ B_1, B_{12}, B_{13}, B_{123} \\ C_1, C_{12}, C_{13}, C_{123} \end{array} \right\} \quad Z_2 = \left\{ \begin{array}{c} A_2, A_{12}, A_{23}, A_{123} \\ B_2, B_{12}, B_{23}, B_{123} \\ C_2, C_{12}, C_{23}, C_{123} \end{array} \right\} \quad Z_3 = \left\{ \begin{array}{c} A_3, A_{13}, A_{23}, A_{123} \\ B_3, B_{13}, B_{23}, B_{123} \\ C_3, C_{13}, C_{23}, C_{123} \end{array} \right\}. \tag{2.8}$$

The cache placement phase is entirely decentralized as the users do not need to consider the number of other users in the system or their cache contents while storing file fragments in their caches. Next, we consider the delivery procedure of the decentralized caching scheme. The system is characterized based on the worst possible rate over the shared link. Thus we consider a request $(F_{d_1}, F_{d_2}, F_{d_3}) = (A, B, C)$. The server responds by transmitting the reply $X_{(A,B,C)}$. Let the set $\mathcal{S} \subseteq \{1, 2, 3\} : |\mathcal{S}| = s$ for $s = 3, 2, 1$. Then we have $X_{(A,B,C)} = \left\{\mathcal{K}_{\mathcal{S}} \oplus_{k\in\mathcal{S}} F_{d_k,\mathcal{S}\setminus\{k\}} : k = 1, 2, 3\right\}_{s=1}^{3}$, where $F_{d_k,\mathcal{S}\setminus\{k\}}$ corresponds to the fraction of the file F_{d_k}, requested by user k which is not present in user k's cache but is present in the cache of the other $s - 1$ users in $\mathcal{S}$. Thus, for $K = 3$ users in the system, the coded multicast delivery procedure has 3 phases for each of $s = 3, 2, 1$.

- *For* $s = 3$: We have $|\mathcal{S}| = 3 \Rightarrow \mathcal{S} = \{1, 2, 3\}$ and $|\mathcal{S} \setminus \{k\}| = 2$. The transmission is $\{A_{23} \oplus B_{13} \oplus C_{12}\}$. In this case, each sub-file is zero padded to the size of the largest sub-file in the set. Considering user 1, we see that Z_1 contains B_{13}, C_{12}. Thus user 1 can XOR out A_{23} from

the transmission. It can be seen that the same holds for users 2 and 3. Thus the transmission is useful for all users. For $s = 3$, there is only one transmission of the size of each of these sub-files. Thus, using (2.7), the rate over the shared link for this transmission is:

$$\left(\frac{M}{3}\right)^2 \left(1 - \frac{M}{3}\right) B. \tag{2.9}$$

- *For* $s = 2$: We have $|\mathcal{S}| = 2 \Rightarrow \mathcal{S} \in \{1,2\},\{2,3\},\{1,3\}$ and $|\mathcal{S} \setminus \{k\}| = 1$. The transmission for each subset $\mathcal{S}$ is $\{\{A_2 \oplus B_1\},\{B_3 \oplus C_2\},\{A_3 \oplus C_1\}\}$. Again for user 1, we can see that Z_1 contains B_1, C_1. Thus it can extract A_2, A_3 from this transmission. Similarly the other users can extract fragments of their requested files. In this case, there are three transmissions, each of the size of file fragment, say, A_2. Thus the rate of this transmission is:

$$3 \cdot \left(\frac{M}{3}\right) \left(1 - \frac{M}{3}\right)^2 B. \tag{2.10}$$

- *For* $s = 1$: We have $|\mathcal{S}| = 1 \Rightarrow \mathcal{S} \in \{1\},\{2\},\{3\}$ and $|\mathcal{S} \setminus \{k\}| = 0$. The transmission for each subset $\mathcal{S}$ is $\{A_\phi, B_\phi, C_\phi\}$. These transmissions are sent to individual users, containing the residual fragments not stored in each user. The size of each transmission is equal to the size of the file fragments A_ϕ, B_ϕ, C_ϕ. Thus the rate of this transmission is:

$$3 \cdot \left(1 - \frac{M}{3}\right)^3 B. \tag{2.11}$$

Again considering user 1, we can see that the fragments of A not present in its cache i.e., A_ϕ, A_2, A_3, A_{23} are extracted from the entire transmission. The same holds true for the other users. The rate for the composite transmission $X_{(A,B,C)}$ is obtained by summing (2.9), (2.10) and (2.11):

$$\begin{aligned} R_{\mathsf{dec}}(M)B &= B\left(\frac{M}{3}\right)^2 \left(1 - \frac{M}{3}\right) + 3B\left(\frac{M}{3}\right)\left(1 - \frac{M}{3}\right)^2 + 3B\left(1 - \frac{M}{3}\right)^3 \\ &= 3\left(1 - \frac{M}{3}\right) \frac{3}{3(M)} \left(1 - \left(1 - \frac{M}{3}\right)^3\right) B, \end{aligned} \tag{2.12}$$

which is the expression given in (2.2) for $N = K = 3$. □

The following theorem presents an information theoretic lower bound on the optimal rate $R^*(M)$ for the single-server cache-aided system.

Theorem 2 (*Theorem 2, [97]*)**.** *For N files and K users, each having a cache size $1 \leq M \leq N$,*

$$R^*(M) \geq \max_{s \in \{1,\ldots,\min\{N,K\}\}} \left(s - \frac{sM}{(\lfloor N/s \rfloor)}\right). \tag{2.13}$$

The proof of Theorem 2 is based on a cut-set argument [174]. The lower bound for the case of $N = K = 3$ is shown by the black curve in Fig. 2.3(b). Further, it was shown in [97, Theorem 3] and [98, Theorem 2] that the cut-set lower bound is within a constant multiplicative gap of 12 from the upper bound for the centralized and decentralized cases respectively.

Caching with D2D-Assisted Delivery

In the information theoretic modeling of cache-aided systems, the nomenclature of centralized/decentralized refers only to the cache storage or pre-fetching phase. Note that the multicast content delivery thereafter is implemented by the central server and is therefore *centralized*. In [105] Ji et. al. studied an alternate device-to-device (D2D) assisted content delivery mechanism, whereby, after the cache storage phase, the content delivery is performed in a distributed manner by leveraging only the contents of the users' caches. The major additional constraint placed on the system is that the users must each posses a minimum cache storage such that the entire library of files can be pre-stored in their collective device caches during the pre-fetching phase. Interestingly, the authors showed that even for such distributed delivery, the multicasting gains are preserved and the scaling behavior of the achievable rate is similar to the case for centralized delivery. The following theorem gives the achievable rate for D2D-assisted content delivery.

Theorem 3 (*Theorem* 1, [105])**.** *For any N files and K users, each with a cache size of $M \in \frac{N}{K}\{1, 2, \ldots, K\}$, such that $KM \geq N$, a delivery rate of*

$$R_{\mathsf{d2d}}(M) \triangleq \min\left\{\frac{N}{M}\left(1 - \frac{M}{N}\right), N\right\} \tag{2.14}$$

is achievable. For any $M \in (N/K, N]$, the lower convex envelope of these points is achievable.

The authors additionally also provide a cut-set lower bound for D2D-assisted content delivery [105, Theorem 2].

Remark 1 (*Looseness of Cut-Set Bounds for Caching*)**.** The lower bounds in literature for both the centralized as well D2D-assisted content delivery methods are derived using a cut-set argument. It was however shown in [97] that the cut-set based lower bound could be potentially loose in general. To this end, the authors characterized the capacity region for the case of $N = K = 2$ and showed that the cut-set bound was loose for this specific example. This looseness stems from the fact that the cut-set arguments do not capture the correlation between the multicast transmissions $X_{(d_1,\ldots,d_K)}$ and the cache contents $Z_1, \ldots, Z_k$ explicitly. Furthermore, the cut-set bound for the D2D-assisted delivery in [105, Theorem 2] is loose even for the minimum cache storage of $M = N/K$. □

Although approximate characterizations of the optimal storage-rate trade-off for both the cache-aided system models are in existence, the *complete characterization* is still an open problem and a

multiplicative gap exists between the upper and lower bounds. There are potentially two reasons for the existence of the gap namely (i) looseness of upper bound and (ii) looseness of the lower bound. Tightening the upper bound entails the design of better achievable schemes and this has been a topic of active research as evidenced by the work in [126–131]. In this work, we instead choose to focus on the converse with an aim to derive tighter lower bounds which fundamentally improve the storage-rate trade-off from an information theoretic standpoint.

Chapter 3

Information Theoretic Limits of Caching

In this chapter, we study the fundamental information theoretic limits of single-server cache-aided systems under uniform file popularity. To this end, we first consider a system where each user has the same cache size and can demand *multiple* files at each transmission interval. For this system, we present new information theoretic lower bounds for centralized as well as D2D-assisted content delivery. Leveraging the structure of the bounds, we present an improved approximation of the fundamental storage vs. rate trade-off. Next, we consider a more practical system setting where every user can have a potentially different cache storage. For this *heterogeneous* setting, we propose a novel layered caching scheme as well as an information theoretic lower bound on the optimal rate. For some system settings of practical interest we show that the proposed layered heterogeneous caching scheme is order-optimal.

3.1 New Outer Bounds on Storage Rate Trade-off for Caching with Multiple Demands

The single-server cache aided system studied by Maddah-Ali and Niesen [97–100] considers the case when each user has the same cache storage and users demand only one file at every transmission interval assuming uniform file popularity i.e., every file is equally likely to be requested. For servicing these requests, a centralized content delivery method is used, whereby the central server uses multicast transmissions to jointly deliver content to users (c.f. Chapter 2). The authors used cut-set based arguments to derive an information theoretic lower bound on the optimal storage-rate trade-off for this system and characterized it to within a constant multiplicative factor of 12 for worst-case user demands. The case when users demand *multiple* files at each transmission interval was initially studied in [171], where the authors proposed the first known cut-set based lower bound for this setting.

In contrast to the centralized delivery model, a distributed device-to-device (D2D) assisted delivery

model was studied in [105] whereby the delivery phase was relegated to the users instead of a centralized server in order to further reduce backhaul load. The main difference between the centralized content delivery studied in [97, 98] and the D2D-assisted delivery studied in [105] is the *distributed nature of multicast transmissions*. In the centralized delivery model of Maddah-Ali and Niesen, the multicast can be any arbitrary function of all the files in the library. Instead, for D2D-assisted delivery, the outgoing multicast from each user *can only depend* on the local cache content of that device. Furthermore, for the case of D2D-assisted delivery the devices must have enough cache storage such that the entire library of files can be stored within the collective caches of the devices. In [105], Ji et.al. presented new storage/delivery mechanisms for D2D-assisted delivery for the case when each user demands a single file at every transmission interval. The results in [105] show that even for D2D-assisted delivery, when the devices can use inter-device coded multicast transmissions to satisfy the demands of other users, order-wise improvements in terms of delivery rate can be achieved as compared to uncoded delivery. The authors also presented a cut-set based lower bound on the storage-rate trade-off. However, the general case when each user can demand *multiple files* at each transmission interval with D2D-assisted delivery has not been considered in literature.

3.1.1 Main Contributions

In this work, we consider the worst-case delivery rate for cache-aided systems under uniform file popularity as in [97, 98, 105, 171] and present fundamental results on the storage vs. rate trade-off for centralized as well as D2D-assisted file delivery. The main contributions are summarized as follows.

- We develop a *new technique* for characterizing information theoretic lower bounds on the *worst-case* storage-rate trade-off for cache-aided systems under centralized and D2D-assisted content delivery for the general case when users can demand multiple files at each transmission interval under the assumption of uniform file popularity.
- The new lower bounds are shown to be *generally tighter* than the cut-set bounds in [97, Theorem 2] and [171, Theorem 2] for all values of problem parameters. The proposed technique also yields the *first known* information-theoretic converse for the case of D2D-assisted delivery when each user demands multiple files at each transmission interval.
- Using the new lower bounds, we characterize the optimal storage-rate trade-off for cache-aided networks to within a constant multiplicative factor of 11 for centralized delivery and 10 for D2D-assisted delivery.

Notation: For any two integers a, b with $a \leq b$, we define $[a:b] \triangleq \{a, a+1, \ldots, b\}$. $b \in [a, c]$ denotes $a \leq b \leq c$ and $b \in (a, c]$ denotes $a < b \leq c$. $Y_{[a:b]}$ denotes the set of random variables $\{Y_i : i = [a:b]\}$ and $Y_{[a,b]}$ denotes the set $\{Y_i : i = a, b\}$. $\mathbb{N}^+$ denotes the set of positive integers; the function $(x)^+ = \max\{0, x\}$; $\lceil x \rceil$, $\lfloor x \rfloor$ are the ceil, floor functions respectively.

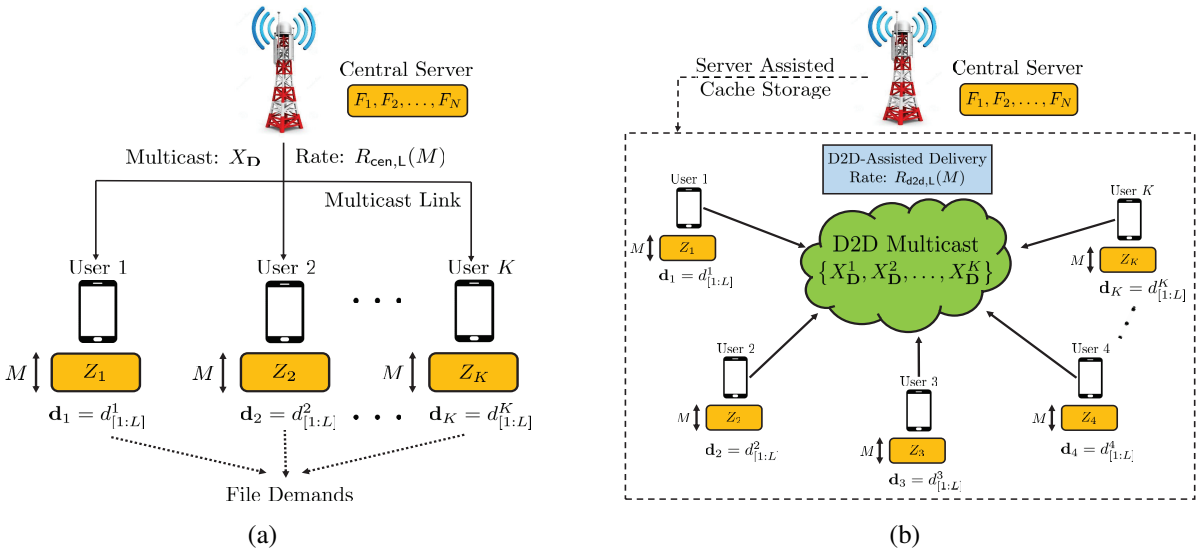


Figure 3.1: System Model for cache-aided network with (a) centralized content delivery where the requested content is delivered via multicast transmission by the central server; and (b) D2D-assisted content delivery where each device multicasts to all the other devices using the contents placed in the device cache by the central server.

3.1.2 System Model

In this section, we introduce the system model for file storage and delivery in cache-aided systems. We consider a cache-aided network (see Fig. 3.1) with K users and a library of N files, $F_{[1:N]}$, where each file is of size B bits, for $B \in \mathbb{N}^+$. Formally, the files F_n are i.i.d. and distributed as:

$$F_n \sim \text{Unif}\left\{1, 2, 3, \ldots, 2^B\right\}, \quad \forall n \in [1:N]. \tag{3.1}$$

Next, we define the key operational phases and the related performance metric for content storage and delivery in cache-aided systems.

Definition 1 (*Cache Storage*)**.** The cache storage phase consists of K caching functions which map the files $F_{[1:N]}$ into the cache content

$$Z_k \triangleq \phi_k\Big(F_{[1:N]}\Big), \tag{3.2}$$

for each user $k \in [1:K]$. For cache-aided systems with centralized content delivery the cache storage constraint is such that $H(Z_k) \in [0, MB]$[1]. For the case of D2D-assisted delivery, an additional storage constraint is that all caches should be collectively capable of storing the entire library

[1]Here $H(Z_k)$ denotes the entropy of the content Z_k stored in the cache of user $k \in [1:K]$ and represents the total size of Z_k in bits i.e., the cache can store at most M files of size B bits each. Similarly, $H(X_{\mathbf{D}})$ denotes the the size in bits of the multicast transmission $X_{\mathbf{D}}$.

$F_{[1:N]}$ i.e., $KM \geq N$ and $H(Z_k) \in [NB/K, MB]$[2]. The cache placement phase generally occurs over a larger time-scale encompassing multiple user demand phases or *transmission intervals*. As a result, the caching functions are agnostic to user demands.

Definition 2 (*File Delivery*)**.** The file delivery phase occurs in each transmission interval in response to user demands with each user requesting $L \in [1:N]$ files. The user demands are denoted by $\mathbf{D} = \mathbf{d}_{[1:K]}$, where each users' demand vector consists of L distinct files $\mathbf{d}_k = d^k_{[1:L]} \in [1:N]$ for $k \in [1:K]$. For the case of centralized delivery, the central server uses N^{KL} encoding functions to map the library of files $F_{[1:N]}$ to the multicast transmission

$$X_{\mathbf{D}} \triangleq \psi_{\mathbf{D}}\big(F_1, \ldots, F_N\big), \tag{3.3}$$

over the shared link with a rate not exceeding RB bits i.e., $H(X_{\mathbf{D}}) \leq RB$. For D2D-assisted delivery, the encoding function $\psi_{\mathbf{D}}$ is composed of K functions, $\psi^k_{\mathbf{D}}$, one for each user. The K users encode the contents of their respective caches into a composite D2D multicast transmission

$$X_{\mathbf{D}} = \Big\{ \big(X^1_{\mathbf{D}}, X^2_{\mathbf{D}}, \ldots, X^K_{\mathbf{D}}\big) : X^k_{\mathbf{D}} = \psi^k_{\mathbf{D}}(Z_k),\ \forall k \in [1:K] \Big\}. \tag{3.4}$$

Each multicast transmission $X^k_{\mathbf{D}}$ has a rate not exceeding $R_k B$ bits i.e., $H\left(X^k_{\mathbf{D}}\right) \leq R_k B$ and the composite multicast has a rate not exceeding the sum-rate of the device multicasts i.e.,

$$H(X_{\mathbf{D}}) \leq \sum_{k=1}^{K} H\left(X^k_{\mathbf{D}}\right) \leq \sum_{k=1}^{K} R_k B \leq RB. \tag{3.5}$$

Definition 3 (*File Decoding*)**.** Once the multicast transmission is received by the users, KN^{KL} decoding functions map the received signal $X_{\mathbf{D}}$ and the local cache content Z_k to the estimates of the L requested files $F_{\mathbf{d}_k}$ for user $k \in [1:K]$ as

$$\widehat{F}_{\mathbf{d}_k} \triangleq \mu_{\mathbf{D},k}\Big(X_{\mathbf{D}}, Z_k\Big). \tag{3.6}$$

The probability of error in file delivery (unreliable delivery) is defined as

$$P_e \triangleq \max_{\mathbf{D},\ k\in[1:K],\ d\in\mathbf{d}_k} \mathbb{P}\left(\hat{F}_d \neq F_d\right), \tag{3.7}$$

which is the worst-case probability of error evaluated over all possible demand vectors and across all users for any number of per-user demands L.

Definition 4 (*Storage-Rate Trade-off*)**.** The storage-rate pair $(M, R_{\mathsf{cen,L}})$ for centralized delivery or $(M, R_{\mathsf{d2d,L}})$ for D2D-assisted delivery is *achievable* if, for any $\epsilon > 0$, there exists a caching and delivery scheme, for which $P_e \leq \epsilon$, where ϵ is an arbitrarily small constant. The optimal storage-rate trade-offs are defined as

$$R^*_{\mathsf{cen,L}}(M) \triangleq \inf\left\{R_{\mathsf{cen,L}} : (M, R_{\mathsf{cen,L}}) \text{ is achievable}\right\}; \tag{3.8}$$

$$R^*_{\mathsf{d2d,L}}(M) \triangleq \inf\left\{R_{\mathsf{d2d,L}} : (M, R_{\mathsf{d2d,L}}) \text{ is achievable}\right\}. \tag{3.9}$$

[2]The lower bound follows from the fact that each cache needs to store at least N/K files.

3.1.3 Preliminary Results

In this section, we present existing achievability results which yield upper bounds on the optimal storage-rate trade-off for cache-aided systems under centralized as well as D2D-assisted delivery for the case of L (≥ 1) demands per user.

3.1.3.1 Centralized Delivery with Multiple Demands

An achievable scheme for caching with centralized delivery was first proposed in [97] for the case of single $(L = 1)$ user requests (c.f. Chapter 2). An extension to the case when each user can make multiple $(L > 1)$ demands at any given transmission interval is given by the following lemma.

Lemma 1. *For any N files and K users, with each user having cache storage of $M \in \frac{Nt}{K}$ files for any $t \in [0 : K]$, an achievable content delivery rate which upper bounds the optimal rate is given by:*

$$R^*_{\mathsf{cen,L}}(M) \leq R_{\mathsf{cen,L}}(M) = KL\left(1 - \frac{M}{N}\right) \min\left(\frac{1}{1 + KM/N}, \frac{N}{KL}\right), \tag{3.10}$$

for the case when each user requests any $L \in [1 : N]$ files at every transmission interval.

Proof. The delivery rate in (3.10) can be achieved by a strategy which treats each of the L sets of user demands *independently* and uses the coded multicast delivery scheme proposed in [97, Theorem 1] for each set of demands. The second term inside the $\min(\cdot)$ function is derived from the unicasting of $\min\{N, KL\}$ files for the cases when multicasting cannot improve on the unicast rate. □

3.1.3.2 D2D-assisted Delivery with Multiple Demands per Device

For the case of D2D-assisted delivery, Ji et. al. proposed an order-optimal caching and delivery scheme in [105] for case of single $(L = 1)$ user demands. An extension to the case of multiple $(L > 1)$ demands per user, is given by the following lemma.

Lemma 2. *For any N files and K users, each having storage size $M \in \frac{Nt}{K}$ files for any $t \in [0 : K]$ with $KM \geq N$, an achievable rate for D2D-assisted content delivery is given by*

$$R_{\mathsf{d2d,L}}(M) \leq \min\left\{\frac{LN}{M}\left(1 - \frac{M}{N}\right), N\right\}, \tag{3.11}$$

for the case when each user requests any $L \in [1 : N]$ files at every transmission interval.

Proof. The delivery rate in (3.11) can be achieved by a strategy which treats each of the L sets of user demands independently and uses the distributed coded multicast delivery scheme proposed in [105, Theorem 1] for each set of demands. The second term inside the $\min(\cdot)$ function is again derived from the multicasting of all N files, which is possible since the storage constraint for D2D-assisted delivery ensures that $KM \geq N$. □

In [171], Ji et. al presented a graph-coloring based index coded delivery scheme which showed that coding across files as well as demands can improve the centralized delivery rate compared to the approach in Lemma 1, while D2D-assisted delivery schemes specifically for multiple $(L > 1)$ demands has not been studied in literature. In this work, we address the following question - *are the schemes which treat multiple sets of user demands independently order-optimal, thereby foregoing the need for more complex approaches?* An answer in the affirmative is provided in Section 3.1.4, where we leverage the proposed lower bounds in conjunction with the upper bounds presented here to improve the approximation of the storage vs. rate trade-off, which in turn proves the order-optimality of treating multiple demands sets independently.

3.1.4 Main Results and Discussion

In this section, we present new converse bounds for centralized and D2D-assisted content delivery in cache-aided networks with multiple ($L \geq 1$) demands per user.

3.1.4.1 Centralized Content Delivery

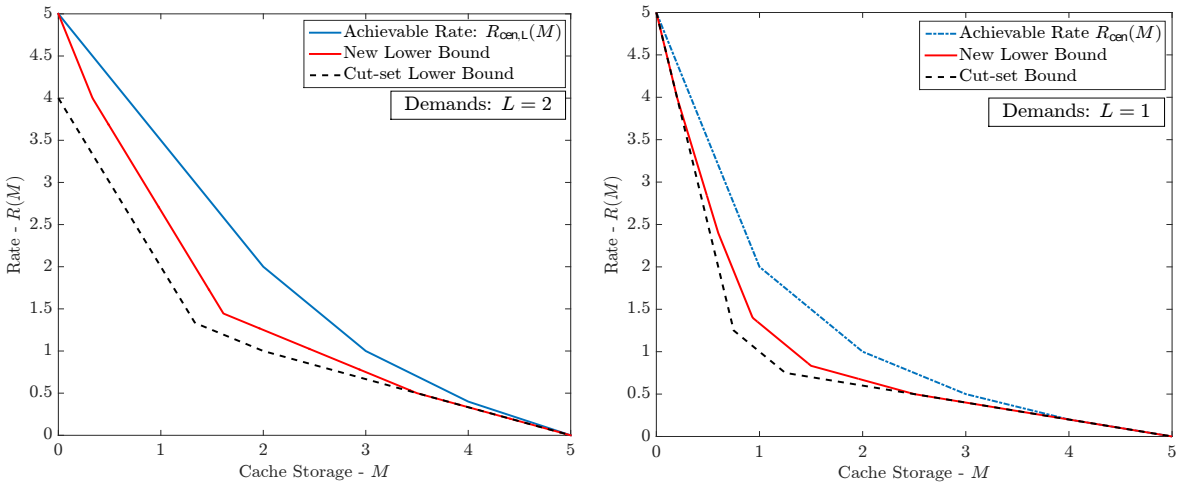
We next present our first main result which gives a new lower bound on the optimal storage-rate trade-off for cache-aided systems with centralized content delivery.

Theorem 4. *For any N files and K users, each having a cache size of $M \in [0, N]$, the optimal centralized content delivery rate $R^*_{\mathsf{cen,L}}(M)$ is lower bounded as*

$$R^*_{\mathsf{cen,L}}(M) \geq \max_{\substack{s\in[1:\min\{\lceil N/L\rceil, K\}],\\ \ell\in[1:\lceil N/(Ls)\rceil]}} \frac{1}{\ell}\left\{N - sM - \frac{\mu(N - L\ell s)^+}{s+\mu} - (N - KL\ell)^+\right\}, \tag{3.12}$$

for the case when each user demands $L \in [1 : N]$ files at every transmission interval. The parameter $\mu = \left(\min\left(\lceil N/(L\ell)\rceil, K\right) - s\right)$, $\forall s, \ell$.

The proof of Theorem 4 is given in Appendix A.1. The expression in Theorem 4 has two parameters, namely (i) the parameter s, which is related to the number of user caches; and (ii) the parameter ℓ, which is related to multicast transmissions. Compared to the cut-set bounds presented in [171, Theorem 2], the additional parameter ℓ adds further flexibility to the lower bound expression and accounts for file decoding through the interaction of caches and transmissions,



(a) (b)

Figure 3.2: Storage-rate trade-off for centralized content delivery with $N = K = 5$ and (a) $L = 2$ demands per user; and (b) $L = 1$ demand per user.

yielding a generally tighter lower bound for the case of centralized content delivery with multiple demands per user. The main difference between the cut-set bound and the proposed lower bound is based on the fact that the new bounds better utilize the possible correlation between caches by carefully bounding the joint and conditional entropy of subsets of cache storages by utilizing Han's inequality on subsets (see Section 3.1.5.1 for more details). The cut-set based lower bound of [171, Theorem 2] is tight only for very large values of cache size M. As shown in the sequel, for such values of M, the proposed bound yields the cut-set bounds for specific choices of s and ℓ and is generally tighter for all other values. This is illustrated in Fig. 3.2(a) where, in addition to the achievable rate from Section 3.1.3, we show that the proposed bound in strictly tighter than the cut-set bound.

We next present our second main result which shows that an improved approximation of the optimal storage-rate trade-off can be obtained by use of the proposed lower bound.

Theorem 5. *For any N files and K users, each with a cache size of $M \in [0, N]$, and each user requesting $L(\leq N)$ files at each transmission interval, we have:*

$$\frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq 11. \tag{3.13}$$

The proof of Theorem 5 is provided in Appendix A.2. This result improves on the gap of 18 between the achievable scheme and the cut-set bound in [171, Theorem 2]. Furthermore, the result shows that treating each of the L sets of user demands independently as a single demand case (as in Lemma 1) is in fact order-optimal, thereby precluding the need for more complex schemes as in [171] which use coding across demands.

Corollary 1. *For any N files and K users, each having a cache size of $M \in [0, N]$, the optimal centralized content delivery rate $R^*_{\mathsf{cen}}(M)$ for the case when each user requests $L = 1$ file at every transmission interval, is lower bounded by:*

$$R^*_{\mathsf{cen}}(M) \geq \max_{s\in[1:K],\ \ell\in[1:\lceil N/s\rceil]} \frac{1}{\ell}\left\{N - sM - \frac{\mu(N-\ell s)^+}{s+\mu} - (N-K\ell)^+\right\}, \tag{3.14}$$

where $\mu = \left(\min\left(\lceil N/\ell\rceil, K\right) - s\right),\ \forall s, \ell$.

Corollary 1 follows by setting $L = 1$ in Theorem 4. The new bounds strictly improve on the cut-set lower bounds presented in [97, Theorem 2] as shown in Fig. 3.2(b). The achievable rate from [97, Theorem 1] is also shown in Fig. 3.2(b). Using (3.14), the approximation of the optimal storage-rate trade-off can be improved as follows.

Theorem 6. *Let $R_{\mathsf{cen}}(M)$ be the achievable rate of the centralized caching scheme given in [97, Theorem 1]. Then, for any K users, N files, and user cache storage in the range $M \in [0, N]$, we have:*

$$\frac{R_{\mathsf{cen}}(M)}{R^*_{\mathsf{cen}}(M)} \leq 8. \tag{3.15}$$

The proof of Theorem 6 is provided in Appendix A.3. The result improves on the gap of 12 yielded by the cut-set bound in [97, Theorem 3][3] and tightens the gap compared to Theorem 5 for the case when $L = 1$.

3.1.4.2 D2D-Assisted Content Delivery

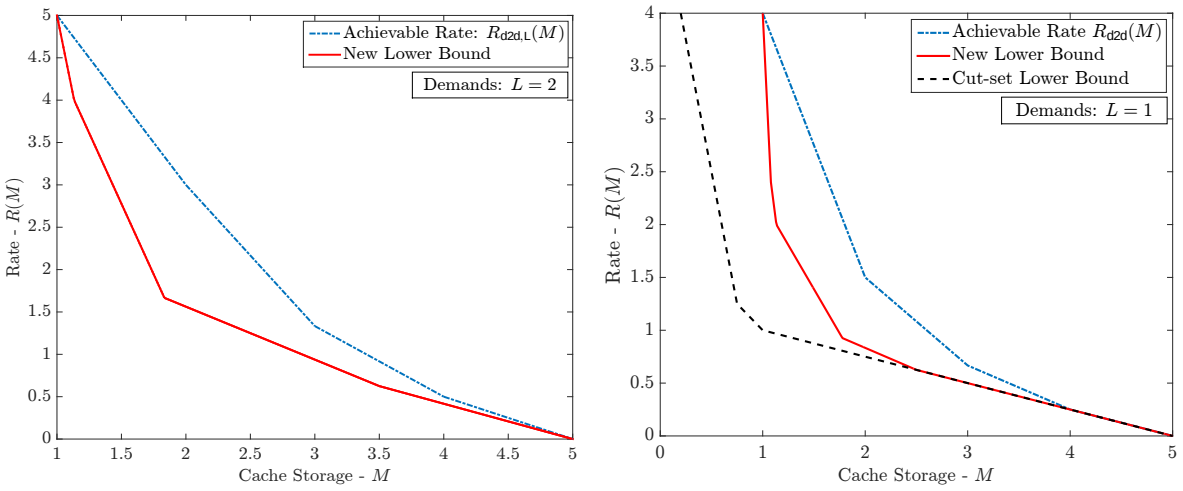
In this section, we consider the case of D2D-assisted content delivery with each user demanding multiple files in each transmission interval. The next theorem presents our main result which gives the first-known lower bound on the optimal storage-rate trade-off.

Theorem 7. *For any N files and K users, each having a cache size of $M \in [N/K, N]$, the optimal D2D-assisted content delivery rate $R^*_{\mathsf{d2d,L}}(M)$ is lower bounded as*

$$R^*_{\mathsf{d2d,L}}(M) \geq \max_{s\in[1:\min\{\lceil N/L\rceil, K\}],\ \ell\in\left[1:\left\lceil\frac{N}{Ls}\right\rceil\right]} \left\{\frac{N - sM - \frac{\mu}{s+\mu}(N-L\ell s)^+}{\ell\left(\frac{K-s}{K}\right)}\right\}, \tag{3.16}$$

for the case when each user demands $L \in [1 : N]$ files at each transmission interval. The parameter $\mu = \left(\min\left(\lceil N/(L\ell)\rceil, K\right) - s\right),\ \forall s, \ell$.

[3]The results presented in this chapter also hold for the case of *decentralized cache placement* as in [98] since the converse makes no assumption on the nature of content placement.



(a) (b)

Figure 3.3: Storage-rate trade-off for D2D-assisted content delivery with $N = K = 5$ and (a) $L = 2$ demands per user; and (b) $L = 1$ demand per user.

The proof of Theorem 7 is presented in Appendix A.4. Similar to Theorem 4, the parameters s and ℓ yield a family of lower bounds by exploiting the correlation between the caches and transmissions by use of Han's Inequality. Fig. 3.3(a) shows the lower bound in (3.16) and the upper bound $R_{\mathsf{d2d,L}}(M)$ given in (3.11). Leveraging the proposed lower bound, we present our second main result in the following theorem.

Theorem 8. *For any N files and K users, each having a cache size of $M \in [N/K, N]$, and with each user requesting $L(\leq N)$ files at each transmission interval, we have*

$$\frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq 10. \tag{3.17}$$

The proof of Theorem 8 is presented in Appendix A.5. The result shows that treating each of the L sets of user demands as a single demand case as outlined in Lemma 2 is in fact order-optimal and yields a constant factor approximation of the storage-rate trade-off for D2D-assisted content delivery with multiple demands per user.

Corollary 2. *For any N files and K users, each having a cache size of $M \in [N/K, N]$, the optimal D2D-assisted content delivery rate $R^*_{\mathsf{d2d}}(M)$, for the case when each user requests $L = 1$ file at every transmission interval, is lower bounded by:*

$$R^*_{\mathsf{d2d}}(M) \geq \max_{s\in[1:K],\ \ell\in[1:\lceil N/s\rceil]} \left\{ \frac{N - sM - \left(\frac{\mu}{s+\mu}\right)(N-\ell s)^+}{\ell\left(\frac{K-s}{K}\right)} \right\}, \tag{3.18}$$

where $\mu = \left(\min\left(\lceil N/\ell \rceil, K\right) - s\right)\ \forall s, \ell$.

Corollary 2 follows by setting $L = 1$ in Theorem 7 and was originally presented in [179].

Remark 2. Compared to the cut-set bound in [105, Theorem 2], we note that the proposed bound in Corollary 2 is always tighter owing to the additional parameter ℓ and the factor $(K - s)/K \leq 1$ in the denominator of (3.18). Furthermore, the bound in [105] is tight only for large values of device storage size M. The new bound is tighter for smaller values of M and yields the existing bound as a special case for large values of M. In fact, for the smallest allowable cache size of $M = N/K$, the lower bound in (3.18) is tight and yields the achievable rate in [105, Theorem 1] as shown in Fig. 3.3(b). □

Using the lower bound in Corollary 2, the optimal storage-rate trade-off for the case of single demands per user can be approximated as follows.

Theorem 9. *For any $K \in \mathbb{N}^+$ user devices, $N \in \mathbb{N}^+$ files, and device storage in the range $M \in \left[\frac{N}{K}, N\right]$, we have:*

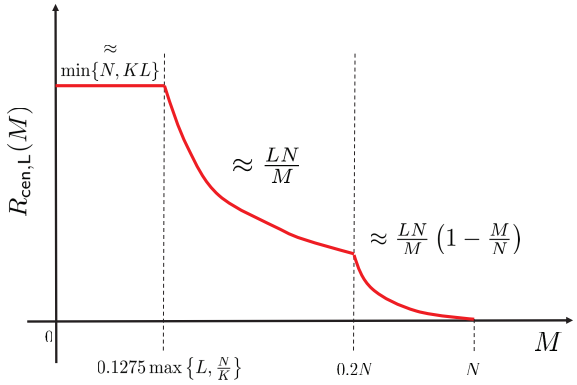
$$\frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \begin{cases} = 1 & M = N/K \\ \leq 3 & M \in (N/K, 2/3] \\ \leq 6 & M \in (2/3, 1] \\ \leq 8 & 1 \leq M \leq N \end{cases}. \tag{3.19}$$

The proof of Theorem 9 is presented in Appendix A.6. The result highlights the fact that for the smallest allowable cache size of $M = N/K$, the lower bound in (3.18) is tight and yields the achievable rate in [105, Theorem 1]. This is also shown in Fig. 3.3(b) for the case of $N = K = 5$ and $L = 1$.

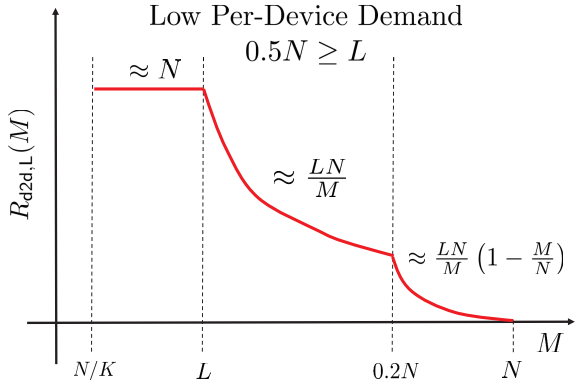
Remark 3. To prove the order-optimality of the schemes which treat each of the L sets of K user demands independently as a single per-user demand case as shown in Theorems 5 and 8, we use approximations to the achievable rates presented in Lemmas 1 and 2. These approximations are highlighted in Fig. 3.4. For the case of centralized content delivery, three regimes of cache storage are considered and for very low cache storage, it is approximately optimal to unicast all requested files as seen in Fig. 3.4(a). For higher cache storage, a linear dependance of the rate on L/M is established. For the case of D2D-assisted delivery, we see that when users demand less than half the library, three regimes of cache storage need to be considered, while for the case of high per-device demands, only 2 regimes suffice and for storage as high as a third of the library, it is approximately optimal for all users to broadcast all N files from their local caches. Further details are provided in Appendix A.2 and A.5. □

3.1.5 Case Studies

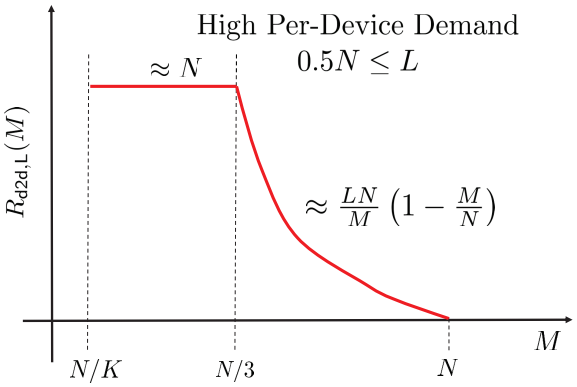
In this section, we present two case studies to illustrate the new techniques used to obtain the lower bounds in Theorems 4 and 7. For ease of exposition, we consider the special case of $L = 1$



(a)



(b)



(c)

Figure 3.4: A representation of the order-optimal approximations to the delivery rate for schemes which treat each of the L sets of K user demands independently as a single per-user demand case for (a) centralized content delivery, which is used in the proof of Theorem 5; and $(b)-(c)$ for D2D-assisted content delivery with low and high per-device demands, which are used in the proof of Theorem 8.

since the results easily extend to any $L > 1$. We show that our technique yields additional bounds as compared to the cut-set techniques in literature and present discussions behind the principal intuitions in applying our method.

3.1.5.1 Centralized Content Delivery: Intuition Behind The Proof of Theorem 4

We consider $N = 3$ files, denoted by A, B, C and $K = 3$ users, each with a cache storage M files. For the case of $L = 1$, Corollary 1 yields the following lower bounds for different values of the parameters s, ℓ:

$$3R^*_{\mathsf{cen}} + 6M \geq 8, \quad s = 2, \ \ell = 1 \tag{3.20}$$
$$4R^*_{\mathsf{cen}} + 2M \geq 5, \quad s = 1, \ \ell = 2 \tag{3.21}$$
$$R^*_{\mathsf{cen}} + 3M \geq 3, \quad s = 3, \ \ell = 1 \tag{3.22}$$
$$3R^*_{\mathsf{cen}} + M \geq 3, \quad s = 1, \ \ell = 3. \tag{3.23}$$

The existing lower bounds from [97, Theorem 2] are given by (3.22)-(3.23). The proposed approach provides the additional bounds in (3.20)-(3.21), thereby yielding tighter lower bounds than [97, Theorem 2] as shown in Fig. 3.5(a). Next, we detail the derivation of the first bound in (3.21) highlighting the new aspects and techniques.

To this end, we consider two consecutive requests $(d_1, d_2, d_3) = (A, B, C)$ and $(d_1, d_2, d_3) = (B, C, A)$. It is clear that the first $s = 2$ caches $Z_{[1,2]}$ along with two corresponding transmissions X_{ABC}, X_{BCA} from the central server suffice to decode all the 3 files. We upper bound the entropy of $\ell = 1$ multicast transmission by the optimal rate R^*_{cen} and use the other transmission's decoding capability with the caches to derive the following bound

$$\begin{aligned}
3B &\leq H(Z_{[1,2]}, X_{ABC}, X_{BCA}) \\
&\leq H(Z_{[1,2]}) + H(X_{ABC}, X_{BCA}|Z_{[1,2]}) \\
&\leq 2MB + H(X_{ABC}) + H(X_{BCA}|Z_{[1,2]}, X_{ABC}) \\
&\overset{\text{(a)}}{\leq} 2MB + R^*_{\mathsf{cen}}B + H(X_{BCA}|Z_{[1,2]}, X_{ABC}, A, B) \\
&\leq 2MB + R^*_{\mathsf{cen}}B + H(X_{BCA}, Z_3|Z_{[1,2]}, X_{ABC}, A, B) \\
&\leq 2MB + R^*_{\mathsf{cen}}B + H(Z_3|Z_{[1,2]}, X_{ABC}, A, B) + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B) \\
&\leq 2MB + R^*_{\mathsf{cen}}B + H(Z_3|Z_{[1,2]}, A, B) + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B, C) \\
&\overset{\text{(b)}}{\leq} 2MB + R^*_{\mathsf{cen}}B + H(Z_3|Z_{[1,2]}, A, B),
\end{aligned} \tag{3.24}$$

where step (a) follows from the fact that $Z_{[1,2]}$ along with X_{ABC} can decode files A, B and step (b) follows from the fact that $H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B, C) = 0$ since each transmission is a deterministic function of the files. Considering the term $H(Z_3|Z_{[1,2]}, A, B)$ in (3.24), we have:

$$H(Z_3|Z_{[1,2]}, A, B) = H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{3.25}$$

Using (3.25) in (3.24), we have:

$$3B \leq 2MB + R^*_{\mathsf{cen}}B + H(Z_{[1:3]}|A,B) - H(Z_{[1,2]}|A,B). \tag{3.26}$$

Now considering all possible subsets of $Z_{[1:3]}$ with cardinality 2, in the RHS of (3.26), we have:

$$3B \leq 2MB + R^*_{\mathsf{cen}}B + H(Z_{[1:3]}|A,B) - H(Z_{[2,3]}|A,B) \tag{3.27}$$

$$3B \leq 2MB + R^*_{\mathsf{cen}}B + H(Z_{[1:3]}|A,B) - H(Z_{[1,3]}|A,B). \tag{3.28}$$

Summing (3.26)-(3.28), and normalizing by 3, we have:

$$3B \leq 2MB + R^*_{\mathsf{cen}}B + H(Z_{[1:3]}|A,B) - \sum_{i,j=1,\ i\neq j}^{3} \frac{H(Z_{[i,j]}|A,B)}{3}. \tag{3.29}$$

We next state Han's Inequality [174, Theorem 17.6.1] on subsets of random variables, which we use for further upper bounding (3.29) in order to derive the proposed lower bound.

Han's Inequality: Let $Y_{[1:n]}$ denote a set of random variables. Further, let $\left(\mathsf{Y}_{[m]}, \mathsf{Y}_{[r]}\right) \subseteq Y_{[1:n]}$ denote subsets of cardinality m, r with $m \leq r$. Han's Inequality states that

$$\frac{1}{\binom{n}{r}} \sum_{\mathsf{Y}_{[r]}:\left|\mathsf{Y}_{[r]}\right|=r} \frac{H\left(\mathsf{Y}_{[r]}\right)}{r} \leq \frac{1}{\binom{n}{m}} \sum_{\mathsf{Y}_{[m]}:\left|\mathsf{Y}_{[m]}\right|=m} \frac{H\left(\mathsf{Y}_{[m]}\right)}{m}, \tag{3.30}$$

where the sums are over all subsets of cardinality r, m respectively. Next, from (3.29), consider the set of random variables $Z_{[1:3]}$ and its subsets $\left(Z_{[1,2]}, Z_{[1,3]}, Z_{[2,3]}\right)$ of cardinality 2. Applying Han's Inequality for these random variables, using $n = r = 3$ and $m = 2$ in (3.30), we have:

$$\frac{2H\left(Z_{[1:3]}|A,B\right)}{3} \leq \sum_{i,j=1,i\neq j}^{3} \frac{H\left(Z_{[i,j]}|A,B\right)}{3}. \tag{3.31}$$

Substituting (3.31) into (3.29), we have:

$$\begin{aligned} 3B &\leq 2MB + R^*_{\mathsf{cen}}B + H(Z_{[1:3]}|A,B) - \frac{2}{3}H(Z_{[1:3]}|A,B) \\ &\leq 2MB + R^*_{\mathsf{cen}}B + \frac{1}{3}H(Z_{[1:3]}|A,B) \leq 2MB + R^*_{\mathsf{cen}}B + \frac{1}{3}H(Z_{[1:3]},C|A,B) \\ &\leq 2MB + R^*_{\mathsf{cen}}B + \frac{1}{3}\left(\underbrace{H(C|A,B)}_{\leq B} + \underbrace{H(Z_{[1:3]}|A,B,C)}_{=0}\right) \leq 2MB + R^*_{\mathsf{cen}}B + \frac{1}{3}B. \end{aligned} \tag{3.32}$$

Rearranging (3.32), we get the new lower bound given by the first inequality in (3.21). The second bound in (3.21) can be obtained similarly by considering $s = 1$ cache and bounding the entropy of $\ell = 2$ transmissions by the optimal rate R^*_{cen} and following steps similar to (3.24)-(3.32).

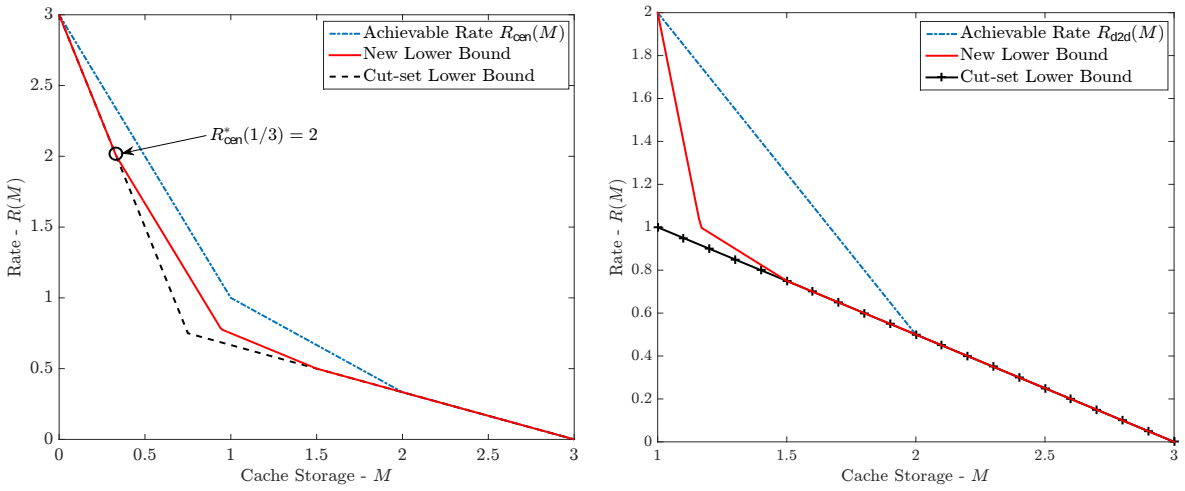


Figure 3.5: Storage-rate trade-off for $N = K = 3$ and $L = 1$ with (a) centralized content delivery and (b) D2D-assisted content delivery.

Remark 4. We note that the key distinction from the cut-set bounds is the mechanism of bounding the joint entropy of random variables representing the multicast transmissions and the stored contents. Specifically, considering the first inequality in (3.24), a naive upper bound on the term $H(X_{BCA}|Z_{[1,2]}, X_{ABC})$ would be R^*_{cen}, which would lead to $3 \leq 2M + 2R^*_{\text{cen}}$, which is a loose bound. The main idea is to first observe that given $Z_{[1,2]}$ and the multicast transmission X_{ABC}, the files A, B can be recovered. Hence, we expect a dependence between X_{BCA} and the random variables in the conditioning. In order to capture this dependency, we consider multiple such requests over time, allowing us to write (3.27), and (3.28), similar to (3.26). This symmetrization argument directly leads to the use of Han's inequality and subsequently to the new lower bound. This is the key approach behind Corollary 1 and Theorem 4 which is a general result and holds for all problem parameters. □

Remark 5. Recently [125,126,128] proposed caching and delivery schemes which improve upon the original multicasting scheme presented in [97, Theorem 1]. Specifically, [125] showed that for $K \geq N$, in the small buffer region of $M = 1/K$, the achievable rate is given by $N(1 - M)$ which improves on the achievable rate in [97, Theorem 1]. For $N = K = 3$, the new achievable point $(M, R) = (1/3, 2)$ is highlighted in Fig. 3.5(a). The lower bound in [97, Th, 2] is shown to be tight only in the regime $0 \leq M \leq 1/K$ for $K \geq N$ in [125]. The lower bound presented in Corollary 1 shows that this is indeed the case and that the new converse is tighter than the cut-set based lower bound for $M > 1/K$ as shown in in Fig. 3.5(a). □

Remark 6. In [97], the authors characterize the optimal storage-rate trade-off for the case of $N = K = 2$ and show that the cut-set lower bound, given by $R^*_{\text{cen}} + 2M \geq 2$ and $2R^*_{\text{cen}} + M \geq 2$, is indeed loose by deriving an additional tighter lower bound $2R^*_{\text{cen}} + 2M \geq 3$ using an alternate approach based on symmetric requests which decode the same file with different combinations

of caches. Our proposed technique also yields this additional bound, making it tighter than cut-set bounds and characterizes the optimal rate for the case of $N = K = 2$. Note however, that the alternate method proposed in [97, Appendix] is discussed only for the case of $N = K = 2$, whereas our approach is a more general one for any N, K. □

3.1.5.2 D2D-assisted Content Delivery: Intuition Behind The Proof of Theorem 7

We next follow up the discussion in the previous section with an additional example to highlight our proposed techniques for the case of D2D-assisted content delivery with $L = 1$ demand per user. To this end, consider again a system with $N = 3$ files (A, B, C) and $K = 3$ users, each with a cache storage of $M \geq 1$. The proposed lower bound in Corollary 2 gives following bounds for different values of parameters s, ℓ:

$$R^*_{\mathsf{d2d}} + 6M \geq 8, \quad s = 2,\ \ell = 1 \tag{3.33}$$

$$8R^*_{\mathsf{d2d}} + 6M \geq 15, \quad s = 1,\ \ell = 2 \tag{3.34}$$

$$2R^*_{\mathsf{d2d}} + M \geq 3, \quad s = 1,\ \ell = 3, \tag{3.35}$$

where (3.35) also recovers the cut set bound in [105, Theorem 2]. Fig. 3.5(b) shows that the additional bounds yielded by the proposed technique outperform the cut-set bounds from literature. To facilitate the derivation of the new bounds, we first consider the request vectors $(d_1, d_2, d_3) = (A, B, C)$ and $(d_1, d_2, d_3) = (B, C, A)$ and two composite transmissions $X_{ABC} = \{X^1_{ABC}, X^2_{ABC}, X^3_{ABC}\}, X_{BCA} = \{X^1_{BCA}, X^2_{BCA}, X^3_{BCA}\}$. From the sum-rate constraint of the multicast transmissions in (3.5), we have

$$H(X_{ABC}) \leq \sum_{k=1}^{3} H(X^k_{ABC}) \leq R^*_{\mathsf{d2d}} B, \text{ and } H\left(X^k_{ABC}\right) \leq R^*_{\mathsf{d2d}} B/3,\ \forall k \in \{1, 2, 3\}, \tag{3.36}$$

where the second inequality follows by symmetry, assuming each device has the same transmission rate. We first note that, given the first $s = 2$ cache contents $Z_{[1,2]}$, the two transmissions X^3_{ABC}, X^3_{BCA} from the *third* user device are able to decode all 3 files. We upper bound the entropy of $\ell = 1$ transmission and use the other transmission's decoding capability, in conjunction with the cache contents $Z_{[1,2]}$, to derive a tighter bound as follows.

$$\begin{aligned}
3B &\leq H(Z_{[1,2]}, X_{ABC}, X_{BCA}) \\
&\leq H(Z_{[1,2]}) + H(X_{ABC}, X_{BCA}|Z_{[1,2]}) \\
&\leq H(Z_{[1,2]}) + H(X_{ABC}|Z_{[1,2]}) + H(X^3_{BCA}|Z_{[1,2]}, X_{ABC}) \\
&\overset{(a)}{\leq} 2MB + H(X^3_{ABC}) + H(X_{BCA}|Z_{[1,2]}, X_{ABC}) \\
&\leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(X_{BCA}|Z_{[1,2]}, X_{ABC}, A, B) \\
&\leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(X_{BCA}, Z_3|Z_{[1,2]}, X_{ABC}, A, B)
\end{aligned}$$

$$\leq 2MB + R^*_{\mathsf{d2d}}B/3 + H(Z_3|Z_{[1,2]}, X_{ABC}, A, B) + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B)$$
$$\overset{\text{(b)}}{\leq} 2MB + R^*_{\mathsf{d2d}}B/3 + H(Z_3|Z_{[1,2]}, A, B), \tag{3.37}$$

where step (a) follows from the fact that in X_{ABC}, the transmissions from devices 1 and 2 are functions of the cache contents $Z_{[1,2]}$ within the conditioning in the second term; step (b) follows from the fact that $H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B, C) = 0$ since X_{BCA} is a function of the *cache contents* $Z_{[1:3]}$. Considering the term $H(Z_3|Z_{[1,2]}, A, B)$, we have:

$$H(Z_3|Z_{[1,2]}, A, B) = H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{3.38}$$

Using (3.38) in (3.37), we have:

$$3B \leq 2MB + R^*_{\mathsf{d2d}}B/3 + H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{3.39}$$

Again, considering all possible subsets of $Z_{[1:3]}$ having cardinality 2, in the RHS of (3.39), we have

$$3B \leq 2MB + R^*_{\mathsf{d2d}}B/3 + H(Z_{[1:3]}|A, B) - H(Z_{[2,3]}|A, B). \tag{3.40}$$
$$3B \leq 2MB + R^*_{\mathsf{d2d}}B/3 + H(Z_{[1:3]}|A, B) - H(Z_{[1,3]}|A, B). \tag{3.41}$$

Symmetrizing over the inequalities in (3.39)-(3.41), we have:

$$3B \leq 2MB + R^*_{\mathsf{d2d}}B/3 + H(Z_{[1:3]}|A, B) - \sum_{i,j=1,i\neq j}^{3} \frac{H(Z_{[i,j]}|A, B)}{3}. \tag{3.42}$$

Next, considering the set of caches $Z_{[1:3]}$ and its subsets $Z_{[1,2]}, Z_{[1,3]}Z_{[2,3]}$ of cardinality 2 and applying Han's Inequality (as in (3.30)), we have from (3.39)

$$3B \leq 2MB + R^*_{\mathsf{d2d}}B/3 + H(Z_{[1:3]}|A, B) - \frac{2H(Z_{[1:3]}|A, B)}{3}$$
$$\leq 2MB + R^*_{\mathsf{d2d}}B/3 + \frac{H(Z_{[1:3]}, C|A, B)}{3} \leq 2MB + R^*_{\mathsf{d2d}}B/3 + B/3. \tag{3.43}$$

Rearranging (3.43), we get the new lower bound in (3.33). Next, we consider $s = 1$ device cache, Z_1, and three request vectors $(d_1, d_2, d_3) = (A, B, C)$, $(d_1, d_2, d_3) = (B, C, A)$ and $(d_1, d_2, d_3) = (C, A, B)$ along with the multicast transmissions $X_{ABC}, X_{BCA}, X_{CAB}$ which are capable of decoding all 3 files. In this case, we upper bound the entropy of $\ell = 2$ composite transmissions with the sum-rate $2R^*_{\mathsf{d2d}}/3$ which is due to the fact that given Z_1 the composite transmissions are simply functions of transmissions from devices $2, 3$. Following similar steps as the previous case leads us to the lower bound in (3.34). Finally, considering again, $s = 1$ device storage content, Z_1, and three request vectors $(d_1, d_2, d_3) = (A, B, C)$, $(d_1, d_2, d_3) = (B, C, A)$ and $(d_1, d_2, d_3) = (C, A, B)$ along with three transmissions $X_{ABC}, X_{BCA}, X_{CAB}$ which are capable of decoding all 3 files. We upper bound the entropy of $\ell = 3$ device transmissions by their sum-rate $2R^*_{\mathsf{d2d}}/3$ as before, thereby recovering the cut set bound in (3.35). The new converse is strictly tighter than the cut set bounds. Furthermore, the proposed converse is tight at the point $M = N/K = 1$. Setting $M = 1$ in (3.33) and comparing with the upper bound from [105, Theorem 1] yields $R^*_{\mathsf{d2d}}(1) = 2$ i.e., the achievable scheme proposed in [105] is optimal at $M = 1$.

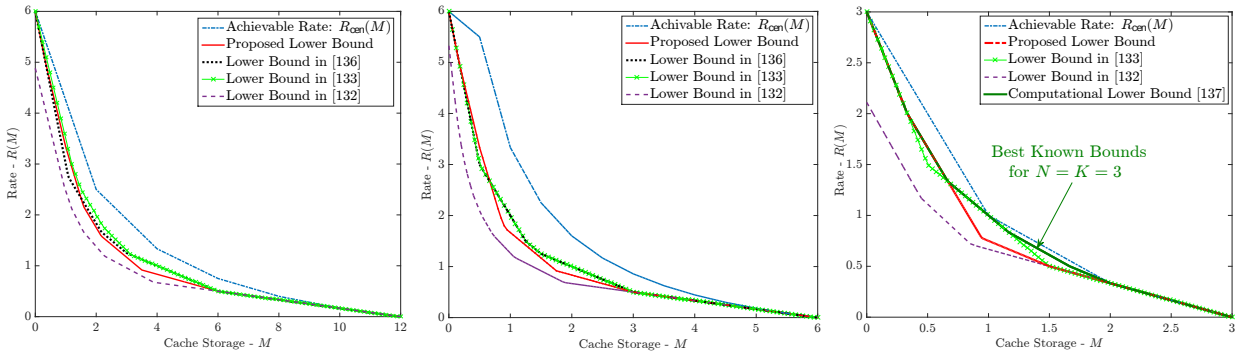


Figure 3.6: Comparisons with parallel results for the case of centralized content delivery with $L = 1$ for a cache-aided system with (a) $N = 12, K = 6$; (b) $N = 6, K = 12$ and (c) $N = K = 3$.

3.1.6 Comparisons with Independent Parallel Results

We acknowledge the recent independent contributions from [132,133,136,137,139,140] on developing converse results for cache-aided systems. The authors in [139, 140] derive a new converse bound based on symmetry of cache placement and principles of index coding with side information respectively, for the case of centralized content delivery with $L = 1$, which shows that the achievable scheme in [97] is optimal if *uncoded cache placement* is assumed. Improvements over the cut-set bound are also obtained for the case when $L = 1$ [136] and $L \geq 1$ [133] for centralized delivery and for the case of $L = 1$ for D2D-assisted delivery in [133], through different approaches than ours. The lower bounding approach adopted in these papers are inspired by the method adopted in [97, Appendix] for deriving a tighter lower bound for the specific case of $N = K = 2$. While a direct comparison is analytically intractable, especially owing to the algorithm based approach of [133], we present some numerical comparisons to show that our bounds supersede these bounds in certain regimes of cache storage M for the single demand case while in some cases [133] yields a better bound. To this end, in Fig. 3.6(a) and 3.6(b), we plot the results in [133, 136] for $L = 1$. It can be seen that our bounds are better than [136] for the case of low cache storage for both cases and supercedes [133] in the second case, again for low cache storage. Note however, that unlike the simple form of our bound, the algorithm used in [133] to evaluate the lower bound has significant complexity with increasing number of users. Finally, we note that a *holistic lower bound* for centralized content delivery with $L = 1$ is obtained only by combination of all lower bounding approaches in literature and maximizing over the bounds yielded by each method. The authors in [136] do not derive a constant gap result, however, the authors in [133] show a constant gap of 4 to the achievable rate in [97, Theorem 1]. We emphasize here that the analyses to obtain multiplicative gaps (as in Theorems 5 and 8) are essentially approximations. Thus, deriving lower bounds geared towards tightening this analysis does not guarantee the best known bounds. To this end, we consider the lower bounds presented in [132]. The proposed lower bounds are generally always looser than the cut-set bounds for the case of centralized content de-

livery with $L = 1$ and by extension than the bounds presented in this chapter as shown in Fig. 3.6. However, the authors leverage the structure of the bounds to approximate the storage-rate trade-off to within a constant multiplicative factor of 4.7. We note here that the analysis presented in this chapter is solely for the purpose of proving the sub-optimality of cut-set bounds in a more general problem setting, i.e., $L \geq 1$, and that the gap to the optimal can be numerically tightened to 3.5 for centralized delivery with $L = 1$, which shows that the bounds are similar to those in [132, 133] in terms of approximately characterizing the optimal storage-rate trade-off.

Finally, Tian [137] has recently obtained improvements for the specific case of $N = K = 3$ for centralized content delivery with $L = 1$, using a novel computer aided approach as shown in Fig. 3.6(c). Our proposed method recovers the bound $6M+3R^*_{\mathsf{cen}} \geq 8$, while the approach in [133,136] recovers the bound $M + R^*_{\mathsf{cen}} \geq 2$. However, it is unclear whether the bounds $12M + 18R^*_{\mathsf{cen}} \geq 29$ and $3M + 6R^*_{\mathsf{cen}} \geq 8$ can be tractably obtained via analytical methods. Therefore, obtaining the numerical bounds for the $N = K = 3$ system with centralized delivery remains an open problem.

In the next section, we consider the problem of caching and centralized content delivery for a system with heterogeneous cache storage.

3.2 Caching for Heterogeneous Storage

A common underlying assumption in information theoretic modeling of cache-aided systems is that each user is equipped with the same cache storage [97–100, 104–106, 125–128, 133, 134, 136, 140, 176, 179]. In practice, however, different types of devices and users in the system might possess different storage capabilities which motivates the study of a network with heterogeneous cache storage. In this work we introduce the *heterogeneous caching problem*, where each user $k \in \{1, 2, \ldots, K\}$, has a potentially different cache storage of size $M_k B$ bits. For this heterogeneous caching problem, we propose the Layered Heterogeneous Caching (LHC) scheme which has two key ingredients: *set partitioning* and *cache layering*. We first partition the set of K users into disjoint sub-sets, where each sub-set of users is dealt with separately. For each sub-set of users, we propose a layered caching scheme which works as follows: each layer is dedicated to the storage/delivery of a specific fraction of the files, and this fraction is selected based on the level of storage heterogeneity within the users in the sub-set. Our proposed model is a structured approach to exploit maximal multicasting opportunities in a cache-aided system with heterogeneous storage and can be applied to both centralized cache storage (where the storage phase is designed by the central server) and decentralized storage (where users are allowed to randomly cache content). We show that the proposed scheme provides significant improvements over the naive extensions of the homogeneous scheme presented in [97] to the heterogeneous case. Recent parallel work in [180,181] also studied the heterogeneous caching problem but only for the case of decentralized caching. Through numerical results, we show that our scheme can improve on the schemes in [180, 181]. We derive an information theoretic lower bound on the optimal rate of the heterogeneous caching problem and show that the proposed scheme is order optimal for systems with 2 and 3

levels of heterogeneity. We highlight interesting aspects of the impact of heterogeneity on the achievable rate of the proposed scheme such as the reduction of multicasting gain and usefulness of set partitioning with increase in heterogeneity.

3.2.1 Main Contributions

The main contributions of this work as are as follows:

- We consider the *heterogeneous caching problem* where each user in the system has a different cache storage. For this problem, we propose the novel Layered Heterogeneous Caching (LHC) algorithm with two main components namely (i) set partitioning and (ii) cache layering with each partition. The proposed scheme exploits maximal multicasting opportunities for a system with heterogeneous cache storage and can be used for both centralized and decentralized caching.
- We derive an information theoretic lower bound for and leverage it to show that the achievable rate of the proposed LHC scheme is within a constant multiplicative gap of the information theoretic optimal rate for systems with two and three levels of heterogeneity.
- Using numerical simulations, we show that the proposed LHC scheme improves significantly on the naive extensions of homogeneous schemes from [97]. Furthermore, we compare our scheme to recent work in [180, 181] and show that LHC can improve on the schemes in literature for decentralized caching and delivery.

3.2.2 System Model

We consider the same system model from Section 3.1.2 with K users and a library of N files of equal popularity, each of size B bits, for some $B \in \mathbb{N}^+$. However, for the heterogeneous caching system, each user $k \in \{1, \ldots, K\}$ has a cache storage of $M_k B$ bits i.e., the maximum allowable size of each user's cache content Z_k is $M_k B$ such that $H(Z_k) \leq M_k B$. We define an ordered set of K heterogeneous caches $\mathcal{M} := \{M_1, M_2, \ldots, M_K\}$ where $M_1 \leq M_2 \leq \ldots \leq M_K$. Fig. 3.7 illustrates the system model. The key components of the system i.e., *cache storage, file delivery and decoding* are defined similar to Section 3.1.2 for the set of heterogeneous caches $\mathcal{M}$. The fundamental storage-rate trade-off is defined as follows.

Definition 5 (*Storage vs. Rate Trade-off*)**.** The storage-rate pair $(\mathcal{M}, R_{\mathsf{het}})$ is *achievable* if, for any $\epsilon > 0$, there exists an $(\mathcal{M}, R_{\mathsf{het}})$ caching scheme for which $P_e \leq \epsilon$. The optimal storage vs. rate trade-off is defined as:

$$R^*_{\mathsf{het}}(\mathcal{M}, N, K) \triangleq \inf \{R_{\mathsf{het}} : (\mathcal{M}, R_{\mathsf{het}}) \text{ is achievable}\}. \tag{3.44}$$

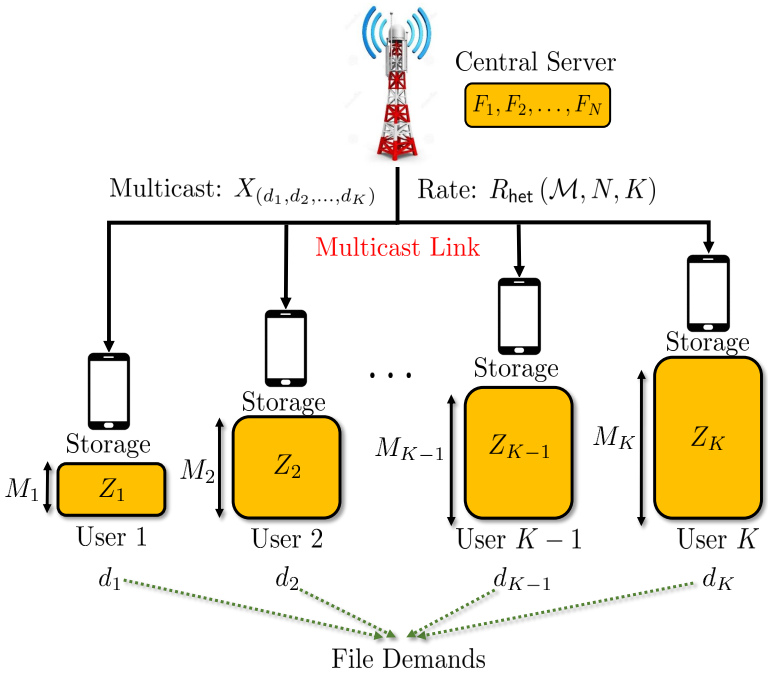


Figure 3.7: System Model for Caching with Heterogeneous Storage.

3.2.3 Existing Results and Preliminaries

In this section, we present the relevant existing results on homogeneous caching and their extensions to the heterogeneous setting. The extensions are naive baselines which are then used to characterize the performance of systems with heterogeneous cache set $\mathcal{M}$.

3.2.3.1 Existing Results for Centralized Homogeneous Caching

For N files and K users with homogeneous cache storage where $M_1 = M_2 = \ldots = M_K = M$ such that $M \in (0, N]$, a novel centralized caching and coded delivery scheme was presented in [97] which yields an achievable rate of

$$R_{\text{hom}}(M, N, K) = \min\left\{R^u_{\text{hom}}(M, N, K), R^m_{\text{hom}}(M, N, K)\right\},$$

which is a minimum of the conventional unicast rate

$$R^u_{\text{hom}}(M, N, K) = \min\{N, K\}\left(1 - \frac{M}{N}\right), \tag{3.45}$$

where each user's demand is served with individual point-to-point transmissions, and the multicast rate

$$R^m_{\text{hom}}(M, N, K) = \frac{1}{t+2}\left[2K - t - \frac{K+1}{t+1} \cdot \frac{KM}{N}\right] \tag{3.46}$$

with $t = \left\lfloor \frac{KM}{N} \right\rfloor$, where multiple users' demands are jointly serviced with multicast transmissions. The caching and delivery scheme proposed in [97] has its roots in the problem of index coding with side information. The multicast delivery achieves a global caching gain which leads to order-wise improvements in the storage-rate trade-off compared to traditional unicast delivery.

3.2.3.2 Naive Extensions to Heterogeneous Setting

We next introduce the preliminaries for heterogeneous caching by extending known homogeneous schemes to the heterogeneous setting.

- *Heterogeneous Unicast*: Let users cache bits of files in a sequential manner i.e., user k caches the first $\frac{M_k B}{N}$ bits from each of the N files. Under this sequential caching, it is enough to deliver the largest complementary fragments of common requested content. Given the set of ordered caches $\mathcal{M}$, the heterogeneous unicast rate is given by:

$$R^u_{\mathsf{het}}(\mathcal{M}, N, K) = \min\{N, K\} - \frac{\sum_{i=1}^{\min\{N,K\}} M_i}{N}. \tag{3.47}$$

- *Heterogeneous Multicast*: Given a set of caches $\mathcal{M}$, the storage and delivery is designed based on the lowest cache storage in the set and the achievable rate in this case is given by

$$R^m_{\mathsf{het}}\left(\mathcal{M}, N, K\right) = R^m_{\mathsf{hom}}\left(\min_{M_i}\{\mathcal{M}\}, N, K\right), \tag{3.48}$$

  where $\min\{\mathcal{M}\}$ is the smallest storage in $\mathcal{M}$ and R^m_{hom} is given in (3.46).

Considering the heterogeneous unicast scheme, although every user's storage is completely utilized, the transmissions are point-to-point and the global caching gain due to multicast is lost. Conversely, the naive multicasting scheme is limited by the lowest storage in the system. Further, as heterogeneity increases, the strategy leads to *cache wastage* i.e., for any user k, $(M_k - \min\{\mathcal{M}\})$ amount of storage is not utilized. The ideal heterogeneous caching and delivery scheme should combine the complete utilization of each users' storage while also leveraging multicasting opportunities. This forms the basis of the proposed *Layered Heterogeneous Caching* (LHC) scheme discussed in the sequel.

3.2.4 Main Results and Discussion

In this section we present new upper and lower bounds on the optimal rate for heterogeneous caching. The following theorem gives our main result, which is an upper bound on the optimal rate, $R^*_{\mathsf{het}}(\mathcal{M}, N, K)$, of the heterogeneous caching problem based on the proposed Layered Heterogeneous Caching (LHC) scheme.

Theorem 10. *For any N files and K users with heterogeneous cache storage $\mathcal{M} := \{M_1, M_2, \ldots, M_K\} \in (0, N]$,*

$$R^*_{\mathsf{het}}(\mathcal{M}, N, K) \leq \min_{\mathcal{G} \in \mathcal{P}_K} \sum_{g \in \mathcal{G}} R_{\mathsf{het}}(\mathcal{M}_g, N, K_g), \tag{3.49}$$

where $\mathcal{P}_K$ is the set of all possible partitions of the set of caches $\mathcal{M}$ and $\mathcal{G} \in \mathcal{P}_K$, with cardinality $G = |\mathcal{G}|$, is any valid partitioning of $\mathcal{M}$ into non-overlapping ordered subsets $\mathcal{M}_g \subseteq \mathcal{M}$ with K_g users for $g \in \{1, 2, \ldots, G\}$. The achievable rate for cache set $\mathcal{M}_g$ is given by

$$R_{\mathsf{het}}(\mathcal{M}_g, N, K_g) = \begin{cases} R^u_{\mathsf{het}}(\mathcal{M}_g, N, K_g), & \text{if } K_g = 1 \\ R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M}_g, N, K_g, \vec{\alpha}^*_g), & \text{if } K_g > 1 \end{cases} \tag{3.50}$$

where $R^{\mathsf{LHC}}_{\mathsf{het}}$ is the rate achieved by the LHC scheme outlined in Algorithm 1 and $\vec{\alpha}^*_g = \left\{\alpha^*_1, \ldots, \alpha^*_{K_g}\right\}$ is the optimal file splitting strategy for any ordered cache set $\mathcal{M}_g$.

The achievable rate in Theorem 10 results from two main concepts namely, (i) *set partitioning* of the ordered cache set $\mathcal{M}$ and (ii) *cache layering* for every ordered subset of heterogeneous caches in a given partition. We next elaborate on these two strategies and outline the proposed LHC scheme detailed in Algorithm 1.

3.2.4.1 Set Partitioning

Set partitioning is used to determine the best grouping of caches in order to maximize the achievable rate of the heterogeneous caching scheme. Let the cache set $\mathcal{M}$ be *partitioned* [182] into a set of disjoint subsets. Let one such partition be $\mathcal{G} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_G\}$ for some integer $G = |\mathcal{G}| \leq K$, such that $\mathcal{M}_g \subseteq \mathcal{M}$ is an ordered subset of K_g users' caches $\forall g \in \{1, 2, \ldots, G\}$. For any valid partitioning $\mathcal{G} \in \mathcal{P}_K$ we also have $\{\mathcal{M}_i \cap \mathcal{M}_j\} = \emptyset$ for $i \neq j$, $\forall i, j \in \{1, 2, \ldots, G\}$ and $\bigcup_{g=1}^{G} \mathcal{M}_g = \mathcal{G}$. Based on such a partitioning, for any subset of caches $\mathcal{M}_g$, if $K_g = 1$, i.e., only one user is present in the group $\mathcal{M}_g$, sequential caching and unicast transmission is used to deliver content to this user. For all other groups $\mathcal{M}_g$ such that $K_g \geq 2$, multicasting opportunities can be exploited. The following example illustrates the concept of partitioning.

Example 3. For a system with N files and $K = 5$ users with cache set $\mathcal{M} := \{M_1, M_2, \ldots, M_5\}$, a valid partition can be $\mathcal{G} = \{\{M_1\}, \{M_2, M_3\}, \{M_4, M_5\}\}$. Here $G = 3$ and the ordered subsets $\mathcal{M}_1 = \{M_1\}$, $\mathcal{M}_2 = \{M_2, M_3\}$, $\mathcal{M}_3 = \{M_4, M_5\}$. Based on this partition, unicast delivery is used for the group $\mathcal{M}_1$, while for $\mathcal{M}_2$ and $\mathcal{M}_3$, the storage and delivery is based on the proposed LHC scheme. □ □

The proposed LHC scheme is a principled approach to exploit all possible multicast and unicast opportunities for a given subset of caches $\mathcal{M}_g$ for any partitioning of $\mathcal{M}$. Note that, when $G =$

Algorithm 1 LAYERED HETEROGENEOUS CACHING

1: INITIALIZE: Split caches in $\mathcal{M}_g$ into K_g layers $\mathcal{L}_{1:K_g}$, such that $\mathsf{m}_\ell = M_\ell - M_{\ell-1}, \ell = 1, 2, \ldots, K_g$. Split each file into fragments $\alpha_1, \alpha_2, \ldots, \alpha_{K_g}$.
2: **for** each layer $\mathcal{L}_\ell$, with $\ell = 1, 2, \ldots, K_g$ **do**
3: CACHE STORAGE: Use centralized cache storage scheme from [97] for N files, $K_g - \ell + 1$ users, each with a cache storage of $\mathsf{m}_\ell B/\alpha_\ell$ bits. At layer $\mathcal{L}_{K_g}$, store m_{K_g}/N fraction of the fragment α_K of each file which have not yet been stored in the cache of user K.
4: FILE DELIVERY:
(a) Deliver $\alpha_\ell B$ bits of requested files of users $\{\ell, \ell+1, \ldots, K_g\}$ via multicast transmission.
(b) Deliver the remaining $\left(1 - \sum_{i=1}^{\ell} \alpha_i\right)$ bits of user ℓ's requested file via unicast.
(c) At the last layer $\mathcal{L}_{K_g}$, deliver the remaining $\left(\alpha_{K_g} - \mathsf{m}_{K_g}/N\right) B$ bits of user K_g's requested file via unicast.
5: **end for**

1, all users are grouped together and served via LHC, while for $G = K$, every user is served individually via unicast. For K users, consider the set of all possible partitions $\mathcal{P}_K$, of the cache set $\mathcal{M}$. The total number of such partitions i.e., the cardinality of $\mathcal{P}_K$ is given by the K-th Bell Number [182]. Minimizing over the achievable unicast and LHC rates for all possible partitions yields an upper bound on the optimal rate in (3.49). The optimal partition $\mathcal{G}^{\mathsf{opt}}$ is then given by

$$\mathcal{G}^{\mathsf{opt}} = \underset{\mathcal{G} \in \mathcal{P}_K}{\arg\min} \sum_{g \in \mathcal{G}} R_{\mathsf{het}}\left(\mathcal{M}_g, N, K_g\right). \tag{3.51}$$

The optimal partition is evaluated based on the application of the LHC scheme within any subset of caches of any given partition. In the next section, we introduce the LHC scheme and analyze the achievable rate.

3.2.4.2 Layered Heterogeneous Caching (LHC) Scheme

In this section, we propose the Layered Heterogeneous Caching (LHC) scheme based on a novel cache layering and file splitting strategy as follows:

- *Cache Layering*: For any partitioning of the cache set $\mathcal{M}$, consider the subset of ordered caches $\mathcal{M}_g$ with K_g users. The caches in $\mathcal{M}_g$ are then divided into K_g *layers*, $\mathcal{L}_1, \ldots, \mathcal{L}_{K_g}$. Layer $\mathcal{L}_1$ consists of all K_g users, each with a storage of $\mathsf{m}_1 = M_1$. Layer $\mathcal{L}_2$ consists of users 2 to K_g (i.e., $K_g - 1$ users) each with a cache storage of $\mathsf{m}_2 = (M_2 - M_1)$. In general, layer $\mathcal{L}_\ell$, $\forall \ell \in \{1, 2, \ldots, K_g\}$ has $K_g - \ell + 1$ users with a per-user storage of $\mathsf{m}_\ell = (M_\ell - M_{\ell-1})$.
- *File Splitting*: Next, each file F_n, is split into K_g non-overlapping fragments of size $(\alpha_1, \alpha_2, \ldots, \alpha_{K_g})B$ bits such that $\sum_{i=1}^{K_g} \alpha_i = 1$. The LHC scheme is based on the premise

that the layer $\mathcal{L}_\ell$ is used to deliver α_ℓ fragment of the files requested by $K_g - \ell + 1$ users via multicast transmission. Resultantly, the ℓ^{th} user receives $\alpha_1 + \alpha_2 + \ldots + \alpha_\ell$ fraction of its requested file via ℓ multicasts. The remaining $\left(1 - \sum_{i=1}^{\ell} \alpha_i\right)$ fraction is delivered via unicast transmission.

The cache layering and file splitting strategies are shown in Fig. 3.8(a) for $K = 3$ users. The overall proposed LHC scheme using these ingredients is presented in Algorithm 1.

Achievable Rate of LHC Scheme

The achievable rate of the LHC scheme for the cache group $\mathcal{M}_g$ is the sum of the rates over the K_g layers. Focusing on the ℓ-th layer, note that the transmission has two components:

1. *Unicast* of $(1 - \sum_{i=1}^{\ell} \alpha_i)$ fraction of the file requested by the ℓ-th user.
2. *Multicast* of α_ℓ fraction of files requested by the set of $(K_g - \ell + 1)$ users, each user having a storge of $\mathsf{m}_\ell = M_\ell - M_{\ell-1}$.

We next separately analyze the unicast and multicast rates for the LHC scheme. To analyze the unicast rate, note that at layer $\mathcal{L}_\ell,\ \forall \ell \in \{1, 2, \ldots, K_g - 1\}$, the unicast rate for user ℓ is given by $1 - \sum_{i=1}^{\ell} \alpha_i = \sum_{i=\ell+1}^{K_g} \alpha_i$. For the final layer $\mathcal{L}_{K_g}$, the unicast rate of the K_g-th user is given by $\left(\alpha_{K_g} - \frac{\mathsf{m}_{K_g}}{N}\right)$. The unicast rate for all K_g layers is given as:

$$\text{Unicast Rate} = \sum_{i=2}^{K_g} (i-1)\alpha_i + \left(\alpha_{K_g} - \frac{\mathsf{m}_{K_g}}{N}\right). \tag{3.52}$$

For multicast delivery, the centralized caching scheme [97] is used in each layer $\mathcal{L}_\ell,\ \forall \ell \in \{1, 2, \ldots, K_g - 1\}$ to deliver α_ℓ fragment of each file. The following lemma gives an achievable *multicast* rate for each layer in the LHC scheme.

Lemma 3. *For any N files and K users, each with cache storage of MB bits, the achievable multicast rate for delivering $\alpha B \in (0, B]$ part of all requested files, is given by:*

$$R(\alpha, M, N, K) = \alpha R^m_{\mathsf{hom}}\left(M/\alpha, N, K\right), \tag{3.53}$$

where $R^m_{\mathsf{hom}}(M, N, K)$ is the achievable rate given in (3.46).

The proof of Lemma 3 is presented in Appendix A.7 along with an illustrative example. Using Lemma 3, the multicast rate for the ℓ-th layer is given by $\alpha_\ell R^m_{\mathsf{hom}}\left(\frac{\mathsf{m}_\ell}{\alpha_\ell}, N, K_g - \ell + 1\right)$. Thus, we have:

$$\text{Multicast Rate} = \sum_{\ell=1}^{K_g - 1} \alpha_\ell R^m_{\mathsf{hom}}\left(\frac{\mathsf{m}_\ell}{\alpha_\ell}, N, K_g - \ell + 1\right). \tag{3.54}$$

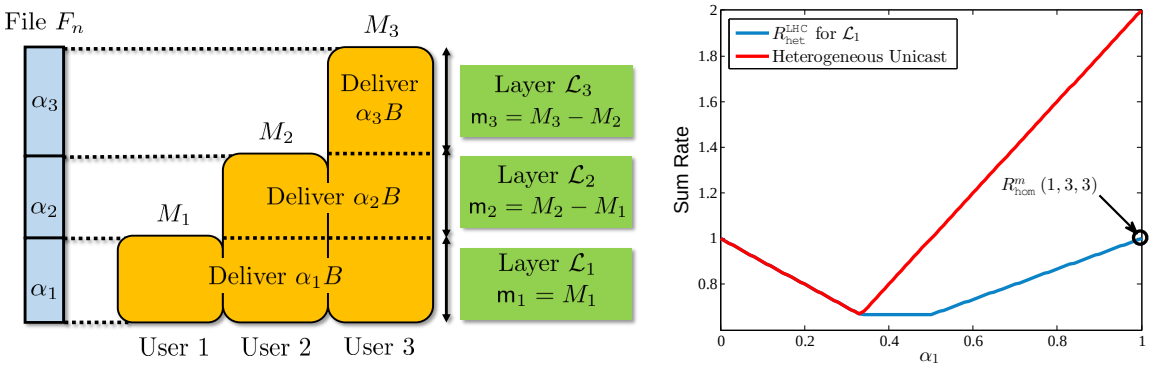


Figure 3.8: (a) Layered Heterogeneous Caching for $K = 3$ users; (b) Scaling of $R_{\mathsf{het}}^{\mathsf{LHC}}$ with α_1 for $N = K = 3$ and $M = 1$.

Thus, for a given $\vec{\alpha}_g = \{\alpha_1, \alpha_2 \ldots, \alpha_{K_g}\}$, the achievable rate of LHC for the cache set $\mathcal{M}_g$ is given by:

$$R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M}_g, N, K_g, \vec{\alpha}_g\right) = \underbrace{\sum_{i=2}^{K_g}(i-1)\alpha_i + \left(\alpha_{K_g} - \frac{\mathsf{m}_{K_g}}{N}\right)}_{\text{Unicast Rate}} + \underbrace{\sum_{\ell=1}^{K_g-1} \alpha_\ell R_{\mathsf{hom}}^{m}\left(\frac{\mathsf{m}_\ell}{\alpha_\ell}, N, K_g - \ell + 1\right)}_{\text{Multicast Rate}}. \tag{3.55}$$

The achievable rate in (3.55) can be minimized by choosing the optimal $\vec{\alpha}_g^*$ as follows:

$$\vec{\alpha}_g^* = \arg\min_{\vec{\alpha}_g} \; R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M}_g, N, K_g, \vec{\alpha}_g\right) \quad \text{s.t.} \quad \sum_{i=1}^{K_g} \alpha_i = 1, \; 0 \leq \alpha \leq 1. \tag{3.56}$$

The solution yields the best achievable LHC rate $R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M}_g, N, K_g, \vec{\alpha}_g^*\right)$ for the cache set $\mathcal{M}_g$. The rate, when evaluated over all subsets $\mathcal{M}_g$ such that $g \in \mathcal{G}^{\mathsf{opt}}$ i.e., over the optimal partitioning of $\mathcal{M}$, yields the upper bound on the optimal rate for heterogeneous caching in (3.50).

Remark 7 (*Optimal File Splitting*)**.** Fig. 3.8(b) illustrates the intuition behind the choice of optimal α for any given layer in the LHC. Consider the system in Fig. 3.8(a) for $N = 3$ files and a storage of $M_1 = 1$ in layer $\mathcal{L}_1$. Fig. 3.8(b) shows the heterogeneous unicast and LHC rates as a function of the split α_1. The rates are minimized at $\alpha_1 = 1/3$. However, a larger fraction, $\alpha_1 = 0.5$, can be delivered in this layer without increasing the sum-rate when LHC is used. This ensures maximum utilization of storage at layer $\mathcal{L}_1$ and transmission with the minimum possible rate: $\frac{1}{6} + (1 - \alpha_1) = 0.67$ (see Appendix A.7). For a homogeneous system, with $M_1{=}M_2{=}M_3{=}1$, however, $\alpha_1 = 1$ is the optimal choice since the entire file needs to be delivered using multicast in layer $\mathcal{L}_1$. □

Remark 8 (*Decentralized Layered Heterogeneous Caching*)**.** In the proposed LHC scheme, within each layer, the cache storage phase is centralized. However, due to the generality of the layering approach, a *decentralized cache storage* scheme based on random caching [98], can be used within each set of caches $\mathcal{M}_g$. In this case, each user $k \in \{1, 2, \ldots, K_g\}$ can randomly store any M_k/N bits in its cache. The delivery phase is then designed by the server similar to the scheme presented in [98]. To this end, the server layers the caches into layers $\mathcal{L}_1, \ldots, \mathcal{L}_{K_g}$. It then divides the cache content of each user $k \in \{1, \ldots, K_g\}$ into k non-overlapping fragments of size $\mathsf{m}_1, \mathsf{m}_2, \ldots, \mathsf{m}_k$ such that $\sum_{\ell=1:k} \mathsf{m}_\ell = M_k$ and each cache fragment m_ℓ corresponds to a layer $\mathcal{L}_\ell$. The server uses the multicast delivery scheme in [98] for each layer and optimizes over $\vec{\alpha}$ to determine the fraction of files to deliver in each layer. The achievable rate of LHC in (3.55) is then based on the rate of decentralized delivery [98] for homogeneous caching within each layer with

$$R^m_{\mathsf{hom}}(M, N, K) = K\left(1 - \frac{M}{N}\right)\left\{\frac{N}{KM} \cdot \left(1 - \left(1 - \frac{M}{N}\right)^K\right)\right\}. \tag{3.57}$$

for N files, K users and any cache size $M \in (0, N]$. Note that Lemma 3 also holds for this case and hence the achievable rate of decentralized LHC for any subset of caches $\mathcal{M}_g$ can be evaluated from (3.55) by using the decentralized multicast rate R^u_{hom} from (3.57). □

3.2.4.3 Information Theoretic Lower Bound

We next present an information theoretic lower bound on the optimal rate of the heterogeneous caching problem and leverage the bound to show that the proposed LHC scheme is order optimal for some system settings. The next theorem presents q lower bound on the optimal rate for the heterogeneous caching problem.

Theorem 11. *For any N files and K users with heterogeneous cache storage $\mathcal{M}$, we have*

$$R^*_{\mathsf{het}}(\mathcal{M}, N, K) \geq \max_{\substack{s \in 1, \ldots, \min\{N,K\} \\ M_i \in \mathcal{M}, \forall i \in [s]}} \left(\frac{N - \sum_{i=1}^s M_i}{\lceil N/s \rceil}\right). \tag{3.58}$$

The proof of Theorem 11 is presented in Appendix A.8 and follows from a cut-set argument [174].

3.2.4.4 Order-Optimality of LHC

Leveraging the achievable rate in Theorem 10 and the lower bound in Theorem 11, the next theorem characterizes the optimal rate of the heterogeneous caching from to within a constant multiplicative factor for systems with 2 and 3 distinct cache sizes across all users.

Theorem 12. *For any N files and K users with heterogeneous cache storage $\mathcal{M} := \{M_1, M_2, \ldots, M_K\}$ such that the system has only two distinct cache sizes i.e., $M_i \in$*

$\{M_L, M_H\}$, $\forall i \in \{1, 2, \ldots, K\}$, *we have*

$$\frac{R_{\mathsf{het}}(\mathcal{M}, N, K)}{R^*_{\mathsf{het}}(\mathcal{M}, N, K)} \leq 19; \tag{3.59}$$

and for a system where there are only three distinct cache sizes i.e., $M_i \in \{M_L, M_I, M_H\}$, $\forall i \in \{1, 2, \ldots, K\}$, *we have*

$$\frac{R_{\mathsf{het}}(\mathcal{M}, N, K)}{R^*_{\mathsf{het}}(\mathcal{M}, N, K)} \leq 28. \tag{3.60}$$

The proof of Theorem 12 is presented in Appendix A.9. The theorem shows that for $2, 3$ levels of heterogeneity, the proposed heterogeneous caching scheme is order-optimal.

Bi-Criteria Approximation

For a general system with K users, each having a potentially different different cache size, we present a bi-criteria approximation result for the LHC algorithm. To this end, let $\overline{\mathcal{M}}_g$ denote the mean cache storage for the set of caches in $\mathcal{M}_g$. Assume that a new cache set $\mathfrak{M}_g$ is formed by replacing all caches in $\mathcal{M}_g$ which are smaller than $\overline{\mathcal{M}}_g$ by the mean cache size. For any user $k \in \{1, 2, \ldots, K_g\}$ with cache storage $M_k \in \mathcal{M}_g$, we define the ratio of the additional cache storage in $\mathfrak{M}_g$, to the mean cache in each group as $\gamma_k = \left(\frac{\overline{\mathcal{M}}_g - M_k}{\overline{\mathcal{M}}_g}\right)^+$, where the function $(x)^+ = \max\{0, x\}$ ensures that the ratio is non-zero for only the caches which have storage lower than the mean. Assume further, that in for each subset of caches $g \in \{1, 2, \ldots, G\}$, of the partition $\mathcal{G}$, a total of $\overline{K}_g (\leq K_g)$ users have cache storage less than the mean $\overline{\mathcal{M}}_g$. The following theorem provides a bi-criteria approximation guarantee for the proposed algorithm.

Theorem 13. *For any N files and K users with heterogeneous cache storage $\mathcal{M} := \{M_1, M_2, \ldots, M_K\} \in (0, N]$,*

$$R^*_{\mathsf{het}}(\mathcal{M}, N, K) \geq \frac{1}{12G} \sum_{g=1}^{G} R_{\mathsf{het}}\left(\mathfrak{M}_g, N, K_g\right), \tag{3.61}$$

where $R_{\mathsf{het}}\left(\mathfrak{M}_g, N, K_g\right)$ is the rate achieved by the proposed LHC algorithm for a cache set $\mathfrak{M}_g$, where caches sizes lower than $\overline{\mathcal{M}}_g$ have been replaced by the mean. This is a $\left(12G, \left[1 + \frac{G}{K} \sum_{g=1}^{G} \sum_{k=1}^{\overline{K}_g} \gamma_k\right]\right)$ − bi-criteria approximation for the LHC algorithm.

The proof of Theorem 13 is provided in Appendix A.10.

Remark 9. Bi-criteria approximations are common in caching literature, particularly [99, Theorem 3], which provides a bi-criteria approximation algorithm for the problem of caching with

non-uniform file popularities. In our approximation guarantee, it is to be noted that for a given ordered cache set $\mathcal{M}$, the constant gap to the achievable rate increases when the number of subsets G in the partition $\mathcal{G}$ is large. But in general, we expect $G \ll K$ thus the constant does not grow arbitrarily large. On the other hand, when large number of groups are chosen, it is expected that the mean cache sizes of the groups are closer to their minimum cache sizes compared to the case when $G = 1$ i.e., all users are grouped together. Thus, for any heterogeneous cache set $\mathcal{M}$ and large G, the values of $\gamma_k,\ \forall k \in \{1, 2, \ldots, K_m\}$ are smaller i.e., smaller cache violations are incurred. Thus choosing an appropriate partition $\mathcal{G} \in \mathcal{P}_K$ presents a trade-off for the bi-criteria approximation. □

We next present numerical results to analyze the performance of the proposed LHC scheme under different levels of system heterogeneity.

3.2.5 Illustration of Numerical Results

To tractably characterize the system heterogeneity and present numerical results, we define the following parameters for heterogeneity. The heterogeneity of a set of ordered caches $\mathcal{M}$ is characterized by $\vec{\eta} = [\eta_1, \eta_2, \ldots, \eta_K]$ such that $\eta_k = M_k/M_1, \forall k \in \{1, 2, \ldots, K\}$, and $\beta = M_1/N$. Furthermore, we also consider an exponential model for heterogeneity similar to [181] where we have

$$M_k = \frac{\gamma^{K-k}}{\gamma^{K-1}} M_1, \ \ \gamma \in (0, 1]. \tag{3.62}$$

The parameter γ characterizes the heterogeneity of the system with $\gamma = 1$ being a homogeneous caching system. For the exponential model, we have $\eta_k = (\gamma^{K-k})/(\gamma^{K-1})$. To characterize the storage-rate trade-off, the sum-rate of the system is plotted against increasing β for a given heterogeneity $\vec{\eta}$ for different γ.

3.2.5.1 Achievable Rate of LHC for Different Heterogeneity

First, we characterize the effect of heterogeneity on the achievable rate for a system with the same aggregate cache size but with the caches distributed with different heterogeneity. Fig. 3.9 shows that the achievable rate of the LHC scheme with $N = K = 5$ for the optimal partition $\mathcal{G}^{\mathsf{opt}}$. It can be seen that for the same aggregate cache size, as system heterogeneity increases, the achievable rate of the LHC scheme increases. This can be attributed to the fact that the increase in heterogeneity leads to decrease in available multicasting opportunities and the unicast rate in (3.55) begins to dominate the achievable rate.

Remark 10 (*Homogeneous Caching*)**.** LHC is modeled on cache layering and file-splitting with each layer operating independently. In the case that multiple users $k, k+1, \ldots, k^{'}$ have the same storage, the layers corresponding to users $k+1, k+2, \ldots, k^{'}$ have zero storage and hence

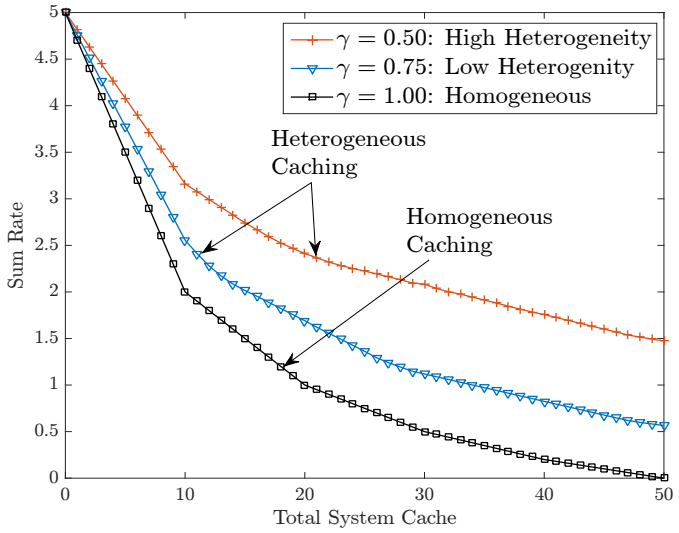
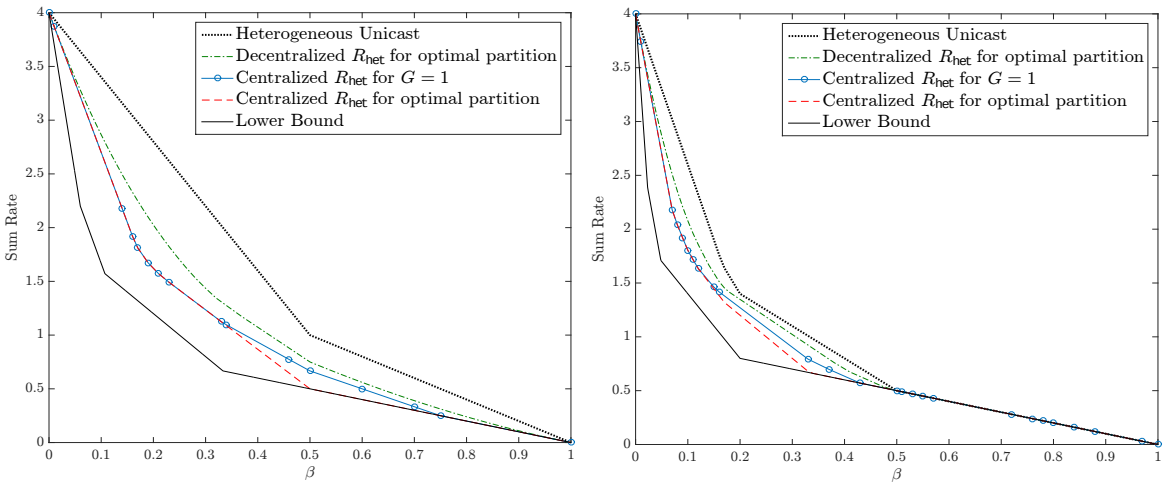


Figure 3.9: $R_{\mathsf{het}}(\mathcal{M})$ trade-off for $N = K = 5$ under varying system heterogeneity.

$\alpha_{k+1}, \alpha_{k+2}, \ldots, \alpha_{k'} = 0$. Thus LHC, when applied to a system with homogeneous caching, uses only one layer with the entire cache storage and all users, to deliver requested files yielding the multicast rate in [97]. All layers other than $\mathcal{L}_1$ in this case have zero storage and $\alpha_1 = 1$. The optimal partition, $\mathcal{G}^{\mathsf{opt}}$, is either the entire set $\{1, 2, \ldots, K\}$ (multicast to all users) with $G = 1$ or the partition with $G = K$ subsets with one user in each subset (unicast to all users) depending on the values of the parameters M, N, K. □

Fig. 3.10(a)-3.10(b) show the achievable rate for heterogeneous unicast along with centralized and decentralized LHC schemes for $N = 10$ files and $K = 4$ users with different levels of heterogeneity. Fig. 3.10(b) shows that, as heterogeneity increases the LHC rate approaches the unicast rate. This is due to the fact that with increase in heterogeneity, layers with more users i.e., with more multicasting opportunities, have low cache storage leading to lesser multicast gains. Therefore the first term in LHC rate in (3.55) dominates. Furthermore, decentralized caching faces a rate loss due random storage which also decreases with heterogeneity due to lack of multicast gains. Fig. 3.10(a)-3.10(b) also show that the rate of Theorem 10 converges to the lower bound of Theorem 11 for large β.

Remark 11 (*Cache Partitioning*)**.** For low heterogeneity, using a single partition $\{1, 2, \ldots, K\}$ with $G = 1$, provides similar performance compared to optimal partitioning with $G \geq 1$. However with increasing heterogeneity, partitioning better captures the disparity in cache sizes to provide maximal multicasting gains by grouping users together optimally. □



(a) (b)

Figure 3.10: $R_{\text{het}}(\mathcal{M})$ trade-off for (a) $N = 10, K = 4$, with $\vec{\eta} = [1\ 1\ 2\ 2]$; (b) $N = 10, K = 4$, with $\vec{\eta} = [1\ 2\ 5\ 6]$.

3.2.5.2 Comparison with Decentralized Schemes from Literature

Next, we compare the achievable rates of the proposed LHC for the case of decentralized caching with results from independent parallel work [180]. First, we consider the case when $N < K$. We can from Fig. 3.11(a)-3.11(b) that for this case, the proposed LHC scheme outperforms the scheme from [180] especially for low values of cache storage. An interesting insight is that optimal partitioning outperforms all schemes for the entire regime of cache storage. However, when using only $G = 1$ i.e., grouping all users together, the LHC scheme loses multicasting opportunities and therefore is outperformed marginally by the scheme from [180]. For the case of $N > K$, Fig. 3.11(c)-3.11(d) show that the proposed LHC scheme with optimal partitioning outperforms other schemes. We would however like to acknowledge that the modified decentralized caching scheme proposed in [181] does achieve an improvement over LHC for the case of decentralized caching with $N < K$ for small values of cache storage.

Next, we consider the case when N, K is large. In this case, evaluating the optimal partition for the LHC scheme is computationally complex since the Bell number grows double exponentially with K. Instead we resort to using $G = 1$ for the LHC scheme. It can be seen from Fig. 3.12(a) that the decentralized LHC is marginally outperformed by the scheme from [180] for $\gamma = 0.75$, while the centralized LHC outperforms all other schemes. Note also that when heterogeneity decreases, Fig. 3.12(b) shows that the decentralized LHC has similar performance to the scheme in [180] while the loss due random caching is also higher. This can be attributed to the fact that centralized placement and delivery can exploit better multicasting opportunities when the caches are similar in size. Furthermore, in these simulations, we have only used $G = 1$. We can always evaluate the LHC rate over more partitions (but less than the optimal Bell Number of partitions) to minimize the

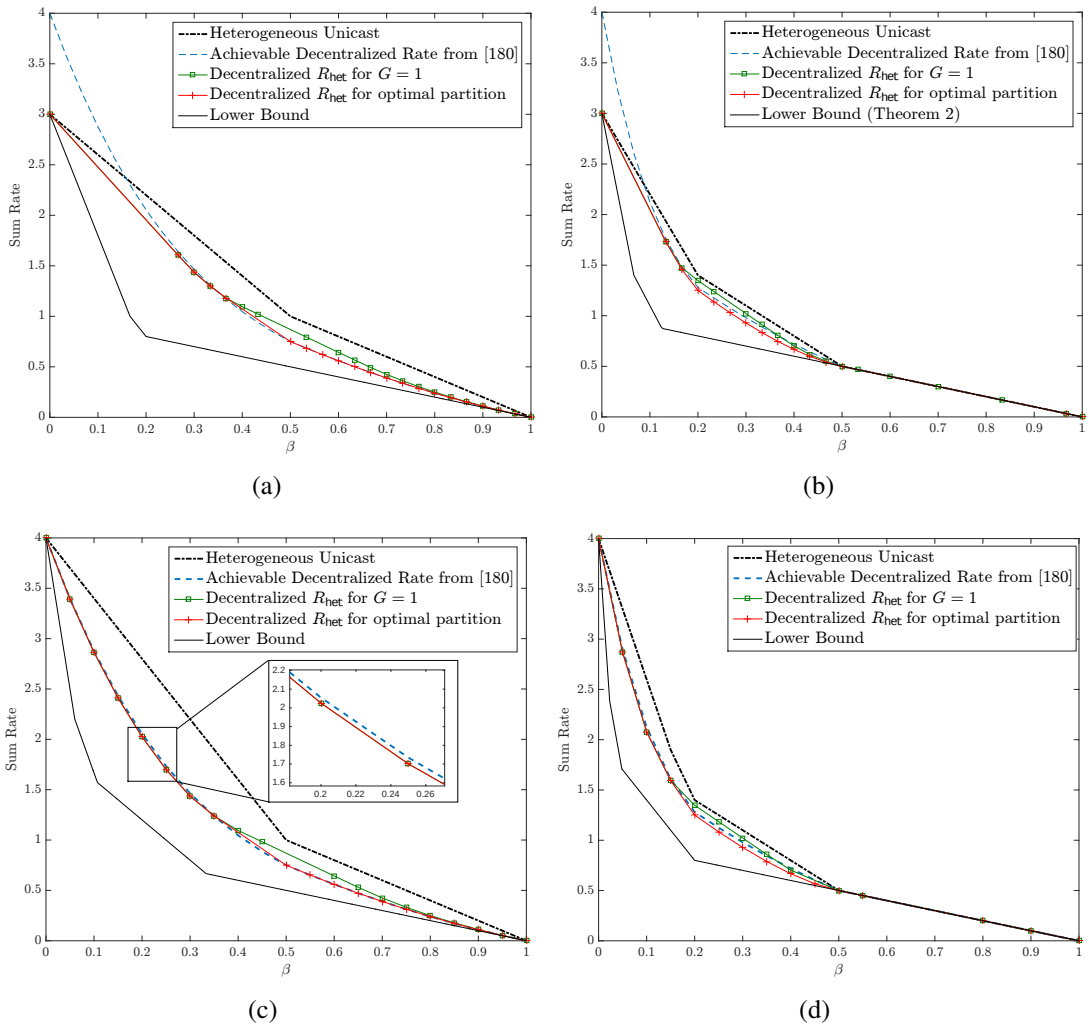


Figure 3.11: $R_{\text{het}}(\mathcal{M})$ trade-off for Decentralized LHC with (a) $N = 3,\ K = 4$ and $\vec{\eta} = [1\ 1\ 2\ 2]$ (a) $N = 3,\ K = 4$ and $\vec{\eta} = [1\ 2\ 5\ 6]$ (c) $N = 10, K = 4$, with $\vec{\eta} = [1\ 1\ 2\ 2]$; (d) $N = 10, K = 4$, with $\vec{\eta} = [1\ 2\ 5\ 6]$.

LHC rate and potentially further improve over the existing schemes in literature. Note that LHC is a general framework allowing for both centralized and decentralized storage while the schemes in [180, 181] works only for decentralized storage.

In the next section we give a summary of our contribution towards gaining a better understanding of the fundamental limits of single-server cache-aided systems and highlight some directions of future work in this paradigm.

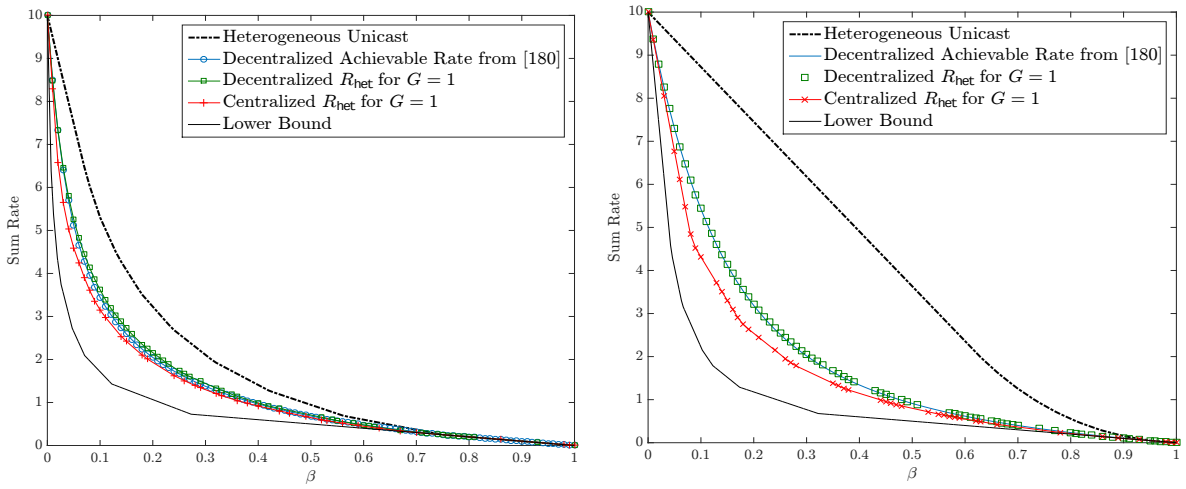


Figure 3.12: $R_{\text{het}}(\mathcal{M})$ trade-off for LHC with $N = 50,\ K = 10$ and (a) $\gamma = 0.75$ (b) $\gamma = 0.96$.

3.3 Directions for Future Research

The study of information theoretic modeling of cache-aided systems has been a topic of a vast body of recent research which aims to gain fundamental understanding of the practical gains afforded by such systems. Based on the work presented in this chapter, some open problems and directions of new research in this paradigm are highlighted next:

- *Optimal Storage-Rate trade-off For Homogeneous Coded Caching*: For cache-aided systems without any constraints on storage, the characterization of the optimal storage-rate trade-off is still an open problem.

- *Non-linear Caching and Delivery Schemes*: Based on recent results, including those presented in this chapter, it is envisioned that to achieve optimality, tighter upper bounds i.e., improved achievable schemes are required. To this end, we note that most of the schemes existing in literature are linear placement and delivery schemes. Non-linear and coded cache placement and delivery schemes might be able to improve on known results and remain an area of active research.

- *Optimal Storage-Rate Trade-off for Heterogeneous Caching*: For the problem of caching with heterogeneous storage, based on the proposed framework, an interesting direction of future work is to design a low-complexity (polynomial time) algorithm for finding or approximating the optimal cache partition $\mathcal{G}^{opt}$. Furthermore, verifying the order-optimality of LHC for any K levels of heterogeneity remains an open problem.

3.4 Summary

In this chapter, we investigated the fundamental information theoretic limits of single-server cache-aided systems and presented improved approximations to the storage-rate trade-off for systems with homogeneous as well heterogeneous cache storage. To this end, we presented a new technique for deriving information theoretic lower bounds for single-server cache-aided systems with centralized as well as D2D-assisted content delivery for the general case when users can demand multiple files at each transmission interval. We leveraged Han's Inequality to better model the interaction of user caches and file decoding capabilities of multicast transmissions to derive lower bounds which are strictly tighter than existing cut-set based bounds. Leveraging the proposed lower bounds, we showed that, for the case of multiple demands per user, treating each set of user demands independently for multicast content delivery, is in fact order-optimal for both delivery settings. Furthermore, we provided an approximate characterization of the fundamental storage-rate trade-off for centralized content delivery to within a constant multiplicative factor of 11 and for D2D-assisted content delivery to within a factor of 10 for all possible values of problem parameters, thereby improving on the existing results in both paradigms. Furthermore, we introduced the heterogeneous caching problem and presented a novel achievable scheme based on set partitioning and cache layering. The proposed Layered Heterogeneous Caching scheme utilizes the system storage maximally and delivers content via a combination of multicast and unicast delivery. The framework is general and can be applied to both centralized and decentralized cache placement. We also presented an information theoretic lower bound on the optimal heterogeneous caching rate. We showed that as heterogeneity in storage increases, the multicasting gain reduces and usefulness of optimal user-partitioning increases. Furthermore, leveraging the lower bound, we showed that the rate of LHC scheme is within a constant multiplicative gap to the information theoretic optimal rate for systems with 2 and 3 levels of heterogeneity respectively.

Chapter 4

Fundamental Limits of Caching with Secure Delivery

In this chapter, we introduce the *secure caching problem* with the goal of minimizing information leakage to an external wiretapper in a single-server cache-aided system while servicing the legitimate users with the minimum possible rate. Firstly, the fundamental cache storage vs. transmission rate trade-off of the secure caching problem is characterized for the case of uniform file popularity. Rather surprisingly, the results show that security can be introduced at a negligible cost, particularly for large number of files and users. It is also shown that the rate achieved by the proposed caching scheme with secure delivery is within a constant multiplicative factor from the information-theoretic lower bound for most parameter values of practical interest. These results are then extended to the case of files with non-uniform popularity profile.

4.1 Caching with Secure Delivery

In the information theoretic formulation of the caching problem in [97, 98], parts of popular files are often shared across caches based on the available per-user cache storage in the system. Further, the content delivery phase also treats groups of users together in order to deliver content via multicast transmissions. As a result, security is a concern in such systems wherein unauthorized access needs to be limited. Furthermore, an external adversary (rogue subscriber) may be able to obtain access to sensitive (unauthorized) data by intercepting enough multicast and unicast transmissions. In this work, we investigate the fundamental *security* aspects of the caching problem in the presence of an external adversary (wiretapper). To this end, we introduce the *secure caching problem* in which the multicast communication between the central server and the users (delivery phase) occurs over a *public (insecure) channel*. The defining feature of this problem is to capture the trade-off between the multicast rate of the insecure link and the size of the cache storage. To the best of our knowledge, none of the works on cache storage and placement design deal with

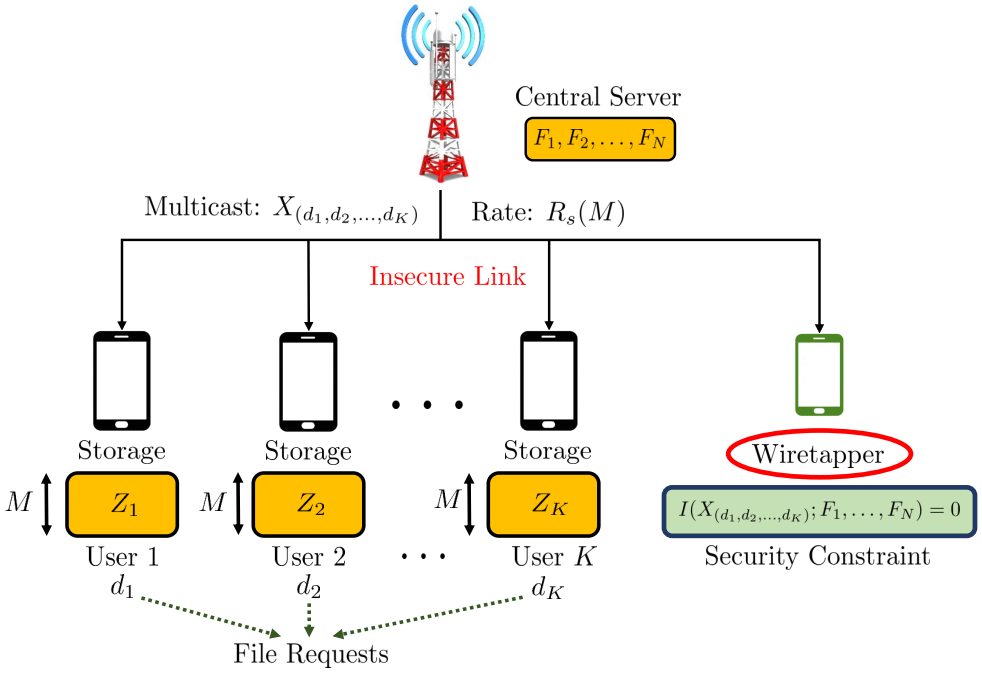


Figure 4.1: System Model for Secure Caching.

security issues.

We consider the coded caching problem of Maddah-Ali and Niesen (introduced in Chapter 2), in the presence of an external wiretapper which can observe the multicast communication $X_{(d_1,\ldots,d_K)}$ i.e., the communication from the central server to the users occurs over an *insecure* link. The wiretapper is considered to be strictly out-of-network and is thus able to observe only the multicast delivery which happens over a broadcast channel. Thus, besides satisfying the users' demands, we require that $X_{(d_1,\ldots,d_K)}$ must not reveal any information about $(F_1, \ldots, F_N)$ i.e., $I\left(X_{(d_1,\ldots,d_K)}; F_1, \ldots, F_N\right) = 0$. As is shown, the additional security constraint necessitates introducing randomness in the form of keys, which occupy a part of the cache of each user. Subsequently, these keys are used in the delivery phase to the keep the delivery information theoretically secure using a one-time-pad scheme [183]. In our system model, the placement phase occurs over unicast channels to individual users and can be secured with the help of individual keys e.g., secure unicast communications using a system similar to code-division-multiple-access (CDMA). As a result, security is considered to be inherent in the placement phase. Thus, in this work, we consider the security of only the *delivery phase* and not the cache *placement phase*. For this problem, a storage-rate pair (M, R_s) is *securely achievable* if, for a cache size of MF and a transmission of rate R_sF bits, it is possible for each user to decode its requested file and the communication over the shared link reveals no information about any file. Fig. 4.1 shows the caching system in the presence of a wiretapper. Let $R_s^*(M)$ denote the smallest R_s such that (M, R_s) is achievable. Thus, the function $R_s^*(M)$ is the fundamental storage-rate trade-off for the *secure* caching prob-

lem. We investigate both the centralized cache placement as well as the decentralized placement with secure file delivery without any assumptions on user demands and file popularity.

4.1.1 Main Contributions

The aim of this work is an approximate characterization of the storage-rate trade-off for caching with secure delivery in the presence of an external wiretapper. The main contribution of this chapter are summarized as follows:

- We design centralized and decentralized caching algorithms which make use of coded multicast delivery to extract global caching gain. The system has uniformly distributed orthogonal keys which are stored across users for secure multicast delivery.
- We present novel upper and lower bounds on $R_s^*(M)$ and show that these bounds are within a constant multiplicative gap. Indeed, for a fixed M, it is intuitively clear that $R_s^*(M) \geq R^*(M)$, i.e., the minimum rate in presence of a wiretapper must be, in general, larger than in the absence of a wiretapper. From our results, we show, rather surprisingly, that the cost for incorporating security in both the centralized and decentralized caching schemes is negligible when the number of users and files are large.
- Finally, we also present an approximate characterization of secure storage-rate trade-off for the case of non-uniform file popularity and show that the cost of security for this setting is also negligible for a large number of files and users.

4.2 System Model

Let $(F_1, F_2, \ldots, F_N)$ be N independent random variables each uniformly distributed over

$$F_n \sim \text{Unif}\{1, 2, \ldots, 2^F\},\ \forall n \in \{1, 2, \ldots, N\} \tag{4.1}$$

for some $B \in \mathbb{N}$. Each F_n represents a file of size F bits. A (M, R_s) secure caching scheme comprises of K *random* caching functions, N^K *random* encoding functions and KN^K decoding functions. The K *random* caching functions map the files $(F_1, \ldots, F_N)$ into the cache content:

$$Z_k \triangleq \phi_k\big(F_1, \ldots, F_N\big) \tag{4.2}$$

for each user $k \in \{1, 2, \ldots, K\}$ during the storage (or placement) phase. The maximum allowable size of the contents of each cache Z_k is MB bits. The N^K *random* encoding functions map the files $(F_1, \ldots, F_N)$ to the input

$$X_{(d_1,\ldots,d_K)} \triangleq \psi_{(d_1,\ldots,d_K)}\big(F_1, \ldots, F_N\big) \tag{4.3}$$

of the shared link in response to the requests $(d_1, \ldots, d_K) \in \{1, 2, \ldots, N\}^K$ during the delivery phase. Finally, the KN^K decoding functions map the received signal over the *insecure* shared link $X_{(d_1,\ldots,d_K)}$ and the cache content Z_k to the estimate

$$\hat{W}_{(d_1,\ldots,d_K),k} \triangleq \mu_{(d_1,\ldots,d_K),k}\Big(X_{(d_1,\ldots,d_K)}, Z_k\Big) \tag{4.4}$$

of the requested file F_{d_k} for user $k \in \{1, 2, \ldots, K\}$. The probability of error is defined as:

$$P_e \triangleq \max_{(d_1,\ldots,d_K)\in\{1,2,\ldots,N\}^K} \max_{k\in\{1,2,\ldots,K\}} \mathbb{P}(\hat{W}_{(d_1,\ldots,d_K),k} \neq F_{d_k}). \tag{4.5}$$

The information leaked at the wiretapper is defined as:

$$L \triangleq \max_{(d_1,\ldots,d_K)\in\{1,2,\ldots,N\}^K} I\left(X_{(d_1,\ldots,d_K)}; F_1, \ldots, F_N\right). \tag{4.6}$$

Definition 6 (*Secure Storage vs. Rate Trade-off*)**.** The pair (M, R_s) is *securely achievable* if for any $\epsilon > 0$ and every large enough file size B, there exists a (M, R_s) secure caching scheme with $P_e \leq \epsilon$ and $L \leq \epsilon$. We define the secure storage-rate trade-off

$$R_s^*(M) \triangleq \inf\{R_s : (M, R_s) \textit{ is securely achievable}\}. \tag{4.7}$$

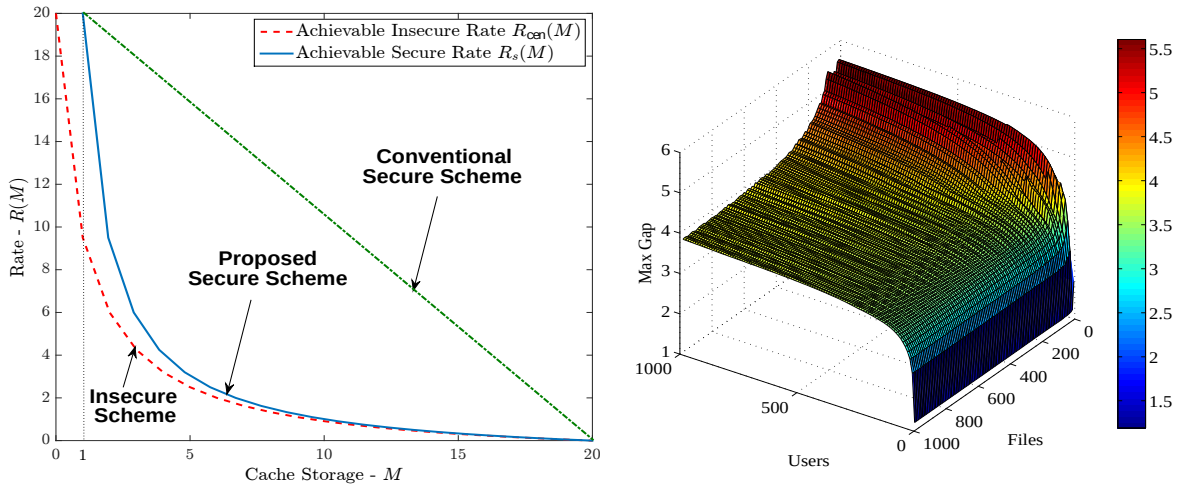
4.3 Centralized Caching with Secure Delivery

The first result gives an achievable rate which upper bounds the optimal storage-rate trade-off $R_s^*(M)$ for the centralized caching scheme with secure delivery. Security is incorporated by introducing randomness in the storage and delivery phase of the achievable scheme in form of a set of uniformly distributed orthogonal keys (independent of the data) stored in the cache of each user. The total cache storage (of size MB bits) is divided into two parts - data storage (of size M_DB bits) and key storage (of size M_KB bits) such that $M = M_D + M_K$. The server uses the keys stored at the users' caches to encode the delivery signal $X_{(d_1,\ldots,d_K)}$ such that the transmission is secure from the wiretapper.

Theorem 14. *For N files and K users, each with a cache size of $M \in \frac{(N-1)}{K} \cdot t + 1$, for $t \in \{0, 1, 2, \ldots, K\}$ we have*

$$R_s^*(M) \leq R_{s,\mathsf{cen}}(M) \triangleq K \cdot \left(1 - \frac{M-1}{N-1}\right) \left\{\frac{1}{1 + K \cdot \frac{M-1}{N-1}}\right\} \tag{4.8}$$

i.e., the rate $R_{s,\mathsf{cen}}(M)$ is securely achievable. For any $1 \leq M \leq N$, the lower convex envelope of these points is achievable.



(a) (b)

Figure 4.2: (a) Centralized Secure vs Non-Secure Bounds $N = K = 20$; (b) Multiplicative gap between $R_s^C(M)$ and lower bound on $R_s^*(M)$.

The algorithm achieving the rate in Theorem 14 is presented in Algorithm 2. Similar to [97], the achievable rate in (4.8) consists of three factors. The first factor K is the worst case rate in the case when no data is cached $(M_D = 0)$. The second factor in (4.8) is $\left(1 - \frac{M-1}{N-1}\right)$. This is the *secure local caching gain* and is relevant whenever M is of the order of N. The third factor in (4.8) is $1/\left(1 + K \cdot \frac{M-1}{N-1}\right)$, which is the *secure global caching gain*. Comparing Theorem 14 to (Th.1, [97]), we observe that the terms $\frac{M}{N}$ in (Th.1, [97]) have been replaced by $\frac{M-1}{N-1}$. However, the combination of the global and local gains leads to the rate in (4.8) being higher than the rate in (Th.1, [97]) for a given value of M, N. This is the cost paid for the security in the system. However, as K, N become large, the secure rate is asymptotically equal to the non-secure case. When $N = K = 20$, it can be seen from Fig. 4.2(a) that the secure and non-secure bounds almost coincide i.e., security from a wiretapper can be achieved at *almost negligible cost* for a large number of files and users. Consider the case of conventional unicast content delivery to each user. In contrast to the insecure scheme in [97], to make the delivery phase secure, however, each user has to store a unique key (of the same size as a single file). During delivery, the server encodes the user's requested file with its key and transmits it. Thus, even with no data storage in cache, the cache size has to be at least B bits to store a key $(M_K = 1)$ i.e., in the secure problem, $M = 0$ is infeasible. The worst case rate is achieved at $M = 1$ and the $(M, R_{s,\text{cen}})$ pair $(1, K)$ is achievable. At the other extreme when $M = N$ i.e., the case where all files are stored in the user's cache and no content delivery is required. In this case $M_D = N, M_K = 0$ and the $(M, R_{s,\text{cen}})$ pair $(N, 0)$ is achievable. We refer to a scheme which achieves points on the line joining $(1, K)$ and $(N, 0)$ as the *conventional secure scheme*, where each user stores one unique key and encrypted files are unicast to each user based on their request. On the other hand, the proposed scheme in Algorithm 2 jointly designs the placement of data and keys in the users' caches such that *coded secure multicasting* can be achieved among users. Next, we present a lower bound on $R_s^*(M)$ stated in the following

Algorithm 2 SECURE CENTRALIZED CACHING ALGORITHM

Centralized Cache Placement: for files $F_1, \ldots, F_N$
$t = K(M-1)/(N-1)$
for $n \in \{1, 2, \ldots, N\}$ **do**
    Split file F_n into equal sized fragments $F_{n,\mathcal{T}} : \mathcal{T} \subseteq \{1, 2, \ldots, K\}, |\mathcal{T}| = t$
end for
Generate keys $\mathcal{K}_{\mathcal{T}_k}$ such that $\mathcal{T}_k \subseteq \{1, 2, \ldots, K\}, |\mathcal{T}_k| = t + 1$
for $k \in \{1, 2, \ldots, K\}$ **do**
    for $n = 1, 2, \ldots, N$ **do**
        File $F_{n,\mathcal{T}}$ is place in cache, Z_k, of user k if $k \in \mathcal{T}$
        Key $\mathcal{K}_{\mathcal{T}_k}$ is placed in cache, Z_k, of user k if $k \in \mathcal{T}_k$
    end for
end for
Coded Delivery:
for $\mathcal{S}$ such that $\mathcal{S} \subseteq \{1, 2, \ldots, K\}, |\mathcal{S}| = t + 1$ **do**
    Server sends $\left\{\mathcal{K}_{\mathcal{S}} \oplus_{k \in \mathcal{S}} F_{d_k, \mathcal{S}\setminus\{k\}}\right\}$
end for

theorem.

Theorem 15. *For N files and K users, each having a cache size $1 \leq M \leq N$,*

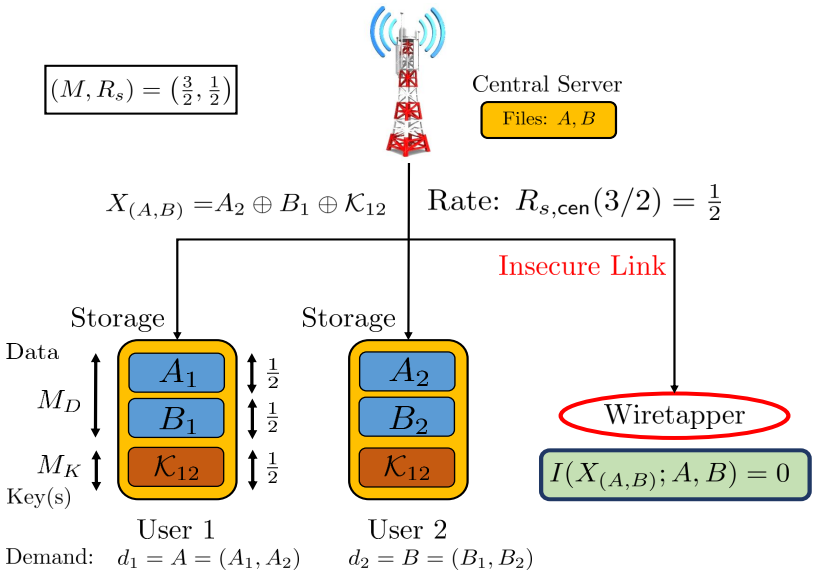
$$R_s^*(M) \geq \max_{s \in \{1,\ldots,\min\{N,K\}\}} \left(s - \frac{s(M-1)}{\left(\lfloor \frac{N}{s} \rfloor - 1\right)} \right). \tag{4.9}$$

The proof of Theorem 15 is presented in Appendix B.2. Next, we compare the achievable rate from Theorem 14 and the lower bound on the optimal rate in Theorem 15, and show that a constant multiplicative gap exists between $R_s^*(M)$ and the achievable rate $R_{s,\mathsf{cen}}(M)$.

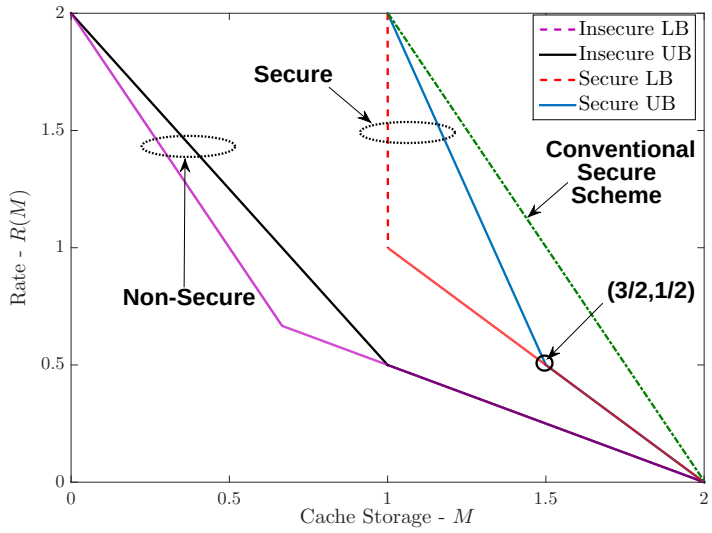
Theorem 16. *For N files and K users, each having a cache size $\max\left\{\frac{(K-N)(N-1)}{KN} + 1, 1\right\} \leq M \leq N$,*

$$1 \leq \frac{R_{s,\mathsf{cen}}(M)}{R_s^*(M)} \leq 17. \tag{4.10}$$

The proof of Theorem 16 is presented in Appendix B.3. The gap is unbounded and scales with K only for the case of $K > N$ in the regime $1 \leq M < \frac{(K-N)(N-1)}{KN} + 1$, which is negligibly small for large K, N as discussed in Appendix B.3. While the analytical constant of 17 is large for practical purposes, the gap can tightened numerically. Fig. 4.2(b) shows the maximum value of the multiplicative gap between $R_{s,\mathsf{cen}}(M)$ and the lower bound on $R_s^*(M)$ for values for N, K ranging from 1 to 1000 and all feasible values of M in each case. It can be seen that the gap is generally less than 4 when $K < N$. However for $K > N$, and for small N, the gap is larger i.e., around 6.



(a)



(b)

Figure 4.3: (a) Secure Caching Scheme and (b) $(M, R_{s,\text{cen}})$ trade-off for $N = K = 2$.

4.3.1 Intuition behind Theorem 14 (Achievability)

We next present a series of examples to explain the intuition behind the achievable rate in Theorem 14 and highlight the interesting features of the proposed secure delivery scheme.

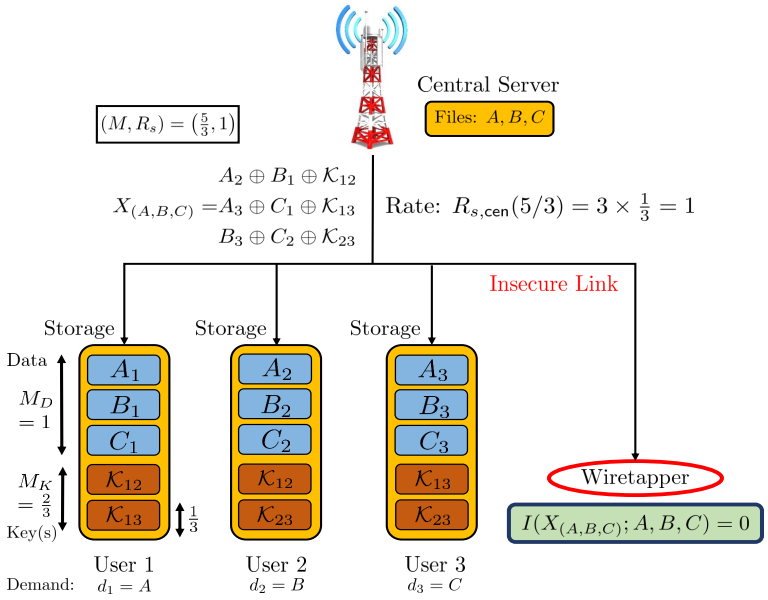
Example 4. We illustrate the achievable scheme in Theorem 14 for the case of $N = 2$ files and $K = 2$ users. From Theorem 14 we have $M \in \frac{2-1}{2}\{0, 1, 2\} + 1 = \{1, \frac{3}{2}, 2\}$ are the possible cache

sizes for each user. Let the two files be $F_1 = A$ and $F_2 = B$. The bounds in Theorems 14 and 15 are shown in Fig. 4.3(b) along with the bounds for the non-secure case from [97]. We start with the upper bound in Theorem 14. Considering the extreme point $M = 1$, the cache of both users Z_1, Z_2 only stores two unique keys $\mathcal{K}_1, \mathcal{K}_2$ and the server transmits both the files A, B over the shared link XOR-ed with a key. Given the worst-case demand $(d_1, d_2) = (A, B)$, the server can transmit $X_{(A,B)} = \{A \oplus \mathcal{K}_1, B \oplus \mathcal{K}_2\}$. This system satisfies every possible request with rate $R = 2$ and it is easily verified that $I\left(X_{(A,B)}; A, B\right) = 0$. Thus $(M, R_{s,\mathsf{cen}}) = (1, 2)$ is *securely* achievable. At the other extreme, when $M = 2$, each user can cache both files and no transmission is necessary. Hence the $(M, R_{s,\mathsf{cen}}) = (2, 0)$ is *securely* achievable.

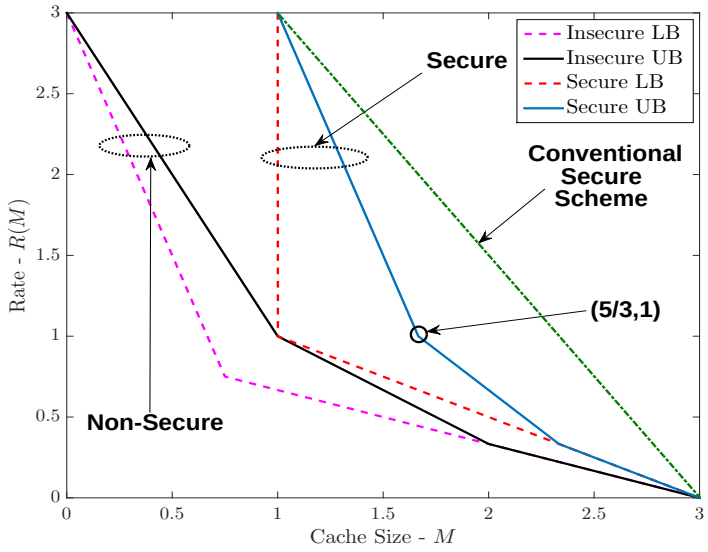
Now we consider the intermediate case in which $M = 3/2$. The scheme for this scenario is depicted in Fig. 4.3(a). Both the files are split into 2 equal parts: $A = (A_1, A_2)$ and $B = (B_1, B_2)$, where A_1, A_2, B_1, B_2 are each of size $B/2$ bits. We also generate a key $\mathcal{K}_{12} \sim \text{unif}\{1, \ldots, 2^{(B/2)}\}$, which is independent of both the files A, B and has the same size as the sub-files i.e., $B/2$ bits. In the storage phase, the server fills the caches as follows: $Z_1 = (A_1, B_1, \mathcal{K}_{12})$ and $Z_2 = (A_2, B_2, \mathcal{K}_{12})$ i.e., each user stores one exclusive part of each file and the key. Thus $M_D = 1/2 + 1/2 = 1$ and $M_K = 1/2$. Now, consider the worst case request $(d_1, d_2) = (A, B)$. In order to satisfy this request, user 1 requires the file fragment A_2 while user 2 requires the file fragment B_1. In this case, the server transmits $X_{(A,B)} = \{A_2 \oplus B_1 \oplus \mathcal{K}_{12}\}$ which is of rate $1/2$. User 1 can obtain A_2 by XOR-ing out $B_1, \mathcal{K}_{12}$ while user 2 can get B_1 by XOR-ing out $A_2, \mathcal{K}_{12}$ from $X_{(A,B)}$. A wiretapper, on the other hand, would gain no knowledge of either file from the transmission since $I\left(X_{(A,B)}; A, B\right) = 0$ which follows from the fact that the key $\mathcal{K}_{12}$ is uniformly distributed. Thus, $(M, R_{s,\mathsf{cen}}) = (3/2, 1/2)$ is *securely* achievable. This can be seen in the secure upper bound in Fig. 4.3(b). Given that the points $(1, 2)$,$(3/2, 1/2)$ and $(2, 0)$ are achievable, the lines joining pairs of these points are also achievable. Thus, this proves the achievability of the secure upper bound in Fig 4.3(b). The gap between the insecure and secure achievable bounds results from the storage of the key in the users' cache. □

In the two user example, there is only a single key $\mathcal{K}_{12}$ in the system. Thus, if the key is compromised, the security of the entire system fails. The scheme proposed in Theorem 14 for general values of (N, K), however is more robust in its key management when the number of files and users increase. We next illustrate this point through an example.

Example 5. We consider the case for $N = K = 3$. For this case, from Theorem 14, $M \in \{1, \frac{5}{3}, \frac{7}{3}, 3\}$. The system and bounds for this case are illustrated in Fig. 4.4(a) and 4.4(b). We consider the case of $M = 5/3$ and three files A, B, C. Each file is split into 3 equal parts i.e., $A = (A_1, A_2, A_3)$, $B = (B_1, B_2, B_3)$, $C = (C_1, C_2, C_3)$. We also have 3 keys in the system, $\mathcal{K}_{12}, \mathcal{K}_{13}, \mathcal{K}_{23}$. In this case, each sub-file and each key is of size $B/3$ bits. In general, the key $\mathcal{K}_{ij}$ is placed in the caches of users i and j. The keys are chosen combinatorially and a general strategy is discussed in Appendix B.1. The overall cache placement is as follows: $Z_1 = \{A_1, B_1, C_1, \mathcal{K}_{12}, \mathcal{K}_{13}\}$, $Z_2 = \{A_2, B_2, C_2, \mathcal{K}_{12}, \mathcal{K}_{23}\}$ and $Z_3 = \{A_3, B_3, C_3, \mathcal{K}_{13}, \mathcal{K}_{23}\}$. Thus each cache has size $M = 5 \times (1/3) = 5/3$, where $M_D = 1, M_K = 2/3$. Now considering a worst case request where all users request different files, $(d_1, d_2, d_3) = (A, B, C)$, the server can



(a)



(b)

Figure 4.4: (a) Secure Caching Scheme and (b) $(M, R_{s,\text{cen}})$ trade-off for $N = K = 3$.

make the transmission, $X_{(A,B,C)} = \{\{A_2 \oplus B_1 \oplus \mathcal{K}_{12}\}, \{A_3 \oplus C_1 \oplus \mathcal{K}_{13}\}, \{B_3 \oplus C_2 \oplus \mathcal{K}_{23}\}\}$, such that everyone can securely retrieve their requested files. Thus $(M, R_{s,\text{cen}}) = (5/3, 1)$ is *securely* achievable since $I(X_{(A,B,C)}; A, B, C) = 0$ i.e., a wiretapper would gain no information about the files from the transmission. It can be seen from the cache contents that there are multiple keys in the system thereby avoiding a single point of failure. In general, if we choose operating

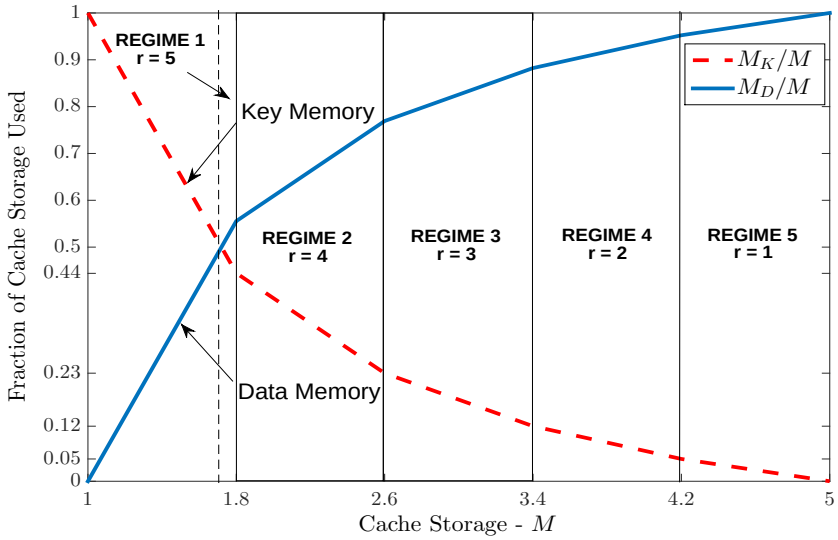


Figure 4.5: M_K vs. M_D trade-off for $N = K = 5$

points $(M, R_{s,\text{cen}})$ such that $M_K > 1/K$, single points of failure in the system can be avoided. Thus the scheme forms an interesting storage-rate trade-off based on users' security constraints which is elaborated subsequently in Remark 12. □

Remark 12 (*Key Storage vs. Data Storage Trade-off*)**.** The trade-off between the fraction of cache storage occupied by the data and the keys in the secure caching system is shown in Fig. 4.5 for $N = 5$ files and $K = 5$ users. Consider the cache storage constraint in Theorem 14 i.e., $M \in \frac{N-1}{K}t + 1,\ \forall t \in \{0, 1, 2, \ldots, K\}$. Now, since $M = M_D + M_K$, from Appendix B.1, we have $M_K = 1 - t/K$ and $M_D = Nt/K$. From Fig. 4.5, it can be seen that M_K dominates at lower values of M. Formally, $M \geq 2N/(N + 1)$, data storage dominates key storage i.e., $M_D > M_K$. From Appendix B.1, we have $\binom{K}{t+1}$ unique keys in the system. Thus the case for there being only one unique key in the system corresponds to $t = K - 1$ i.e., $M_K = 1/K$. Thus for avoiding one shared key across all users i.e., a single point of failure in the system, we need $M_K > 1/K \Rightarrow t \leq K - 1$, which corresponds to $M \leq (N - 1)(K - 1)/K + 1$. It is also undesirable that new keys be redistributed to the entire system each time a user leaves. The proposed scheme avoids this scenario by sharing keys. In case a user leaves or is compromised, only the keys contained in that user's cache need to be replaced, leaving the others untouched. Thus, a desirable region of operation would be:

$$\frac{2N}{(N+1)} \leq M \leq \frac{(N-1)(K-1)}{K} + 1.$$

In general, a close inspection of Algorithm 2 reveals that when $t > (K - r)$ i.e., when $M > (N - 1)(K - r)/K + 1$, a wiretapper can obtain all the keys in the system if it gains access to

any r of the K user caches. This means that if r users are compromised, system security will be violated. It is a trivial fact that at $t = 0$, $M = 1$ and each user has one unique key. In this case, the wiretapper will need access to all caches in order to violate the security of the system.

From Fig. 4.5, we can see that Regime 5, i.e., when $r = 1$, is the weakest regime from the security perspective as there is only one key in the system. Thus operation in Regimes 1-4 is desirable for the case of $N = K = 5$. Now, considering the *conventional secure scheme*, it is seen that there is no sharing of keys as each transmission is useful to only one user. Thus each user stores an unique key of size $|\mathcal{K}| = (1 - \frac{M-1}{N-1})B$ bits. This scheme thus requires the wiretapper to have access to all the caches for the system security to be compromised. Comparing the conventional and proposed schemes from a security perspective, we see that the proposed scheme is a *trade-off* between security and minimization of the rate over the shared link. While the conventional scheme is more difficult to compromise for $M \in \mathbb{N}$, the proposed scheme is able to improve on the transmission rate significantly while still providing security. □

4.3.2 Intuition behind Theorem 15 (Converse)

We next present the main idea behind the proof of the converse stated in Theorem 15 through a novel extension of the cut-set bound to incorporate the security constraint. To this end, we focus on the caching system with $N = 2$ files (denoted by A and B) and $K = 2$ users (with cache contents denoted by Z_1 and Z_2). Consider the scenario where user 1 demands file A and user 2 demands file B, i.e., the demand vector is $(d_1, d_2) = (A, B)$. It is easy to check that using the communication $X_{(A,B)}$ from the central server along with the two caches Z_1, Z_2, both files (A, B) can be recovered. This implies the following constraint:

$$H\left(A, B|X_{(A,B)}, Z_1, Z_2\right) \leq \epsilon. \tag{4.11}$$

Next, for the communication $X_{(A,B)}$ to be secure, we also require the following security constraint to hold:

$$I\left(A, B; X_{(A,B)}\right) \leq \epsilon. \tag{4.12}$$

Using these two constraints, we next show that for any scheme, $M \geq 1$ must necessarily hold. From the constraints (4.11)-(4.12), we have the following sequence of inequalities:

$$\begin{aligned}
2B &\leq H(A, B) = I\left(A, B; X_{(A,B)}, Z_1, Z_2\right) + H\left(A, B|X_{(A,B)}, Z_1, Z_2\right) \\
&\overset{(4.11)}{\leq} I\left(A, B; X_{(A,B)}, Z_1, Z_2\right) + \epsilon \\
&= I\left(A, B; X_{(A,B)}\right) + I\left(A, B; Z_1, Z_2|X_{(A,B)}\right) + \epsilon \\
&\overset{(4.12)}{\leq} I\left(A, B; Z_1, Z_2|X_{(A,B)}\right) + 2\epsilon \\
&\leq H\left(Z_1, Z_2|X_{(A,B)}\right) + 2\epsilon \leq H(Z_1, Z_2) + 2\epsilon \\
&\leq H(Z_1) + H(Z_2) + 2\epsilon \leq 2MB + 2\epsilon.
\end{aligned}$$

This implies

$$M \geq 1 - \frac{\epsilon}{B}. \tag{4.13}$$

Taking the limit $\epsilon \to 0$, we arrive at the proof of $M \geq 1$. Now consider the fact that given the transmissions from the server $X_{(A,B)}$ for demands $(d_1, d_2) = (A, B)$, $X_{(B,A)}$ for demands $(d_1, d_2) = (B, A)$ and one cache Z_1, both the files A, B can be recovered. Again, we have the following constraints for file retrieval and security:

$$H\left(A, B | X_{(A,B)}, X_{(B,A)}, Z_1\right) \leq \epsilon \tag{4.14}$$

$$I\left(A, B; X_{(A,B)}\right) \leq \epsilon. \tag{4.15}$$

Thus we have,

$$\begin{aligned}
2B &\leq H(A,B) = I\left(A,B; X_{(A,B)}, X_{(B,A)}, Z_1\right) + H\left(A,B|X_{(A,B)}, X_{(B,A)}, Z_1\right) \\
&\overset{(4.14)}{\leq} I\left(A,B; X_{(A,B)}, X_{(B,A)}, Z_1\right) + \epsilon \\
&= I\left(A,B; X_{(A,B)}\right) + I\left(A,B; X_{(B,A)}, Z_1 | X_{(A,B)}\right) + \epsilon \\
&\overset{(4.15)}{\leq} I\left(A,B; X_{(B,A)}, Z_1 | X_{(A,B)}\right) + 2\epsilon \\
&\leq H\left(X_{(B,A)}, Z_1 | X_{(A,B)}\right) + 2\epsilon \\
&\leq H\left(X_{(B,A)}\right) + H(Z_1) + 2\epsilon \\
&\leq R_s^* B + MB + 2\epsilon.
\end{aligned}$$

This implies that

$$R_s^* + M \geq 2 - \frac{2\epsilon}{B}. \tag{4.16}$$

Taking the limit $\epsilon \to 0$, we arrive at the proof of $R_s^* + M \geq 2$. We can see that both (4.13) and (4.16) hold for all achievable (M, R_s) pairs. Thus we have, $R_s^*(M) \geq 2 - M$ and $M \geq 1$ which gives the lower bound in Fig. 4.3(b).

4.4 Decentralized Caching with Secure Delivery

In this section, we extend the secure caching problem to a decentralized caching scheme as discussed in [98]. In the decentralized caching scheme, each user is allowed to cache any random $\frac{M-1}{N-1}$ bits of each of the N files in the system. In the coded delivery scheme, the central server maps the contents of individual users' caches to fragments (which contain non-overlapping combination of bits) in each file. The fragments reflect which user (or set of users) has cached bits contained in the given fragment. This phase is followed by a centralized key placement procedure where the server stores shared keys in each user's cache. The key placement needs to be centralized to maintain key integrity and to secure the files from an external wiretapper. In the delivery phase, the server receives a request $(d_1, \ldots, d_K)$ and forms coded multicast transmissions to extract global caching

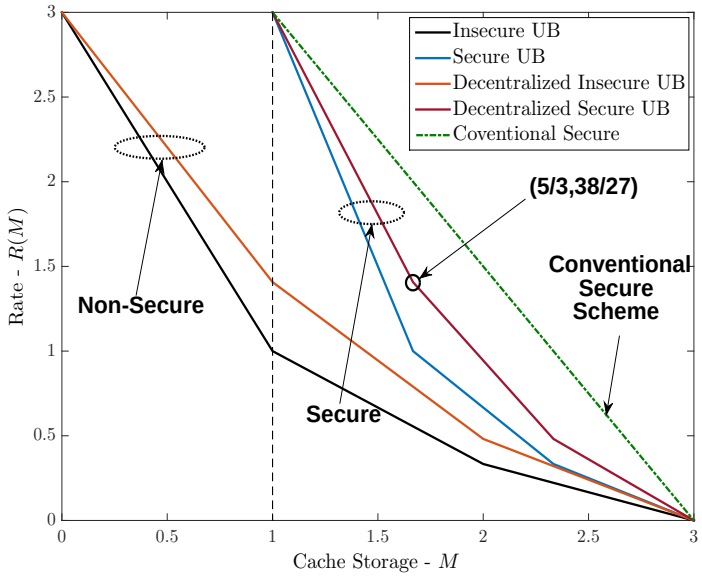


Figure 4.6: $(M, R_{s,\mathsf{dec}})$ trade-off for $N = K = 3$

gain from the system. It then encodes the transmissions with the shared keys and transmits them over the multicast link. The decentralized algorithm is presented in Algorithm 3. In the case of decentralized caching, similar to the centralized case, the *conventional secure scheme* is one which stores only one unique key per user and exploits only the local caching gain by using encrypted unicast delivery. The transmission rate in this case is given by $K(1-\frac{M-1}{N-1})$. After the cache placement, the server chooses the scheme which provides the minimum rate over the shared link. The secure rate is then characterized by the following theorem.

Theorem 17. *For N files and K users, each with a cache size of $M \in \frac{N-1}{N} \cdot t + 1$, for $t \in (0, N]$,*

$$R_{s,\mathsf{dec}}(M) \triangleq K\left(1-\frac{M-1}{N-1}\right) \cdot \min\left\{\frac{N-1}{K(M-1)} \cdot \left(1-\left(1-\frac{M-1}{N-1}\right)^{K}\right), 1\right\} \tag{4.17}$$

is securely achievable. For any $1 < M \leq N$, the lower convex envelope of these points is achievable.

The proof of Theorem 17 is given in Appendix B.4. The variable $t = M_D$, represents the part of the cache storage used to store data at each user (as detailed in Appendix B.4). Theorem 17 is defined for $t > 0$. At $t = 0$, $M = 1$ i.e., the caches store a single key of the size of each file. Entire files, XOR-ed with the keys, are then transmitted over the shared link. Thus the rate in this case is $R_{s,\mathsf{dec}}(1) \triangleq K$. As before, the same argument holds for the infeasibility of the secure scheme for $M = 0$. The following example illustrates the caching scheme which achieves the rate in Theorem 17.

Algorithm 3 SECURE DECENTRALIZED CACHING ALGORITHM

Decentralized Cache Placement:

for $k \in \{1, \ldots, K\}, n \in \{1, \ldots, N\}$ **do**
    User k randomly caches $\frac{M-1}{N-1}F$ bits of file n.
end for

Delivery Procedure for request $(d_1, \ldots, d_K)$

CENTRALIZED KEY PLACEMENT:

Central server maps the cache contents to fragments in the files $W_1, \ldots, W_N$ and generates keys

for $i = 0, 1, 2, \ldots, K$ **do**
    for $n = 1, 2, \ldots, N$ **do**
        $W_n = \{W_{n,\mathcal{T}}\}, \ \mathcal{T} \subseteq \{1, \ldots, K\} : |\mathcal{T}| = i$ such that $W_{n,\mathcal{T}}$ is cached at user k, if $k \in \{\mathcal{T}\}$
    end for
end for
for $s = 1, 2, \ldots, K$ **do**
    for $\mathcal{S} \subseteq \{1, \ldots, K\} : |\mathcal{S}| = s$ **do**
        Key $\mathcal{K}_{\mathcal{S}}$ is generated
        $\mathcal{K}_{\mathcal{S}}$ is placed in cache of user k if $k \in \{\mathcal{S}\}$
    end for
end for

CODED SECURE DELIVERY:

for $s = K, K-1, \ldots, 1$ **do**
    for $\mathcal{S} \subseteq \{1, \ldots, K\} : |\mathcal{S}| = s$ **do**
        Server sends $\left\{\mathcal{K}_{\mathcal{S}} \oplus_{k \in \mathcal{S}} W_{d_k, \mathcal{S}\setminus\{k\}}\right\}$
    end for
end for

Conventional Delivery Procedure for request $(d_1, \ldots, d_K)$

Server places individual keys of size $(1 - \frac{M-1}{N-1})F$ bits at each user's cache
for $n \in \{0, \ldots, N\}$ **do**
    Server sends enough random linear combinations of bits in file n XOR-ed with individual keys for the all users requesting it
end for

Example 6. We consider the case for $N = 3$ files and $K = 3$ users, each with a cache of size MB bits. Let the three files be denoted as $(F_1, F_2, F_3) = (A, B, C)$. Fig. 4.6 shows the rate achieved by the secure decentralized caching scheme given by Theorem 17, the rate of the insecure decentralized scheme from [98] and the corresponding centralized bounds. In the decentralized placement phase, each of the 3 users caches a subset of $(M-1)B/2$ bits of each file independently at random. Thus, each bit of a file is cached by a specific user with probability $(M-1)/2$. Considering the file A, the server maps the storage of fragments of file A at the different users' caches into splits, $A_{\mathcal{T}}$, such that $\mathcal{T} \subseteq \{1, 2, 3\}$, $|\mathcal{T}| = i$ for $i = 0, 1, 2, 3$. Thus there are $\sum_{i=0}^{3} \binom{3}{i} = 2^3 = 8$ splits of file A: $(A_\phi, A_1, A_2, A_3, A_{12}, A_{13}, A_{23}, A_{123})$, where A_ϕ consists of bits of A which are not stored in

any users' cache. On the other hand, A_{123} has bits which are stored in all users cache. In general, bits in $A_{\mathcal{T}}$ are stored in user k's cache if $k \in \mathcal{T}$. By law of large numbers, we have:

$$|A_{\mathcal{T}}| \approx \left(\frac{M-1}{2}\right)^{|\mathcal{T}|} \left(1-\frac{M-1}{2}\right)^{3-|\mathcal{T}|} B \text{ bits} \tag{4.18}$$

with probability approaching one for large enough file size B. The same analysis holds for files B, C. Next, we consider the generation of keys $\mathcal{K}_{\mathcal{S}}$ for $\mathcal{S} \subseteq \{1,2,3\}, |\mathcal{S}| = j$ for $j = 1,2,3$. Thus the keys generated in the system are: $\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3, \mathcal{K}_{12}, \mathcal{K}_{13}, \mathcal{K}_{23}, \mathcal{K}_{123}$. It can be seen that there are $2^K - 1 = 7$ unique keys in the system. Next we look at the cache contents from the central server's perspective after the centralized key placement phase and before the delivery procedure begins. The cache placement for $N = K = 3$ is:

$$Z_1 = \left\{ \begin{array}{c} A_1, A_{12}, A_{13}, A_{123} \\ B_1, B_{12}, B_{13}, B_{123} \\ C_1, C_{12}, C_{13}, C_{123} \\ \mathcal{K}_1, \mathcal{K}_{12}, \mathcal{K}_{13}, \mathcal{K}_{123} \end{array} \right\} \quad Z_2 = \left\{ \begin{array}{c} A_2, A_{12}, A_{23}, A_{123} \\ B_2, B_{12}, B_{23}, B_{123} \\ C_2, C_{12}, C_{23}, C_{123} \\ \mathcal{K}_2, \mathcal{K}_{12}, \mathcal{K}_{23}, \mathcal{K}_{123} \end{array} \right\} \quad Z_3 = \left\{ \begin{array}{c} A_3, A_{13}, A_{23}, A_{123} \\ B_3, B_{13}, B_{23}, B_{123} \\ C_3, C_{13}, C_{23}, C_{123} \\ \mathcal{K}_3, \mathcal{K}_{13}, \mathcal{K}_{23}, \mathcal{K}_{123} \end{array} \right\}. \tag{4.19}$$

The cache placement phase is entirely decentralized as the users do not need to consider the number of other users in the system or their cache contents while storing file fragments in their caches. Next, we consider the delivery procedure of the decentralized caching scheme. The system is characterized based on the worst possible rate over the shared link. Thus we consider a request $(F_{d_1}, F_{d_2}, F_{d_3}) = (A, B, C)$. The server responds by transmitting the reply $X_{(A,B,C)}$. Let the set $\mathcal{S} \subseteq \{1,2,3\} : |\mathcal{S}| = s$ for $s = 3,2,1$. Then we have $X_{(A,B,C)} = \left\{\mathcal{K}_{\mathcal{S}} \oplus_{k\in\mathcal{S}} F_{d_k,\mathcal{S}\setminus\{k\}} : k = 1,2,3\right\}_{s=1}^{3}$, where $F_{d_k,\mathcal{S}\setminus\{k\}}$ corresponds to the fraction of the file F_{d_k}, requested by user k which is not present in user k's cache but is present in the cache of the other $s-1$ users in $\mathcal{S}$. Thus, for $K = 3$ users in the system, the coded secure multicast delivery procedure has 3 phases for each of $s = 3,2,1$.

- *For* $s = 3$: We have $|\mathcal{S}| = 3 \Rightarrow \mathcal{S} = \{1,2,3\}$ and $|\mathcal{S} \setminus \{k\}| = 2$. The transmission is $\{A_{23} \oplus B_{13} \oplus C_{12} \oplus \mathcal{K}_{123}\}$. It can be seen that $\mathcal{K}_{123}$ is associated with sub-files A_{23}, B_{13}, C_{12}. Thus the size of the key is $|\mathcal{K}_{123}| = \max\{|A_{23}|, |B_{13}|, |C_{12}|\}$. In this case, each sub-file is zero padded to the size of the largest sub-file in the set. Considering user 1, we see that Z_1 contains B_{13}, C_{12} and $\mathcal{K}_{123}$. Thus user 1 can XOR out A_{23} from the transmission. It can be seen that the same holds for users 2 and 3. Thus the transmission is useful for all users and the key makes it secure from the wiretapper. For $s = 3$, there is only one transmission of the size of each of these sub-files. Thus, using (4.18), the rate over the shared link for this transmission is:

$$\left(\frac{M-1}{2}\right)^2 \left(1-\frac{M-1}{2}\right) B. \tag{4.20}$$

- *For* $s = 2$: We have $|\mathcal{S}| = 2 \Rightarrow \mathcal{S} \in \{1,2\},\{2,3\},\{1,3\}$ and $|\mathcal{S} \setminus \{k\}| = 1$. The transmission for each subset $\mathcal{S}$ is $\{\{A_2 \oplus B_1 \oplus \mathcal{K}_{12}\}, \{B_3 \oplus C_2 \oplus \mathcal{K}_{23}\}, \{A_3 \oplus C_1 \oplus \mathcal{K}_{13}\}\}$. Again

for user 1, we can see that Z_1 contains $B_1, C_1, \mathcal{K}_{12}, \mathcal{K}_{13}$. Thus it can extract A_2, A_3 from this transmission. Similarly the other users can extract fragments of their requested files. In this case, there are three transmissions, each of the size of file fragment, say, A_2. Thus the rate of this transmission is:

$$3 \cdot \left(\frac{M-1}{2}\right)\left(1-\frac{M-1}{2}\right)^2 B. \tag{4.21}$$

- *For* $s=1$: We have $|\mathcal{S}|=1 \Rightarrow \mathcal{S} \in \{1\},\{2\},\{3\}$ and $|\mathcal{S}\backslash\{k\}|=0$. The transmission for each subset $\mathcal{S}$ is $\{\{A_\phi \oplus \mathcal{K}_1\},\{B_\phi \oplus \mathcal{K}_2\},\{C_\phi \oplus \mathcal{K}_3\}\}$. These transmissions are sent to individual users, containing the residual fragments not stored in each user. The size of each transmission is equal to the size of the file fragments A_ϕ, B_ϕ, C_ϕ. Thus the rate of this transmission is:

$$3 \cdot \left(1-\frac{M-1}{2}\right)^3 B. \tag{4.22}$$

Again considering user 1, we can see that the fragments of A not present in its cache i.e., A_ϕ, A_2, A_3, A_{23} are extracted from the entire transmission. The same holds true for the other users. The rate for the composite transmission $X_{(A,B,C)}$ is obtained by summing (4.20), (4.21) and (4.22):

$$\begin{aligned} &R_{s,\mathsf{dec}}(M)B \\ &= B\left(\frac{M-1}{2}\right)^2\left(1-\frac{M-1}{2}\right) + 3B\left(\frac{M-1}{2}\right)\left(1-\frac{M-1}{2}\right)^2 + 3B\left(1-\frac{M-1}{2}\right)^3 \\ &= 3\left(1-\frac{M-1}{2}\right)\frac{2}{3(M-1)}\left(1-\left(1-\frac{M-1}{2}\right)^3\right)B, \end{aligned} \tag{4.23}$$

which is the expression given in Theorem 17 for $N=K=3$. Now, we have $M \in \frac{N-1}{N}\{1,2,\ldots,N\}+1 = \left\{\frac{5}{3},\frac{7}{3},3\right\}$. Considering the point $M=5/3$, we have $R_{s,\mathsf{dec}}(M)=38/27$. Thus the pair $(M, R_{s,\mathsf{dec}}) = (5/3, 38/27)$, is securely achievable. This is seen from the $(M, R_{s,\mathsf{dec}})$ trade-off in Fig. 4.6. Similarly other points on the trade-off curve can be evaluated using other feasible values of M. All points on the lines joining the achievable $(M, R_{s,\mathsf{dec}})$ points are also achievable. □

Next, we consider the centralized and decentralized trade-off for a large number of files and users. Fig. 4.7 illustrates the case for $N=K=20$. Compared to Fig. 4.6, we can see that as the number of files and users increase, the decentralized scheme approaches the centralized caching. Thus for large number of files and users, the rates are *asymptotically equal*. This also implies that in the decentralized case, similar to the centralized case, that the cost for security is *almost negligible* when number of files and users increase. The following theorem and corollary compares the rate of the achievable secure decentralized scheme given in Theorem 17 to the lower bound on the rate of the optimal secure scheme given in Theorem 15 and the rate of the achievable secure centralized caching scheme given in Theorem 14.

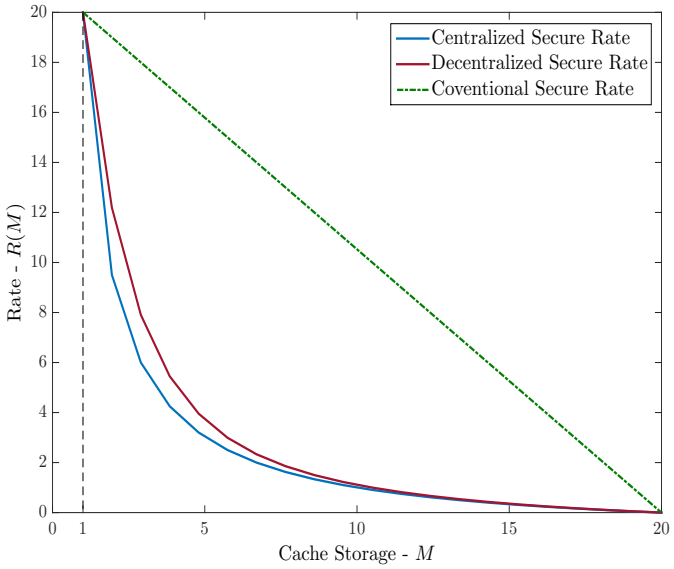


Figure 4.7: Centralized vs Decentralized Secure Bounds for $N = K = 20$.

Theorem 18. *Given $R_{s,\mathsf{dec}}(M)$ be the rate of the secure decentralized caching scheme given by Algorithm 3 and $R_s^*(M)$ be the rate of the optimal secure caching scheme, for N files and K users, each having a cache size $\frac{N-1}{N} + 1 \leq M \leq N$,*

$$\frac{R_{s,\mathsf{dec}}(M)}{R_s^*(M)} \leq 17. \tag{4.24}$$

The proof sketch of Theorem 18 is given in Appendix B.5. Theorem 18 implies that no scheme, regardless of complexity can improve by more than a constant factor upon the secure decentralized caching scheme presented in Algorithm 3 for the given regime of M. The gap is unbounded only for the case of $K > N$ in the regime $1 \leq M \leq \frac{N-1}{N} + 1$, which is negligibly small for large N, K as discussed in Appendix B.5.

Corollary 3. *Let $R_{s,\mathsf{cen}}(M)$ be the rate of the secure centralized caching scheme given in Theorem 14 and $R_{s,\mathsf{dec}}(M)$ be the rate of the secure decentralized caching scheme given in Theorem 17. For N files and K users, for $\frac{N-1}{N} + 1 \leq M \leq N$, we have*

$$\frac{R_{s,\mathsf{dec}}(M)}{R_{s,\mathsf{cen}}(M)} \leq 17. \tag{4.25}$$

Corollary 3 is a direct outcome of Theorems 16 and 18. It shows that the decentralized scheme is at most a constant factor 17 worse than the secure centralized scheme in the given regime of M.

4.5 Secure Caching with Non-Uniform Demands

The results on secure delivery presented so far consider the *peak* rate of the system given all the files are of uniform popularity. In this section we consider the *expected* rate of the system in a more realistic setting where the files popularities vary by several orders of magnitude. Similar to the scheme in [99], the files in $\mathbb{N}$ are partitioned into sets of approximately uniform popularities. Without loss of generality, the files in $\mathbb{N}$ can be relabeled such that $p_1 > p_2 > \ldots > p_N$. We partition the N files into L groups $\{\mathbb{N}_\ell : \ell = 1, \ldots, L\}$ where $N_\ell \ (= |\mathbb{N}_\ell|)$ is the number of files in the $\ell-$th group. We have

$$\sum_{\ell=1}^{L} N_\ell = N. \tag{4.26}$$

The groups are chosen as follows: for any two files in the same group, $F_n, F_{n'} \in \mathbb{N}_\ell$, the popularities $p_n, p_{n'}$ differ by at most a factor p i.e.,

$$p_n \geq p_{n+N_\ell-1} \geq p_n/\mathsf{p} \ \text{ and } \ p_{n+N_\ell} < p_n/\mathsf{p}.$$

The files in $\mathbb{N}$ are maximally partitioned to within a popularity factor of p into L sets $\mathbb{N}_1, \mathbb{N}_2, \ldots, \mathbb{N}_L$. Each set $\mathbb{N}_\ell$ is allocated a portion M_ℓ of the cache storage M at each user such that the total storage constraint is satisfied i.e.,

$$\sum_{\ell=1}^{L} M_\ell = M \tag{4.27}$$

In the placement phase of the proposed secure caching scheme, each user is allowed to cache $\frac{M_\ell - 1}{N_\ell - 1}$ bits of each file from the group $\mathbb{N}_\ell$ for $\ell = 1, \ldots, L$. Now, from the analysis of Algorithm 3 in Appendix B.4, the amount of cache storage allocated to data and keys for the $\ell-$th group at each user is:

$$M_{D,\ell} = N_\ell \cdot \frac{M_\ell - 1}{N_\ell - 1}; \quad M_{K,\ell} = 1 - \frac{M_\ell - 1}{N_\ell - 1}$$

Thus the cache storage at each user for the L groups is

$$\sum_{\ell=1}^{L} (M_{D,\ell} + M_{K,\ell}) = \sum_{\ell=1}^{L} \left(N_\ell \cdot \frac{M_\ell - 1}{N_\ell - 1} + 1 - \frac{M_\ell - 1}{N_\ell - 1} \right) = \sum_{\ell=1}^{L} M_\ell = M \tag{4.28}$$

which satisfies the storage constraint. Next we discuss the proposed secure content delivery.

4.5.1 Secure Content Delivery

For the delivery phase of the secure caching scheme, the delivery phase in Algorithm 3 is applied independently to each of the L groups which have uniform popularity to within a factor of p. For

the $\ell-$th group, the server first maps the contents of individual users' caches to fragments (which contain non-overlapping combination of bits) in each file in $\mathbb{N}_\ell$. The fragments reflect which user (or set of users) has cached bits contained in the given fragment. We denote by $\mathbb{K}_\ell$, the users who request a file in the group $\mathbb{N}_\ell$. Thus $\mathbb{K}_1, \mathbb{K}_1, \ldots, \mathbb{K}_L$ partitions the users into L groups. Since choice of groups is dependent on the the random user requests, which in turn depend on the file popularities, the cardinality K_ℓ $(= |\mathbb{K}_\ell|)$ of each group is a random variable. Thus, given N files in the system, K_ℓ models the non-uniform user demands in the system.

Key Placement

The mapping phase is followed by a centralized key placement procedure where the server stores shared keys in each users' cache following the procedure in Algorithm 3 for each group of files $\mathbb{N}_\ell$ and corresponding users $\mathbb{K}_\ell$, who request files in $\mathbb{N}_\ell$. For each subset $\mathcal{S} \subseteq \{1, 2, \ldots, \mathsf{K}_\ell\}$ such that $|\mathcal{S}| = s$, for each $s = 1, 2, \ldots, \mathcal{K}_\ell$, a key is generated according to

$$\mathcal{K}_{\mathcal{S}}^{\ell} \sim \text{unif}\left\{1, 2, 3, \ldots, 2^{Bq_\ell^{s-1}(1-q_\ell)^{\mathsf{K}_\ell - s+1}}\right\}, \tag{4.29}$$

where $q_\ell = (M_\ell - 1)/(N_\ell - 1)$. The interested reader is referred to (Appendix E, [176]) for rigorous analysis of key generation and distribution within each group.

Secure Achievability

Using the above notion of key distribution across groups of files and corresponding users, we next show that the delivery procedure of the proposed scheme is information theoretically secure. From Theorem 17, it can be seen that, within each of the L groups, the rate $R_{s,\mathsf{dec}}(M_\ell, N_\ell, \mathsf{K}_\ell)$ is *securely achievable* i.e., for each group of files and requesting users, $(\mathbb{N}_\ell, \mathbb{K}_\ell)$, the content delivery to the K_ℓ users in $\mathbb{K}_\ell$ is information theoretically secure. Thus, it remains to be proved that the delivery process maintains information theoretic security even across the L groups. The achievable rate for the proposed scheme over all L groups is given by

$$R_s(M, N, K) = \sum_{\ell=1}^{L} R_{s,\mathsf{dec}}(M_\ell, N_\ell, \mathsf{K}_\ell). \tag{4.30}$$

Consider the coded delivery process of Algorithm 3. Given a request $(d_1, d_2, \ldots, d_K)$, it can be partitioned into L sets i.e., $(d_{\mathbb{K}_1}, d_{\mathbb{K}_2}, \ldots, d_{\mathbb{K}_L})$ where $d_{\mathbb{K}_\ell}$ defines the request of users in the group $\mathbb{K}_\ell$. The composite transmission which is sent by the central server can be written as

$$X_{(d_1,\ldots,d_K)} = \left\{\left\{X^s_{(d_{\mathbb{K}_\ell})}\right\}_{s=1}^{\mathsf{K}_\ell}\right\}_{\ell=1}^{L}, \tag{4.31}$$

where, for each group of files and users, $(\mathbb{N}_\ell, \mathbb{K}_\ell)$, $X^s_{(d_{\mathbb{K}_\ell})}$ consists of $\binom{\mathsf{K}_\ell}{s}$ transmissions, one for each possible sub-set $\mathcal{S} \subset \mathbb{N}_\ell$ of size s where $s = 1, 2, \ldots, \mathsf{K}_\ell$ i.e.,

$$X^s_{(d_{\mathbb{K}_\ell})} = \left\{ \mathcal{K}^\ell_{\mathcal{S}} \oplus_{k\in\mathcal{S}} F_{d_k,\mathcal{S}\setminus\{k\}} : |\mathcal{S}| = s \right\}. \tag{4.32}$$

$F_{d_k,\mathcal{S}\setminus\{k\}}$ denotes the part of the file $F_{d_k} \in \mathbb{N}_\ell$ requested by user $k \in \mathbb{K}_\ell$ which is present in the caches all the users in set $\mathcal{S}$ except in the cache of user k. The key $\mathcal{K}^\ell_{\mathcal{S}}$ is associated with the transmission $\oplus_{k\in\mathcal{S}} F_{d_k,\mathcal{S}\setminus\{k\}}$. Furthermore, from the design of the key placement, the key $\mathcal{K}^\ell_{\mathcal{S}}$ is available in the cache of all the s users in the sub-set $\mathcal{S}$ since $\mathcal{S} \subset \mathbb{N}_\ell$. To prove that the delivery is information theoretically secure across all groups, we need to show that

$$I\left(X_{(d_1,\ldots,d_K)}; \mathbb{N}\right) = 0 \tag{4.33}$$

where $F_1, \ldots, F_N \in \mathbb{N}$. We have,

$$I\left(X_{(d_1,\ldots,d_K)}; \mathbb{N}\right) = H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(X_{(d_1,\ldots,d_K)}|\mathbb{N}\right) \tag{4.34}$$

Using the fact that $H(A,B) \leq H(A) + H(B)$, we have:

$$\begin{aligned}
H\left(X_{(d_1,\ldots,d_K)}\right) &= H\left(\left\{\left\{X^s_{(d_{\mathbb{K}_\ell})}\right\}_{s=1}^{\mathsf{K}_\ell}\right\}_{\ell=1}^{L}\right) \leq \sum_{\ell=1}^{L}\sum_{s=1}^{\mathsf{K}_\ell} H\left(X^s_{(d_{\mathbb{K}_\ell})}\right) \\
&\leq \sum_{\ell=1}^{L}\sum_{s=1}^{\mathsf{K}_\ell}\sum_{i=1}^{\binom{\mathsf{K}_\ell}{s}} H\left(\mathcal{K}^\ell_{\mathcal{S}_i} \oplus_{k\in\mathcal{S}_i} F_{d_k,\mathcal{S}_i\setminus\{k\}} : |\mathcal{S}_i| = s\right) \\
&\leq \sum_{\ell=1}^{L}\sum_{s=1}^{\mathsf{K}_\ell}\sum_{i=1}^{\binom{\mathsf{K}_\ell}{s}} \log_2\left(B q_\ell^{s-1}(1-q_\ell)^{\mathsf{K}_\ell - s + 1}\right) \\
&= \sum_{\ell=1}^{L}\sum_{s=1}^{\mathsf{K}_\ell} \binom{\mathsf{K}_\ell}{s} \log_2\left(B q_\ell^{s-1}(1-q_\ell)^{\mathsf{K}_\ell - s + 1}\right).
\end{aligned} \tag{4.35}$$

On the other hand, we have:

$$\begin{aligned}
H\left(X_{(d_1,\ldots,d_K)}|\mathbb{N}\right) &= H\left(\left\{\left\{X^s_{(d_{\mathbb{K}_\ell})}\right\}_{s=1}^{\mathsf{K}_\ell}\right\}_{\ell=1}^{L}|\mathbb{N}\right) \\
&= H\left(\left\{\left\{\left\{\mathcal{K}^\ell_{\mathcal{S}} \oplus_{k\in\mathcal{S}} F_{d_k,\mathcal{S}\setminus\{k\}} : |\mathcal{S}| = s\right\}\right\}_{s=1}^{\mathsf{K}_\ell}\right\}_{\ell=1}^{L}|\mathbb{N}\right) \\
&= H\left(\left\{\left\{\left\{\mathcal{K}^\ell_{\mathcal{S}} : |\mathcal{S}| = s\right\}\right\}_{s=1}^{\mathsf{K}_\ell}\right\}_{\ell=1}^{L}|\mathbb{N}\right) \\
&= H\left(\left\{\left\{\left\{\mathcal{K}^\ell_{\mathcal{S}} : |\mathcal{S}| = s\right\}\right\}_{s=1}^{\mathsf{K}_\ell}\right\}_{\ell=1}^{L}\right)
\end{aligned}$$

where the last equality follows from the fact that the keys are uniformly distributed as shown in (4.29) and are independent of all the files in $\mathbb{N}$. This implies that the keys generated in each group are also independent of keys in the remaining $L-1$ groups and to all the files in all the groups. Even though the keys in each group have variable length, orthogonality can be maintained by using codes like Orthogonal Variable Spreading Factor (OSVF) codes, derived from Walsh codes, which can generate orthogonal sequences of variable length. Thus we have

$$\begin{aligned} H\left(\left\{\left\{\left\{\mathcal{K}_{\mathcal{S}}^{\ell}:|\mathcal{S}|=s\right\}\right\}_{s=1}^{\mathsf{K}_\ell}\right\}_{\ell=1}^{L}\right) &= \sum_{\ell=1}^{L}\sum_{s=1}^{\mathsf{K}_\ell} H\left(\left\{\mathcal{K}_{\mathcal{S}}^{\ell}:|\mathcal{S}|=s\right\}\right) \\ &= \sum_{\ell=1}^{L}\sum_{s=1}^{\mathsf{K}_\ell}\sum_{i=1}^{\binom{\mathsf{K}_\ell}{s}} H\left(\mathcal{K}_{\mathcal{S}_i}:|\mathcal{S}_i|=s\right) \\ &= \sum_{\ell=1}^{L}\sum_{s=1}^{\mathsf{K}_\ell}\sum_{i=1}^{\binom{\mathsf{K}_\ell}{s}} \log_2\left(Bq_\ell^{s-1}(1-q_\ell)^{\mathsf{K}_\ell-s+1}\right) \\ &= \sum_{\ell=1}^{L}\sum_{s=1}^{\mathsf{K}_\ell}\binom{\mathsf{K}_\ell}{s}\log_2\left(Bq_\ell^{s-1}(1-q_\ell)^{\mathsf{K}_\ell-s+1}\right), \end{aligned} \tag{4.36}$$

where the equality in (4.36) follows from the fact that the keys are orthogonal to each other and they are uniformly distributed as in (4.29). Substituting (4.35) and (4.36) into (4.34), we have:

$$I\left(X_{(d_1,\ldots,d_K)};\mathbb{N}\right)\leq 0 \tag{4.37}$$

Using the fact that for any X, Y, $I(X;Y)\geq 0$, we have:

$$I\left(X_{(d_1,\ldots,d_K)};\mathbb{N}\right)= 0 \tag{4.38}$$

which proves that the rate $R_s(M,N,K)$ is *securely* achievable for non-uniform file popularity distributions.

4.5.2 Bounds on Optimal Expected Secure Rate

Next, we analyze the achievable rate of the secure caching scheme for large file size B. The achievable rate given by (4.30) is an instantaneous rate for the proposed scheme, given a cache storage allocation of M_ℓ at each user for the group of files $\mathbb{N}_\ell$. In the next theorem the expected rate of the secure scheme over all possible storage allocations, $\{M_\ell\}:\sum_{\ell=1}^{L}M_\ell=M$, is analyzed. This rate yields an upper bound on the optimal expected rate $R_s^*(M,\mathbb{N},K)$ of the secure caching problem.

Theorem 19. *For N files $\mathbb{N}$ partitioned to within popularity factor of p into L groups $\{\mathbb{N}_\ell:\ell=1,\ldots,L\}$ and K users, each with normalized cache size M, we have*

$$R_s^*(M,\mathbb{N},K)\leq \min_{\{M_\ell:\sum_\ell M_\ell=M\}}\sum_{\ell=1}^{L}\mathbb{E}\left[R_{s,\mathsf{dec}}(M_\ell,N_\ell,\mathsf{K}_\ell)\right] \tag{4.39}$$

where $R_{s,\mathsf{dec}}(M,N,K)$ is defined in (4.17), $N_\ell = |\mathbb{N}_\ell|$ and $\mathsf{K}_\ell = |\mathbb{K}_\ell|$. All expectations are with respect to the random number of users K_ℓ who request files in the set $\mathcal{N}_\ell$.

The proof of Theorem 19 is given in Appendix B.6. Similar to [99], each term in the sum in 4.39 corresponds to the rate of serving the users in each group $\mathbb{K}_\ell$. The sum is minimized over the choice of all possible storage allocations M_ℓ. The amount of storage allocated in the users' cache to each file as also the size of the keys in each group depends on the choice of storage allocation. A simple example could is allocating $M_\ell = M/L$ amount of storage to each group of files $\mathbb{N}_\ell$. However, even if each group is allocated the same amount of storage, the storage allocated to each file within a group is $M/(N_\ell L)$, which is a function of ℓ. Since each group contains different number of files, even this simple allocation ensures that most popular files are allocated more storage in the users' caches. The next theorem gives a lower bound on the optimal expected secure rate.

Theorem 20. *For N files $\mathbb{N}$ partitioned to within popularity factor of p into L groups $\{\mathbb{N}_\ell : \ell = 1,\dots,L\}$ and K users, each with normalized cache size M, we have*

$$R_s^*(M,\mathbb{N},K) \geq \frac{1}{cL}\sum_{\ell=1}^{L}\mathbb{E}\left[R_{s,\mathsf{dec}}(M,N_\ell,\mathsf{K}_\ell)\right] \tag{4.40}$$

where c is a strictly positive constant and $R_s(M,N,K)$ is defined in (4.17).

The proof of Theorem 20 is given in Appendix B.7. The theorem states that if the normalized cache storage size at each user is increased to ML and the proposed secure scheme is applied, then the achievable expected rate is at most cL times larger than the optimal expected rate for the original secure caching problem where each user has a cache size of M. For the case of uniform file popularities, we have $L = 1$ and Theorems 19 and 20 imply that

$$\frac{1}{c}R_{s,\mathsf{dec}}(M,N,K) \leq R_s^*(M,\mathbb{N},K) \leq R_{s,\mathsf{dec}}(M,N,K), \tag{4.41}$$

which in turn implies that the peak and expected rates in this case are approximately equal for the secure caching problem.

4.5.3 Empirical Results on Performance

We compare the performance of the proposed secure scheme with the insecure scheme of Maddah-Ali et.al [99] and the classical Least-Frequently-Used (LFU) Caching scheme. In LFU, each user caches the M most popular files in its cache and any request outside these M files is delivered via unicast to the requesting user. Thus, the expected rate of the LFU scheme is equal to the expected number of users with a request outside of the M most popular files. We consider a secure version of the LFU scheme where each user stores an unique key of size B bits and $M-1$ of the most popular files. The requests outside of the $M-1$ most popular files are coded with the key of the

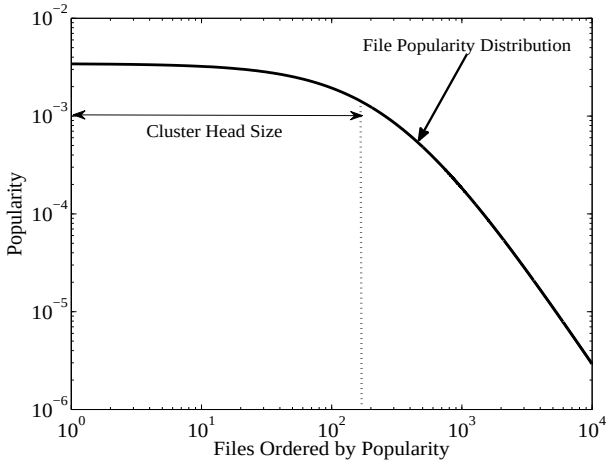


Figure 4.8: File popularities p_n for $N = 10000$ files. The M-Zipf distribution has a flattened head for the first approximately 800 files followed by a power law tail with an exponent of -2.

requesting user and delivered via unicast. For N files with popularity $p_1 > p_2 > \ldots > p_N$ and K users, each with a cache size of M, the expected rate of the secure LFU scheme is given by

$$R_s^{\text{LFU}} = K \sum_{n=M}^{N} p_n. \tag{4.42}$$

The popularity of files in a network can be empirically modeled using the Mandelbrot-Zipf distribution as in [99, 184]. The MZipF distribution is a heavy-tailed probability distribution whose pdf is given by

$$f(R; N; q; \gamma) = \frac{\frac{1}{(R+q)^\gamma}}{\sum_{i=1}^{N} \frac{1}{(i+q)^\gamma}},$$

where R indicates the $R-$th most popular file, q is the size of the *head*, where all the files belonging to the head are the most popular (and are of approximately equal popularity), and γ is a roll-off factor which determines how rapidly the file popularity drops past the head of the distribution. This is highlighted in Fig. 4.8, where a sample popularity distribution for $N = 10,000$ files is shown. Similar to the popularity distribution of files in the Netflix database, shown in [99], the distribution has a head consisting roughly of 800 most popular files and a power law tail with an exponent of -2. For an ordered set of files, the MZipF distribution gives the probability of a file being requested by an user, with the most popular files having highest request probability.

The comparison of the proposed secure scheme with the insecure scheme and secure LFU are shown in Fig. 4.9(a) and 4.9(b). For the proposed scheme we show an upper bound on the expected rate using Jensen's inequality as in [99] since $R(M, N, K)$ is s a concave function of K. The upper bound is given by

$$\min_{\{M_\ell\}:\sum_{\ell=1}^{L} M_\ell = M} \sum_{\ell=1}^{L} R\left(M_\ell, N_\ell, \mathbb{E}(\mathsf{K}_\ell)\right), \quad \text{where } \mathbb{E}(\mathsf{K}_\ell) = K \sum_{n \in \mathbb{N}_\ell} p_n.$$

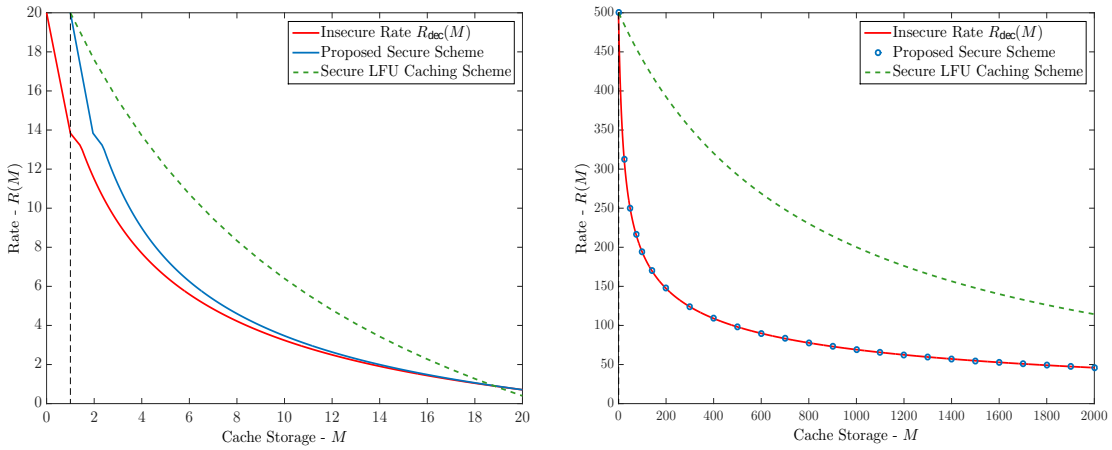


Figure 4.9: (M, R_s) trade-off for (a) $N = K = 20$ and p_n modeled by the M-Zipf distribution; (b) $N = 10,000$ files and $K = 500$ users and popularity distribution from Fig. 4.8

From Fig. 4.9(a), it can be seen that the rate of the proposed secure scheme outperforms the secure LFU scheme and is only marginally worse than the insecure scheme for a small number of files and users $N = K = 20$. For a large number of files, $N = 10,000$ and users, $K = 500$, and a file popularity distribution as shown in Fig. 4.8, the rate of the proposed scheme is asymptotically equal to that of the insecure scheme, as evidenced in Fig. 4.9(b). Thus we can see that the proposed scheme introduces security at a negligible cost, particularly for a large number of files ans users. The choice of the popularity factor p is left as a design choice. A reasonable choice is $\mathsf{p} = 2$, which was used in the simulations. Note that if $p_1 > p_2 \ldots > p_N$ and $\mathsf{p} = p_1/p_N$, then the system reduces to $L = 1$ i.e., a system which precludes the use of file popularities to allocate storage. Thus for practical applications, p should ideally be magnitudes of order smaller than p_1/p_N.

4.6 Directions of Future Research

Based on the results presented in the previous sections on the secure caching problem, some recent work has extended our initial results in this new paradigm e.g., *secrecy*[1] in cache-aided networks in the form of *secretive coded caching* [185] and secure D2D-assisted content delivery [186]. The following highlights a few research areas which are of practical interest in this paradigm.

- *Extension to Multiple Demands Per User*: In this work, we consider the secure caching prob-

[1] *Secrecy* is distinguished from *privacy* in that, for secretive communications, the data needs to secured even among the legitimate users in the system.

lem for the case of single requests from users at every transmission interval. However, the generalization to the case when users can demand multiple, say $L \geq 1$, files is an open problem. However, our results on the corresponding non-secure problem make it an avenue of interesting future work.

- *Extension to Multiple Requests Over Time*: Another area for future work is the case of security in delivering content for multiple requests over time i.e., security for an online coded caching scheme similar to the one in [100] which would require a key generation technique such that collection of keys over time by an eavesdropper cannot lead to information leakage.

- *Noisy Links and Multiple Eavesdroppers*: In the current treatment of the security problem, it is also interesting to note that the presence of multiple eavesdroppers would not alter the presented results since each eavesdropper would view the same multicast transmission which leaks no information about the files. This is due to the fact that we consider noiseless delivery in this model. The analysis of the problem for multiple eavesdroppers in the presence of noisy links is a direction of future research.

- *Security in D2D-assisted Delivery*: Another possible extension of the secure caching results would be to the case of caching with secure D2D-assisted delivery. Preliminary results were presented in [186]. However, the general problem of approximately characterizing the storage-rate trade-off for D2D-assisted secure delivery remains an open problem.

- *Closing the Gap to Optimal in Small Buffer Case*: Finally closing the gap between the achievable rate and the information theoretic optimal secure rate for $K > N$ in the regime $1 < M < \frac{(K-N)(N-1)}{KN} + 1$ for the centralized scheme and $1 < M < \frac{N-1}{N} + 1$ for the decentralized scheme, is an interesting open problem.

4.7 Summary

In this chapter, we have analyzed the problem of *secure* caching in the presence of an external wiretapper for both *centralized* and *decentralized* cache placement. We have proposed a key based secure caching strategy which is robust to compromise of users and keys. We have approximated the information theoretic optimal rate of the secure caching problem with novel upper and lower bounds. It has been shown that there is a constant multiplicative gap between the optimal and the achievable rates for the given scheme in case of both centralized and decentralized caching scenarios for most parameters of practical interest. We have shown that for large number of files and users, the secure bounds approach that of the non-secure case i.e., the cost of security in the system is negligible when the number of files and users increase. The results were then extended to the case of non-uniform file popularity where it was shown that similar advantages exist. Finally, scope of future work and extensions were also discussed.

Chapter 5

Fundamental Limits of Cloud and Cache-Aided Wireless Networks

In this chapter, we extend our analysis of cache-aided networks to the wireless domain under a multi-server setting. To this end, we study a cloud and cache-aided wireless network architecture in which edge-nodes (ENs), such as base stations, are connected to a cloud processor via dedicated fronthaul links, while also being endowed with caches. Cloud processing enables the centralized implementation of cooperative transmission strategies at the ENs, albeit at the cost of an increased latency due to fronthaul transfer. In contrast, the proactive caching of popular content at the ENs allows for the low-latency delivery of the cached files, but with generally limited opportunities for cooperative transmission among the ENs. The interplay between cloud processing and edge caching is addressed from an information-theoretic viewpoint by investigating the fundamental limits of a high Signal-to-Noise-Ratio (SNR) metric, termed normalized delivery time (NDT), which captures the worst-case latency for delivering any requested content to the users. The NDT is defined under the assumption of either serial or pipelined fronthaul-edge transmissions, and is studied as a function of fronthaul and cache capacity constraints. Transmission policies that encompass the caching phase as well as the transmission phase across both fronthaul and wireless, or edge, segments are proposed, with the aim of minimizing the NDT for given fronthaul and cache capacity constraints. Information-theoretic lower bounds on the NDT are also derived. Achievability arguments and lower bounds are leveraged to characterize the minimal NDT in a number of important special cases, including systems with no caching capability, as well as to prove that the proposed schemes achieve optimality within a constant multiplicative factor of 2 for all values of the problem parameters.

5.1 Cloud and Cache-Aided Wireless Networks

In this work, we present a latency centric analysis of cloud and cache-aided wireless edge networks. Moving the location of the caches closer to the *edge of the network* has the advantage of reducing the latency required for accessing and delivering users' requests. In particular, caching at the edge nodes (ENs), such as at macro or small-cell base stations, allows the delivery of content to mobile users with limited need for backhaul usage to connect to a remote content server (see [6] and references therein).

While potentially reducing delivery latency and backhaul load, edge caching generally limits the operation of ENs to non-cooperative transmission strategies. This is because, with edge caching, each EN formats its transmitted signal based only on its local cached content, which may only partially overlap with that of other ENs, hence preventing cooperative transmission schemes such as joint beamforming. The *localized* processing afforded by edge caching is in contrast to the *centralized* processing that is instead possible in network architectures in which the ENs are controlled by a *cloud* processor. An important example of this class of networks is the Cloud Radio Access Network (C-RAN) architecture, in which the ENs are connected to a cloud processor by means of so called *fronthaul* links. In a C-RAN, the signals transmitted by the ENs are produced at the cloud based on a direct connection to the content server and forwarded to the ENs on the fronthaul links. As such, cloud processing in C-RAN enables the implementation of cooperative transmission strategies across the ENs, but at the cost of a potentially large latency, owing to the time required for fronthaul transfer (see, e.g., [172, 187]).

Motivated by the complementary benefits highlighted above between cloud-based and edge-based architectures, in this work we consider a *cloud and cache-aided* wireless network architecture, which we term *Fog Radio Access Network* (F-RAN). In an F-RAN, as seen in Fig. 5.1, the ENs are connected to a cloud processor via dedicated fronthaul links, while also being endowed with caches that can be used to proactively store popular content [188]. Within this architecture, cloud processing enables the centralized implementation of cooperative transmission strategies by the ENs, albeit at the cost of an increased latency due to fronthaul transfer. In contrast, edge caching allows for the low-latency delivery of the cached files, but with generally limited cooperation among the ENs.

The design of F-RAN systems involves two key design questions: (*i*) *What to cache at the ENs?*; and (*ii*) *How to deliver the requested content across the fronthaul and wireless, or edge, segments?*. The two questions pertain to network functions, namely caching and delivery, that operate at nested time scales: while caches are updated only at the time scale over which popular content is expected to change, e.g., every night, delivery is performed in each transmission interval in order to satisfy the current users' requests from the content library. Nevertheless, the two questions are strongly intertwined since delivery strategies need to operate by leveraging the existing cached content, as well as cloud processing.

In order to address the design of F-RAN, in this work, we adopt as a performance metric the

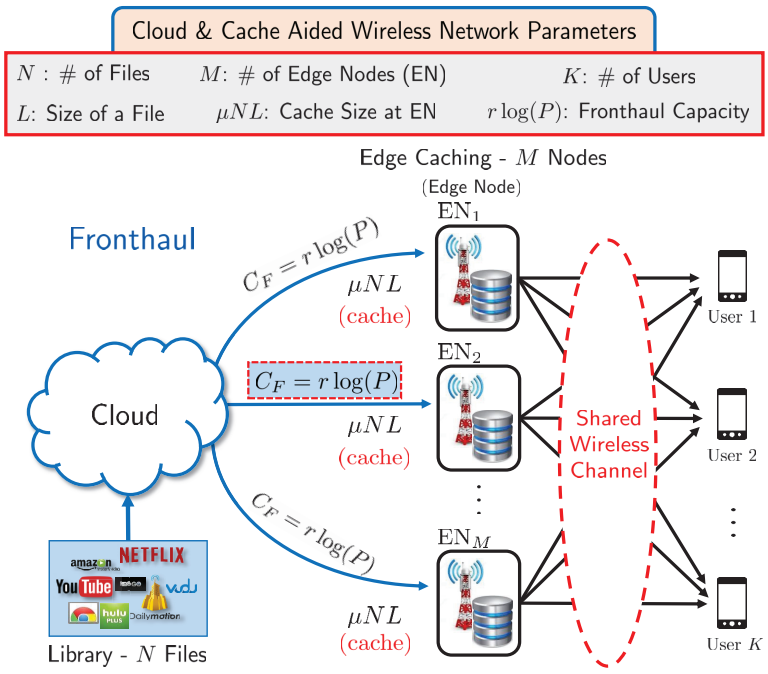
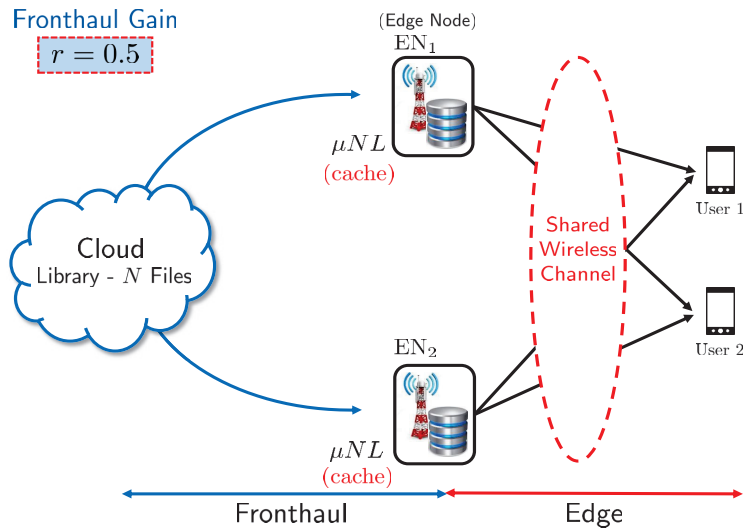


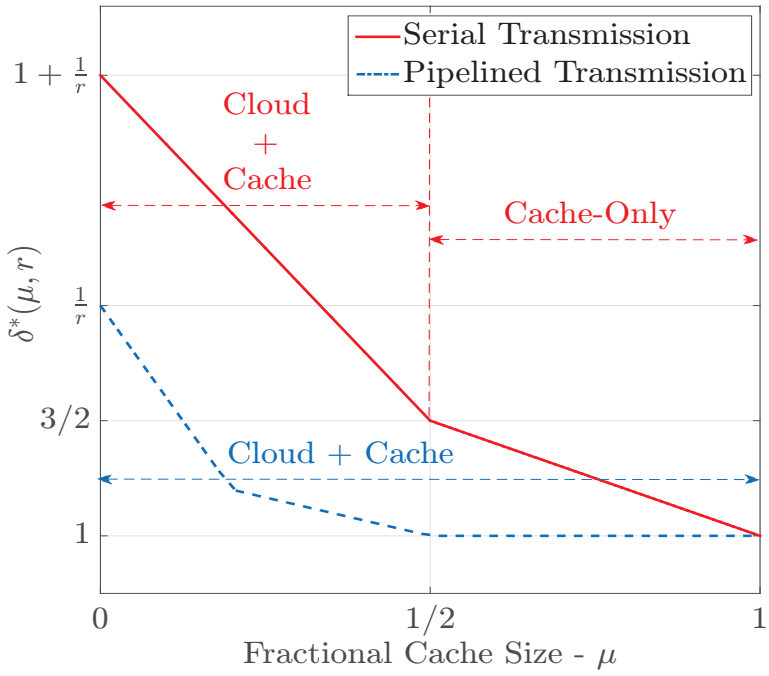
Figure 5.1: Information-theoretic model for a cloud and cache-aided wireless system, referred to as Fog-Radio Access Network (F-RAN).

worst-case latency accrued when serving any set of users' requests in a given transmission interval. We aim to characterize optimal caching and delivery strategies that minimize the delivery latency. To this end, we first focus on delivery strategies in which fronthaul and wireless channels are operated in a *serial* manner, so that the overall latency is the sum of the time spent for fronthaul communication between cloud and ENs and of the time required for wireless transmission from ENs to users. To enable analytical insights, we specifically propose a latency metric, termed *Normalized Delivery Time* (NDT), which captures the high signal-to-noise-ratio (SNR) ratio of the latency achievable in an F-RAN, with given fronthaul and caching limitations, as compared to that of an *ideal* system with unlimited caching capability and interference-free links to the users. We then extend the analysis to characterize the NDT for systems using delivery strategies in which fronthaul and wireless channels are operated in a *pipelined* (parallel) manner, so that fronthaul and wireless transmissions can take place at the same time (see, e.g., [189]).

Example 7. To exemplify the analysis put forth in this chapter, we briefly illustrate here the F-RAN set-up of Fig. 5.2(a), in which two ENs (labeled as EN_1 and EN_2) are deployed to serve two users over a shared wireless channel. The ENs are connected to the cloud via fronthaul links whose capacity scales with the SNR P of the wireless edge links as $r\log(P)$, with $r \geq 0$ being a parameter that characterizes the fronthaul capacity. We assume that there is a library of $N \geq 2$ popular files, each of a given size, and that each EN can cache at most a fraction $\mu \in [0,1]$ of the library content, where μ is defined as the *fractional cache size*. Full Channel State Information (CSI) is assumed as needed at all nodes. For this example, the *information-theoretically optimal* trade-off $\delta^*(\mu, r)$ between the NDT and the fractional cache size μ is shown in Fig. 5.2(b) for



(a)



(b)

Figure 5.2: (a) Information-theoretic model for an F-RAN with $M = 2$ ENs serving $K = 2$ users and a fronthaul gain $r = 0.5$; (b) Trade-off between the normalized delivery time (NDT) and the fractional cache size μ in the presence of full CSI at ENs, users and the cloud.

$r = 0.5$ for serial as well as for pipelined fronthaul-edge transmission. The NDT captures the worst-case latency required by the F-RAN to deliver all files requested by the users across both fronthaul and wireless segments. An NDT $\delta^*(\mu, r)$ indicates that an F-RAN with fractional cache

size μ and fronthaul gain r requires a total latency that is $\delta^*(\mu, r)$ times the time required by the mentioned ideal system with unlimited caching and no interference.

Among other conclusions, as illustrated in Fig. 5.2, the analysis presented in this chapter reveals that, for serial fronthaul-edge transmission, in the regime of low fronthaul capacity ($r \leq 0.5$), the latency due to fronthaul transfer makes cloud processing not useful in reducing the overall delivery latency if the cache capacity is large enough ($\mu \geq 1/2$). In contrast, for pipelined fronthaul-edge transmission, cloud processing is instrumental in obtaining the minimum delivery latency for all values of μ, even when the fronthaul capacity is small. This is because, with pipelined transmission, the ENs need not wait for the fronthaul transmission to be completed before communicating to the users on the edge links. For the same reason, pipelined fronthaul-edge transmission generally improves the NDT compared to serial transmission. In particular, even with partial caching, that is, with $\mu < 1$, the ideal NDT $\delta^* = 1$ is achievable with pipelined fronthaul-edge transmission, while this is not the case with serial transmission. More details can be found in Sections 5.6.1 and 5.7.4. □

Related Work

The line of work pertaining to the information-theoretic analysis of cache-aided wireless communication systems can be broadly classified into studies that consider caching at the end-users' devices or at the ENs. This research direction was initiated by [97, 98] (c.f. Chapter 2) for a set-up that consists of a multicast link with cache-aided receivers. This work demonstrates that coded multicasting enables global caching gains to be reaped, as opposed to the conventional local caching gains of uncoded transmission. Follow-up papers on related models with receiver-end caching include [99, 100, 104, 105, 123, 124, 128, 132–134, 136, 140, 176, 190]. The work in this chapter is instead inscribed in the parallel line of work that concerns caching at the ENs of a wireless network. A pioneering effort on this subject is [152], in which "femto-caching", that is caching at small-cell base stations, is introduced as a means to reduce backhaul usage and delivery latency. This and follow-up papers, including [106, 120, 121, 153, 159, 162, 178, 191], assume that cache-aided ENs are not allowed to cooperate on the basis of the cached content to mitigate or cancel mutual interference. In contrast, references [146, 192] investigate, from an information-theoretic viewpoint, an interference-limited wireless system with cache-aided ENs that can carry out coordinated transmission strategies, such as interference alignment, as well as cooperative transmission schemes, such as joint beamforming. Specifically, in [146, 192] an upper bound on the worst-case delivery latency, which is formulated in terms of the inverse of the degrees-of-freedom metric, is derived for $M = 3$ ENs and $K = 3$ users by proposing a specific caching and delivery policy. Upper and lower bounds on this metric are obtained in [147] by accounting for caching at both ENs and users, under the assumption of delivery strategies based on linear precoding on the wireless channel. Related works that focus on the optimization of signal processing strategies at the ENs can be found in [193–197]. This work was partially presented in [149, 198–200] and a brief informal summary was provided in [150]. Reference [201] extends the lower bounds in [198] to the case with caching also at the receivers.

5.1.1 Main Contributions

The main contributions of this chapter are summarized as follows.

- An information-theoretic model of a cloud and cache-aided system, termed F-RAN, is presented, along with a novel latency metric, namely the *normalized delivery time* (NDT). The NDT measures the worst-case latency required to deliver an arbitrary vector of requests to the users in the high-SNR regime, as compared to an ideal system with full caching and no interference on the wireless channel. In the network operation assumed for most of the chapter, fronthaul and wireless segments are operated in a serial manner, with fronthaul transmission preceding transmission on the wireless channel (see Fig. 5.1 and Fig. 5.3). As a result, the delivery latency has two components, namely the fronthaul latency incurred due to fronthaul transfer from the cloud to the ENs and the edge latency required for transmission from the ENs to the users over the shared wireless channel.

- Under the assumptions of uncoded inter-file caching (but allowing for arbitrary intra-file coding) and full CSI at all nodes, we develop general information-theoretic lower bounds on the minimum NDT for an F-RAN with any number of ENs and users as a function of the caching and fronthaul limitations as defined by the parameters μ and r, respectively. The lower bounds are derived by adopting cut-set arguments that are tailored to the set-up at hand that includes both fronthaul and wireless segments.

- We present a general upper bound on the NDT of an arbitrary F-RAN by leveraging file-splitting between cloud-aided and cache-aided transmission strategies. For the cloud-aided scheme, we consider a novel *soft-transfer* fronthauling approach, inspired by the standard operation of C-RAN [172], which is based on the transmission of quantized encoded signals on the fronthaul links. For cache-aided strategies, we leverage both coordination via interference alignment and cooperation via joint beamforming at the ENs based on cached content. A number of alternative strategies are also considered for reference, including the conventional *hard-transfer* of uncached content on the fronthaul links.

- The proposed achievable schemes are shown to achieve the minimum NDT to within a factor of 2 for all values of the system parameters.

- The minimum NDT is characterized exactly in a number of important special cases. These include: cloud-only F-RANs, also known as C-RAN; cache-only F-RANs, that is the cache-aided wireless system studied in [146, 192] for extremal values of fractional cache size μ; and general F-RAN models with both cloud processing and caching for the case when the number of users exceeds the number of ENs in the low fronthaul regime.

- We present specific case studies for the 2×2 F-RAN, where the minimum NDT is completely characterized by the proposed bounds (see Fig. 5.2(b)), and for the 3×3 F-RAN, where the minimum NDT is partially characterized by leveraging the proposed lower bounds and achievability results presented in [146, 192].

- We define and investigate an F-RAN model in which the fronthaul and wireless edge segments can be operated in a pipelined, or parallel, manner. We show that, in comparison to serial transmission, pipelined fronthaul-edge transmission can improve the NDT by a multiplicative factor of at most 2.
- We present a general lower bound on the minimum NDT for the the pipelined fronthaul-edge transmission model as well as achievable schemes which leverage block-Markov encoding along with file-splitting between cloud and cache-aided transmission strategies.
- We characterize the minimum NDT for cloud-only F-RAN with pipelined fronthaul-edge transmission. Furthermore, for a general $M \times K$ F-RAN with pipelined fronthaul-edge transmission, the proposed schemes are shown to achieve the minimum NDT to within a factor of 2 for all values of system parameters. We present the case study for the 2×2 F-RAN for which the minimum NDT is completely characterized by the proposed bounds (see Fig. 5.2(b)).

The remainder of the chapter is organized as follows. Section 5.2 presents the information-theoretic model for a general $M \times K$ F-RAN and introduces the NDT metric for serial fronthaul-edge transmission. Lower bounds on the NDT for an F-RAN are derived in Section 5.3, while achievable schemes are proposed in Section 5.4. In Section 5.5, we present the mentioned finite-gap and exact characterization of the minimum NDT. Section 5.6 elaborates on two use cases, namely the 2×2 and 3×3 F-RAN models. Section 5.7 discusses the F-RAN model with pipelined fronthaul-edge transmissions. General upper and lower bounds on the minimum NDT for this model are presented along with a finite-gap characterization of the minimum NDT. Section 5.8 highlights some of the open problems and directions for future work, while Section 5.9 concludes the chapter.

Notation: For any two integers a and b with $a \leq b$, we define the notation $[a : b] \triangleq \{a, a+1, \ldots, b\}$. We also use the notation $b \in [a, c]$ to imply that b lies in the interval $a \leq b \leq c$ for any a, b, c. Furthermore, $b \in (a, c]$ denotes $a < b \leq c$. We use the notation $x \in \{a, b, \ldots, c\}$ to denote that the variable x takes the values in the set $\{a, b, \ldots, c\}$. We define the function $(x)^+ \triangleq \max\{0, x\}$. The set of all positive integers is denoted by $\mathbb{N}^+$ and the set of all complex numbers is denoted by $\mathbb{C}$.

5.2 System Model and Performance Metrics

In this section, we first present a model for the cloud and cache aided F-RAN system under study. Then, we introduce the normalized delivery time (NDT) metric, along with a number of remarks to provide additional context on the adopted model and performance metric.

5.2.1 System Model

We consider an $M \times K$ F-RAN, shown in Fig. 5.1, where M ENs serve a total of K users through a shared wireless channel. The ENs can cache content from a library of N files, $F_1, \ldots, F_N$, where each file is of size L bits, for some $L \in \mathbb{N}^+$. Formally, the files F_n are independent and identically distributed (i.i.d.) as:

$$F_n \sim \text{Unif}\left\{1, 2, \ldots, 2^L\right\}, \ \ \forall n \in [1:N]. \tag{5.1}$$

Each EN is equipped with a cache in which it can store μNL bits, where the fraction μ, with $\mu \in [0,1]$, is referred to as the *fractional cache size* and can be interpreted as the fraction of each file which can be cached at an EN. The cloud has full access to the library of N files, and each EN is connected to the cloud by a fronthaul link of capacity of C_F bits per symbol, where a symbol refers to a channel use of the downlink wireless channel.

In a transmission interval, each user $k \in [1:K]$ requests one of the N files from the library. The demand vector is denoted by $\mathbf{D} \triangleq (d_1, \ldots, d_K) \in [1:N]^K$. This vector is known at the beginning of a transmission interval by both cloud and ENs, which attempt to satisfy the users' demands within the lowest possible latency. As illustrated in Fig. 5.1, we assume a serial operation over the fronthaul and wireless segments, whereby the cloud first communicates to the ENs and then the ENs transmit on the shared wireless channel to the users. As a result, the total latency is the sum of fronthaul and edge latencies (see Remark 16 for additional discussion on this point).

All the nodes have access to the global CSI about the wireless channels $\mathbf{H} = \left\{ \{h_{km}\} : \substack{k=[1:K] \\ m=[1:M]} \right\}$, where $h_{km} \in \mathbb{C}$, denotes the wireless channel between user $k \in [1:K]$ and EN_m, $m \in [1:M]$. The coefficients are assumed to be drawn independent and identically distributed (i.i.d.) from a continuous distribution and to be time-invariant within each transmission interval.

As mentioned, the design of the system entails the definition of caching and delivery policies, which are formalized next for the case of serial fronthaul-edge transmission. Various generalizations of the definition below are presented in Sections 5.3, 5.4 and 5.7.

Definition 7 (*Policy*)**.** A caching, fronthaul, edge transmission, and decoding policy $\pi = (\pi_c, \pi_f, \pi_e, \pi_d)$ is characterized by the following functions.

a) Caching Policy π_c: The caching policy at each edge node EN_m, $m \in [1:M]$, is defined by a function $\pi_c^m(\cdot)$ that maps each file F_n to its cached content $S_{m,n}$ as

$$S_{m,n} \triangleq \pi_c^m\left(F_n\right), \ \ \forall n \in [1:N]. \tag{5.2}$$

The mapping is such that $H(S_{m,n}) \leq \mu L$ in order to satisfy the cache capacity constraints. The overall cache content at EN_m is given by $S_m = (S_{m,1}, S_{m,2}, \ldots, S_{m,N})$. Note that the caching policy π_c allows for arbitrary coding within each file, but it does not allow for inter-file coding. Furthermore, the caching policy is kept fixed over multiple transmission intervals and is thus agnostic to the demand vector $\mathbf{D}$ and the global CSI $\mathbf{H}$.

b) Fronthaul Policy π_f: A fronthaul policy is defined by a function $\pi_f(\cdot)$, which maps the set of files $F_{[1:N]}$, the demand vector $\mathbf{D}$ and CSI $\mathbf{H}$ to the fronthaul message

$$\mathbf{U}_m^{T_F} = (U_m[t])_{t=1}^{T_F} = \pi_f^m \left(\{F_{[1:N]}\}, \mathbf{D}, \mathbf{H}\right), \tag{5.3}$$

which is transmitted to EN_m via the fronthaul link of capacity C_F bits per symbol. Here, T_F is the duration of the fronthaul message. In keeping with the definition of fronthaul capacity C_F, all time intervals, including T_F, are normalized to the symbol transmission time on the downlink wireless channel. Thus, the fronthaul message cannot exceed $T_F C_F$ bits.

c) Edge Transmission Policy π_e: After fronthaul transmission, each edge node EN_m follows an edge transmission policy $\pi_e^m(\cdot)$ to map the demand vector $\mathbf{D}$ and global CSI $\mathbf{H}$, along with its local cache content and the received fronthaul message, to output a codeword

$$\mathbf{X}_m^{T_E} = (X_m[t])_{t=1}^{T_E} = \pi_e^m\Big(S_m, \mathbf{U}_m^{T_F}, \mathbf{D}, \mathbf{H}\Big), \tag{5.4}$$

which is transmitted to the users on the shared wireless link. Here, T_E is the duration of the transmission on the wireless channel, on which an average power constraint of P is imposed for each codeword $\mathbf{X}_m^{T_E}$. Note that the fronthaul policy, π_f and the edge transmission policy, π_e, can adapt to the instantaneous demands and CSI at each transmission interval, unlike the caching policy, π_c, which remains unchanged over multiple transmission intervals.

d) Decoding Policy π_d: Each user $k \in [1:K]$, receives a channel output given by:

$$\mathbf{Y}_k^{T_E} = (Y_k[t])_{t=1}^{T_E} = \sum_{m=1}^{M} h_{km}\mathbf{X}_m^{T_E} + \mathbf{n}_k^{T_E}, \tag{5.5}$$

where the noise $\mathbf{n}_k^{T_E} = (n_k[t])_{t=1}^{T_E}$ is such that $n_k[t] \sim \mathcal{CN}(0,1)$ is i.i.d. across time and users. Each user $k \in [1:K]$, implements a decoding policy $\pi_d(\cdot)$, which maps the channel outputs, the receiver demands and the channel realization to the estimate

$$\widehat{F}_{d_k} \triangleq \pi_d^k\Big(\mathbf{Y}_k^{T_E}, d_k, \mathbf{H}\Big) \tag{5.6}$$

of the requested file F_{d_k}. The caching, fronthaul, edge transmission and decoding policies together form the policy $\pi = (\pi_c^m, \pi_f^m, \pi_e^m, \pi_d^k)$ that defines the operation of the F-RAN system. The probability of error of a policy π is defined as

$$P_e = \max_{\mathbf{D}} \max_{k\in[1:K]} \mathbb{P}\left(\widehat{F}_{d_k} \neq F_{d_k}\right), \tag{5.7}$$

which is the worst-case probability of decoding error measured over all possible demand vectors $\mathbf{D}$ and over all users $k \in [1:K]$. A sequence of policies, indexed by the file size L, is said to be *feasible* if, for almost all channel realizations $\mathbf{H}$, i.e., with probability 1, we have $P_e \to 0$ when $L \to \infty$.

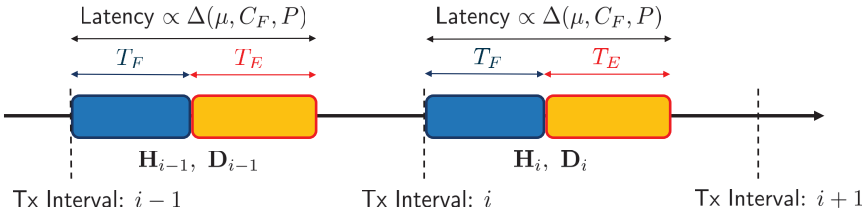


Figure 5.3: Illustration of the delivery latency within each transmission interval for the F-RAN under study with serial fronthaul-edge transmission.

5.2.2 Performance Metric: Normalized Delivery Time

We next define the proposed performance metric of normalized delivery time (NDT) by first introducing the notion of delivery time per bit.

Definition 8 (*Delivery time per bit*)**.** A *delivery time per bit* $\Delta(\mu, C_F, P)$ is achievable if there exists a sequence of feasible policies such that

$$\Delta(\mu, C_F, P) = \limsup_{L\to\infty} \frac{T_F + T_E}{L}. \tag{5.8}$$

The delivery time per bit accounts for the latency within each transmission interval as illustrated in Fig. 5.3. Specifically, the total latency is given by the sum of the fronthaul and edge contributions, namely T_F and T_E, respectively. In order to obtain a vanishing probability of error, as required by Definitions 7 and 8, the latencies T_F and T_E need to scale with L, and it is this scaling that is measured by (5.8). We also observe that the definition of delivery time per bit in (5.8) is akin to the completion time studied in [202, 203] for standard channel models, such as broadcast and multiple access channels.

While $\Delta(\mu, C_F, P)$ generally depends on the power level P, as well as on the fronthaul capacity C_F and fractional cache size μ, we next define a more tractable metric that reflects the latency performance in the high SNR regime. To this end, we let the fronthaul capacity scale with the SNR parameter P as $C_F = r\log(P)$, where r measures the multiplexing gain of the fronthaul links.

Definition 9 (*NDT*)**.** For any achievable $\Delta(\mu, C_F, P)$, with $C_F = r\log(P)$, the *normalized delivery time* (NDT), is defined as

$$\delta(\mu, r) = \lim_{P\to\infty} \frac{\Delta(\mu, r\log(P), P)}{1/\log P}. \tag{5.9}$$

Moreover, for any given pair (μ, r), the minimum NDT is defined as

$$\delta^*(\mu, r) = \inf\left\{\delta(\mu, r) : \delta(\mu, r) \text{ is achievable}\right\}. \tag{5.10}$$

Remark 13 (*Operational significance of NDT*)**.** In (5.9), the delivery time per bit (5.8) is normalized by the term $1/\log P$. The latter is the delivery time per bit in the high SNR regime for

an ideal baseline system with no interference and unlimited caching, in which each user can be served by a dedicated EN which has locally stored all the files. An NDT of δ^* hence indicates that the worst-case time required to serve any possible request vector $\mathbf{D}$ is δ^* times larger than the time needed by this ideal baseline system. □

Remark 14 (*Cache-Only F-RAN and Cloud-Only F-RAN*)**.** Throughout this chapter, we will often consider separately the two important special cases of cache-only F-RAN and cloud-only F-RAN. The former corresponds to the case in which the fronthaul capacity is zero, i.e., $r = 0$, while the latter, which amounts to a C-RAN system (see Section 4.1), is obtained by setting $\mu = 0$. We observe that, in a cache-only F-RAN, as studied in [146, 192], it is required that the collective cache size of the M ENs be large enough to completely store the entire library of N files in order to obtained a finite worst-case delivery latency. This requires the condition $M \times \mu NL \geq NL$, i.e., $\mu \geq 1/M$, holds. Therefore, for this case, it suffices to focus on the range $\mu \in [1/M, 1]$ of fractional cache capacity. □

Remark 15 (*NDT vs. DoF*)**.** For the specific case of a cache-only F-RAN, the NDT in (5.10) is proportional to the inverse of the more conventional degrees of freedom (DoF) metric DoF(μ) defined in [146, 192]. Specifically, we have the relationship $\delta^*(\mu, 0) = K/\text{DoF}(\mu)$. □

We show next that the NDT is convex in the fractional size μ for any value of the fronthaul gain $r \geq 0$. The proof follows from a *file-splitting and cache-sharing* argument, whereby files are split into two fractions, with the two fractions being served by different policies that share the cache resources and whose delivery times add up to yield the overall NDT.

Lemma 4 (*Convexity of Minimum NDT*)**.** *The minimum NDT, $\delta^*(\mu, r)$, is a convex function of μ for every value of $r \geq 0$.*

Proof. Consider any two feasible policies π_1 and π_2, where policy π_i requires a fractional cache capacity and fronthaul gain pair (μ_i, r) and achieves an NDT of $\delta(\mu_i, r)$ for $i = 1, 2$. Given an F-RAN system with cache storage capacity $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$ and fronthaul gain r for some $\alpha \in [0, 1]$, we consider the following policy. Each file is split into two parts of sizes αL and $(1-\alpha)L$, respectively, where the first is delivered by using policy π_1 and the second by using policy π_2. Note that a fractional cache capacity μ is sufficient to support the operation of this policy. The NDT achieved by this policy can be computed as $\delta(\mu, r) = \alpha\delta(\mu_1, r) + (1-\alpha)\delta(\mu_2, r)$ since, by (5.8) and (5.9), the NDT is proportional to the file size. Applying this argument to two policies that achieve minimum NDTs $\delta^*(\mu_i, r)$ for $i = 1, 2$ proves the inequality

$$\delta^*\left(\alpha\mu_1 + (1-\alpha)\mu_2, r\right) \leq \alpha\delta^*(\mu_1, r) + (1-\alpha)\delta^*(\mu_2, r), \tag{5.11}$$

since the right-hand side of (5.11) is achievable by file-splitting. This shows the joint convexity of the minimum NDT $\delta^*(\mu, r)$ as a function of μ for every value of $r \geq 0$. □

Remark 16 (*Pipelined Fronthaul-Edge Transmission*)**.** As discussed, the system model presented in this section adopts a serial delivery model, whereby fronthaul transmission is followed by

edge transmission as seen in Fig. 5.3. Alternatively, a pipelined delivery model could be considered in which the ENs can simultaneously receive on fronthaul links and transmit on the wireless channel. In this case, each edge node EN_m starts transmitting at the beginning of the transmission interval using an edge transmission policy $\pi^m_{\mathsf{P},e}(\cdot)$, such that, at any time instant t, the EN maps the demand vector $\mathbf{D}$, the global CSI $\mathbf{H}$, the local cache content S_m and the fronthaul messages received up to time $t-1$, to the transmitted signal at time t as

$$X_m[t] = \pi^m_{\mathsf{P},e}\Big(S_m, \left(U_m[1], U_m[2], \ldots, U_m[t-1]\right), \mathbf{D}, \mathbf{H}\Big), \;\; t \in [1:T] \tag{5.12}$$

The overall latency is given by T and the NDT can be defined in a manner analogous to Definition 9, namely

$$\delta_{\mathsf{P}}(\mu, r) = \lim_{P\to\infty} \limsup_{L\to\infty} \frac{T}{L/\log P}. \tag{5.13}$$

We observe that the serial fronthaul-edge transmission policies in Definition 7 are included as special cases in the class of pipelined fronthaul-edge transmission schemes. As a result, the minimum NDT $\delta^*_{\mathsf{P}}(\mu, r)$ under pipelined operation can be no larger than that under serial operation. Furthermore, following the same arguments as in Lemma 4, the minimum NDT $\delta^*_{\mathsf{P}}(\mu, r)$ can be seen to be a convex function of μ for any $r \geq 0$. We provide a detailed study of the pipelined delivery model in Section 5.7. □

5.3 Lower Bound on minimum NDT

In this section, we provide a general lower bound on the minimum NDT for the $M \times K$ F-RAN described in the previous section. The main result is stated in the following theorem.

Theorem 21 (*Lower Bound on Minimum NDT*)**.** *For an F-RAN with M ENs, each with a fractional cache size $\mu \in [0,1]$, K users, a library of $N \geq K$ files and a fronthaul capacity of $C_F = r\log(P)$ bits per symbol, the minimum NDT is lower bounded as*

$$\delta^*(\mu, r) \geq \delta_{LB}(\mu, r), \tag{5.14}$$

where $\delta_{LB}(\mu, r)$ is the minimum value of the following linear program (LP)

$$\begin{aligned} \text{minimize} \;\; & \delta_F + \delta_E && (5.15)\\ \text{subject to}: \;\; & \ell\delta_E + (M-\ell)^+ r\delta_F \geq K - (M-\ell)^+(K-\ell)^+\mu, && (5.16)\\ & \delta_F \geq 0, \; \delta_E \geq 1, && (5.17) \end{aligned}$$

where (5.16) *is a family of constraints with $\ell \in [0 : \min\{M, K\}]$.*

Proof. The proof of Theorem 21 is presented in Appendix C.1. □

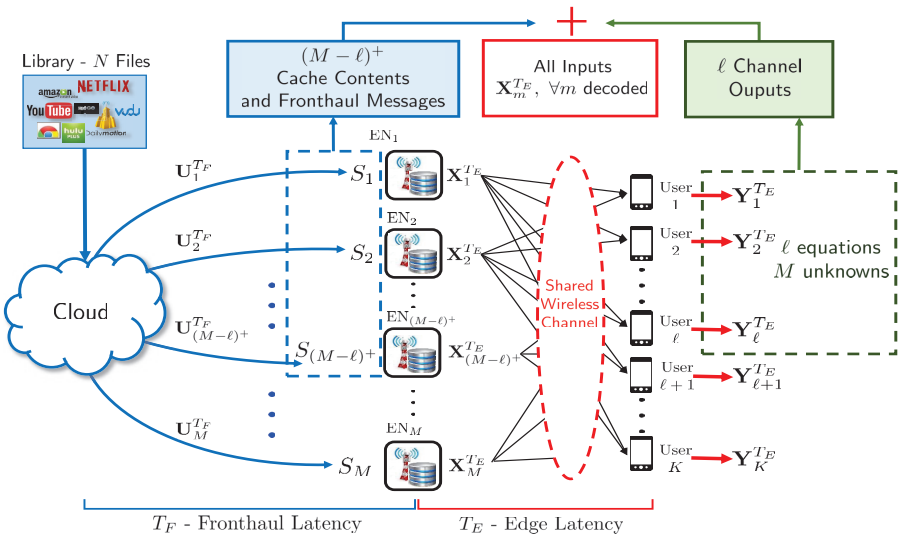


Figure 5.4: Illustration of the proof of Theorem 21.

In Theorem 21, and henceforth, we refer to *fronthaul-NDT* as the normalized delivery time for fronthaul transmission, that is,

$$\delta_F = \lim_{P\to\infty} \frac{T_F \log(P)}{L}, \tag{5.18}$$

and *edge-NDT* as the normalized delivery time for edge transmission, that is,

$$\delta_E = \lim_{P\to\infty} \frac{T_E \log(P)}{L}. \tag{5.19}$$

Note that the NDT (5.9) is the sum of fronthaul-NDT and edge-NDT i.e., $\delta = \delta_F + \delta_E$. Theorem 21 hence provides a lower bound on the minimum NDT by means of bounds on linear combinations of fronthaul and edge NDTs.

The proof of the main bound (5.16) is based on a cut-set-like argument, which is illustrated in Fig. 5.4. Specifically, it can be argued that, for all sequence of feasible policies guaranteeing a vanishing probability of error, in the high-SNR regime, any K requested files must be decodable with low error probability from the received signal of ℓ users along with the cache contents and fronthaul messages of the remaining $(M-\ell)^+$ ENs. This is because, any $\ell \leq \min\{M, K\}$ received signals $\mathbf{Y}_{[1:\ell]}^{T_E}$ are functions of M channel inputs $\mathbf{X}_{[1:M]}^{T_E}$, which in turn are functions of the M user caches and their corresponding fronthaul messages $\mathbf{U}_{[1:M]}^{T_F}$. Thus, using these ℓ signals and the contents of $(M-\ell)^+$ caches, $S_{[1:(M-\ell)^+]}$ and associated fronthaul messages $\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}$, all the inputs can be almost surely decoded using the invertible linear system of the form of (5.5), neglecting the noise in the high-SNR regime. The theorem is proved by carefully bounding the joint entropy of these random variables, which upper bounds the amount of information that can be reliably conveyed in the given time intervals T_E and T_F or edge-NDT δ_E and fronthaul-NDT δ_F.

We next present a sequence of corollaries that specialize the lower bound of Theorem 21 to the settings of cache-only and cloud-only F-RANs (see Remark 14).

Corollary 4 (*Lower Bound for Cache-Only F-RAN*)**.** *For an $M \times K$ cache-only F-RAN ($r = 0$) with $\mu \in [1/M, 1]$, the NDT is lower bounded as*

$$\delta^*(\mu, 0) \geq \max_{\ell \in [1:\min\{M,K\}]} \frac{K - (M-\ell)^+(K-\ell)^+\mu}{\ell}. \tag{5.20}$$

Proof. The proof of Corollary 4 follows directly by substituting $r = 0$ in constraint (5.16) in Theorem 21 and noting that any lower bound on the optimal value of the LP in Theorem 21 is also a valid lower bound on the NDT. Varying the parameter $\ell \in [1 : \min\{M, K\}]$ leads to the family of lower bounds in Corollary 4. □

Corollary 5 (*Lower Bound for Cloud-Only F-RAN*)**.** *For an $M \times K$ cloud-only F-RAN ($\mu = 0$), the NDT is lower bounded as*

$$\delta^*(0, r) \geq \frac{K}{\min\{M, K\}} + \frac{K}{Mr}. \tag{5.21}$$

Proof. Summing the constraints obtained from (5.16) by setting $\ell = M$ and $\ell = 0$ yields the following lower bound on the optimal value of the LP:

$$\delta^*(0, r) \geq \delta_E + \delta_F \geq \frac{K}{M} + \frac{K}{Mr}. \tag{5.22}$$

Instead, summing the constraint in (5.16) with $\ell = 0$ and the constraint $\delta_E \geq 1$ in (5.17) yields the following lower bound:

$$\delta^*(0, r) \geq \delta_E + \delta_F \geq 1 + \frac{K}{Mr}. \tag{5.23}$$

Combining the bounds in (5.22) and (5.23) yields

$$\delta^*(0, r) \geq \max\left(\frac{K}{M} + \frac{K}{Mr}, 1 + \frac{K}{Mr}\right) = \frac{K}{\min\{M, K\}} + \frac{K}{Mr}, \tag{5.24}$$

which concludes the proof. □

5.4 Upper Bounds on the Minimum NDT

In this section, we expound on upper bounds on the minimum NDT by considering the performance of specific policies. We proceed by first investigating cache-aided and cloud-aided transmission strategies separately, which are then combined to obtain a cloud and cache-aided policy by means of file-splitting and cache-sharing (see Lemma 4).

5.4.1 Cache-Aided Policies

We consider first cache-aided policies that do not use cloud resources and hence operate even when there is no fronthaul infrastructure, i.e., when $r = 0$. We specifically focus on the two extremal scenarios in which $\mu = 1$, so that all ENs can cache the entire library of files, and $\mu = 1/M$, so that the library can be fully cached as long as different portions of it are stored at distinct ENs.

Example 8 (*Cache-Aided EN Cooperation via Zero-Forcing Beamforming* ($\mu = 1$))**.** Assume that we have an equal number of ENs and users, i.e., $M = K$, that the number of files is $N \geq K$, and that $\mu = 1$ so that every EN can store the entire file library. Under these assumptions, given the worst-case request vector in which all K users request different files, the resulting system can be treated as a multi-antenna broadcast channel with M co-located transmit antennas. This is because all ENs share any set of requested files. Therefore, given that $N \geq K$, transmitter cooperation in the form of zero-forcing (ZF) beamforming can be carried out with high probability with respect to the channel realizations, yielding interference-free transmission to the $K = M$ users. As a result, the delivery latency is the same as in the interference-free ideal system and hence an NDT equal to 1 is achievable. □

Generalizing the cache-aided cooperative approach described in the example, the following lemma provides an upper bound on the minimum NDT for the case $\mu = 1$.

Lemma 5 (*Achievable NDT with Cache-Aided EN Cooperation*)**.** *For an F-RAN with fractional cache size $\mu = 1$ and any $r \geq 0$, the NDT is upper bounded as $\delta^*(\mu = 1, r) \leq \delta_{\mathsf{Ca-ZF}}$, where*

$$\delta_{\mathsf{Ca-ZF}} = \frac{K}{\min\{M, K\}} \tag{5.25}$$

is achieved by means of ZF-beamforming based on the cached files.

Proof. Following Example 8, the ENs employ ZF-beamforming to serve the users' requests. Note that the worst-case demand can be easily seen to be any vector $\mathbf{D}$ of distinct files. In fact, any other vector that contains the same file for multiple users can always be delivered with the same latency by treating the files as being different. Using ZF, a sum-rate of $\min\{M, K\}\log(P)$, neglecting $o(\log(P))$ terms, can be achieved [204]. Thus, the delivery time per bit (5.8) achieved by this scheme is approximately, that is, neglecting $o(\log(P))$ terms, given by

$$\Delta\left(\mu = 1, 0, P\right) = \frac{K/\log(P)}{\min\{M, K\}}, \tag{5.26}$$

which, by definition of the NDT (Definition 9), yields an achievable NDT $\delta_{\mathsf{Ca-ZF}} = K/\min\{M, K\}$, hence concluding the proof. □

Example 9 (*Cache-Aided EN Coordination via Interference Alignment*)**.** We consider now the other extreme case in which each EN has fractional cache capacity $\mu = 1/M$. To fix the ideas,

we focus in this example on a system with $M = 3$ ENs, $K = 3$ users and $N = 3$ files, namely $\{A, B, C\}$, each of size L bits. With $\mu = 1/3$, each EN can store one full file or L bits from the library in its cache. Different caching policies can be put in place. A first, naive, approach would be to cache the same file, say A, at each EN. However, this approach yields an infinite latency for any request vector that contains files other than A and hence also for the worst-case vector. A better solution would be to place each file in a different cache, e.g., files A, B and C in the caches of EN_1, EN_2 and EN_3, respectively. In this case, for the worst-case request vector in which users request different files, the wireless channel can be operated as a 3×3 user interference channel, for which a sum-DoF of $3/2$ can be achieved [79, 173, 205], yielding an NDT equal to 2 (see Remark 15).

As pointed out in [146, 192], more sophisticated caching strategies in which files are split into multiple subfiles are generally are able to outperform the reference schemes discussed thus far. Specifically, divide each file into three non-overlapping subfiles of equal length, e.g., for file A we have $A = (A_1, A_2, A_3)$. Now, the cache placement at the three ENs is as follows

$$S_1 = (A_1, B_1, C_1); \;\; S_2 = (A_2, B_2, C_2); \;\; S_3 = (A_3, B_3, C_3).$$

Under this placement scheme, each EN has one fragment from a file requested by a user under any request vector $\mathbf{D}$. For the worst-case demand vector in which each user requests a different file, the edge transmission policy can follow the interference alignment scheme for an X-channel of [80, 146, 173, 192]. We recall that the X-channel refers to a model in which each transmitter intends to communicate a dedicated independent message to each receiver under interference from the other transmitters. This yields a sum-DoF of $9/5$ and therefore an NDT of $5/3 < 2$. □

Following the example above, the following lemma provides an upper bound on the minimum NDT that is obtained by means of cache-aided coordination strategies based on interference alignment for the case $\mu = 1/M$.

Lemma 6 (*Achievable NDT with Cache-Aided EN Coordination*)**.** *For an F-RAN with fractional cache size $\mu = 1/M$ and any $r \geq 0$, the NDT is upper bounded as $\delta^*(\mu, 0) \leq \delta_{\mathsf{Ca-IA}}$, where*

$$\delta_{\mathsf{Ca-IA}} = \frac{M + K - 1}{M} \tag{5.27}$$

is achievable by means of interference alignment.

Proof. Following the example above, which is inspired by [146, 192], each file is split into M non-overlapping fragments $F_n = (F_{n,1}, F_{n,2}, \ldots, F_{n,M})$, each of size L/M bits. The fragment $F_{n,m}$ is stored in the cache of EN_m for $n \in [1 : N]$. Thus, the cache storage for each EN is NL/M bits and $\mu = 1/M$. For any file d_k is requested by a user k, each of the ENs has a fragment $F_{d_k,m}$ to transmit to the user. For the worst-case demand vector in which all users request different files (see proof of Lemma 5), the $M \times K$ system then becomes an X-channel for which a reliable sum-rate of $(MK/(M+K-1)) \log(P)$, neglecting $o(\log(P))$ terms, is achievable by interference

alignment [80, 173]. Thus, the achievable delivery time per bit, in Definition 8, is approximately given by

$$\Delta\left(\mu=\frac{1}{M},0,P\right)=\frac{M+K-1}{M\log(P)}, \tag{5.28}$$

yielding an NDT equal to $\delta_{\mathsf{Ca-IA}}=(M+K-1)/M$. This concludes the proof of the Lemma. □

5.4.2 Cloud-Aided Policies

We now move to considering cloud-aided policies that neglect the caches at the ENs and hence operate even in the case in which the ENs have no storage capabilities, that is, when $\mu=0$. We first discuss a more conventional hard-transfer fronthauling approach, whereby the fronthaul is used to send the requested files in raw form to the ENs. Then, we elaborate on the the soft-transfer scheme that is typical of C-RAN, in which quantized coded signals are transferred on the fronthaul links.

Example 10 (*Cloud-Aided Hard-Transfer Fronthauling*)**.** Consider an F-RAN with $M=3$ ENs and $K=3$ users with a library of $K=3$ files $\{A,B,C\}$, each of size L bits. We are interested in developing delivery strategies that only rely on cloud processing and fronthaul transfer, while neglecting the use of caches. We focus again on the worst-case in which each user requests a different file, i.e., $\mathbf{D}=(d_1,d_2,d_3)=(A,B,C)$. With hard-transfer fronthaul, the cloud sends files, or subfiles, over the fronthaul links to each EN, which then encodes the signal to be transmitted on the shared wireless channel. A first approach would be to send all three files and hence $3L$ bits, to each EN, so as the enable the ENs to perform cooperative ZF-beamforming on the wireless channel. Using the fact that the fronthaul capacity is $C_F=r\log(P)$ bits per symbol, the fronthaul delivery time is $T_F=3L/(r\log(P))$, yielding a fronthaul-NDT equal to $\delta_F=3/r$. Since, with ZF, the edge-NDT is $\delta_E=1$ as discussed in Example 8, the overall NDT achieved by this strategy is $\delta=1+3/r$. Alternatively, the cloud can divide each file into three fragments as discussed in Example 9 and send the fragments (A_i,B_i,C_i) to EN_i over the corresponding fronthaul link. In this case, the fronthaul delivery time is $T_F=L/(r\log(P))$ yielding a fronthaul-NDT of $\delta_F=1/r$. The ENs then transmit on wireless channel using interference alignment for an X-channel, achieving an edge-NDT of $\delta_E=5/3$, as seen in Example 9. Thus the achievable NDT with this approach is $\delta=5/3+1/r$. Based on the available fronthaul gain r, the cloud can choose the policy which yields the minimum NDT. In this case, when $r\leq 3$ the interference alignment-based scheme should be utilized, while the ZF-based strategy is to be preferred otherwise. □

Generalizing the previous example, the following theorem gives an upper bound on the minimum NDT, which can be achieved by the use of hard-transfer fronthauling.

Theorem 22 (*Achievable NDT with Cloud-Aided Hard-Transfer Fronthauling*)**.** *For an $M\times K$ F-RAN with each EN having a fractional cache size $\mu\in[0,1]$ and a fronthaul gain of $r\geq 0$, the*

NDT is upper bounded as $\delta^*(\mu, r) \leq \delta_{\mathsf{Cl-Hf}}$*, where*

$$\delta_{\mathsf{Cl-Hf}} = \min\left\{\frac{K}{\min\{M,K\}} + \frac{K}{r},\ \frac{M+K-1}{M} + \frac{K}{Mr}\right\}, \tag{5.29}$$

which is achieved by means of hard-transfer fronthauling.

Proof. Following the discussion in Example 10, we consider the selection between two different strategies to prove Theorem 22.

5.4.2.1 Cloud-Aided EN Cooperation via ZF Beamforming

In the first strategy, the cloud transmits all the requested files to each EN over the fronthaul links. Thus, for any request vector $\mathbf{D}$, the cloud needs to transmit KL bits to each EN. Since the fronthaul links have capacity $C_F = r\log(P)$ each, the fronthaul delivery time is $T_F = KL/(r\log(P))$, yielding a fronthaul-NDT of $\delta_F = K/r$. Furthermore, ZF-based EN cooperation achieves an edge-NDT of $\delta_E = K/\min\{M,K\}$ as shown in Lemma 5. Thus, the achievable NDT with this strategy is

$$\delta_F + \delta_E = \frac{K}{\min\{M,K\}} + \frac{K}{r}. \tag{5.30}$$

5.4.2.2 Cloud-Aided EN Coordination via Interference Alignment

With this second strategy, for the K requested files $F_{d_1}, F_{d_2}, \ldots, F_{d_K}$, the cloud splits each file into M non-overlapping fragments $F_{d_k} = (F_{d_k,1}, F_{d_k,2}, \ldots, F_{d_k,M})$, for $k \in [1:K]$, where each fragment is of size L/M bits. The fragments $F_{[d_1,\ldots,d_K],m}$ are transmitted to EN_m for $m \in [1:M]$. Thus, for a fronthaul capacity of $C_F = r\log(P)$ bits per symbol, the fronthaul delivery time is $T_F = KL/(Mr\log(P))$, yielding a fronthaul-NDT of $\delta_F = K/(Mr)$. As seen in Lemma 6, using an X-channel interference alignment scheme achieves an edge-NDT of $\delta_E = (M+K-1)/M$. Thus, the NDT

$$\delta_F + \delta_E = \frac{M+K-1}{M} + \frac{K}{Mr}, \tag{5.31}$$

is achievable via interference alignment. For a given fronthaul gain of r, the cloud then chooses the transmission strategy which achieves the minimum NDT between (5.30) and (5.31), which yields (5.29), hence concluding the proof. □

We now move to the consideration of the soft-transfer fronthauling approach.

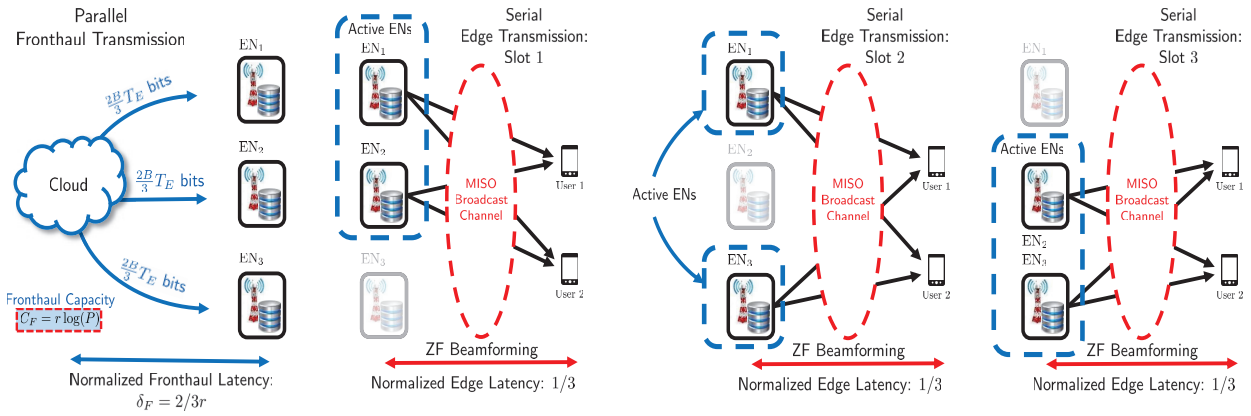


Figure 5.5: Illustration of the proposed cloud-aided soft-transfer fronthauling acheme with $M = 3$ ENs and $K = 2$ users.

Example 11 (*Cloud-Aided Soft-Transfer Fronthauling*)**.** Consider an F-RAN with M ENs and $K = M$ users. With soft-transfer fronthauling, as first proposed in [206], the cloud implements ZF-beamforming and quantizes the resulting encoded signals. Using a resolution of $\log(P)$ bits per downlink baseband sample, it can be shown that the effective SNR in the downlink scales proportionally to the power P (see Appendix C.2 and [206]). As a result, this scheme entails a fronthaul transmission time T_F that equals the edge transmission time T_E of the ZF-beamforming scheme, namely $T_E = L/(\log(P))$, multiplied by the time needed to carry each baseband sample on the fronthaul link, namely $\log(P)/(r\log(P))$, yielding the NDT $\delta_{\mathsf{Cl-Sf}} = 1 + 1/r$. Comparing with the NDT obtained in Example 10 by means of hard-transfer fronthauling, we see that soft-transfer fronthaul yields a lower NDT. □

The following theorem generalizes the previous example to give an upper bound on the minimum NDT, which is achieved by a cloud-aided policies using soft-transfer fronthauling.

Theorem 23 (*Achievable NDT with Cloud-Aided Soft-Transfer Fronthauling*)**.** *For an $M \times K$ F-RAN with each EN having a fractional cache size $\mu \in [0, 1]$ and a fronthaul gain $r \geq 0$, the NDT is upper bounded as $\delta^*(\mu, r) \leq \delta_{\mathsf{Cl-Sf}}$, where*

$$\delta_{\mathsf{Cl-Sf}} = \frac{K}{\min\{M, K\}} + \frac{K}{Mr}, \tag{5.32}$$

which can be achieved by means of soft-transfer fronthauling.

Proof. The formal proof of Theorem 23 is presented in Appendix C.2. A proof sketch outlining the main ideas is provided below. □

For the case $M \leq K$, the main arguments follow in a manner similar to Example 11. The case $M \geq K$ instead requires a novel delivery approach that is based on the parallel transmission on the fronthaul links of quantized encoded signals that are sent using time-sharing on the wireless

channel. We explain the scheme at hand with an example for an F-RAN with $M = 3$ ENs serving $K = 2$ users, which is illustrated in Fig. 5.5. We first list all possible $\binom{3}{2} = 3$ clusters of 2 ENs. Each of the 3 clusters of ENs is scheduled to transmit sequentially for $1/3$ of the total edge delivery time T_E on the wireless channel. The signals to be transmitted by each cluster on the wireless channels are sent in parallel on the fronthaul links by the cloud by means of a soft-transfer fronthauling strategy. Specifically, each EN participates in 2 clusters and hence it needs to receive only $2T_E/3$ quantized samples from the cloud on the fronthaul link. Thus, using a resolution of $B = \log(P)$ bits per sample as in Example 11, a fronthaul latency of $T_F = 2BT_E/(3C_F) = 2T_E/3r$ is achieved. This yields a fronthaul-NDT of $\delta_F = 2\delta_E/(3r)$ for a total achievable NDT of $\delta_{\mathsf{Cl-Sf}} = 1 + 2/(3r)$.

Remark 17 (*Hard vs. Soft-Transfer Fronthaul*)**.** Comparing the NDT of soft-transfer fronthauling in Theorem 23 with the achievable NDT for hard-transfer fronthauling in Theorem 22, we see that the achievable NDT in Theorem 23 is strictly lower, demonstrating that soft-transfer fronthauling is to be preferred when the goal is to minimize the NDT. □

5.4.3 Cache and Cloud-Aided Policies

Here, we propose a general upper bound on the minimum NDT for an F-RAN with M ENs, K users and $N \geq K$ files, which is attained by combining the cache-aided strategy discussed in Section 5.4.1 and the cloud-aided soft-transfer fronthaul policy of Section 5.4.2 by means of file-splitting and cache-sharing (see Lemma 4). Note that the choice of soft-transfer fronthauling over hard-transfer fronthauling is motivated by Remark 17.

Theorem 24 (*Achievable NDT via Cloud and Cache-Aided Policies*)**.** *For an $M \times K$ F-RAN with a fronthaul gain of $r \geq 0$, the NDT is upper bounded as $\delta^*(\mu, r) \leq \delta_{\mathsf{Ach}}(\mu, r)$, where, for fractional cache size $\mu \in [0, 1/M]$ we define*

$$\delta_{\mathsf{Ach}}(\mu, r) = \min \left\{ \begin{array}{c} (M + K - 1)\mu + (1 - \mu M) \left[\dfrac{K}{\min\{M, K\}} + \dfrac{K}{Mr} \right], \\ \dfrac{K}{\min\{M, K\}} + \dfrac{(1 - \mu)K}{Mr} \end{array} \right\}, \tag{5.33}$$

and for fractional cache size $\mu \in [1/M, 1]$ we have

$$\delta_{\mathsf{Ach}}(\mu, r) = \min \left\{ \begin{array}{c} \dfrac{K}{\min\{M, K\}} \left(\dfrac{\mu M - 1}{M - 1} \right) + (1 - \mu) \dfrac{M + K - 1}{M - 1}, \\ \dfrac{K}{\min\{M, K\}} + \dfrac{(1 - \mu)K}{Mr} \end{array} \right\}. \tag{5.34}$$

Proof. The theorem is proved by considering the NDT of a policy that performs file-splitting and cache-sharing, as described in the proof of Lemma 4, between cache-aided and cloud-aided

schemes. Specifically, or $\mu \in [0, 1/M]$, we use the cache-aided policy described in Lemma 6, yielding $\delta_{\mathsf{Ca-IA}}$, for a fraction of the files equal to μM and the cloud-aided soft-transfer fronthauling policy described in Theorem 23, yielding $\delta_{\mathsf{Cl-Sf}}$, for the remaining $(1 - \mu M)$ fraction of the files. This requires a fractional cache capacity of $\mu M \times (1/M) + (1 - \mu M) \times 0 = \mu$, since the two schemes at hand use fractional cache size $1/M$ and 0 respectively. Moreover, the achievable NDT is

$$\delta^{'}_{\mathsf{Ach}}(\mu, r) = (\mu M)\delta_{\mathsf{Ca-IA}} + (1 - \mu M)\delta_{\mathsf{Cl-Sf}}, \tag{5.35}$$

which equals the first term in (5.33). In a similar manner, for $\mu \in [1/M, 1]$, we use the cache-aided policy described in Lemma 6, yielding $\delta_{\mathsf{Ca-IA}}$, for a fraction $M(1-\mu)/(M-1)$ of the files and the cache-aided policy described in Lemma 5, yielding $\delta_{\mathsf{Ca-ZF}}$, for the remaining $(\mu M - 1)/(M - 1)$ fraction of files. This requires a fractional cache size of $M(1 - \mu)/(M - 1) \times (1/M) + (\mu M - 1)/(M-1) \times 1 = \mu$ since the schemes at hand use fractional cache size of $1/M$ and 1 respectively. The achievable NDT is

$$\delta^{''}_{\mathsf{Ach}}(\mu, r) = \frac{M(1 - \mu)}{(M - 1)}\delta_{\mathsf{Ca-IA}} + \frac{(\mu M - 1)}{(M - 1)}\delta_{\mathsf{Ca-ZF}}, \tag{5.36}$$

which equals the first term in (5.34). Finally, for fractional cache size $\mu \in [0, 1]$, the NDT

$$\delta^{'''}_{\mathsf{Ach}}(\mu, r) = \mu\delta_{\mathsf{Ca-ZF}} + (1 - \mu)\delta_{\mathsf{Cl-Sf}} \tag{5.37}$$

is achieved by file-splitting between the cache-aided policy described in Lemma 5, yielding $\delta_{\mathsf{Ca-ZF}}$, for a fraction μ of the files and the cloud-aided soft transfer fronthaul policy of Theorem 23, yielding $\delta_{\mathsf{Cl-Sf}}$, for the remaining $(1 - \mu)$ fraction of the files. Note that this requires a fractional cache size of $\mu \times 1 + (1 - \mu) \times 0 = \mu$ since the schemes at hand use fractional cache size of 1 and 0 respectively. The NDT (5.37) equals the second term in both (5.33) and (5.34). Choosing the minimum NDT among $\delta^{'}_{\mathsf{Ach}}, \delta^{''}_{\mathsf{Ach}}, \delta^{'''}_{\mathsf{Ach}}$ yields the upper bound $\delta_{\mathsf{Ach}}(\mu, r)$ in (5.33)-(5.34). This completes the proof of Theorem 24. □

5.5 Characterization of the Minimum NDT

Based on the lower and upper bounds presented in Sections 5.3 and 5.4, in this section, we show that the proposed achievable schemes in Section 5.4 are optimal in a number of important special cases, including cloud-only F-RANs, also known as C-RAN; cache-only F-RANs for extremal values of fractional cache size μ; and general F-RAN models with both cloud processing and caching for the case when the number of users exceeds the number of ENs in the low fronthaul regime. Furthermore, we present a constant factor approximation of the minimum NDT, $\delta^*(\mu, r)$, for all values of problem parameters, which shows that the proposed achievable schemes are approximately optimal to within a factor of at most 2. To proceed we first consider separately cache-only and cloud-only F-RAN and then study the general F-RAN model.

5.5.1 Minimum NDT for Cache-Only F-RAN

The following theorem characterizes the minimum NDT for a cache-only F-RAN ($r = 0$) for extremal values of the fractional cache size i.e., for $\mu \in \{1/M, 1\}$. We recall that with $\mu \leq 1/M$, the minimum NDT is unbounded (see Remark 14).

Theorem 25 (*Minimum NDT for Cache-Only F-RAN*)**.** *For an $M \times K$ F-RAN with a fronthaul gain $r = 0$, the minimum NDT given by*

$$\delta^*(\mu, 0) = \begin{cases} \delta_{\mathsf{Ca-IA}} & \text{for } \mu = 1/M, \\ \delta_{\mathsf{Ca-ZF}} & \text{for } \mu = 1, \end{cases} \tag{5.38}$$

where $\delta_{\mathsf{Ca-IA}}$ can be achieved by means of EN coordination via interference alignment (see (5.27)*) and $\delta_{\mathsf{Ca-ZF}}$ can be achieved by EN cooperation via ZF-beamforming (see* (5.25)*).*

Proof. The proof of Theorem 25 is provided in Appendix C.3. □

The result indicates that, in a cache-only F-RAN, the proposed converse in Corollary 4 is tight at extremal values of fractional cache size μ, and that cache-aided EN cooperation and coordination as described in Examples 8 and 9, are optimal for $\mu = 1$ and $\mu = 1/M$, respectively.

5.5.2 Minimum NDT for Cloud-Only F-RAN

The following theorem gives the minimum NDT for a cloud-only F-RAN ($\mu = 0$), showing the optimality of soft-transfer fronthauling.

Theorem 26 (*Minimum NDT for Cloud-Only F-RAN*)**.** *For an $M \times K$ F-RAN with $\mu = 0$, the minimum NDT is characterized as*

$$\delta^*(0, r) = \delta_{\mathsf{Cl-Sf}} \tag{5.39}$$

for $r \geq 0$ which can be achieved by soft-transfer fronthauling (see (5.32)*).*

Proof. The proof follows directly from the lower bound on the minimum NDT for cloud-only F-RANs, presented in Corollary 5 and from the achievable NDT presented in Theorem 23 that uses soft-transfer fronthauling. □

5.5.3 Approximate Characterization of the Minimum NDT for a Cache and Cloud-Aided F-RAN

We next provide an approximate characterization of the minimum NDT for a general $M \times K$ F-RAN by showing that the lower bound in Theorem 21 and the upper bound in Theorem 24, are

within a constant multiplicative gap equal to 2, independent of problem parameters for all regimes of fractional cache size μ and fronthaul gain r.

Theorem 27 (*Minimum NDT for a General F-RAN*)**.** *For a general $M \times K$ F-RAN, we have*

$$\frac{\delta_{\mathsf{Ach}}(\mu, r)}{\delta^*(\mu, r)} \leq 2, \tag{5.40}$$

for $\mu \in [1/M, 1]$ when $r = 0$ (cache-only F-RAN) and for $\mu \in [0, 1]$ when $r > 0$ (cloud and cache-aided F-RAN).

Proof. The proof of Theorem 27 is given in Appendix C.4. □

We finally provide another exact characterization of the NDT, in addition to the results in Theorems 25 and 26 for $r = 0$ and $\mu \in \{1/M, 1\}$ and for $\mu = 0$ respectively. Specifically, in the low cache memory regime in which $\mu \in [0, 1/M]$, when the number of ENs is smaller than the number of users, i.e., $M \leq K$, and the fronthaul gain is small i.e., $r \leq 1/(M-1)$, the following theorem gives the minimum NDT.

Theorem 28 (*Minimum NDT for F-RAN with Low Fronthaul and Cache Size*)**.** *For an $M \times K$ F-RAN with $M \leq K$ and with each EN having a fractional cache size $\mu \in [0, 1/M]$ and a fronthaul gain of $r \in (0, 1/(M-1)]$, the minimum NDT is given as*

$$\delta^*(\mu, r) = (M + K - 1)\mu + \frac{K\left(1 - \mu M\right)}{M}\left(1 + \frac{1}{r}\right). \tag{5.41}$$

Proof. The proof of Theorem 28 is provided in Appendix C.5. □

Remark 18. When considering a cache-aided F-RAN ($r = 0$), the system studied in this chapter becomes a special case of the system considered in [147], which is a cache-aided system with caching at both ENs and users. The authors in [147] show that, under the constraint of linear precoding strategies for transmission over the wireless channel, the optimal sum-DoF can be characterized to within a factor of 2. Theorem 27 shows that the factor 2 approximation of the minimum NDT, and hence of the sum-DoF, as seen in Remark 15, holds over a larger class of precoding schemes, including non-linear transmission strategies, and that it extends to cloud and cache-aided F-RANs. □

Remark 19 (*Sub-Packetization of Files*)**.** The caching and delivery schemes designed for the cache-only systems studied in [146] and [147] are based on techniques which require splitting each file into a number of sub-packets which increases *exponentially* in the number of ENs. In contrast, in this work, we use file-splitting strategies between only two of the schemes discussed in Section 5.4.1. Since the schemes require either no sub-packetization or in the case of the X-Channel based EN coordination scheme, file-splitting into M fragments, the schemes in this chapter require a number of file splits which is *linear* in the number of ENs. □

5.6 Case Studies

In this section, we elaborate on two specific examples of F-RANs with $M = K = 2$ and $M = K = 3$, for which we provide conclusive or approximate characterizations of the NDT. The discussion here highlights the proposed achievable schemes that obtain different operating points on the minimum NDT trade-off curve.

5.6.1 2×2 F-RAN

In this section, we provide the complete characterization of the minimum NDT of an F-RAN with $M = 2$ ENs and $K = 2$ users and we offer insights on optimal delivery policies.

Corollary 6. *For an F-RAN with $M = 2$ ENs, $K = 2$ users and $N \geq 2$ files, the minimum NDT is characterized as*

• *Cache-Only F-RAN* $(r = 0)$*:*

$$\delta^*(\mu, r) = 2 - \mu. \tag{5.42}$$

• *Low Fronthaul* $(r \in (0, 1])$*:*

$$\delta^*(\mu, r) = \max\left(1 + \mu + \frac{1 - 2\mu}{r}, 2 - \mu\right). \tag{5.43}$$

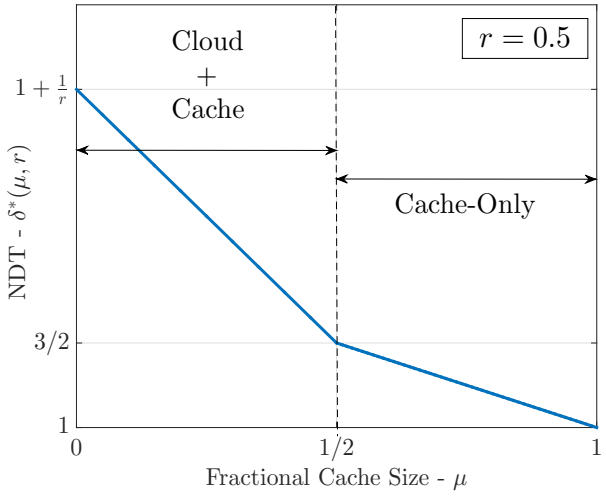
• *High Fronthaul* $(r > 1)$*:*

$$\delta^*(\mu, r) = 1 + \frac{1 - \mu}{r}. \tag{5.44}$$

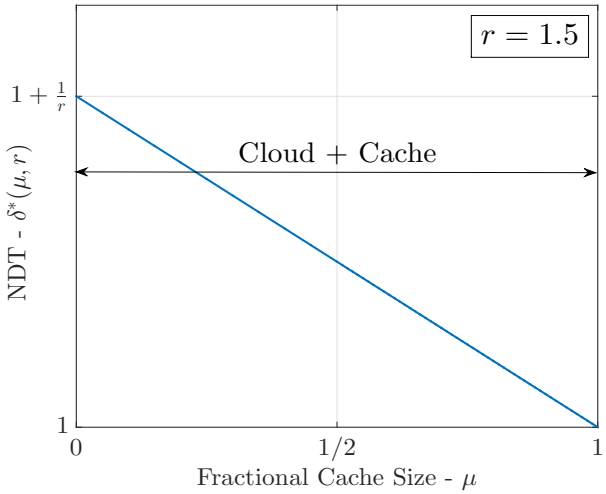
Proof. For the converse, please see Appendix C.6. The achievability is discussed below. □

Corollary 6 presents a complete characterization of the minimum NDT for a 2×2 F-RAN for all regimes of fractional cache size μ and fronthaul gain r, which is illustrated in Fig. 5.6(a)-5.6(b). Corollary 6 identifies three distinct regimes for the fronthaul gain r, namely cache-aided F-RAN i.e., $r = 0$, *a low fronthaul regime* with $r \in (0, 1]$ and a *high fronthaul regime* with $r \geq 1$. In the high fronthaul regime, the use of both cloud and caching resources are necessary to achieve the minimum NDT, while in the low fronthaul regime, if the cache size is sufficiently large, namely if $\mu \geq 1/2$, it is enough to leverage cache-only resources to achieve the minimum NDT. To further discuss these points, we next detail the policies that can achieve the minimum NDT.

Cache-Only F-RAN $(r = 0)$: For the cache-only 2×2 F-RAN, the minimum NDT is achieved by file-splitting and cache-sharing between the cache-aided EN coordination scheme that yields $\delta_{\mathsf{Ca-IA}}$



(a)



(b)

Figure 5.6: Minimum NDT for an F-RAN with $M = K = 2$: (a) low fronthaul regime, here $r = 0.25$; and (b) high fronthaul regime, here $r = 1.5$. The labels "Cache" and "Cloud" refer to the achievable schemes.

in (5.27), which requires $\mu = 1/2$, and the cache-aided EN cooperation strategy that achieves $\delta_{\mathsf{Ca-ZF}}$ in (5.25) and requires $\mu = 1$. Therefore, from (5.36), we have

$$\delta^*(\mu, r) = \delta_{\mathsf{Ach}}(\mu, 0) = 2 - \mu. \tag{5.45}$$

Low Fronthaul ($r \in (0, 1]$): In the low fronthaul regime, when the fractional cache size satisfies $\mu \leq 1/2$, the minimum NDT is achieved by file-splitting and cache-sharing between the cloud-aided soft-transfer fronthaul scheme that yields $\delta_{\mathsf{Cl-Sf}}$ in (5.32) with $\mu = 0$ and the cache-aided

EN coordination strategy that yields $\delta_{\mathsf{Ca-IA}}$ in (5.27) with $\mu = 1/2$. Therefore, from (5.35), using $M = K = 2$, we have

$$\delta^*(\mu, r) = \delta_{\mathsf{Ach}}(\mu, r) = 1 + \mu + \frac{1-2\mu}{r}. \tag{5.46}$$

Instead, for the high cache memory regime, $\mu \in [1/2, 1]$, the minimum NDT is given by (5.45) and is achieved by the cache-aided EN coordination.

High Fronthaul ($r \geq 1$): In this regime, the minimum NDT can be achieved by file-splitting and cache-sharing between the cloud-aided soft-transfer fronthaul scheme that yields $\delta_{\mathsf{Cl-Sf}}$ in (5.32) with $\mu = 0$, and the cache-aided EN cooperation strategy that yields $\delta_{\mathsf{Ca-ZF}}$ in (5.25) with $\mu = 1$. Therefore, from (5.37), we have

$$\delta^*(\mu, r) = \delta_{\mathsf{Ach}}(\mu, r) = 1 + \frac{1-\mu}{r}. \tag{5.47}$$

5.6.2 3×3 F-RAN

Here, we provide a partial characterization of the minimum NDT of an F-RAN with $M = 3$ ENs and $K = 3$ users by leveraging the results presented in the previous sections and the achievable schemes presented in [146, 192]. In a similar manner to Corollary 6, the following corollary distinguishes different fronthaul regimes, namely cache-only ($r = 0$), low fronthaul ($r \in (0, 1/2]$), intermediate fronthaul ($r \in [1/2, 2]$) and high fronthaul ($r > 2$).

Corollary 7. *For an F-RAN with $M = 3$ ENs, $K = 3$ users and $N \geq 3$ files, the minimum NDT is characterized as:*

• *Cache-Only F-RAN ($r = 0$):*

$$\delta^*(\mu, r) = \begin{cases} 5/3 & \text{for } \mu = 1/3, \\ 3/2 - \mu/2 & \text{for } \mu \in [2/3, 1], \end{cases}$$

$$\delta^*(\mu, r) \begin{cases} \geq \max\left(3 - 4\mu, \dfrac{3-\mu}{2}\right) \\ \leq 13/6 - 3\mu/2 \end{cases} \quad \text{for } \mu \in [1/3, 2/3]. \tag{5.48}$$

• *Low Fronthaul ($r \in (0, 1/2]$):*

$$\delta^*(\mu, r) = \begin{cases} 1 + 2\mu + \dfrac{1-3\mu}{r} & \text{for } \mu \in [0, 1/3], \\ 3/2 - \mu/2 & \text{for } \mu \in [2/3, 1], \end{cases}$$

$$\delta^*(\mu, r) \begin{cases} \geq \max\left(3 - 4\mu, \dfrac{3-\mu}{2}\right) \\ \leq 13/6 - 3\mu/2 \end{cases} \quad \text{for } \mu \in [1/3, 2/3]. \tag{5.49}$$

• *Intermediate Fronthaul 1* ($r \in [1/2, 6/7]$)*:*

$$\delta^*(\mu, r) \begin{cases} \geq 1 + \frac{2}{3}\mu + \frac{3-7\mu}{r} & \\ \leq 1 + 2\mu + \frac{1-3\mu}{r} & \text{for } \mu \in [0, 1/3], \end{cases}$$

$$\delta^*(\mu, r) \begin{cases} \geq \max\left(1 + \frac{2}{3}\mu + \frac{3-7\mu}{3r}, \frac{3-\mu}{2}\right) & \\ \leq 13/6 - 3\mu/2 & \text{for } \mu \in [1/3, 2/3], \end{cases}$$

$$\delta^*(\mu, r) = 3/2 - \mu/2, \qquad \text{for } \mu \in [2/3, 1]. \tag{5.50}$$

• *Intermediate Fronthaul 2* ($r \in [6/7, 2]$)*:*

$$\delta^*(\mu, r) \begin{cases} \geq \max\left(1 + \frac{2}{3}\mu + \frac{3-7\mu}{3r}, \frac{3-\mu}{2}\right) & \\ \leq 1 + \frac{\mu}{4} + \frac{2-3\mu}{2r} & \text{for } \mu \in [0, 2/3], \end{cases}$$

$$\delta^*(\mu, r) = 3/2 - \mu/2, \qquad \text{for } \mu \in [2/3, 1]. \tag{5.51}$$

• *High Fronthaul* ($r \geq 2$)*:*

$$\delta^*(\mu, r) = 1 + \frac{1-\mu}{r}, \qquad \text{for } \mu \in [0, 1]. \tag{5.52}$$

Proof. The proof is provided in Appendix C.7. We note here that the achievability leverages the scheme proposed in [146, 192] that requires $\mu = 2/3$ and $r = 0$. □

Corollary 7 provides a partial characterization of the minimum NDT of a 3×3 F-RAN by identifying upper and lower bounds for all values of μ and r, as well as conclusive results for specific regimes of the parameters. To aid the interpretation of the main results in Corollary 7, Fig. 5.7 shows the bounds on the NDT presented in Corollary 7 for four values of r, namely $\{0.25, 0.75, 1.25, 2\}$, which lie in the different regimes defined in Corollary 7. The figures partition the values of μ into two distinct intervals: for smaller values of μ, the policy used in the achievability proof of Corollary 7 leverages both cloud and cache resources, whereas for larger values of μ, cache-only resources are employed for transmission on the wireless channel as briefly discussed next.

As illustreated in Fig. 5.7(a), in the *low fronthaul* regime of $r \leq 1/2$, file-splitting between cache- and cloud-based schemes is optimal for $\mu \leq 1/3$ as proved in Theorem 28. Instead, for larger cache storage, using cloud resources in addition to cache resources may only provide a marginal decrease of the NDT. A similar behavior is observed also for *intermediate fronthaul*, as seen in

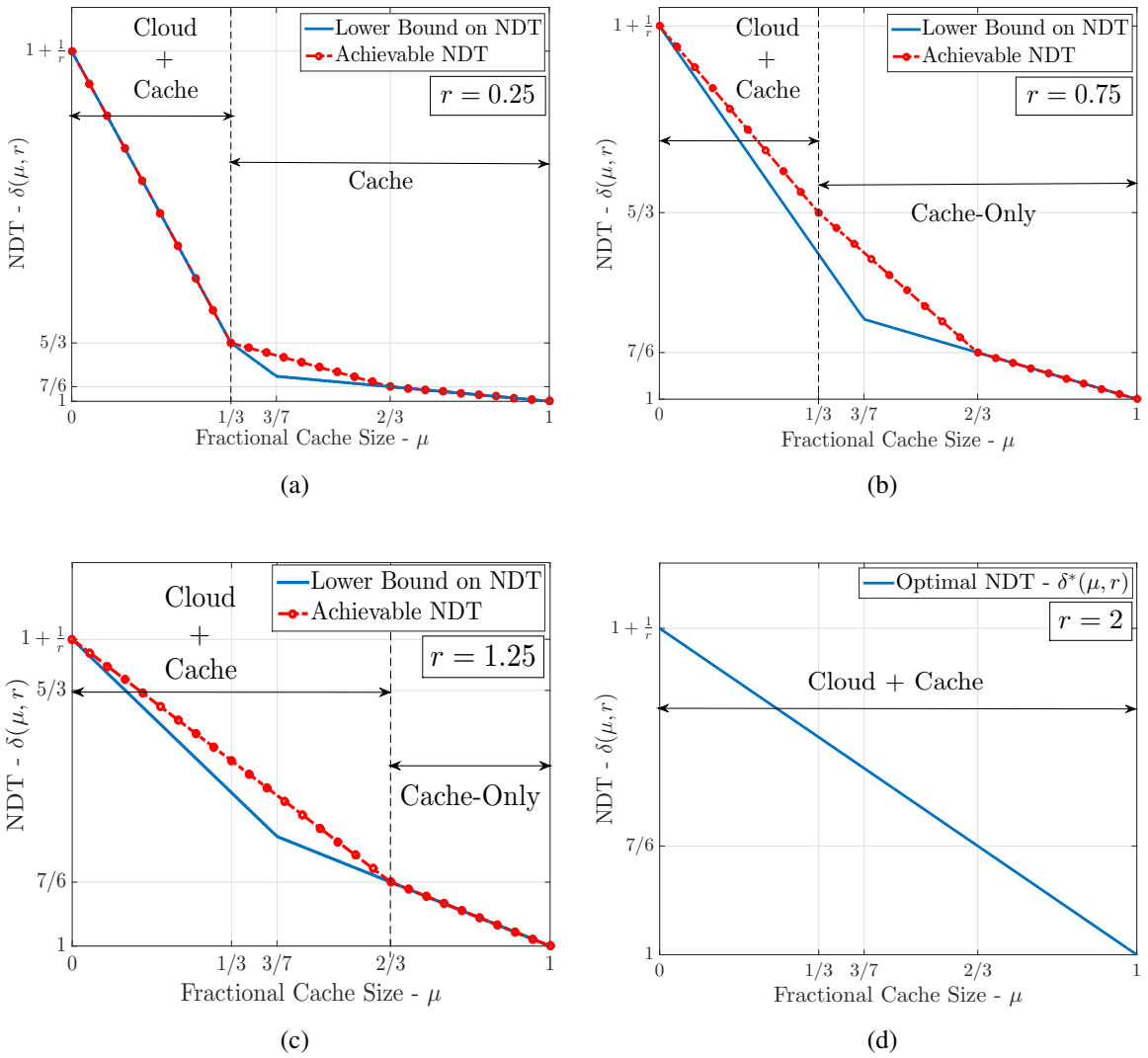


Figure 5.7: NDT bounds for an F-RAN with $M = K = 3$: (a) low fronthaul, here $r = 0.25$, (b) intermediate fronthaul, here $r = 0.75$, (c) intermediate fronthaul, here $r = 1.25$ and (d) high fronthaul, here $r = 2$. The labels "Cache" and "Cloud" refer to the achievable schemes.

Fig. 5.7(b), 5.7(c) for $r = \{0.75, 1.25\}$, which falls in the second and third intervals described in Corollary 7. In particular, for $\mu \geq 2/3$, optimal F-RAN operation does not require the use of the cloud. Finally, in the *high fronthaul* regime, where $r \geq 2$, achieving the minimum NDT requires the use of both cloud and caching resources as seen in Fig. 5.7(d). Details on the achievable schemes can be found in Appendix C.7.

5.7 Pipelined Fronthaul-Edge Transmission

In this section, we elaborate on the F-RAN model with pipelined fronthaul-edge transmission introduced in Section 5.2.2 (see Remark 16). The following lemma bounds the improvement in NDT that can be achieved by the use of pipelined fronthaul-edge transmission as compared to serial fronthaul-edge transmission.

Lemma 7 (*Pipelined vs. Serial Fronthaul-Edge Transmission*)**.** *For an $M \times K$ cloud and cache-aided F-RAN, pipelined fronthaul-edge transmission can improve the minimum NDT as compared to serial transmission by a factor of at most 2, i.e.,*

$$\delta^*_{\mathsf{P}}(\mu, r) \geq \frac{\delta^*(\mu, r)}{2}. \tag{5.53}$$

Proof. For the case of pipelined fronthaul-edge transmission, consider an optimal policy π^*_{P} that achieves the minimum NDT $\delta^*_{\mathsf{P}}(\mu, r)$. We use this policy π^*_{P} to construct a policy π for serial fronthaul-edge transmission model as follows: the caching and fronthaul policies for π are the same as for π^*_{P}; and the edge-transmission policy for π is the same as for π^*_{P} with the caveat that the ENs start transmitting only after the fronthaul transmission is complete. The NDT $\delta(\mu, r)$ achieved by the serial policy π is no larger than $2\delta^*_{\mathsf{P}}(\mu, r)$ since the durations of fronthaul and edge transmission for π^*_{P} are by definition of the NDT (5.13), both limited by $\delta^*_{\mathsf{P}}(\mu, r)$ when normalized by $L/\log(P)$ in the limit of large L and P. This concludes the proof. □

We next derive a lower bound on the minimum NDT $\delta^*_{\mathsf{P}}(\mu, r)$ based on Theorem 21 and an upper bound that relies on the fronthaul and edge transmission strategies discussed in Section 5.4. Since the results concerning cache-only F-RANs ($r = 0$) coincide with those presented thus far, we focus here only on the case of $r > 0$.

5.7.1 Lower Bound on the Minimum NDT

Here, we provide a general lower bound on the minimum NDT for the $M \times K$ F-RAN with pipelined fronthaul-edge transmission. The main result is stated in the following corollary which can be derived based on Theorem 21.

Corollary 8 (*Lower Bound on the Minimum NDT for Pipelined Fronthaul-Edge Transmission*)**.** *For an F-RAN with M ENs, each with a fractional cache size $\mu \in [0, 1]$, K users, a library of $N \geq K$ files and a fronthaul capacity of $C_F = r \log(P)$ bits per symbol, the minimum NDT for pipelined fronthaul-edge transmission is lower bounded as*

$$\delta^*_{\mathsf{P}}(\mu, r) \geq \max\left\{ \max_{\ell \in [0:\min\{M,K\}]} \frac{K - (M-\ell)^+(K-\ell)^+\mu}{\ell + (M-\ell)^+ r}\ ,\ 1 \right\}. \tag{5.54}$$

Proof. The corollary is proved via the same steps as in the proof of Theorem 21 (see Appendix C.1) with the following caveat. For pipelined fronthaul-edge transmission, the vectors $\mathbf{U}_m^T, \mathbf{X}_m^T, \mathbf{Y}_k^T$ and $\mathbf{n}_k^T$ corresponding to the fronthaul messages and transmitted signal for each EN_m, and the received signal and channel noise for each user k, respectively, have T entries, as per (5.12), where T is the overall transmission latency. This is because pipelining allows for parallel fronthaul-edge transmissions. Using these definitions, along with (5.13), and following the same steps as in (C.2)-(C.7) in Appendix C.1, the first term in the lower bound can be derived. The second term follows in a similar manner from (C.8) in Appendix C.1. □

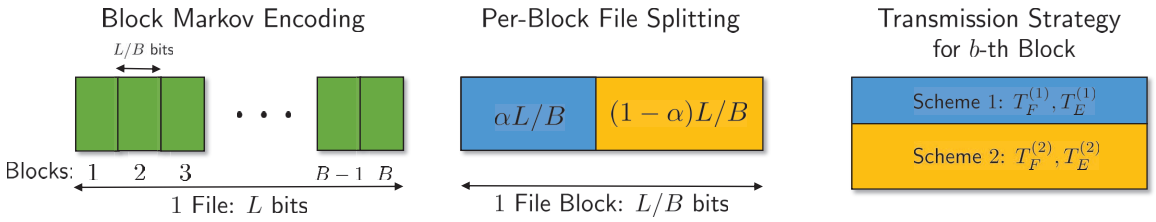
To provide some intuition on the lower bound (5.54) in relation to Theorem 21, we note that, for an F-RAN with pipelined fronthaul-edge transmission, the fronthaul and edge transmission intervals generally overlap and hence the fronthaul-NDT δ_F and the edge-NDT δ_E, which may be defined as in (5.18) and (5.19), satisfy $\max\{\delta_F, \delta_F\} \leq \delta$, where δ is the overall NDT. Therefore, from constraint (5.16) of Theorem 21, by setting $\delta_E, \delta_F \leq \delta$ and maximizing over all ℓ we obtain the first term inside the $\max(\cdot)$ function. The second term follows in a similar manner from (5.17). We also observe that the lower bound (5.54) is strictly smaller than the lower bound (5.14) derived under serial operation in accordance with the discussion in Remark 16. Next, we consider achievable schemes that yield upper bounds on the minimum NDT for the pipelined fronthaul-edge transmission model.

5.7.2 Upper Bounds on the Minimum NDT

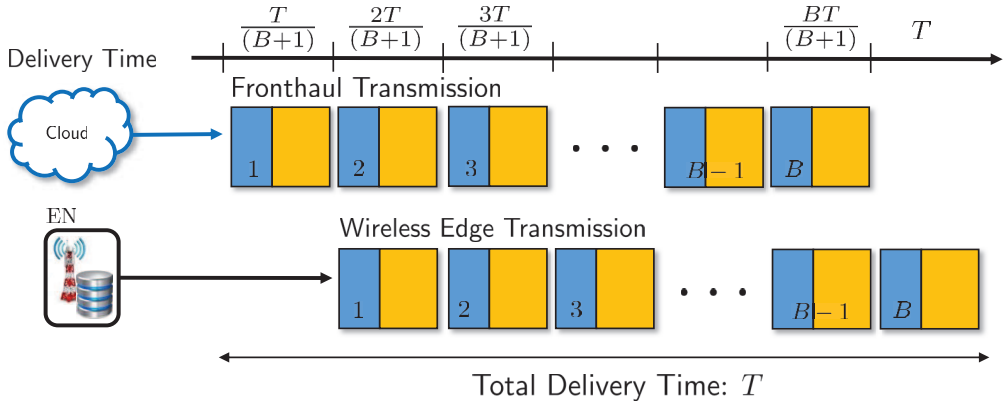
The proposed achievable scheme for pipelined fronthaul-edge transmission leverages *block-Markov encoding* to convert serial transmission policies discussed in Section 5.4 to pipelined policies. We further integrate block-Markov encoding with *per-block file splitting* to time-share between two transmission policies within each block.

- *Block-Markov Encoding*: To convert a serial policy into a pipelined policy, we split each file in the library into B blocks, so that each block is of size L/B bits. Correspondingly, we also divide the total delivery time T into $B+1$ slots, each of duration $T/(B+1)$. In each slot $b \in [1:B]$, the cloud operates the fronthaul according to the serial policy to deliver the bth blocks of the requested files, while the ENs apply the corresponding edge delivery policy to deliver the $(b-1)$th blocks of the requested files, as illustrated in Fig. 5.8(b).

  Let $T_F^{(B)}$ denote the per-block fronthaul time and $T_E^{(B)}$ denote the per-block edge time required by the selected policies in each block. These times are related to the total fronthaul and edge delivery times T_F and T_E of the serial policy as $T_F^{(B)} = T_F/B$ and $T_E^{(B)} = T_E/B$, since in each block, only a fraction L/B of a file is transmitted. The total delivery time per bit is hence



(a)



(b)

Figure 5.8: Pipelined F-RAN operation: (a) File-splitting and block Markov encoding using B blocks; file-splitting enables the use of two constituent schemes to deliver content; (b) pipelined transmission where a serial transmission strategy is used within each block.

given by

$$\Delta_{\mathsf{P}}(\mu, C_F, P) = \limsup_{L\to\infty} \frac{(B+1)\max\left(T_F^{(B)}, T_E^{(B)}\right)}{L} = \limsup_{L\to\infty} \frac{(B+1)}{B}\frac{\max\left(T_F, T_E\right)}{L}. \tag{5.55}$$

The corresponding NDT (5.10) is computed as

$$\begin{aligned}\delta_{\mathsf{P,Ach}}(\mu, r) &= \lim_{B\to\infty}\lim_{P\to\infty}\limsup_{L\to\infty} \frac{(B+1)}{B}\frac{\max\left(T_F, T_E\right)}{L/\log(P)} \\ &= \max\left(\delta_F, \delta_E\right),\end{aligned} \tag{5.56}$$

where δ_F and δ_E are the fronthaul and edge NDTs of the serial transmission scheme. Thus, under the limit of an arbitrarily large number of blocks B, the achievable NDT under pipelined fronthaul-edge transmission is the *maximum* of the edge and fronthaul NDTs of the serial policy.

- *Per-Block File Splitting*: To further improve the performance of the block-Markov coding, we propose a per-block *file-splitting* strategy in order to time-share between any two serial fronthaul-edge policies. To elaborate, for some $\alpha \in [0,1]$ fraction of each file block (of size L/B bits), a (serial) policy requiring total fronthaul and edge NDTs $\delta_F^{(1)}$ and $\delta_E^{(1)}$ is used, and for the remaining $(1-\alpha)$ fraction of each file block, a (serial) policy requiring NDTs $\delta_F^{(2)}$ and $\delta_E^{(2)}$ is used (see Fig. 5.8(a)). Based on the discussion above, this yields an achievable NDT of

$$\delta_{\mathsf{P,Ach}} = \max\left(\alpha\delta_F^{(1)} + (1-\alpha)\delta_F^{(2)}, \alpha\delta_E^{(1)} + (1-\alpha)\delta_E^{(2)}\right). \tag{5.57}$$

The following theorem gives an achievable NDT by considering a pipelined fronthaul-edge transmission strategy that utilizes cloud-aided soft-transfer fronthauling along with either cache-aided EN coordination via interference alignment or cache-aided EN cooperation via ZF-beamforming (see Section 5.4) as the constituent schemes, as for Theorem 24. We note that, unlike Theorem 24, we do not consider file-splitting between cache-aided schemes (cf. (5.36)), since it can be shown that this would not improve the NDT in the presence of pipelined fronthaul-edge transmission.

Theorem 29 (*Achievable NDT for Pipelined Fronthaul-Edge Transmission*)**.** *For an $M \times K$ F-RAN with a fronthaul gain of $r > 0$, the minimum NDT for pipelined fronthaul-edge transmission is upper bounded as $\delta_{\mathsf{P}}^*(\mu, r) \leq \delta_{\mathsf{P,Ach}}(\mu, r)$, where*

$$\delta_{\mathsf{P,Ach}}(\mu,r) = \begin{cases} \delta_{\mathsf{P-IA}} = \dfrac{(1-\mu M)K}{Mr} & \text{for } \mu \in [0,\mu_1], \\ \delta_{\mathsf{P-FS}} = \dfrac{K}{Mr}\left[1-\mu_2-[\mu_1 M-\mu_2]\left(\dfrac{\mu_2-\mu}{\mu_2-\mu_1}\right)^+\right] & \text{for } \mu \in [\mu_1,\mu_2], \\ \delta_{\mathsf{P-ZF}} = \dfrac{K}{\min\{M,K\}} & \text{for } \mu \in [\mu_2,1], \end{cases} \tag{5.58}$$

and

$$\mu_1 = \left(\frac{K-\max\{M,K\}r}{KM+Mr\left[\min\{M,K\}-1\right]}\right)^+, \quad \mu_2 = \left(1-\frac{Mr}{\min\{M,K\}}\right)^+, \tag{5.59}$$

with $\mu_1 \leq \mu_2 \leq 1$. The NDT $\delta_{\mathsf{P-IA}}$ is achieved by file-splitting between cloud-aided soft-transfer fronthauling and cache-aided EN coordination via X-channel based interference alignment; the NDT $\delta_{\mathsf{P-ZF}}$ is achieved by file-splitting between cloud-aided soft-transfer fronthauling and cache-aided EN cooperation via ZF-beamforming; and the NDT $\delta_{\mathsf{P-FS}}$ is achieved by file-splitting between the schemes achieving $\delta_{\mathsf{P-IA}}$ at $\mu = \mu_1$ and $\delta_{\mathsf{P-ZF}}$ at $\mu = \mu_2$ respectively.

Proof. The proof is presented in Appendix C.9.1. □

As indicated in Theorem 29, the NDT (5.58) is achieved by selecting the best among three block-Markov strategies which use as constituent schemes cloud-aided soft-transfer on the fronthaul and either cache-aided ZF-beamforming or X-channel-based interference alignment on the edge. An illustration will be provided below for a 2×2 F-RAN.

5.7.3 Minimum NDT for a Cloud and Cache-Aided F-RAN

We next provide a partial characterization of the minimum NDT for a general cloud and cache-aided F-RAN with pipelined fronthaul-edge transmission. Specifically, the following theorem gives the minimum NDT for the low cache regime with $\mu \in [0, \mu_1]$; for the high cache regime with $\mu \in [\mu_2, 1]$; and for the high fronthaul regime with $r \geq ((1-\mu)\min\{M, K\})/M$.

Theorem 30 (*Minimum NDT for a General F-RAN with Pipelined Fronthaul-Edge Transmissions*)**.** *For a general $M \times K$ F-RAN, with pipelined fronthaul-edge transmission and with fronthaul gain $r > 0$, we have*

$$\delta^*_{\mathsf{P}}(\mu, r) = \begin{cases} \delta_{\mathsf{P-IA}}, & \textit{for} \quad \mu \in [0, \mu_1], \\ \delta_{\mathsf{P-ZF}}, & \textit{for} \quad \mu \in [\mu_2, 1], \end{cases} \tag{5.60}$$

where $\delta_{\mathsf{P-IA}}$ and $\delta_{\mathsf{P-ZF}}$ are defined in (5.58) *and the fractional cache sizes μ_1, μ_2 are defined in* (5.59)*. Furthermore, for any fractional cache size $\mu \in [0, 1]$, we have*

$$\delta^*_{\mathsf{P}}(\mu, r) = \delta_{\mathsf{P-ZF}}, \quad \textit{for} \quad r \geq \frac{(1-\mu)\min\{M, K\}}{M}. \tag{5.61}$$

Proof. The proof is presented in Appendix C.9.2. □

Remark 20. Theorem 30, along with Theorem 29, demonstrate that, even with partial caching, i.e., with $\mu < 1$, it is possible to achieve the same performance as in a system with full caching or ideal fronthaul, namely $\delta = \delta_{\mathsf{P-ZF}} = K/\min\{M, K\}$. This is the case as long as either the fronthaul capacity is large enough (see (5.61)) or the fronthaul capacity is positive and the cache capacity μ is sufficiently large (see (5.60)). We observe that this is not true for serial fronthaul-edge transmission, in which case no policy can achieve the NDT $\delta = K/\min\{M, K\}$ for $\mu < 1$ and finite fronthaul capacity. The intuition behind this result is that, with pipelined transmission, cloud resources can be leveraged to make up for partial caching by transmitting on the fronthaul while edge transmission takes place (see [189] for practical implications). □

We finally provide an approximate characterization of the minimum NDT for a general $M \times K$ F-RAN with pipelined fronthaul-edge transmission by showing that the lower bound in Corollary 8 and the upper bound in Theorem 29 are within a constant multiplicative gap, independent of problem parameters for any fronthaul gain $r > 0$, in the intermediate cache regime with $\mu \in [\mu_1, \mu_2]$, where the minimum NDT is not characterized by Theorem 30.

Theorem 31 (*Approximate Characterization of Minimum NDT in the Intermediate Cache Regime*)**.** *For a general $M \times K$ F-RAN with pipelined fronthaul-edge transmission and with fronthaul gain $r > 0$, we have*

$$\frac{\delta_{\mathsf{P,Ach}}(\mu, r)}{\delta^*_{\mathsf{P}}(\mu, r)} \leq 2, \quad \textit{for} \ \ \mu \in [\mu_1, \mu_2]. \tag{5.62}$$

Proof. The proof of Theorem 31 is presented in Appendix C.9.3. □

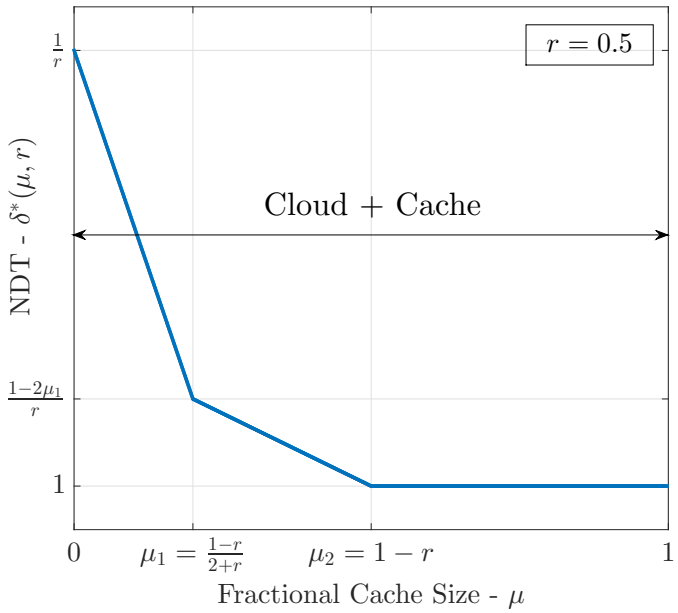


Figure 5.9: Minimum NDT for an F-RAN with $M = K = 2$ and pipelined fronthaul-edge transmissions in the low fronthaul regime, here with $r = 0.5$.

5.7.4 Case Study: 2×2 F-RAN with Pipelined Fronthaul-Edge Transmission

In this section, we provide the complete characterization of the minimum NDT of an F-RAN with $M = 2$ ENs and $K = 2$ users with pipelined fronthaul-edge transmission and we offer insights on optimal delivery policies.

Corollary 9. *For an F-RAN with $M = 2$ ENs, $K = 2$ users, $N \geq 2$ files, and with pipelined fronthaul-edge transmissions, the minimum NDT is characterized as*

• *Low Fronthaul* $(r(0, 1])$*:*

$$\delta^*_{\mathsf{P}}(\mu, r) = \begin{cases} \dfrac{1-2\mu}{r}, & \text{for } \mu \in [0, \mu_1 = (1-r)/(2+r)] \\ \dfrac{2-\mu}{1+r}, & \text{for } \mu \in [\mu_1, \mu_2 = (1-r)] \\ 1, & \text{for } \mu \in [\mu_2, 1] \end{cases} \tag{5.63}$$

• *High Fronthaul* $(r \geq 1)$*:*

$$\delta^*_{\mathsf{P}}(\mu, r) = 1, \quad \text{for } \mu \in [0, 1]. \tag{5.64}$$

Proof. The proof of Corollary 9 is provided in Appendix C.9.4. □

The minimum NDT for a 2×2 F-RAN is shown in Fig. 5.9 in the regime of low fronthaul gain, here with $r = 0.5$. The optimal strategy uses block-Markov encoding with cloud-aided soft transfer fronthaul in conjunction with cache-aided EN cooperation or coordination as for Theorem 29. We observe that, in contrast to serial fronthaul-edge transmission (see Corollary 6), the optimal strategy leverages cloud resources for any given fronthaul gain $r > 0$. Furthermore, in line with the discussion in Remark 20, by using cloud resources, it is possible here to obtain the minimum NDT $\delta^*_{\mathsf{P}}(\mu, r) = 1$ for all $\mu \geq \mu_2$ as well as for $r \geq 1$.

5.8 Directions of Future Research

In this section, we discuss some of the open problems and directions for future work on the topic of cloud and cache-aided content delivery in F-RAN architectures.

5.8.1 Is an Equal Cache Allocation Optimal?

Throughout the chapter, as per Definition 7, we have assumed that each file F_n is cached with the same maximum number of bits, namely μL, at each EN. Here, we aim at understanding if the minimum NDT could be potentially reduced by allocating a different number of bits to each file at the ENs under relaxed constraints

$$\sum_{m=1}^{M} H\left(S_{m,n}\right) \leq M\mu L, \ \ \forall n \in [1:N], \tag{5.65}$$

$$\sum_{n=1}^{N} H\left(S_{m,n}\right) \leq N\mu L, \ \ \forall m \in [1:M], \tag{5.66}$$

where the first constraint imposes the per-file condition that the overall number of bits used to cache file F_n across all ENs cannot exceed $M\mu L$ bits, while the second is the per-EN cache capacity constraint. By using the same arguments as in the proof of Theorem 21 in Appendix C.1, it can be shown that the lower bound in Theorem 21 holds also under the relaxed constraints (5.65)-(5.66). To this end, we first note that the bounds in (5.17) remain unchanged since they do not make use of the cache constraints. For the proof of (5.16), we refer to Appendix C.1. This shows that the strategy of allocating an equal number of bits to each file at every EN as in Definition 7 is in fact information-theoretically optimal under the assumption of uncoded cache placement.

5.8.2 Caching with Inter-File Coding

The results presented in this chapter are developed under the assumption that the caching policy at the ENs do not allow for inter-file coding (as in (5.2)). For schemes with caching only at receivers,

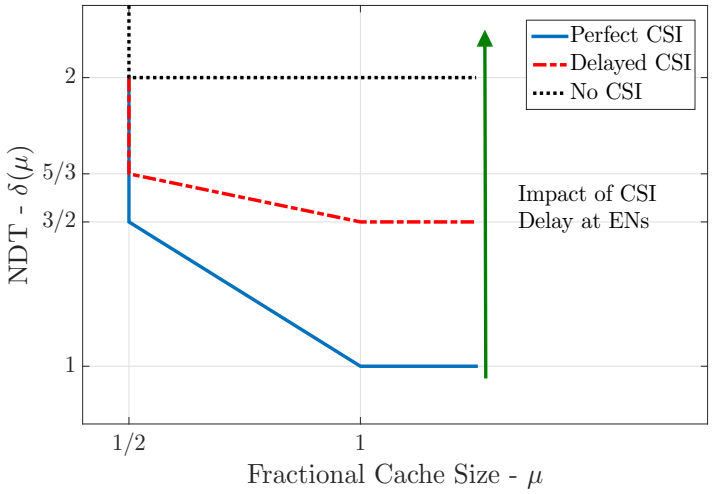


Figure 5.10: Effect of delayed or no CSI on the NDT for $M = K = 2$.

it has been shown that coding across files during cache placement has the potential to improve the system performance [97, 125, 133, 134]. To elaborate on the potential gains of inter-file coding for F-RANs, we observe that, under such more general caching policies, the joint entropy of the cache contents of each EN is generally bounded as $H(S_m) \leq \mu NL$, for all $m \in [1 : M]$, instead of as $H(S_{m,n}) \leq \mu L$ for all $m \in [1 : M]$ and $n \in [1 : N]$ as is the case without inter-file coding. As a result, following the proof of Theorem 21 in Appendix C.1, we can see that, the constraint (5.16) in Theorem 21 is modified as

$$\ell\delta_E + (M - \ell)^+ r\delta_F \geq K - (M - \ell)^+ N\mu, \tag{5.67}$$

which yields strictly lower bounds on the minimum NDT. Whether this lower bound is achievable by caching strategies with inter-file coding remains an open problem.

5.8.3 Imperfect CSI

Another aspect that is left open by this study is the impact of imperfect CSI availability at the ENs on the minimum NDT. To elaborate on this point, we consider a cache-only F-RAN and we assume that within a transmission interval i, the channel $\mathbf{H}_i^t$ varies across every channel use t according to an i.i.d. process. When CSI is delayed, at any time t on the ith transmission interval, the ENs only have access to the CSI of the previous $t - 1$ channel uses, namely $\mathbf{H}_i^1, \mathbf{H}_i^2, \ldots, \mathbf{H}_i^{t-1}$.

For illustration, we consider a cache-only F-RAN with $M = K = 2$ and $N \geq 2$ with $\mu \in [1/2, 1]$. For the case of perfect CSI, the minimum NDT can be characterized as in Corollary 6. Next, we elaborate on the achievable NDT results for the case of delayed and no CSI.

5.8.3.1 Delayed CSI at ENs

For the case of delayed CSI, consider the corner point $\mu = 1/2$ where the system behaves like a 2×2 X-channel (see Example 9). The maximum known sum-DoF for the 2×2 X-channel with delayed CSI is $6/5$ [207]. As a result, an NDT of $\delta_{\mathsf{Ach}}(\mu, 0) = 5/3$ is achievable by Remark 15. Compared to the perfect CSI case, for which the NDT is $3/2$, this achievable NDT thus incurs a loss due to delayed CSI. Next, consider the corner point $\mu = 1$, where the system reduces to a 2×2 broadcast channel with delayed CSI (see Example 8). The maximum sum-DoF for such a system is $4/3$ [85], i.e., a NDT of $\delta_{\mathsf{Ach}}(\mu, 0) = 3/2$ is achievable, which is larger than the NDT of 1 with perfect CSI.

5.8.3.2 No CSI at ENs

In case of no CSI, it is known that the optimal strategy on the edge channel is to transmit using time-division to each user in a separate slot [208]. Therefore a sum-DoF of 1 i.e., an NDT of 2 can be achieved, which is hence optimal for all values of $\mu \in [1/2, 1]$ and shows a significant loss as compared to the cases with full or delayed CSI as shown in Fig. 5.10.

Quantifying the impact of delayed CSI on a general $M \times K$ F-RAN with cloud and cache-aided delivery, as considered in this work, remains an area of future work.

5.9 Summary

In this chapter, we presented a latency-centric study of the fundamental information-theoretic limits of cloud and cache-aided wireless networks, which we referred to as fog radio access networks (F-RANs). To this end, we introduced a new metric, namely the normalized delivery time (NDT), which measures the worst-case end-to-end latency required to deliver requested content to the end users in the high-SNR regime. We developed a converse result for the NDT of a general F-RAN with arbitrary number of ENs and users and then presented achievable schemes which leverage both cache and cloud resources. We characterized the minimum NDT for cloud-only F-RANs for all problem parameters; and for cache-only F-RANs in the regime of extremal values of the fractional cache size. Furthermore, we showed that the proposed achievable schemes are approximately optimal to within a constant factor of 2 for all parameter values for the general F-RAN with fronthaul and edge-caching. We elaborated on two case studies, consisting of F-RANs with two or three ENs and users and (partially) characterized the NDT for these systems using the proposed upper and lower bounds. We also considered an alternative F-RAN model with pipelined fronthaul-edge transmissions. We presented a general lower bound on the NDT and proposed achievable schemes which are shown to be approximately optimal in terms of NDT to within a constant factor of 2. Open problems were finally presented to highlight the richness of the problem introduced in this chapter.

Chapter 6

Learning-Aided Collaborative Caching

In this chapter, we study collaborative caching strategies in a multi-sBS heterogeneous network with unknown file popularities from a reinforcement learning perspective. We present topology-aware uncoded and coded caching strategies under a learning-aided framework where the file popularities are dynamically learned over time by observing user requests. The uncoded collaborative cache placement problem is NP-hard and we propose a novel weighted graph-coloring and local search-based approximation algorithm with an approximation ratio of $(\frac{1}{3} - \epsilon)$ for some $\epsilon > 0$. Alternately, we formulate a coded cache placement strategy which is shown to be a linear program yielding an optimal solution. Through simulations we show that the uncoded approximate caching algorithm performs close to the optimal coded scheme when integrated with the learning framework in the multi-sBS setting. We also show that for network topologies of practical interest, the collaborative caching strategies outperform local caching strategies.

6.1 Introduction

With the proliferation of heterogeneous wireless architectures, caching large volumes of data at the network edge has become a feasible means for load distribution over the network. As a result, benefits of caching in next-generation wireless network settings has been studied extensively in recent literature [97, 102–104, 120, 153, 154, 157–161, 163, 176, 178]. Recent works like [97, 120, 176] study caching at mobile end-users from a novel information theoretic framework which has its roots in index coding and multicast delivery to increase efficiency of content delivery over a single server network. Caching in multi-sBS networks from an interference management perspective was studied in [146, 147, 209] under uniform file (defined as a block of data) popularity distribution. Information theoretic caching for non-uniform file popularity was studied in [99, 107, 117] However, the problem of determining the optimal content to cache in a practical network setting when file popularity is *unknown* is a difficult one. In this work, we consider this practical problem of caching popular content at the network edge augmented by a framework for learning the varying

file popularity profile across users.

Heterogeneous networks, where small-cell base stations (sBSs) like femto or pico-cells are connected to a central cellular base station (BS), lend themselves to caching of data at the network edge by equipping the sBSs with caches for local, low-latency content dissemination. Distributed cache placement in sBSs was studied in [152, 210], with an aim to reduce the latency of file delivery to the end-users while assuming that the popularity of files was known apriori at the sBSs. However, such an assumption is unrealistic and learning based caching frameworks were presented in [153, 157–161, 163, 178] and references therein for the case of unknown file popularity profiles. The authors in [153,160,163] presented a multi-armed bandit [211,212] based reinforcement learning framework for cache placement in a single sBS network. Transfer-learning based approaches for caching in small-cell networks were studied in [157–159], where minimization of backhaul load and the evaluation of time to achieve desired learning accuracy under random caching were the main areas of focus. These works, however, do not consider the problem of caching for a multi-sBSs network with overlapping connectivity to users under unknown file popularity distribution. In this work, we study topology-dependent cache placement in a multi-sBS wireless network from a reinforcement learning perspective. We ask the following fundamental questions:

Given a multi-sBS network topology, per-sBS cache capacity and random file requests from users based on an unknown file popularity profile -

(i) *What is the best learning method with finite-time performance guarantees for estimating the file popularity from observations of user requests?*

(ii) *What is the best collaborative caching strategy, such that a maximum number of requests can be served directly from the sBS caches?*

To address these questions, we formulate the *learning-aided collaborative caching* framework for the multi-sBS environment as shown in Fig. 6.1. We present novel topology-aware caching strategies using a combinatorial multi-armed bandit (CMAB)-based learning framework [212–215]. In real networks, the file popularity profile could change over time and hence it is necessary to learn it dynamically during the caching procedure. The files to be cached are modeled as the arms of a combinatorial multi-armed bandit (CMAB) problem [213] in order to learn their popularity over time. The goal of the caching strategy is to pick the best set of files at any time t based on their estimated popularity so that the requests from users can be directly served by the caches without accessing the core network.

6.1.1 Main Contributions

Based on this network model, the following are the main contributions of our work:

1. *Learning Framework* - We present a CMAB-based learning framework for dynamically learning the file popularity distributions in a multi-sBS network setting. We show that the bound

on the sub-optimality gap of the proposed learning algorithm at each sBS at time t scales as $O(\log(t))$ i.e., the sub-optimality gap scales logarithmically with time.

2. *Collaborative Caching Framework* - We formulate a novel network topology-aware *uncoded caching* strategy where entire files are cached at the sBSs. We show that the uncoded caching problem is NP-hard and propose a novel graph-coloring based algorithm which provides a $\left(\frac{1}{3}-\epsilon\right)$-approximate cache placement solution in polynomial time for some $\epsilon > 0$. We also formulate a *coded caching* strategy which allows fractional file placement. We show that this strategy can be represented as a linear program which can be solved optimally.

3. Through simulations, we present a detailed comparison of the uncoded and coded collaborative caching schemes. We show that for both coded and uncoded schemes, collaborative caching generally outperforms naive strategies that locally optimize the cache content at each sBS without accounting for the network topology. We further show, that in spite of using a greedy approximation algorithm, the uncoded collaborative caching strategy performs similar to the provably optimal coded strategy. This is attributed to the fact that learning popularity using fractional placements is harder than learning from placement of entire files in the multi-sBS setting.

6.2 Network Model

We consider a small cell network with a set of N sBSs, $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\}$, connected to a central Base Station (BS). A set of K users $\mathbb{U} = \{U_1, U_2, \ldots, U_K\}$ are located in an area covered by the small cell network. Users can make random requests from a directory of F files, $f \in \mathbb{F} = \{1, 2, \ldots, F\}$, where file f has size S_f units[1]. The central BS is considered to have enough memory to store the entire file directory $\mathbb{F}$. Each sBS in the network has a cache memory that can store M units of data. We assume that $M \geq \min_{f\in\mathbb{F}} S_f$ i.e., the sBS caches are large enough to store at least one whole file. Each user can be served from the cache of one or more sBSs in the network. If a user's request cannot be serviced from the caches of the sBSs to which it is connected, then it can be downloaded directly from the central BS. This ensures that all the users' requests are serviced. However, the objective is to enable loacalized content delivery with minimal use of backhaul resources. Therefore, the optimal caching policy ensures that maximum number of user requests can be serviced locally from the sBS caches at the network edge, thereby alleviating load at the central BS. An example small cell wireless network architecture with $N = 6$ sBSs and $K = 12$ users is shown in Fig. 3.1. The central BS is connected to sBSs via wireless (or wired) backhaul links.

[1]The unit of file size can be Megabytes or Gigabytes depending upon the networks under consideration. We keep it generic for our exposition,

Table 6.1: Learning-Aided Collaborative Caching: Table of Notations

$\mathbb{S}$	Set of N sBSs
$\mathbb{U}$	Set of K users
$\mathbb{F}$	Set of F files
n	sBS Index
u	User Index
f	File Index
$\mathcal{S}_n$	$n-$th sBS
S_f	Size of $f-$th file
$\mathcal{G}(\mathbb{S},\mathbb{U},\mathbb{E})$	Bipartite connectivity graph
$\mathcal{N}(u)$	sBSs serving user u
$d^t_{f,n}$	Average instantaneous demand for file f at sBS n and time t
$\mathcal{U}(\mathcal{S})$	Users connected to sBS $\mathcal{S}$
$\theta_{f,n}$	True Popularity of file f at sBS n
Θ_n	True Popularity distribution of all files at sBS n
γ	Skewness of ZipF popularity distribution
$\mathcal{C}^\pi(t)$	$F \times N$ joint cache placement matrix with elements $c^\pi_{f,n} \in (0,1]$

6.2.1 Network Connectivity

The connectivity of the users and sBS within the network can be modeled as a bipartite graph $\mathcal{G} = (\mathbb{S},\mathbb{U},\mathbb{E})$, where the edges $(\mathcal{S}_n,u) \in \mathbb{E}$ if there exists a communication link, subject to physical layer constraints (i.e., the user is not in *outage* based on path loss, shadowing etc.), between user u and the n-th sBS $\mathcal{S}_n$. $\mathcal{N}(u) \subseteq \mathbb{S}$ denotes the neighborhood of user u i.e., the sBSs that can serve the user (or conversely, the sBSs to which the user is connected). For example, from Fig. 3.1, $\mathcal{N}(U_1) = \{\mathcal{S}_1\}$ while $\mathcal{N}(U_3) = \{\mathcal{S}_2,\mathcal{S}_4\}$. On the other hand, the neighborhood of the nth sBS i.e., the number of users connected to sBS $\mathcal{S}_n$ is denoted by $\mathcal{U}(\mathcal{S}_n)$. For example, in Fig. 3.1, $\mathcal{U}(\mathcal{S}_1) = \{U_1,U_4,U_5,U_9\}$ and $\mathcal{U}(\mathcal{S}_2) = \{U_2,U_3,U_8\}$. The central BS has complete network knowledge i.e., knowledge about the neighborhoods of all sBS connected to it and thereby the knowledge of the sBS neighborhood of each user.

6.2.2 Model for File Popularity

The popularity of files (e.g., videos) in a multimedia content distribution network is generally modeled as a ZipF distribution [108, 184] where the mean demand (number of requests) for a file f at any sBS $\mathcal{S}_n$ can be modeled as:

$$\theta_{f,n} = \frac{f^{-\gamma}}{\sum_{k=1}^{F} k^{-\gamma}}, \tag{6.1}$$

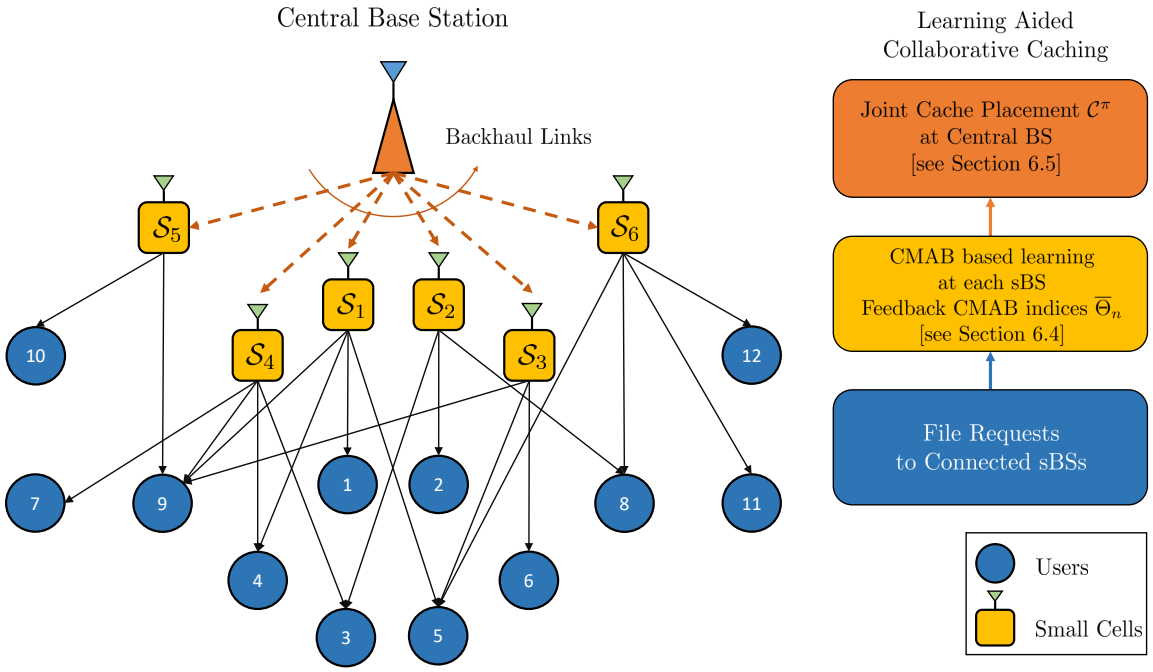


Figure 6.1: Network model for learning-aided collaborative caching in a small cell network with a central BS, $N = 6$ sBSs and $K = 12$ users.

where $\theta_{f,n} \in [0, 1]$ and γ models the skewness of the popularity profile. For example, $\gamma = 0$ models a uniform file popularity profile and as γ increases the skewness increases. The file popularity distribution at $\mathcal{S}_n$ is denoted by $\Theta_n = \{\theta_{1,n}, \theta_{2,n}, \ldots, \theta_{F,n}\}$.

In practical settings, the distribution $\Theta_n,\ \forall n \in \{1, \ldots, N\}$ is unknown. Therefore, in this work, we empirically estimate the mean demand for each file at every sBS based on observations of user requests over a period of time. To this end, let $d_{f,n}^{t,u} \in \{0, 1\}$ be an indicator variable such that $d_{f,n}^{t,u} = 1$ when a user u in the neighborhood of sBS $\mathcal{S}_n$ i.e., $u \in \mathcal{U}(\mathcal{S}_n)$ requests the file f at time t. Let

$$d_{f,n}^{t} = \frac{1}{|\mathcal{U}(\mathcal{S}_n)|} \sum_{u \in \mathcal{U}(\mathcal{S}_n)} d_{f,n}^{t,u}, \tag{6.2}$$

indicate the instantaneous demand for file $f \in \mathbb{F}$, averaged over all users served by sBS $\mathcal{S}_n$. $|\mathcal{U}(\mathcal{S}_n)|$ is the cardinality of the set $\mathcal{U}(\mathcal{S}_n)$. The instantaneous demand $d_{f,n}^{t}$ is an i.i.d random variable with an empirical mean

$$\widehat{\theta}_{f,n} = \mathbb{E}\left[d_{f,n}^{t}\right] \tag{6.3}$$

which is bounded in support $[0, 1]$ and the expectation is over the random user requests at time instants $1, 2, \ldots, t$. In other words, the mean $\widehat{\theta}_{f,n}$ signifies the empirical popularity of the f-th file at the n-th sBS $\mathcal{S}_n$ i.e., the average number of per-user requests for the file f at $\mathcal{S}_n$ (from the users

in $\mathcal{U}(\mathcal{S}_n)$) till time t. Thus, although the true popularity distribution Θ_n is unknown, we estimate the file popularity distribution $\widehat{\Theta}_n = \{\widehat{\theta}_{1,n}, \widehat{\theta}_{2,n}, \ldots, \widehat{\theta}_{F,n}\}$ at $\mathcal{S}_n$ based on the user requests up to time t. In the next section we introduce the learning framework for collaborative caching in a multi-sBS environment.

6.3 Learning-Aided Collaborative Caching

In multimedia content distribution networks, files generally have varying popularities i.e., a few files in the system are highly popular (e.g., viral YouTube videos) while others are requested less frequently [99, 216]. The resulting popularity profile is accurately modeled by the heavy-tailed ZipF distribution [99, 184, 216]. In order to design efficient caching and delivery policies such that majority of the requested content can be delivered locally, reliable estimation of the file popularity over time is required at each sBS. To address this, we introduce a topology-aware learning-aided collaborative caching framework. Over learning iterations, the central BS receives feedback from the sBSs about the *performance* of the caching policy. Performance is measured in terms of a *reward* i.e., the amount of data which is downloaded from an sBS in servicing the requests of connected users. Similarly, the sub-optimality gap of the learning process is defined as *regret*. Traditional learning techniques such as Q-Learning, ϵ-greedy learning etc. provide only asymptotic convergence guarantees [217]. However, in a caching framework, such asymptotic guarantees (when the number of observed user demands goes to infinity) are not sufficient.

To this end, we model the popularity estimation and cache placement as a Multi-Armed-Bandit (MAB) problem [212], allowing us to present finite time guarantees on the learning performance. In this setting, the files $f \in \mathbb{F}$ are treated as the arms and caching a file at a sBS is equivalent to pulling an arm in a MAB framework. The caching of files at each sBS enables estimation of file popularity by observing the rewards obtained in return i.e., by the amount of data which can be used from the local caches to serve user requests. Since each sBS cache can typically store multiple files, it is appropriate to model the problem as a combinatorial MAB (CMAB) problem [153, 160, 161, 163, 178, 213–215]. Under a combinatorial setting, a set of arms, called a *super-arm*, can be pulled at a given time instant. Thus for a cache placement policy which places multiple files in the cache of $\mathcal{S}_n$, the strategy at any time t is equivalent to pulling a CMAB super-arm. A single-sBS CMAB based caching framework was studied in [153]. Our work extends it to a topology-aware multi-sBS network.

For CMAB-based learning, there is a trade-off between the exploration of new arms (i.e., caching new files to estimate their popularity) and the exploitation of the known arms (caching files that are known to have high popularity and hence give higher rewards).

In this work, we utilize the exploration-exploitation trade-off by use of the Combinatorial Upper Confidence Bound (CUCB) algorithm [213, 214], which provides finite time convergence guarantees. We propose a distributed learning framework detailed in Algorithm 4, which has two main parts:

Algorithm 4 LEARNING-AIDED COLLABORATIVE CACHING

DISTRIBUTED POPULARITY ESTIMATION AT SBS:

for Each sBS $\mathcal{S}_n \in \mathbb{S}$ at time t **do**
  for Each file $f \in \mathbb{F}$ **do**
    if File f is cached i.e., $c^{\pi}_{f,n}(t) \geq 0$ **then**

$$T_{f,n} = T_{f,n} + 1$$

where, $T_{f,n}$ is the number of times file f (or a fraction of file f) is cached in $\mathcal{S}_n$ upto t.
    else if
      for $u \in \mathcal{U}(\mathcal{S}_n)$ **do**,
        if $c^{\pi}_{f,n'}(t) \geq 0$, $\exists\, \mathcal{S}_{n'} \in \mathcal{N}(u) \setminus \{n\}$ **then**

$$T_{f,n} = T_{f,n} + 1$$

        end if
      end for
    end if
    Update the mean demand:

$$\widehat{\theta}_{f,n} = \frac{\sum_{i=1}^{t} d^{i}_{f,n}}{t}$$

where, $\sum_{i=1}^{t} d^{i}_{f,n}$ is the cumulative sum of requests for file f from users $u \in \mathcal{U}(\mathcal{S}_n)$ until time instant t.
    Calculate CMAB Index:

$$\overline{\theta}_{f,n} = \widehat{\theta}_{f,n} + \sqrt{\frac{\Psi_{f,n} \log\left(|\mathcal{U}(\mathcal{S}_n)|t\right)}{2T_{f,n}}},$$

where $|\mathcal{U}(\mathcal{S}_n)|$ indicates cardinality of $\mathcal{U}(\mathcal{S}_n)$.
  end for
  Feedback $\overline{\Theta}_n = [\overline{\theta}_{1,n}, \overline{\theta}_{2,n}, \ldots, \overline{\theta}_{F,n}]$ to central BS.
end for

TOPOLOGY BASED COLLABORATIVE CACHING:

Caching Strategy at time $t+1$:

$$[\mathcal{C}^{\pi}(t+1)]_{F\times N} = \text{CCP}\left(\overline{\Theta}_1, \overline{\Theta}_2, \ldots, \overline{\Theta}_N\right)$$

Observe actual user requests $d^{t+1}_{f,n}$ at time $(t+1)$

Return to Step 1 and update $T_{f,n}$'s and $\overline{\Theta}_n$'s for the next time step $(t+1)$

1. *Learning the File Popularity Distribution:* A CMAB based distributed popularity estimation runs at each sBS $\mathcal{S}_n$ to estimate the empirical file popularity distribution $\widehat{\Theta}_n$, associated with its users $\mathcal{U}(\mathcal{S}_n)$. At time t, based on user demands up to $t-1$, it updates the popularity of every file that is cached in the sBS. The updated popularity profile is used in the subsequent cache placement by the central BS. The learning framework is discussed in Section 6.4.

2. *Collaborative Cache Placement (CCP):* The central BS takes the CMAB based popularity profile from the sBSs as input and designs cache placement strategies by jointly maximizing the sum reward of the sBSs. The cache placement can be *uncoded* where entire files are cached at the sBSs or *coded* where files are sub-packetized and fractions of files are allowed to be cached. Details on the design of both uncoded and coded collaborative cache placement are discussed in Section 6.5.

We next present a detailed analysis of the learning-aided collaborative caching algorithm.

6.4 Learning the File Popularity Distribution

In this section, we present the CMAB based framework for learning the file popularity distribution. First, we define the caching model, present the CUCB based learning algorithm and finally, derive finite time performance guarantees for the learning step in Algorithm 4.

6.4.1 The Cache Placement Model

In this work, we consider a Collaborative Cache Placement (CCP) policy π which determines the optimal content to be cached at each of the N sBSs in the network. The objective of the policy π is to place files in the caches of the sBSs based on the network topology and empirical history of user demands (file popularity) such that maximum number of file requests can be downloaded directly from the sBS caches. Let the $F \times N$ binary matrix $\mathcal{C}^{\pi}(t)$, denote the collaborative cache placement by the policy π at time step t. Let

$$[\mathcal{C}^{\pi}(t)]_{F\times N} = [\mathbf{c}_1^{\pi}(t), \mathbf{c}_2^{\pi}(t), \ldots, \mathbf{c}_N^{\pi}(t)], \tag{6.4}$$

where $\mathbf{c}_n^{\pi}(t)$ is the $F \times 1$ cache placement vector for the sBS $\mathcal{S}_n$ and $c_{f,n}^{\pi} \in [0,1]$ are the elements of $\mathbf{c}_n^{\pi}(t)$ such that, at a time step t,

$$c_{f,n}^{\pi}(t) \begin{cases} \geq 0 & \text{if file } f \text{ is cached at } \mathcal{S}_n \\ = 0 & \text{otherwise} \end{cases} \tag{6.5}$$

In our problem formulation, we define *caching* as placement of a *whole file* or any *fraction of a file* in the cache of the sBS i.e., $c_{f,n}^{\pi}$ denotes the fraction of file f cached at $\mathcal{S}_n$. At $t=0$, we consider that the caches of all sBSs are empty i.e., $\mathcal{C}^{\pi}(0) = [\mathbf{0}]_{F\times N}$.

6.4.2 CMAB aided File Popularity Estimation

The CMAB based file popularity estimation at each sBS gets as reward, the amount of data downloaded from the sBS cache to serve the requests of its users. By tracking the reward for cache placement over time, the sBS aims to learn the optimal cache placement policy π. The algorithm is initialized by sequentially placing each file once in the cache of each sBS. At time t, sBS $\mathcal{S}_n$ learns the file popularity distribution i.e., the empirical mean of the instantaneous demands $\widehat{\Theta}_n = \{\widehat{\theta}_{1,n}, \widehat{\theta}_{2,n}, \ldots, \widehat{\theta}_{F,n}\}$ based on the history of instantaneous demands, $d^1_{f,n}, d^2_{f,n}, \ldots, d^t_{f,n}$, upto time t and the cache placement by the CCP policy π at time t i.e., $c^{\pi}_{f,n}(t) \in \mathcal{C}^{\pi}(t), f \in \mathbb{F}$. To this end, a *CMAB index* $\overline{\theta}_{f,n}$ is calculated for each file $f \in \mathbb{F}$. The CMAB index has two components:

1. The *empirical mean* of the instantaneous demand $\widehat{\theta}_{f,n}$ derived from the observation of user requests over time.

2. An additive *perturbation factor*

$$\sqrt{\frac{\Psi_{f,n} \log\left(|\mathcal{U}(\mathcal{S}_n)|t\right)}{2T_{f,n}}}, \tag{6.6}$$

   where $|\mathcal{U}(\mathcal{S}_n)|$ is the number of users connected to $\mathcal{S}_n$ and $T_{f,n}$ denotes the number of times a file $f \in \mathbb{F}$ has been cached in $\mathcal{S}_n$ until time t. The factor

$$\Psi_{f,n} = \frac{\delta}{|\mathcal{U}(\mathcal{S}_n)|} \left(\frac{|\mathcal{U}(\mathcal{S}_n)|S_f}{F^{\gamma}}\right)^2 \tag{6.7}$$

   is directly proportional to $|\mathcal{U}(\mathcal{S}_n)|$ and a roll-off factor δ (which is usually set to 3 [153, 212]), and inversely proportional to the skewness factor γ of the ZipF based model of the file popularity profile. The factor γ can be empirically estimated as in [184] from the observations of the instantaneous file demands $d^t_{f,n}$ over time. We show in Section 6.4.3 that the sub-optimality of the learning process scales linearly with $\Psi_{f,n}$ for the proposed collaborative caching algorithm.

The index perturbation factor in (6.6) promotes exploration and exploitation based on the number of connected users, the size of each file and the Zipf parameter γ. It promotes exploration by forcing the CCP to place less-often-cached sets of files (for which $T_{f,n}$ is low) in the caches by increasing their index value. This promotes the sufficient sampling of lesser requested files in order to accurately evaluate their popularity. It also promotes exploitation when $|\mathcal{U}(\mathcal{S}_n)|$ is large i.e., $\mathcal{S}_n$ has a large user-set and also when the popularity profile is skewed i.e., when γ is large and there are few popular files in the system. Once, the sBSs calculate the local index values $\overline{\Theta}_n$, this information is fed back to the central BS where the CCP policy π determines the caching strategy $\mathcal{C}^{\pi}(t)$.

Remark 21 (*Topology-Aware $T_{f,n}$ Update*)**.** In this work, we use a unique topology-aware $T_{f,n}$ update procedure which helps in capturing the interaction of users who are connected to multiple sBSs. At $\mathcal{S}_n$, let a user u request a file $f \in \mathbb{F}$. If the file is not cached at $\mathcal{S}_n$ i.e., $c^{\pi}_{f,n}(t) = 0$, then we consider the caches of all *other* sBSs in the neighborhood of u: $\mathcal{S}_{n'} \in \mathcal{N}(u) \setminus \{\mathcal{S}_n\}$. If f is cached at any of these sBSs i.e., if the user's request is satisfied by any other sBS, then we update the $T_{f,n}$ for the sBS $\mathcal{S}_n$ as well. As opposed to the single sBS learning in [153], this leads to a topology-aware learning framework which can account for the network connectivity. Note that at any time-step t, $T_{f,n} \leq t$ still holds. □

6.4.3 Upper Bounds on the Regret for Algorithm 4

In order to derive finite time performance guarantees on the learning step of Algorithm 4, we first formally define the concept of reward and regret (sub-optimality gap) for the caching framework under consideration.

6.4.3.1 Reward

For the policy π, with a caching strategy $\mathbf{c}^{\pi}_n(t)$ at sBS $\mathcal{S}_n$ at time step t, we define the *instantaneous reward* for caching a file f in $\mathcal{S}_n$ as:

$$\begin{aligned} r^t_{f,n} &= \sum_{u \in \mathcal{U}(\mathcal{S}_n)} d^{t,u}_{f,n} c^{\pi}_{f,n}(t) S_f \\ &= |\mathcal{U}(\mathcal{S}_n)| d^t_{f,n} c^{\pi}_{f,n}(t) S_f = |\mathcal{U}(\mathcal{S}_n)| d^t_{f,n} \tilde{S}_{f,n}, \end{aligned} \tag{6.8}$$

which indicates that a reward of $\tilde{S}_{f,n} = c^{\pi}_{f,n}(t) S_f$ is obtained when $c^{\pi}_{f,n}(t)$ fraction of a file f requested by a user u is available for download from a local cache. The reward is proportional to the amount of data downloaded from the cache. At time t, the *expected accumulated reward*, $R_{\widehat{\Theta}_n}(\mathbf{c}^{\pi}_n(t))$, is obtained by taking an expectation over the instantaneous rewards up to time t:

$$\begin{aligned} R_{\widehat{\Theta}_n}(\mathbf{c}^{\pi}_n(t)) &\overset{\text{(a)}}{=} \mathbb{E}\left[\sum_{f: c^{\pi}_{f,n}(t) > 0} r^t_{f,n}\right] = \mathbb{E}\left[\sum_{f: c^{\pi}_{f,n}(t) > 0} |\mathcal{U}(\mathcal{S}_n)| d^t_{f,n} \tilde{S}_{f,n}\right] \\ &\overset{\text{(b)}}{=} |\mathcal{U}(\mathcal{S}_n)| \sum_{f: c^{\pi}_{f,n}(t) > 0} \widehat{\theta}_{f,n} \tilde{S}_{f,n}, \end{aligned} \tag{6.9}$$

where, in (a), the expectation is taken over the i.i.d instantaneous file demands $d^t_{f,n}$ and (b) follows from the definition of empirical estimate of mean demand in (6.3). The objective of Algorithm 4 is to maximize the expected reward for all time instants $1, 2, \ldots, t$. We define the *optimal* reward, $R^{\text{opt}}_{\Theta_n}$, as the expected reward obtained when the CCP policy π caches the *optimal* set of files when the *true* popularity profile $\Theta_n = \{\theta_{1,n}, \theta_{2,n}, \ldots, \theta_{F,n}\}$ is perfectly known at the BS.

6.4.3.2 Regret

The regret of the CCP policy π at a time instant t is defined as the difference between the expected accumulated reward obtained by the caching policy and the optimal expected reward $R^{\text{opt}}_{\Theta_n}$. Assuming that policy π runs an $(\alpha, \beta)-$approximation cache placement algorithm i.e., $\Pr\left[R_{\widehat{\Theta}_n}(\mathbf{c}^{\pi}_n(t)) > \alpha \cdot R^{\text{opt}}_{\Theta_n}\right] \geq \beta$, the regret at a finite time horizon $t = T$ is given by:

$$\text{Reg}^{\pi}_{\Theta_n,\alpha,\beta}(T) = T\alpha\beta R^{\text{opt}}_{\Theta_n} - \mathbb{E}\left[\sum_{t=1}^{T} R_{\widehat{\Theta}_n}(\mathbf{c}^{\pi}_n(t))\right], \tag{6.10}$$

where the expectation is over policy π and all the rewards generated by files cached by $\mathbf{c}^{\pi}_n(0), \mathbf{c}^{\pi}_n(1), \ldots, \mathbf{c}^{\pi}_n(T)$.

A caching strategy $\mathbf{c}^{\pi}_n(t)$ is defined to be α-*sub-optimal* if the reward obtained by the strategy is less than α fraction of the optimal reward:

$$R(\mathbf{c}^{\pi}_n(t)) < \alpha \cdot R^{\text{opt}}_{\Theta_n}.$$

Let $\Delta^{f}_{n,\max}$ be the difference in expected reward between the optimal caching strategy and the *worst* α-sub-optimal caching strategy in which file f is cached at $\mathcal{S}_n$ i.e., it is the sub-optimality gap of the strategy which yields the lowest reward when file f is cached at $\mathcal{S}_n$. Similarly, let $\Delta^{f}_{n,\min}$ be the sub-optimality gap of the *best* α-sub-optimal caching strategy i.e., the best strategy with reward less than $\alpha \cdot R^{\text{opt}}_{\Theta_n}$ in which file f is cached at $\mathcal{S}_n$. Also, let

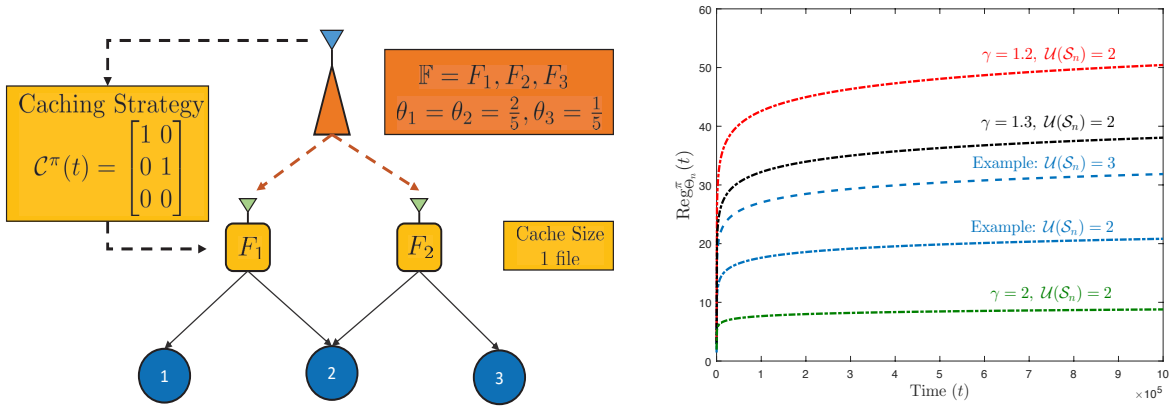
$$\Delta_{n,\max} = \max_{f \in \mathbb{F}} \Delta^{f}_{n,\max}.$$

be the *worst* sub-optimality gap across all files $f \in \mathbb{F}$. Now, define a counter $N_b(t)$ which counts the number of times a α-sub-optimal caching strategy is employed by the CMAB algorithm upto a time instant t. Using techniques similar to [213], we can derive an upper bound on the expected number of α-sub-optimal periods at a finite time horizon $t = T$:

$$N_b(T) \leq (1-\beta)(T-F) + 2F\left[\frac{\zeta(\Psi_{f,n}-1)}{|\mathcal{U}(\mathcal{S}_n)|^{\Psi_{f,n}}} + \Delta_{n,\max}\sum_{f\in\mathbb{F}}\frac{\Psi_{f,n}\log(|\mathcal{U}(\mathcal{S}_n)|T)}{(g^{-1}(\Delta^{f}_{n,\min}))^2}\right], \tag{6.11}$$

where $g(\cdot)$ is a strictly increasing, invertible function[2] such that for any two popularity distributions Θ and Θ', we have $|R_{\Theta}(\mathbf{c}^{\pi}(t)) - R_{\Theta'}(\mathbf{c}^{\pi}(t))| \leq g(\Lambda)$ if $\max_{f\in\mathbf{c}^{\pi}(t)} |\theta_f - \theta'_f| \leq \Lambda$. From (6.11), we can see that the number of α-sub-optimal periods scales logarithmically with time T. The following Lemma gives an upper bound on the regret for the n-th sBS for the Collaborative Caching Algorithm (Algorithm 4).

[2]The function $g(\cdot)$ is a *bounded smoothness function* [213] which is an artifact of the proof of the upper bound on the regret bound in Lemma 8.



(a) (b)

Figure 6.2: (a) A collaborative caching set-up with $N = 2$ sBSs, $K = 3$ users, (b) Upper bound on regret vs. time.

Lemma 8. *At each sBS, the regret at time $t = T$, when using an $(\alpha, \beta)-$approximation CCP algorithm, is upper bounded by:*

$$\mathrm{Reg}^{\pi}_{\Theta_n,\alpha,\beta}(T) \leq \sum_{f\in\mathbb{F},\Delta^f_{n,\min}>0} \frac{2\Psi_{f,n}\log\left(|\mathcal{U}(\mathcal{S}_n)|T\right)}{(g^{-1}(\Delta^f_{n,\min}))^2}\cdot\Delta^f_{n,\max} + \left(\frac{2\zeta(\Psi_{f,n}-1)}{|\mathcal{U}(\mathcal{S}_n)|^{\Psi_{f,n}}}+1\right)\cdot F\cdot\Delta_{n,\max}, \tag{6.12}$$

where $\zeta(x)$ is the Riemann Zeta function.

The proof is presented in Appendix D.1. It can be seen from (6.12), that the regret at each sBS scales as $O\left(\log\left[|\mathcal{U}(\mathcal{S}_n)|T\right]\right)$. Thus, when $T >> |\mathcal{U}(\mathcal{S}_n)|$, the loss due lack of knowledge of the true popularity distribution i.e., the regret, grows slowly with T. It can also be seen that the regret scales linearly with the system parameter $\Psi_{f,n}$. We next present an example to elucidate the regret bound, noting that the result holds for any arbitrary network.

Example 12 (*Regret Bounds for 2 sBS, 3 User System*)**.** We consider the network in Fig. 6.2(a). Let there be a library of 3 files $\mathbb{F} = \{F_1, F_2, F_3\}$ with $S_f = 1, \forall f$ and let each sBS have a cache size of $M = 1$. We also define a file popularity profile $\Theta_n = \{\theta_{n,1}, \theta_{n,2}, \theta_{n,3}\}$ such that $\gamma = 2$ and

$$\theta_{n,1} = \theta_{n,2} = \frac{2}{5} \quad \theta_{n,3} = \frac{1}{5},\ n = 1, 2.$$

For the system under consideration, the scaling factor $\Psi_{f,n}$ at each sBS is given by

$$\Psi_{f,n} = \Psi_n = \frac{3}{2}\times\left(\frac{2\times 1}{3^2}\right)^2 = 0.0741;\quad f = 1, 2, 3. \tag{6.13}$$

We assume a caching strategy where only whole files can be cached i.e., $c^{\pi}_{f,n} \in \{0,1\}$. Since the file F_3 is the least popular in both sBSs, lets assume that an optimal caching policy is as shown in Fig. 6.2(a) i.e., $\mathcal{S}_1$ stores F_1 and $\mathcal{S}_2$ stores F_2. The reward in this case is equal to $\theta_{f,n}$ since files have unit size. In this case, by definition, the only non-zero $\Delta^{f}_{m,\min}, \Delta^{f}_{m,\max}$ are those corresponding to file F_3 and given by

$$\Delta^{3}_{n,\min} = \Delta^{3}_{n,\max} = \frac{2}{5} - \frac{1}{5} = \frac{1}{5}. \tag{6.14}$$

Again for this example, we use $g(x) = x$ or conversely $g^{-1}(x) = x$. Then at time $t = T$, we have

$$\mathrm{Reg}^{\pi}_{\Theta_n}(T) \leq \frac{2\Psi_n \log(2T)}{0.2} + \frac{3}{5}\left(\frac{2\zeta(\Psi_n - 1)}{2^{\Psi_n}} + 1\right). \tag{6.15}$$

The upper bound on the regret is plotted as a function of time t in Fig. 6.2(b). We also plot a bound for a case where the number of users per sBS increases to 3. These are shown by the blue curves. It is seen that the value of regret increases with increasing users which is expected. But for both cases, the regret grows very slowly with large values of t. Thus in both cases, the bound shows that with increasing t, the learning converges and the regret grows only slowly with time. Further we also generate some popularities using the Zipf distribution [184] with varying skewness γ. The optimal placement in this case is to place the most popular file in the cache. It can be seen that as the skewness increases, the regret decreases i.e., as fewer files become highly popular, it gets easier to learn the optimal cache placement. Also, when $\gamma = 0$ i.e., every file is equally popular, the regret grows arbitrarily large and optimal cache placement is difficult to learn. This will be further illustrated in Section 6.6 where we present simulation results. □

Corollary 10 (Overall Regret)**.** *The system regret over all the N sBSs is upper bounded by:*

$$\sum_{n=1}^{N}\left[\sum_{f\in\mathbb{F},\Delta^{f}_{n,\min}>0} \frac{2\Psi_{f,n}\log\left(|\mathcal{U}(\mathcal{S}_n)|t\right)}{(g^{-1}(\Delta^{f}_{n,\min}))^2}\cdot\Delta^{f}_{n,\max} + \left(\frac{2\zeta(\Psi_{f,n}-1)}{|\mathcal{U}(\mathcal{S}_n)|^{\Psi_{f,n}}}+1\right)\cdot F\cdot\Delta_{n,\max}\right] \tag{6.16}$$

From Corollary 10, we can see that the overall regret for the multi-sBS system with N sBSs is the sum of the regrets of the individual sBSs. Thus, if the CCP policy π intelligently places content such that a user u with multiple connections can download its requested content from *any one* of its connections, then the regret for all the sBS in $\mathcal{N}(u)$ also decreases. This is one of main design goals for the CCP solvers as discussed in the sequel. In the next section we detail the design of the Collaborative Cache Placement policy π which chooses the joint caching strategy at the sBSs accounting for the topology of the network.

6.5 Collaborative Cache Placement

In this section, we detail the design of the collaborative cache placement (CCP) policy π which enables the central BS to determine joint caching strategies at the sBSs. The CCP in Algorithm 4

assumes full topological knowledge and takes as input, the CMAB indices, $\overline{\Theta}_1, \overline{\Theta}_2, \ldots, \overline{\Theta}_N$ at time t. The caching strategy $\mathcal{C}^\pi(t)$ determines which files are cached in the sBSs for servicing the user requests at time $t+1$. The CCP in Algorithm 4 produces an $(\alpha, \beta)-$approximate solution i.e., an $\alpha-$approximate solution for at least β fraction of time. Thus, for an *optimal caching scheme*, we have $\alpha = \beta = 1$. However, as discussed in the following subsection, not all caching policies lend themselves to optimal solutions in polynomial time and hence the use of an $(\alpha, \beta)-$approximation algorithm is assumed for generality. It is to be noted that for the regret bound in Lemma 8, the terms $\Delta^f_{n,\max}, \Delta^f_{n,\min}$ are dependent on α while the expected number of α-sub-optimal periods in (6.11) scales with β. Therefore, we should aim to use cache placement algorithms with $\beta = 1$ and α close to 1. We next highlight two different caching strategies considered in this work and comment on the α, β values for both:

1. The first is an *uncoded caching strategy* where whole files are stored in the sBS caches. We show that this is an NP-hard problem and propose a novel graph-coloring based approximate solution.

2. The second is a *coded caching strategy* where files are encoded using rateless-codes [152, 178, 218] thereby enabling fractions of files to be stored in caches. We show that, in contrast to the uncoded caching problem, this strategy leads to a concave reward function and the cache placement for reward maximization yields an optimal linear programming solution.

In the following discussion the indexing on t is dropped for simplicity. We assume that user requests and the data downloaded from the caches to satisfy these requests, at time t, are observed. The demands are then updated for the cache placement phase at the next time instant $t+1$.

6.5.1 Uncoded Collaborative Caching Strategy

We first formulate the optimal uncoded caching strategy where entire files are cached at the sBSs. In order to facilitate learning, the CMAB indices $\overline{\Theta}_n$ are used as a proxy for the true popularity of all files $f \in \mathbb{F}$ at the sBSs $\mathcal{S}_n$. Based on this acquired knowledge, the average reward for each user u in the network (upto the current time step) can be defined as:

$$\overline{R}^f_u = \max_{n \in \mathcal{N}(u)} \overline{\theta}_{f,n} \cdot S_f \cdot c^\pi_{f,n}, \tag{6.17}$$

where $c^\pi_{f,n}$ are constrained to be binary variables i.e., $c^\pi_{f,n} \in \{0, 1\}$. They are the elements of the $F \times N$ cache assignment matrix $\mathcal{C}^\pi$ and $c^\pi_{f,n} = 1$ denotes that the file f is cached at $\mathcal{S}_n$. The reward formulation is such that for a user connected to multiple sBSs, the obtained reward is maximized when the requested file is placed in the sBS where it is *most popular* i.e., the sBS with the highest $\overline{\theta}_{f,n}$. Additionally, in case the file is placed in multiple locations, the $\max(\cdot)$ ensures that only the reward based on the most popular location is counted. As a result, the reward function implicitly encourages the placement of a file f requested by user u at the sBS in $\mathcal{N}(u)$ where the popularity of the file is the highest.

The optimal uncoded cache placement maximizes the sum reward of the users in the network. The cache placement is formulated as the following binary integer program subject to cache memory constraints at each sBS:

$$\textbf{UNC-1:} \quad \max_{\mathcal{C}^\pi} \sum_{u\in\mathbb{U}} \sum_{f=1}^{F} \overline{R}_u^f \qquad \text{s.t.} \quad \sum_{f=1}^{F} c_{f,n}^\pi \cdot S_f \leq M, \ \ \forall n.$$

The optimal cache placement $\mathcal{C}^\pi$ in **UNC-1** that maximizes the sum reward of the multi-sBS network ensures that there is minimal repetitive placement of a file f in the system by virtue of the $\max(\cdot)$ function in the reward. The caveat remains that repetition might be necessary to account for similar requests from users not connected to the same sBSs. The formulation is flexible in this regard by allowing repetition since the overall reward is the sum of rewards for each user in the network. Thus our formulation utilizes the network topology to optimize the cache placement at the sBSs.

Lemma 9. *The binary integer program **UNC-1** is NP-hard.*

Proof. The lemma can be proved by showing that specific instances of the problem is NP-hard. Consider the two network examples presented in Fig. 6.3. Network Example 1 represents a net-

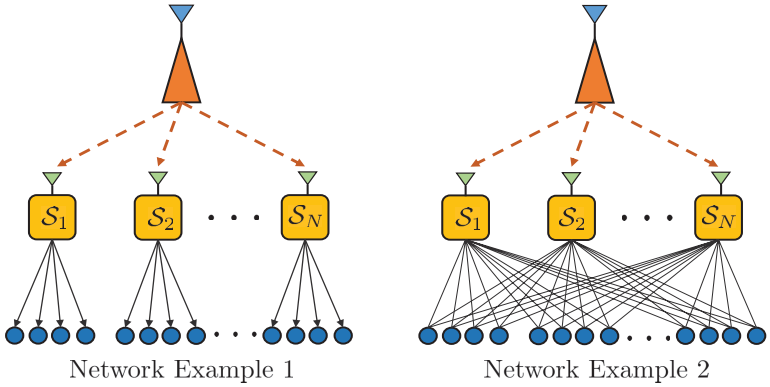


Figure 6.3: Network Topology Examples

work topology where every sBS has its own set of users and each user is connected to only one sBS. Under this topology, the multi-sBS cache placement problem reduces to solving N single sBS problems of the following form subject to the cache storage constraint of **UNC-1**:

$$\max_{\mathcal{C}^\pi} \sum_{n=1}^{N} \sum_{u\in\mathcal{U}(\mathcal{S}_n)} \sum_{f=1}^{F} \overline{\theta}_{f,n} \cdot S_f \cdot c_{f,n}^\pi. \tag{6.18}$$

Each of the single-sBS placement problems are NP-hard Knapsack Problems [219] with values $\overline{\theta}_{f,n}S_f$ and weights S_f. Again, Network Example 2 shows the case of a fully connected network,

where all users are connected to all sBSs. In this case, solving **UNC-1** becomes analogous to solving a single sBS problem but with a cache size of NM units. This is also a Knapsack problem and is NP-hard. Since these two special cases of **UNC-1** are NP-hard, **UNC-1** itself is NP-hard. Specifically, **UNC-1** is an example of a the Generalized Assignment Problem (GAP) [220]. This concludes the proof of the lemma. □

Since **UNC-1** is NP-hard, we next formulate an approximation algorithm which runs in polynomial time. We propose a novel graph coloring based approach for reducing the given cache placement problems to multiple sub-problems where each sub-problem can be cast as a Separable Assignment Problem (SAP) [221] which yield $\left(1-\frac{1}{e}\right)-$approximations but in exponential time. For this work, we use a local-search based $\epsilon-$greedy polynomial time approximation algorithm for each SAP for some $\epsilon > 0$.

6.5.2 An Approximation Algorithm for Uncoded Caching

In this section we present a novel approximation algorithm for **UNC-1**. The algorithm stems from two key ideas namely:

1. The N sBSs are divided into groups such that sBSs within each group have a high number of common users. This reduction helps in grouping sBSs such that each group of sBSs can be treated as a single contiguous cache.
2. Within each group, a variant of the **UNC-1** problem is solved with an added constraint that files cannot be replicated across the sBSs. This converts **UNC-1** to an SAP and was originally studied in [221].

We first present the formulation of the uncoded collaborative caching algorithm and then discuss the two steps in detail.

To this end, we first present a reduction of the original NP-hard GAP problem **UNC-1** to a related problem whose solution upper bounds the solution of **UNC-1** subject to the same constraints:

$$\begin{aligned}\max_{\mathcal{C}^\pi} \sum_{u\in\mathbb{U}}\sum_{f=1}^{F}\max_{n\in\mathcal{N}(u)}\overline{\theta}_{f,n}\cdot S_f\cdot c^\pi_{f,n} &\overset{\text{(a)}}{\leq} \max_{\mathcal{C}^\pi}\sum_{u\in\mathbb{U}}\sum_{n=1}^{\mathcal{N}(u)}\sum_{f=1}^{F}\overline{\theta}_{f,n}\cdot S_f\cdot c^\pi_{f,n}\\ &\overset{\text{(b)}}{\leq} \max_{\mathcal{C}^\pi}\sum_{n=1}^{N}\sum_{u\in\mathcal{U}(\mathcal{S}_n)}\sum_{f=1}^{F}\overline{\theta}_{f,n}\cdot S_f\cdot c^\pi_{f,n}\\ &\overset{\text{(c)}}{=} \max_{\mathcal{C}^\pi}\sum_{n=1}^{N}|\mathcal{U}(\mathcal{S}_n)|\sum_{f=1}^{F}\overline{\theta}_{f,n}\cdot S_f\cdot c^\pi_{f,n},\end{aligned} \tag{6.19}$$

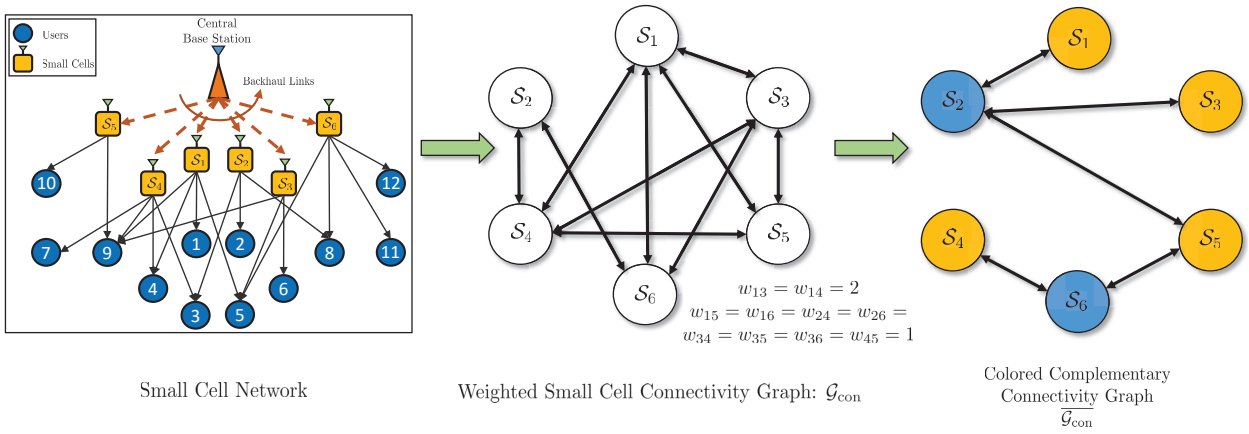


Figure 6.4: Coloring the weighted complementary connectivity graph using the proposed W-DSATUR algorithm.

where, (a) follows from replacing the $\max(\cdot)$ function with the sum over all the sBS $\mathcal{S}_n \in \mathcal{N}(u)$, and (b)-(c) follows from reversing the order of summation in (a). Note that the term within the first summation in (6.19) is the expected accumulated reward from (6.9). The problem in (6.19) is still a GAP and hence NP-hard.

In order to simplify the problem in (6.19), it is divided into sub-problems such that each sub-problem deals with a group of sBSs. We aim to identify sBSs with common users which can be grouped together to act as a single cache. Thus, the summation over n in (6.19) is divided into multiple sums:

$$\max_{\mathcal{C}^\pi} \sum_{g=1}^{G} \sum_{n \in g} |\mathcal{U}(\mathcal{S}_n)| \sum_{f=1}^{F} \overline{\theta}_{f,n} \cdot S_f \cdot c^{\pi}_{f,n},, \tag{6.20}$$

where, G is the total number of groups of sBSs. Sub-problems for each group g can then be solved individually. Now considering such a sub-problem, for sBSs which can be grouped together into a single cache, each file f requested by a user in the group will be placed in *only one* of the caches. Thus assuming a group g of $\tilde{n}$ ($\tilde{n} \leq N$) sBSs, the following problem needs to be solved:

$$\textbf{UNC-2:} \quad \max_{\mathcal{C}^\pi} \sum_{n=1}^{\tilde{n}} |\mathcal{U}(\mathcal{S}_n)| \sum_{f=1}^{F} \overline{\theta}_{f,n} S_f c^{\pi}_{f,n}$$

$$\text{s.t.} \quad \sum_{f=1}^{F} c^{\pi}_{f,n} \cdot S_f \leq M, \ \forall n \in g; \ \sum_{n=1}^{\tilde{n}} c^{\pi}_{f,n} = 1, \ \ \forall f. \tag{6.21}$$

The additional constraint $\sum_{n=1}^{\tilde{n}} c^{\pi}_{f,n} = 1, \ \ \forall f$, ensures the placement of each file in only one sBS within the group. The optimization in (6.21) is a special case of the SAP presented in [221]. We

Algorithm 5 W-DSATUR COLORING ALGORITHM

1: Arrange nodes in $\overline{\mathcal{G}_{\text{con}}}$ by decreasing order of degrees.
2: Color a node of maximal degree with color 1.
3: Choose a node $\mathcal{S}_i$ with a maximal degree of saturation. If there is an equality, choose any node of maximal degree in the uncolored sub-graph.
4: **if** All non-neighbors of $\mathcal{S}_i$ are uncolored **then**
5: Color $\mathcal{S}_i$ with the least possible color.
6: **else**
7: Order all the non-neighbors, $\mathcal{S}_j$, of $\mathcal{S}_i$ in $\overline{\mathcal{G}_{\text{con}}}$ in decreasing order of w_{ij} in $\mathcal{G}_{\text{con}}$.
8: Select the lowest ranked node in this list which is colored with a feasible color. Assign the color of this node to current node.
9: **end if**
10: If all the vertices are colored, stop. Else goto 3.

next present the graph coloring based sBS grouping algorithm and then provide an greedy local search based algorithm for caching within each group.

6.5.2.1 Graph Coloring based sBS Grouping

In order to facilitate the grouping of sBSs, first, a *weighted connectivity graph* $\mathcal{G}_{\text{con}} = (\text{sBS}, E)$ is constructed, where edges between any two sBSs, $(\mathcal{S}_i, \mathcal{S}_j)$ exist if they have common users i.e., $\{\mathcal{U}(\mathcal{S}_i) \bigcap \mathcal{U}(\mathcal{S}_j) \neq \phi\}$. The weight of each edge w_{ij} is given by the number of common users between $\mathcal{S}_i$ and $\mathcal{S}_j$. Since $\mathcal{G}_{\text{con}}$ is an undirected graph, $w_{ij} = w_{ji}$. Next we construct the *complementary connectivity graph* $\overline{\mathcal{G}_{\text{con}}} = (\text{sBS}, E^{'})$ where edge $(\mathcal{S}_i, \mathcal{S}_j) \in E^{'}$ iff the edge between $(\mathcal{S}_i, \mathcal{S}_j) \notin E$. Note that for any node $\mathcal{S}_i$ in $\overline{\mathcal{G}_{\text{con}}}$, the set of non-neighbors forms a set of weighted neighbors in $\mathcal{G}_{\text{con}}$. Fig. 6.4 shows the construction of the weighted connectivity and complementary connectivity graph for the example network shown in Fig. 3.1.

Next, a $k-$coloring for the graph $\overline{\mathcal{G}_{\text{con}}}$ is obtained [222]. A $k-$coloring is the assignment of k colors to the vertices of a graph such that no two adjacent vertices have the same color. Note that $k = N$ signifies that $\overline{\mathcal{G}_{\text{con}}}$ is fully connected i.e., none of the sBSs have common users as in Network Example 1 in Fig. 6.3. Again for Network Example 2 in Fig. 6.3, $\overline{\mathcal{G}_{\text{con}}}$ will have no edges and $k = 1$. Since these two are the extreme examples of topology, a real network is expected to have $1 \leq k \leq N$ colors. For coloring the weighted $\overline{\mathcal{G}_{\text{con}}}$, we propose a modified version of the greedy degree-of-saturation (DSATUR) algorithm [222] which accounts for weighted edges. The *degree-of-saturation* of a node is the maximum number of unique colors that can be found in the neighborhood of the node. In this work, we incorporate the edge weights w_{ij} of the original connectivity graph to form the Weighted-DSATUR (W-DSATUR) coloring algorithm which is presented in Algorithm 5.

The W-DSATUR algorithm assigns integers $1, 2, \ldots, k$ to k colors and initializes by coloring the

sBS with highest degree in $\overline{\mathcal{G}_{\text{con}}}$ with the color numbered 1. In any given iteration, W-DSATUR assigns to an uncolored node, the lowest numbered color which is available i.e., which is not present in its set of neighbors. For any uncolored node $\mathcal{S}_i$, we rank all its non-neighbors, which are colored with an available color, in decreasing order of the weights from $\mathcal{G}_{\text{con}}$ and then choose to assign the color of the node with highest weight ensuring pairing of neighbors with highest common users. The algorithm gives as an output, the set of colors for every node in $\overline{\mathcal{G}_{\text{con}}}$. Nodes with same color e.g., $\mathcal{S}_1, \mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5$ in Fig. 6.4 have a set of common users and for the purpose of collaborative caching, their caches can be grouped together and file placement without repetition can be performed. The total number of colors required to color $\overline{\mathcal{G}_{\text{con}}}$ is given by $\chi(\overline{\mathcal{G}_{\text{con}}})$ i.e., the sBSs can be grouped into $\chi(\overline{\mathcal{G}_{\text{con}}})$ groups.

6.5.2.2 Collaborative Caching within sBS Groups

After grouping, we consider $\chi(\mathcal{G}_{\text{con}})$ different cache placement problems of the form **UNC-2** in (6.21) within each group. The per-group sub-problems form a class of SAP and are NP-hard. In [221], Fleischer et. al provide $\left(1-\frac{1}{e}\right)-$approximation algorithms for this class of problems which run in exponential time. However, in this work, we use such a local search based polynomial time approach [221].

The placement problem for each sBS can be presented as a knapsack problem (refer to proof of Lemma 9). Knapsacks are a well-known class of NP-hard problems which have finite polynomial time approximation algorithms [153, 219, 223] with approximation ratio $\alpha_{\text{KP}} \in (0, 1]$. For the caching problem, $\alpha_{\text{KP}}-$approximate placement yields a reward which is at least α_{KP} fraction of the optimal reward for given CMAB indices $\overline{\Theta}_n$. Based on the existence of an $\alpha_{\text{KP}}-$approximation for the single sBS knapsack problem, a local search based $\epsilon-$greedy cache placement algorithm [221] is presented in Algorithm 6 lines $5-18$ for collaborative caching within each group. For a given color $\mathsf{C} \in \{1, 2, \ldots, \chi(\mathcal{G}_{\text{con}})\}$, with N_c sBSs, the local search algorithm decouples the N_c placement problems. Based on the value of ϵ, the algorithm runs $\frac{1}{\alpha_{\text{KP}}} N_c \ln\left(\frac{1}{\epsilon}\right)$ iterations. The algorithm initializes with an empty cache assignment. Within each iteration, for each sBS, the algorithm calculates a $\mathsf{value}_{f,n}$ for each file by assigning to it, the CMAB index if the file is cached at any other sBS within the group and 0 if it is cached at $\mathcal{S}_n$ or not cached at all. Based on this, a marginal value for each file $m_{f,n}$ is calculated as the difference of the CMAB index $\overline{\theta}_{f,n}$ and $\mathsf{value}_{f,n}$. Using this marginal value, the following single sBS knapsack problem is solved at $\mathcal{S}_n$:

$$\textbf{SKP:} \max_{\mathbf{c}_n^{\pi}} \sum_{f=1}^{F} |\mathcal{U}(\mathcal{S}_n)| m_{f,n} S_f c_{f,n}^{\pi} \text{ s.t. } \sum_{f=1}^{F} c_{f,n}^{\pi} S_f \leq M.$$

An $\alpha_{\text{KP}}-$approximate cache placement $\mathbf{c}_n^{\pi,\mathsf{new}}$ for single sBS **SKP** sub-problem is discussed in the next section. Based on this new cache placement, the marginal value for this solution $\mathcal{M}_n$ is calculated for $\mathcal{S}_n$. Next, difference D_n between the value of current cache placement and the marginal value of new placement is calculated. For the sBS $\mathcal{S}_{n^*}$ with the maximum differential D_{n^*}, the placement is changed to the new one if its $D_{n^*} > 0$ i.e., if the new placement improves

the reward. The approximation ratio of the greedy cache placement algorithm is given by the following lemma.

Lemma 10. *Given an $\alpha_{KP}-$approximation for the single sBS cache placement problem **SKP**, the local search based $\epsilon-$greedy cache placement in lines $5-18$ of Algorithm 6 has an approximation ratio of $\left(\frac{\alpha_{KP}}{\alpha_{KP}+1}-\epsilon\right)$ for some $\epsilon>0$.*

The Lemma directly follows from [221, Theorem 3.1]. Once the greedy search algorithm finishes, a residual cache placement is performed to maximize cache utilization. In this step, for each sBS $\mathcal{S}_n$, if additional cache space is left after initial placement, the most popular files not already in the cache of $\mathcal{S}_n$ are placed there subject to the storage constraint. The reward obtained after this placement can only be greater than or equal to the reward without it. Therefore, the residual cache placement stage potentially improves the approximation ratio of the caching strategy. Next we discuss an $\alpha_{\text{KP}}=0.5$ approximation of the single sBS knapsack problem **SKP**.

6.5.2.3 Single sBS Greedy Knapsack

A greedy approximation for the single-sBS knapsack problem is based on a relaxation of the cache placement variable such that $0\leq c^{\pi}_{f,n}\leq 1$ which reduces the placement problem to a linear program (LP). The solution to the LP is then rounded to give an $\alpha_{\text{KP}}=0.5$ approximate solution [223] for **SKP**. The solution is as follows:

1. Reorder $m_{f,n}$ such that $m_{i,n}\geq m_{j,n}$ if $i<j, \forall i,j\in\mathbb{F}$ i.e., relabel files in decreasing order of their marginal values. For **SKP** with values $v_i=m_{i,n}S_i$ and weights $w_i=S_i,\ \forall i\in\mathbb{F}$, the re-ordering satisfies the *regularity condition* [223]:
$$\frac{v_1}{w_1}\geq\frac{v_2}{w_2}\geq\ldots\geq\frac{v_F}{w_F}\triangleq m_{1,n}\geq m_{2,n}\geq\ldots\geq m_{F,n}.$$
As a result, the following assignment assures $\alpha_{\text{KP}}=0.5$.
2. Place files with highest marginal value $m_{f,n}$ sequentially into the cache of $\mathcal{S}_n$ until capacity of the cache is reached.

The above approximation algorithm is used to solve each **SKP** in Algorithm 6. Furthermore, if $S_f=1,\ \forall f\in\mathbb{F}$, the above placement is optimal with $\alpha_{\text{KP}}=1$. Combining the graph coloring based sBS grouping and the per-group greedy caching yields the uncoded collaborative caching algorithm presented in Algorithm 6.

6.5.2.4 Approximation Ratio

Let R^{opt}_{unc} be the reward of the optimal uncoded caching strategy given by the solution to **UNC-1**. Again, consider the reduction into $\chi(\overline{\mathcal{G}_{\text{con}}})$ number of sub-problems of the form **UNC-2** in (6.21).

Algorithm 6 UNCODED COLLABORATIVE CACHING

GRAPH COLORING BASED SBS GROUPING:

Construct graphs $\mathcal{G}_{\text{con}}$ and $\overline{\mathcal{G}_{\text{con}}}$

Color $\overline{\mathcal{G}_{\text{con}}}$ using Algorithm 5 and find $\chi(\overline{\mathcal{G}_{\text{con}}})$.

LOCAL SEARCH BASED APPROXIMATE CACHING:

for Each color $\mathsf{C} \in \{1, 2, \ldots, \chi(\overline{\mathcal{G}_{\text{con}}})\}$ **do**

  Initialize Cache placement matrix $\mathcal{C}^{\pi} = [\mathbf{0}]_{F \times N_c}$ where N_c is the number of sBSs with color C.

  while `loop_count` $\leq \frac{1}{\alpha_{\text{KP}}} N_c \ln\left(\frac{1}{\epsilon}\right)$ **do**

    Let current cache placement be

$$\mathcal{C}^{\pi}_{\mathsf{curr}} = [\mathbf{c}_1^{\pi,\mathsf{curr}}, \mathbf{c}_2^{\pi,\mathsf{curr}}, \ldots, \mathbf{c}_{N_c}^{\pi,\mathsf{curr}}]$$

    for Each sBS $\mathcal{S}_n$ with color C **do**

      For each file $f \in \mathbb{F}$, let

$$\mathsf{value}_{f,n} = \begin{cases} \overline{\theta}_{f,n'}, & \text{if } c_{f,n'}^{\pi,\mathsf{curr}} = 1 \ \exists\, n' \neq n \\ 0, & \text{if } c_{f,n}^{\pi,\mathsf{curr}} = 1 \text{ or } f \notin C^{\pi}_{\text{curr}} \end{cases}$$

      For each file f, let marginal value

$$m_{f,n} = \overline{\theta}_{f,n} - \mathsf{value}_{f,n}$$

      Solve the **SKP** problem at $\mathcal{S}_n$ to cache files, using value $|\mathcal{U}(\mathcal{S}_n)| m_{f,n} S_f$ and weights S_f.

      Let $\mathbf{c}_n^{\pi,\mathsf{new}}$ be the new cache placement and

$$\mathcal{M}_n = \sum_{f \in \mathbb{F}} m_{f,n}\, S_f\, c_{f,n}^{\pi,\mathsf{new}}$$

      be the marginal value for this solution.

    end for

    For each sBS $\mathcal{S}_n$ with color C, let

$$D_n = \mathcal{M}_n - \sum_{f \in \mathbb{F}} \overline{\theta}_{f,n}\, S_f\, c_{f,n}^{\pi,\mathsf{curr}}$$

    Let $\mathcal{S}_{n^*}$ be the sBS s.t. $n^* = \arg\max_n D_n$.

    if $D_{n^*} > 0$ **then**

      Change the cache placement $\mathbf{c}_{n^*}^{\pi,\mathsf{curr}} \leftarrow \mathbf{c}_{n^*}^{\pi,\mathsf{new}}$

    end if

  end while

end for

RESIDUAL CACHE PLACEMENT:

for Each sBS $\mathcal{S}_n \in \mathbb{S}$ **do**

  for All files $\mathbb{F}_n \notin$ cache of $\mathcal{S}_n$ **do**

    Rank files $f \in \mathbb{F}_n$ in descending order of $\overline{\theta}_{f,n}$.

  end for

  Place files from $\mathbb{F}_n$ with highest $\overline{\theta}_{f,n}$ in the cache subject to storage constraints

end for

Let R_{C}^{opt} be the optimal reward for the $\mathsf{C}-$th sub-problem for $\mathsf{C} \in \{1, 2, \ldots, \chi(\overline{\mathcal{G}_{\text{con}}})\}$. Thus we have:

$$R_{\mathsf{unc}}^{opt} \leq \sum_{\mathsf{C}=1}^{\chi(\overline{\mathcal{G}_{\text{con}}})} R_{\mathsf{C}}^{opt}. \tag{6.22}$$

Let $R_{\mathsf{C}}^{\mathsf{greedy}}$ be the reward of the greedy approximate cache placement in Algorithm 6. Then, we have

$$R_{\mathsf{C}}^{\mathsf{greedy}} \geq \left(\frac{\alpha_{\text{KP}}}{\alpha_{\text{KP}} + 1} - \epsilon \right) R_{\mathsf{C}}^{opt}. \tag{6.23}$$

Using (6.22) and (6.23), we have:

$$\sum_{\mathsf{C}=1}^{\chi(\overline{\mathcal{G}_{\text{con}}})} R_{\mathsf{C}}^{\mathsf{greedy}} \geq \left(\frac{\alpha_{\text{KP}}}{\alpha_{\text{KP}} + 1} - \epsilon \right) R_{\mathsf{unc}}^{opt}, \tag{6.24}$$

which gives an approximation guarantee on the greedy algorithm. Adding the final residual placement stage, the uncoded collaborative cache placement can be represented as an $(\alpha, \beta)-$approximation algorithm with $\alpha = \left(\frac{\alpha_{\text{KP}}}{\alpha_{\text{KP}}+1} - \epsilon \right)$ and $\beta = 1$. Since we have $\alpha_{\text{KP}} = 0.5$, we have an approximation ratio of $\alpha = \left(\frac{1}{3} - \epsilon \right)$. In the next section, we discuss coded collaborative cache placement which is an optimal strategy, given the file popularity profile, with $\alpha = \beta = 1$.

6.5.3 Coded Collaborative Caching Strategy

In this section we present the coded collaborative caching strategy wherein file segments (fractions of files) can be stored in the caches instead of the entire files. This can be achieved in practice by encoding files using a rateless code like Raptor Codes [152, 218]. The output codewords of the rateless code are *coded packets* of the original file which can be stored by the sBSs in their caches. These coded packets enable the recovery of the original file when a user gathers enough number of such packets. The storing of coded packets is modeled as each sBS storing a fraction of a file and the file is assumed to be recovered when fractions summing to 1 are recovered. The coded cache placement matrix C^{π} for a policy π is an $F \times N$ matrix defined in (6.4) with elements $c_{f,n}^{\pi} \in [0, 1]$ for $f \in \mathbb{F}$ and $n = 1, \ldots, N$ as in (6.5). The fraction $c_{f,n}^{\pi}$ is the fraction of coded packets of file f stored at $\mathcal{S}_n$. Note that for the coded caching strategy, $T_{f,n}$ in Algorithm 4 denotes the number of times a *fraction* of file $f \in \mathbb{F}$ has been cached in $\mathcal{S}_n$ upto time t. The CMAB indices, $\overline{\theta}_n, \forall n \in \{1, 2, \ldots, N\}$ in Algorithm 4, are evaluated using this definition of $T_{f,n}$.

The optimal cache placement is again modeled as a reward maximization problem. Based on the per-user reward model from (6.17), we formulate a related per-user reward for the coded caching strategy. The intuition behind the formulation is as follows: For a user u and file f at time t, we first order the sBSs, $\mathcal{S}_n \in \mathcal{N}(u)$, in descending order of CMAB indices upto time $t-1$, with the

first sBS being the one with the highest $\overline{\theta}_{f,n},\ \mathcal{S}_n \in \mathcal{N}(u)$. The amount of data for file f that the user u is able to download from $\mathcal{S}_n$ is denoted by $c^{\pi}_{f,n}S_f$. An expected accumulated reward for this download is given by $c^{\pi}_{f,n}\overline{\theta}_{f,n}S_f$. The expected accumulated reward for a user u that can download file f from the first k sBS's in the ordered list is given by,

$$\begin{aligned}\overline{R}_u^{f,k} &= \sum_{i=1}^{k-1} c^{\pi}_{f,i}\,\overline{\theta}_{f,i}S_f + \left(1 - \sum_{i=1}^{k-1} c^{\pi}_{f,i}\right)\overline{\theta}_{f,k}S_f \\ &= \left[\overline{\theta}_{f,k} - \sum_{i=1}^{k-1} c^{\pi}_{f,i}(\overline{\theta}_{f,k} - \overline{\theta}_{f,i})\right] S_f, \end{aligned} \tag{6.25}$$

where $k \in \{1, 2, \ldots, |\mathcal{N}(u)|\}$. The per-sBS reward function is similar to (6.17) in that it only assures the maximum reward for the case when the user downloads an entire file from the sBS where the file has the highest popularity. Since the sBSs are ordered and fractional storage is allowed, download of an entire file from the k most popular sBSs in the user's neighborhood yields the maximum rewards for the smallest k. Thus implicitly, the reward function also discourages replication of file fragments. The reward function is linear (affine) in terms of the placement variables $c^{\pi}_{f,n}$. The file f can be fully downloaded from the k best caches in the list if and only if $\sum_{i=1}^{k} c^{\pi}_{f,i} \geq 1$. The reward for a user u for downloading the file f is thus a piecewise-defined affine function of the placement variables $c^{\pi}_{f,n}$:

$$\overline{R}_u^f = \begin{cases} \overline{R}_u^{f,1} & \text{if } c^{\pi}_{f,1} \geq 1 \\ \vdots & \\ \overline{R}_u^{f,j} & \text{if } \begin{array}{l}\sum_{i=1}^{j-1} c^{\pi}_{f,i} < 1 \\ \sum_{i=1}^{j} c^{\pi}_{f,i} \geq 1\end{array} \\ \vdots & \\ \overline{R}_u^{f,|\mathcal{N}(u)|} & \text{if } \sum_{i=1}^{|\mathcal{N}(u)|-1} c^{\pi}_{f,i} < 1 \end{cases} \tag{6.26}$$

We aim to maximize the sum of minimum expected accumulated rewards for all users and files to determine the optimal cache placement.

Lemma 11. *The per-user reward, $\bar{R}_u^f$, is a concave function of the placement matrix $\mathcal{C}^{\pi}$.*

Proof. The point-wise minimum of a piece-wise affine function is concave [224]. Thus it suffices to show that the per-user reward in (6.26) can be represented as:

$$\bar{R}_u^f = \min_{k \in \{1,2,\ldots,|\mathcal{N}(u)|\}} \bar{R}_u^{f,k}. \tag{6.27}$$

Suppose that user u downloads the entire file from the first j caches in the list. Then the conditions $\sum_{i=1}^{j-1} c^{\pi}_{f,i} < 1$ and $\sum_{i=1}^{j} c^{\pi}_{f,i} \geq 1$ hold and we have $\bar{R}_u^f = \bar{R}_u^{f,j}$. We have to show that $\bar{R}_u^{f,j} \leq$

$\bar{R}_u^{f,j'}$ $\forall j \neq j'$. Using (6.25), the condition is given by:

$$\overline{\theta}_{f,j} - \sum_{i=1}^{j-1} c_{f,i}^{\pi}(\overline{\theta}_{f,j} - \overline{\theta}_{f,i}) \leq \overline{\theta}_{f,j'} - \sum_{i=1}^{j'-1} c_{f,i}^{\pi}(\overline{\theta}_{f,j'} - \overline{\theta}_{f,i}) \tag{6.28}$$

Considering the case for $j' > j$, we can re-write (6.28) as:

$$\left[\sum_{i=1}^{j} c_{f,i}^{\pi} - 1\right]\left(\overline{\theta}_{f,j} - \overline{\theta}_{f,j'}\right) + \sum_{i=j+1}^{j'-1} c_{f,i}^{\pi}\left(\overline{\theta}_{f,i} - \overline{\theta}_{f,j'}\right) \geq 0 \tag{6.29}$$

Now we have, $\sum_{i=1}^{j} c_{f,i}^{\pi} \geq 1$ and for all $j' > j$, we have $(\overline{\theta}_{f,j} - \overline{\theta}_{f,j'}) \geq 0$. Also since $i < j'$, the factor $(\overline{\theta}_{f,i} - \overline{\theta}_{f,j'}) \geq 0$. Therefore (6.29) is satisfied. Next, considering the case of $j' < j$, we can rewrite (6.28) as:

$$\left[\sum_{i=1}^{j-1} c_{f,i}^{\pi} - 1\right]\left(\overline{\theta}_{f,j'} - \overline{\theta}_{f,j}\right) + \sum_{i=j'+1}^{j-1} c_{f,i}^{\pi}\left(\overline{\theta}_{f,i} - \overline{\theta}_{f,j'}\right) \geq 0. \tag{6.30}$$

We have $\sum_{i=1}^{j-1} c_{f,i}^{\pi} - 1 \leq 0$ and since $i > j'$, the factor $\left(\overline{\theta}_{f,i} - \overline{\theta}_{f,j'}\right) \leq 0$ which ensures that (6.30) is satisfied. This completes the proof. □

The optimal coded cache placement problem can then be formulated as the following convex optimization problem:

$$\begin{aligned} \max_{\mathcal{C}^{\pi}} \quad & \sum_{u=1}^{U}\sum_{f=1}^{F} \min_{k\in\{1,2,\ldots,|\mathcal{N}(u)|\}} \overline{R}_u^{f,k} \\ \text{s.t.} \quad & \sum_{f=1}^{F} c_{f,n}^{\pi} \cdot S_f \leq M, \ \ \forall n \ \text{ and } \ c_{f,n}^{\pi} \in [0,1]. \end{aligned} \tag{6.31}$$

The optimization in (6.31) can be reduced to the following LP by introducing the auxiliary variable y_u^f [152]. The resulting optimization problem is given by,

$$\begin{aligned} \max_{\mathcal{C}^{\pi}} \ & \sum_{u=1}^{U}\sum_{f=1}^{F} \mathrm{y}_u^f \quad \text{s.t.} \quad \mathrm{y}_u^f \leq \overline{R}_u^{f,k}, \ \forall k \in \{1,2,\ldots,|\mathcal{N}(u)|\} \\ & \sum_{f=1}^{F} c_{f,n}^{\pi} \cdot S_f \leq M, \ \ \forall n, \ \text{and } c_{f,n}^{\pi} \in [0,1]. \end{aligned} \tag{6.32}$$

Since the LP in (6.32) can be solved optimally, the coded cache placement problem gives an $(\alpha, \beta)-$approximate placement for the CCP policy π with $\alpha = \beta = 1$ i.e., *given the CMAB*

indices the solution is optimal. The formulation in (6.32) can be considered as a convex relaxation of the uncoded caching problem **UNC-1** in the sense that the $\{0, 1\}$ binary assignment of the uncoded scheme is also a feasible solution to the coded caching problem. As a result, the average accumulated reward obtained with the coded cache placement for a given popularity index $\overline{\theta}_{f,n}$, encompasses the uncoded case. The caveat however, is that the convergence to the optimal reward value is also dependent on the accuracy of the CMAB based distributed learning and the network connectivity. In the next section, numerical results are presented for the proposed collaborative caching framework with discussions on impact of coded and uncoded caching on the learning accuracy.

6.6 Numerical Results

In this section we present simulation results under different network configurations to show the performance of the proposed collaborative caching strategies. In order to compare results, and present insight on multi-sBS edge caching networks, we first outline two baseline caching strategies:

- *Uncoded Local Caching Strategy:* In this strategy, uncoded cache placement is performed locally at each sBS using the local CMAB indices $\overline{\theta}_{f,n}$. In this procedure, every sBS learns the file popularity profile from its own set of connected users without taking into account the overall network connectivity. In other words, the topology-aware $T_{f,n}$ update discussed in Remark 21 is not performed and the exploration-exploitation based perturbation in this baseline scheme is based only on the local cache placement. This scheme entails the solving following knapsack problem at each sBS $\mathcal{S}_n$:

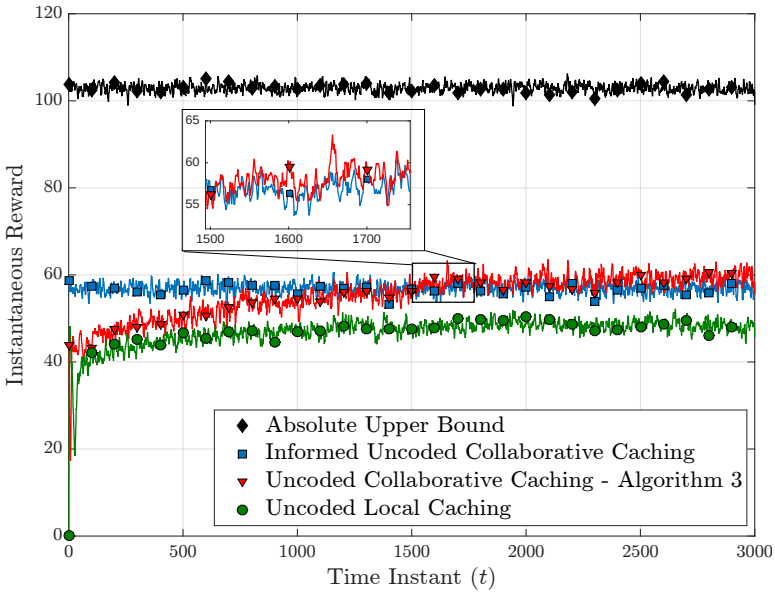
$$\max_{\mathcal{C}^{\pi}\in\{0,1\}} \sum_{f=1}^{F} \overline{\theta}_{f,n} S_f c_f^{\pi} \text{ s.t. } \sum_{f=1}^{F} c_f^{\pi} S_f \leq M, \ \forall n. \tag{6.33}$$

  This scheme was proposed and analysed in [153] for a single sBS caching system.

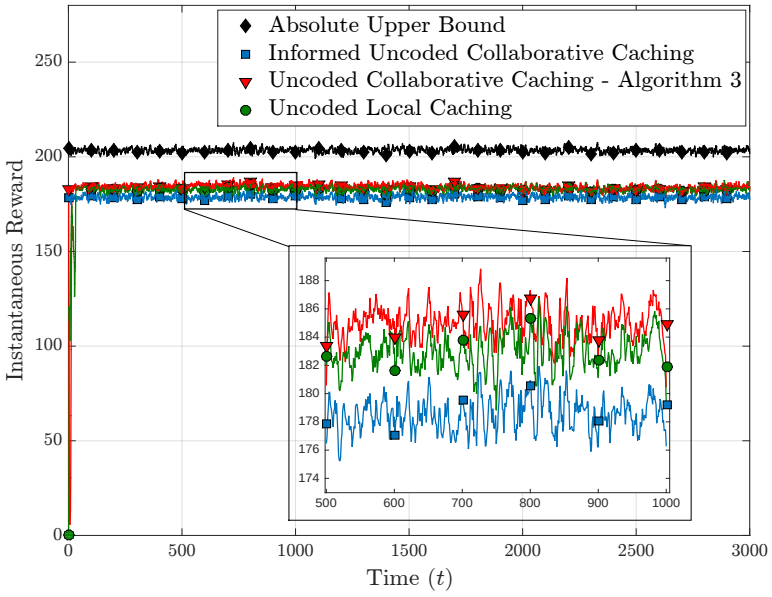
- *Coded Local Caching Strategy:* Coded local caching scheme is a linear relaxation of Uncoded local caching. In this case, the file placement is relaxed such that fractions of files (in the form of rateless encoded packets) can be stored in the caches. The learning process is again topology agnostic and $T_{f,n}$ is updated only based on fractions of files cached in each sBS. This scheme entails the solution of the following linear program at each sBS $\mathcal{S}_n$:

$$\max_{\mathcal{C}^{\pi}\in[0,1]} \sum_{f=1}^{F} \overline{\theta}_{f,n} S_f c_f^{\pi} \text{ s.t. } \sum_{f=1}^{F} c_f^{\pi} S_f \leq M, \ \forall n. \tag{6.34}$$

We compare the performance of the proposed schemes to the baselines to highlight the advantage of collaborative learning over local learning in a multi-sBS setting. In order to obtain results in



(a)



(b)

Figure 6.5: Performance of Uncoded Caching for (a) $\gamma = 0.56$ and (b) $\gamma = 2$.

a reasonable time-frame, we consider a $N = 5$ sBS setting with $K = 30$ users in the system. There is a file library of $F = 30$ files $\mathbb{F}$, with file sizes $S_f \in \{1, 3, 5, 7, 9\}$ units. The entire library has a size of 150 units. The cache size of each sBS is $M = 15$ units, which is 10% of the entire library. We assume that the users randomly request files at each time step t. The requests are generated i.i.d from a Zipf distribution (as in (6.1)) with $\gamma \in \{0.56, 2\}$. Furthermore, we consider

two collaborative caching strategies for comparison -

- *Absolute Upper Bound:* This strategy assumes that the instantaneous user demands are known apriori at the central BS and represents the optimal strategy in terms of reward maximization. It gives an upper bound on our learning based caching strategy.
- *Informed Collaborative Caching:* In this strategy, we consider that the popularity distribution i.e., the Zipf γ value at each sBS is known apriori. Based on this the use of the proposed caching strategies in the previous section with the known popularity distribution, instead of the CMAB indices, as input leads to a distribution optimal caching strategy. The reward of the learning-aided strategy should converge to the reward of this stragety over time.

For comparison of different schemes, we consider the metric of *instantaneous reward* at time t. The instantaneous reward is the total amount of data downloaded from the sBS caches to serve the requests of the users and is a measure of the cache hits at each instant. For the uncoded caching schemes, it is defined as

$$R^t_{\text{unc}} = \sum_{u\in\mathbb{U}} \sum_{f\in\mathbb{F}} \mathbb{1}_{\{f\in\mathcal{N}(u)\}} \cdot d^{t,u}_f \cdot S_f, \tag{6.35}$$

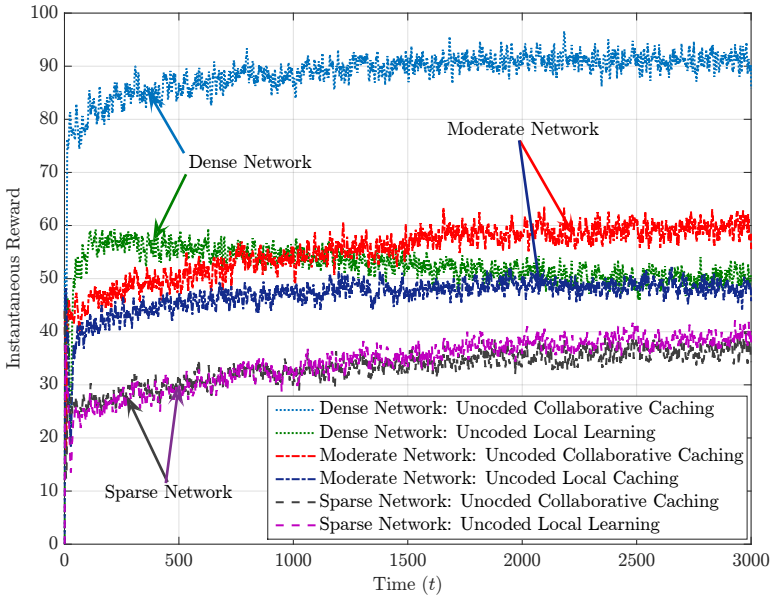
where, $d^{t,u}_f = 1$ for the file f requested by user u and $\mathbb{1}_{\{f\in\mathcal{N}(u)\}}$ is the indicator function and is equal to 1 when file f is cached in a sBS in the neighborhood of user u. For the coded caching strategy, the instantaneous reward is defined as:

$$R^t_{\text{cod}} = \sum_{u\in\mathbb{U}} \sum_{f\in\mathbb{F}} d^{t,u}_f \cdot \max\left\{1, \sum_{n\in\mathcal{N}(u)} c^{\pi}_{f,n}\right\} \cdot S_f, \tag{6.36}$$

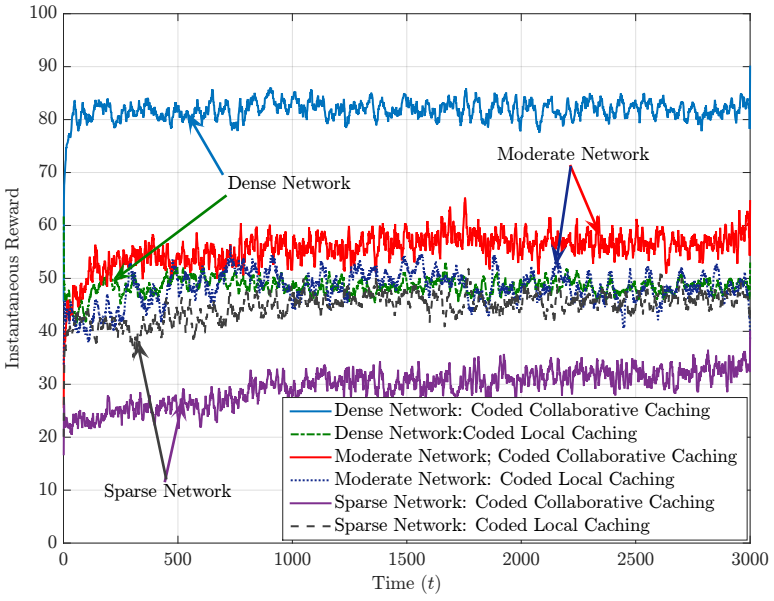
which accounts for the fact that a user u either downloads the entire file or the sum of fractions (≤ 1) placed in the caches of sBSs in its neighborhood $\mathcal{N}(u)$. Using this metric, we study the multi-sBS network under different connectivity profiles and caching strategies. To this end, we define three network topologies namely (i) *sparse connectivity* - every user is connected to at least 1 and at most 2 sBSs; (ii) *moderate connectivity* - every user is connected to at least 2 and at most 3 sBSs and (iii) *dense connectivity* - every user connects to at least 3 and at most 5 sBSs.

6.6.1 Uncoded Caching Strategies

We first simulate a network scenario with moderate connectivity. Fig. 6.5(a) shows the simulation results for uncoded caching for a file popularity distribution with $\gamma = 0.56$ i.e., the distribution is not skewed and there are many popular files in the system. Algorithm 6 is used for the caching strategy at each time step for collaborative caching with an epsilon value of $\epsilon = 0.01$. We also assume a worst-case $\alpha_{\text{KP}} = 0.5$ for the iterations in the greedy placement. It can be seen from



(a)



(b)

Figure 6.6: Performance under different network configurations (a) Uncoded Caching and (b) Coded Caching.

the results, that the collaborative learning based caching converges to the informed collaborative caching strategy i.e., the learning effectively converges to the true file popularity distribution. As expected, the absolute upper bound outperforms all other schemes. Finally, the collaborative caching strategy clearly out-performs the local caching strategy. Fig. 6.5(b) shows the simulation

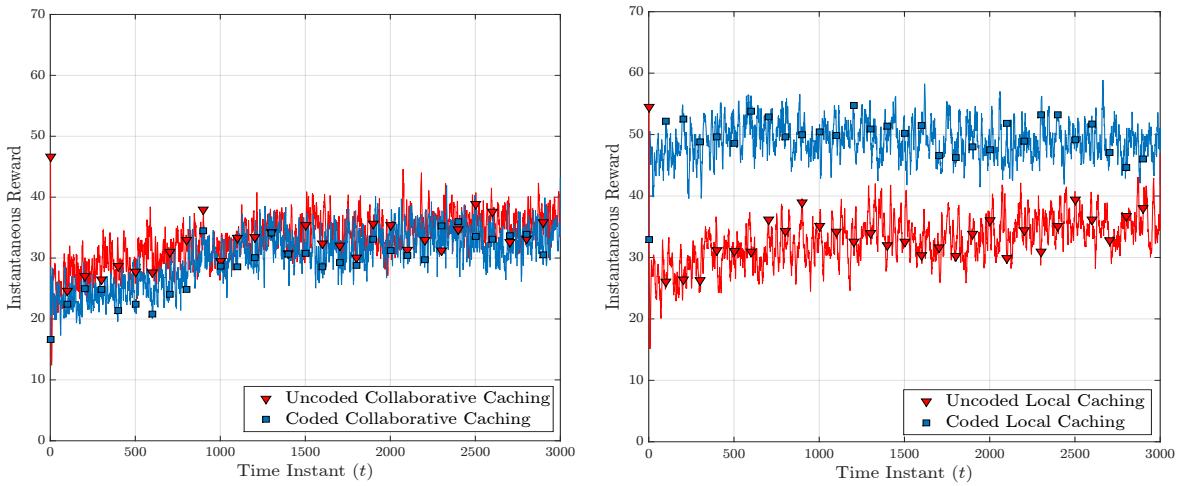
results for uncoded caching for a popularity distribution with $\gamma = 2$ i.e., when there are very few popular files in the system. In this case, the learning converges faster the cumulative reward is higher in comparison with the previous case of $\gamma = 0.56$. The reward is closer to the absolute upper bound i.e., the sub-optimality gap or regret is small as discussed in Example 12 for skewed popularity profiles. Furthermore, in this case, both the local and collaborative caching strategies perform better than the informed upper bound. This is owing to the exploration of new files in the CMAB based learning which allows the placement of new files in the caches, in addition to the most popular ones, which can serve some non-popular demands as well. We observe that, for very few popular files in the system, the performance of the local scheme is almost as good as the collaborative caching scheme. In general, $\gamma = 0.56$ i.e., having a large number of popular files is a more realistic network parameter for cache aided systems [153, 153] and we present further results for this network parameter.

Next, we compare the performance of the uncoded collaborative caching strategies under the different network topologies e.g., sparse, moderate and dense. Fig. 6.6(a) shows the comparison of the collaborative and local learning strategies under the three network settings. It can be seen that the collaborative caching scheme improves upon the local learning based scheme when network connectivity is dense. Under a sparse setting, users are mostly served by single sBSs and the network reduces to a one similar to Network Example 1 in Fig. 6.3. Under such a setting, the collaboration among sBSs to jointly cache content does not offer any added advantage.

6.6.2 Coded Caching Strategies

Next, we study the performance of the coded caching strategies. Fig. 6.6(b) illustrates the rewards obtained by using coded strategies over the same three network configurations as before. In this case, it can be seen that similar performance trends exist i.e., the collaborative caching schemes outperform the local learning schemes under denser network settings. A comparison of the rewards in the coded and uncoded cases shows that the greedy algorithm for the uncoded caching performs similar to the coded caching case.

Remark 22 (*Learning for Coded Caching*)**.** Coded cache placement is a relaxation of the uncoded placement in the sense that uncoded caching is also a viable solution for the coded formulation. Thus, given the same CMAB indices $\overline{\theta}_{f,n}$ at any time instant t, coded caching should outperform the uncoded scheme in terms of sum reward. However, the reward value, to which a strategy converges over time, is not only dependent on the optimality of the cache placement at every instant. It is also dependent on the effectiveness of the CMAB based learning (Algorithm 4) in conjunction with the caching strategy. From our simulations, we observe that both the coded and uncoded caching schemes converge to very similar instantaneous rewards. It is interesting to note that a simpler local search with no sub-packetization (no rateless coding) performs close to a provably optimal coded caching strategy for most network settings of interest. This stems from the fact that learning the popularity profile over time through fractional placement (i.e., sampling fractions of files to learn their popularity distributions) is more difficult than learning based on caching entire



(a) (b)

Figure 6.7: Performance under sparse network: (a) Collaborative Caching and (b) Local learning based Caching.

files. This in turn offsets the reward gains offered by coded caching over the uncoded scheme. □

For coded caching, collaborative learning and placement is particularly detrimental in the sparse setting. In this case, the exploration based on collective topology leads to consistently lower rewards.

Fig. 6.7(a) and 6.7(b) show the performance of the coded and uncoded caching schemes for local and collaborative learning under the sparse network setting. In this case, the collaborative learning leads to very similar performance for both coded and uncoded schemes. However, for the sparse network, local learning suffices. In fact, it can be seen from Fig. 6.7(b), that the coded local caching, which is a direct linear relaxation of the uncoded local learning, outperforms all schemes in this setting. Note that the uncoded local caching is a naive multi-sBS extension of the scheme presented in [153]. The proposed collaborative caching framework lends itself to easy adaptation, based on network configuration, by enabling the sBSs to change the updating of the $T_{f,n}$ parameter in the CMAB index. Local learning based updates can be used when connectivity is sparse while collaborative learning can be used for moderate to dense connectivity. Finally we study the performance of all the schemes under the more realistic moderately connected network setting when the cache size at the sBS changes. We plot the instantaneous reward values (averaged over 500 time steps) to which the algorithms converge after 3000 initial learning time steps[3]. From Fig. 6.8, it can be seen that the collaborative caching schemes outperform the local learning based schemes in terms of reward. However, when the cache size is particularly small ($\leq 5\%$ of the library size), the

[3]The convergence time in a real network application would be of the order of a few hours to a day. Once converged, caching can be performed till popularity changes again when a new learning phase should begin.

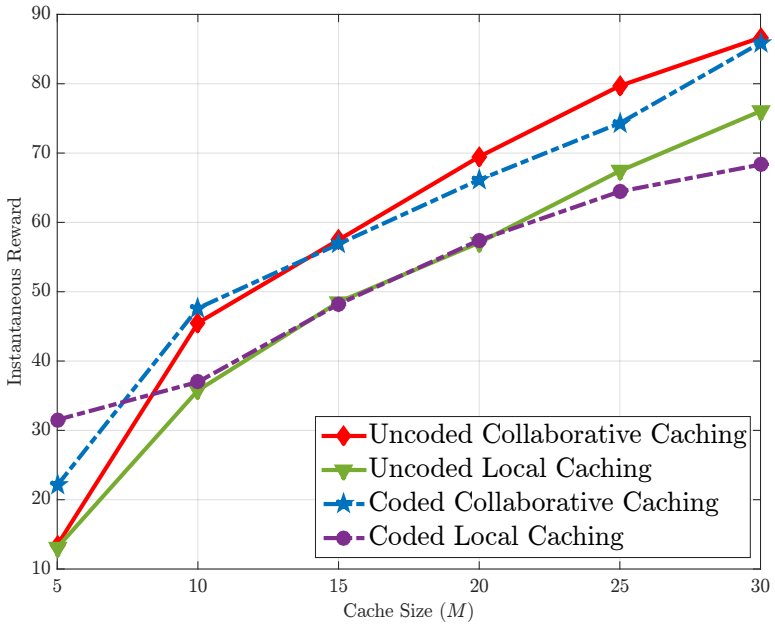


Figure 6.8: Performance in moderate network under varying cache sizes.

coded caching schemes outperform both the uncoded schemes. This is due to the fact that when cache size is very small, the coded schemes offer flexibility by allowing fractional storage of very popular files. Interestingly, at $M = 5$, the coded local scheme outperforms the coded collaborative scheme. This stems from the fact that at such small cache sizes, accounting for topology-aware placement leads to fractional storage of files which might not directly contribute to the sBS's reward. Thus, when cache size is less than 5%, coded local learning offers the highest rewards.

6.7 Directions of Future Research

The paradigm of learning based caching studied in this chapter presents a rich set of interesting problems which can be investigated further and are highlighted next.

- *Utilizing All Available Information*: In the case of cache-aided networks, the mapping of the problem to the CMAB case, although elegant and simple, leads tois information loss. This is due to the fact that when users reveal their requests at each transmission interval, the central BS has the option of observing rewards for all files in the library rather than just the cached files. Under the CMAB setting, this is similar to observing all arms at every round. How to utilize this additional information under the current model is an open problem.

- *Other Learning Approaches*: The previous problem directly leads into the question whether CMAB is the best possible learning framework for the cache-aided small cell network under

consideration. To this end, other solutions like *expert learning* [225] might be viable options to better utilize all knowledge.

- *A Better Reward Function?*: In bandit based reinforcement learning, the formulation of the reward function plays a major role in the learning accuracy and overall system performance. While the reward function we chose was carefully designed to provide the best possible gains, it might yet be possible to design better reward functions and hence it still remains an area of future research.
- *Dependence on Topology*: In this work, we presented novel topology-aware collaborative caching strategies. However, a more formal and theoretical dependence of regret on the system topology is an area of future work.

6.8 Summary

In this chapter, we presented a novel topology-aware collaborative cache placement framework for a multi-sBS small cell network using a reinforcement learning perspective. The multi-armed bandit based learning was shown to have a regret scaling logarithmically with time. We proposed two cache placement strategies: uncoded and coded collaborative caching for the multi-sBS setting. The uncoded caching strategy was shown to be NP-hard and a novel graph coloring based polynomial time approximation algorithm was proposed. A linear relaxation to the uncoded problem, namely, the collaborative coded cache placement problem was formulated and was shown to be optimal for a given popularity distribution. Through numerical simulations, collaborative caching was shown to outperform the naive local learning based schemes for network topologies of interest. It was also demonstrated that the sub-optimal greedy uncoded caching performed close to the optimal coded cache placement due to the fact that learning file popularity through fractional placement was harder as compared to learning by caching entire files.

Chapter 7

Conclusions

In this dissertation, we have considered the approach of leveraging storage at the edge nodes and users to save bandwidth and increase efficiency of wireless network operation. To this end, fundamental information theoretic limits of caching and its impact on wireless networks was studied. The main goal of the work was to evaluate the efficacy of cache-aided networking in the paradigm of content-centric wireless cellular networks and determine its potency as a viable solution for handling the exponential increase wireless data traffic envisioned for the next generation $5\mathrm{G}$ wireless networks.

In order to address the general usefulness of caching, this work concentrated on four major thrust areas namely - (i) Information theoretic limits of caching; (ii) Caching with secure delivery; (iii) Cloud and cache-aided wireless networks; and (iv) Learning-aided collaborative caching in small-cell networks. In each thrust area, we ventured to answer a specific set of questions which aids us in providing a more holistic answer to the question on the advantages of cache-aided networks as posed at the beginning of this dissertation.

1. ***Information Theoretic Limits of Caching:*** In this area of the work our aim was to better understand the fundamental Shannon-type limits of single-server cache-aided systems which form the last hop of a general multi tier heterogeneous network. In this case the server has a library of content and the users are endowed with cache storage. Under this setting, we asked and answered the following questions:

   - *For cache-aided single-server systems, can we improve the information theoretic arguments in literature to better approximate the optimal storage-rate trade-off?*

     We answered this question in the affirmative by proposing a new method for deriving information theoretic lower bounds for single-server cache-aided systems with central server aided cache placement and both centralized as well as D2D-assisted content delivery. The proposed bounds better modeled the correlation between user caches and multicast transmissions to provide tighter bounds than the traditional cut-set bounds from literature. Leverag-

ing the proposed lower bounds we presented an improved approximation of the fundamental storage-rate trade-off for such systems. While exact characterization of the optimal rate is still an open problem, our contributions helped in deepening the understanding of the fundamental limits of cache-aided systems.

- *Can the multicasting gains from systems with homogeneous cache storage be preserved in presence of storage heterogeneity?*

  We showed that multicast gains can still be leveraged in cache aided systems with heterogeneous storage by proposing a novel layered caching architecture which uses set partitioning and cache layering as the main components to maximize multicasting opportunities in the presence of storage heterogeneity. We showed however, that the unicast rate begins to dominate for the case when cache sizes are highly disparate and multicast delivery begins to lose efficacy. For some system settings of practical interest, we showed that the proposed layered heterogeneous caching scheme achieves a rate which is within a constant multiplicative factor of the information theoretic optimal rate.

2. ***Information Theoretic Security in Caching:*** This area of work addressed the concerns about communication privacy in single-server cache-aided systems with multicast joint content delivery. To this end, we introduced the problem of caching with secure delivery and provided an answer to the following question:

   - *Can the multicast gains from non-secure caching schemes be preserved under the strict constraint of information theoretic security from an external wiretapper in cache-aided networks?*

     We again answered the question in the affirmative and proposed a secure caching scheme for centralized and decentralized storage. We further derived an information theoretic lower bound on the secure storage-rate trade-off which helped in showing that the proposed secure caching scheme achieves a rate which is within a constant multiplicative factor from the information theoretically optimal secure rate i.e., the proposed scheme is order optimal. Surprisingly, our results showed that even under the strict constraint of information theoretic security, the cost of security is in fact negligible in terms of rate especially when the number of files and users are large.

3. ***Cloud and Cache-Aided Wireless Networks:*** In this area, we moved our theoretical analysis to a more practical wireless network setting where users are served by multiple edge nodes over a wireless interference channel. The edge nodes in turn are connected to a central cloud server through a rate limited fronthaul link. In this new paradigm of fog radio access networks, we asked the following question:

   - *How can we leverage the interplay between virtualization of cloud processing and localized edge-caching in fog radio access networks to design low latency content delivery schemes?*

To answer this question, we defined a new metric namely the normalized delivery time (NDT) which captures the worst case content delivery latency. We proposed achievable schemes, with an aim to minimize the NDT, which leverage a novel soft-transfer fronthauling from the C-RAN and cooperative transmission over the wireless channel from the edge nodes to achieve low latency delivery. We derived the first known information-theoretic lower bound on the NDT for the fog radio access network. Leveraging the bounds we characterized the minimum NDT for a number of system settings of practical interest and showed that, in general, the achievable NDT of the proposed schemes is within a multiplicative factor of 2 from the optimal NDT. Furthermore, we also studied a pipelined delivery model which showed that pipelining can improve the achievable NDT i.e., it can reduce delivery latency even further compared to a sequential delivery process.

4. ***Caching in Small Cell Networks:*** In this area, we moved to a more practical problem of caching in 2-tier heterogeneous networks with multiple small cells serving a set of users in the case when file popularity is unknown. We studied the problem from a reinforcement learning perspective and the following questions and related answers helped shed light on caching strategies which can be useful in this setting:

   - *Is it better to jointly cache content at multiple small cell base stations in order to maximally serve users?*

     We answered this question mostly in the affirmative with a small caveat that when network topology is extremely sparse, joint caching loses much of its usefulness. In densely and moderately connected networks, joint caching outperforms local storage. However, to cache content jointly in the face of unknown content popularity is a hard problem in general. To this end, we proposed a novel learning-aided collaborative caching framework which presents a topology-aware caching strategy for this network setting.

   - *What impact does the method of cache placement have on the learning process in the collaborative caching framework?*

     This question yielded an interesting insight into the joint learning and cache placement problem studied in this work. We studied two different cache placement methods namely (i) uncoded caching where entire files are cached and (ii) coded caching which allows fractions of files to be cached at each base station. The uncoded caching problem is provably NP-hard and was solved using an approximation algorithm. The coded caching problem on the other hand reduces to a linear program which yields an optimal solution. However, in practice, fractional caching adversely impacts the learning process and hence the provably optimal coded caching method performs similar in terms of servicing user demands compared to the uncoded caching strategy which uses an approximate solution.

Through the course of this dissertation, we moved from a very theoretical treatment of caching in the last hop of the network to a much more practical treatment of learning-aided caching in small cell networks for unknown file popularity. The technical exposition highlights the richness of problems considered in this work and also points to the vast array of research avenues still remaining unexplored in the paradigm of cache-aided wireless systems. However, the results also show unequivocally that caching offers significant gains by leveraging cheap and ubiquitous storage across network elements without the need for drastic changes in radio access technologies. As a result, caching can be a viable and scalable solution to deal with the exponentially increasing capacity demands of the ever evolving wireless communications landscape.

Appendix A

Proofs From Chapter 3

Information Theoretic Limitis of Caching

A.1 Proof of Theorem 4

Consider a cache-aided system with N files, each of size B bits, and K users, each with a cache size of M files. Let s be an integer such that $s \in [1 : \min\{\lceil N/L \rceil, K\}]$. For the case of centralized delivery with $L \in [1 : N]$ demands per user, the demand vector is such that each user demands L distinct files at each transmission interval. Consider the first s caches $Z_{[1:s]}$ and a demand vector

$$\mathbf{D}_1 = \Bigg(\underbrace{\mathbf{d}_{[1:s]}, \mathbf{d}_{[s+1:K]}) = \Big([1:L], [L+1:2L], \ldots, [L(s-1)+1:Ls], \phi\Big)}_{= \mathbf{d}_{[1:s]}} \Bigg), \tag{A.1}$$

where the first s user demands are for Ls unique files and last $K - s$ users' demands can be for any arbitrary $L(K - s)$ files. To service this set of demands, the central server makes a multicast transmission X_1, which along with the $Z_{[1:s]}$ is capable of decoding the files $F_{[1:Ls]}$. Similarly, consider another demand,

$$\mathbf{D}_2 = \Big([Ls+1 : L(s+1)], [L(s+1)+1 : L(s+2)], \ldots, [L(2s-1) : 2Ls], \phi\Big), \tag{A.2}$$

and a resultant multicast transmission X_2, which along with the s caches, are capable of decoding the files $F_{[Ls+1:2Ls]}$. Thus considering the demand vectors $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{\lceil N/(Ls) \rceil}$ and their corresponding multicast transmissions $X_1, X_2, \ldots, X_{\lceil N/(Ls) \rceil}$, along with the first s caches $Z_{[1:s]}$, the whole library of files $F_{[1:N]}$ can be decoded. Considering $B = 1$ without loss of generality. We have:

$$N \leq I\left(F_{1:N}; Z_{[1:s]}, X_{[1:\lceil N/(Ls) \rceil]}\right) \leq H\left(Z_{[1:s]}, X_{[1:\lceil N/(Ls) \rceil]}\right) \leq H\left(Z_{[1:s]}\right) + H\left(X_{[1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}\right)$$

$$
\begin{aligned}
&\leq sM + H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) \leq sM + H\left(X_{[1:\ell]}|Z_{[1:s]}\right) + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}, X_{[1:\ell]}\right) \\
&\overset{\text{(a)}}{\leq} sM + \ell R^*_{\mathsf{cen,L}}(M) + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right) \\
&\overset{\text{(b)}}{\leq} sM + \ell R^*_{\mathsf{cen,L}}(M) + H\big(X_{[\ell+1:\lceil N/(Ls)\rceil]}, Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\big) \\
&\leq sM + \ell R^*_{\mathsf{cen,L}}(M) + \underbrace{H\left(Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right)}_{\triangleq \delta} \\
&\qquad\qquad + \underbrace{H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right)}_{\triangleq \lambda},
\end{aligned} \tag{A.3}
$$

where step (a) results from bounding the entropy of $\ell \in \{1, 2, \ldots, \lceil N/(Ls)\rceil\}$ transmissions given the caches $Z_{[1:s]}$ by $\ell R^*_{\mathsf{cen,L}}(M)$, where each transmission is of rate $R^*_{\mathsf{cen,L}}(M)$. Furthermore, the caches $Z_{[1:s]}$ with transmissions $X_{[1:\ell]}$ can decode files $F_{[1:L\ell s]}$. In step (b), μ number of caches are introduced into the entropy, where μ is the number of remaining caches which along with caches $Z_{[1:s]}$ and transmissions $X_{[1:\ell]}$, can decode the remaining $(N - L\ell s)$ files. It is to be noted that all the remaining $K - s$ caches might not be required for decoding all files. Thus we have:

$$
\mu = \min\left\{\left\lceil \frac{N - L\ell s}{L\ell} \right\rceil, K - s\right\} = \min\left\{\lceil N/(L\ell)\rceil, K\right\} - s, \tag{A.4}
$$

where the last equality follows since s is an integer. Next, we obtain upper bounds on the two terms δ and λ in (A.3).

Upper Bound on δ : We consider the factor δ, from (A.3) and upper bound it as follows:

$$
\begin{aligned}
\delta &= H\left(Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right) \leq H\left(Z_{[s+1:s+\mu]}|Z_{[1:s]}, F_{[1:L\ell s]}\right) \\
&= H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) - H\left(Z_{[1:s]}|F_{[1:L\ell s]}\right).
\end{aligned} \tag{A.5}
$$

Considering all possible subsets of $Z_{[1:s+\mu]}$ having cardinality s, i.e., considering all possible combinations of distinct files in the request vectors and all possible combinations of s caches in (A.3), we can obtain $\binom{s+\mu}{s}$ different inequalities of the form of (A.5). Symmetrizing over all the inequalities, we have:

$$
\delta \leq H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) - \sum_{i=1}^{\binom{s+\mu}{s}} \frac{H\left(\mathsf{Z}^i_{[s]}|F_{[1:L\ell s]}\right)}{\binom{s+\mu}{s}}, \tag{A.6}
$$

where, $\mathsf{Z}^i_{[s]}$ is the i-th subset of $Z_{[1:s+\mu]}$ with cardinality s. Next, consider $Z_{[1:s+\mu]}$ as the set of random variables $\{Z_k : k \in 1, \ldots, s+\mu\}$ and the subsets $\mathsf{Z}^i_{[s]} \subseteq Z_{[1:s+\mu]},\ \forall i = 1, \ldots, \binom{s+\mu}{s}$. Applying Han's Inequality from (3.30), we have:

$$
\frac{s}{s+\mu} H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) \leq \frac{1}{\binom{s+\mu}{s}} \sum_{i=1}^{\binom{s+\mu}{s}} H\left(\mathsf{Z}^i_{[s]}|F_{[1:L\ell s]}\right). \tag{A.7}
$$

Substituting (A.7) into (A.6), we have:

$$\begin{aligned}\delta &\leq H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) - \frac{s}{s+\mu}H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right)\\ &= \frac{\mu}{s+\mu}H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) \leq \frac{\mu}{s+\mu}H\left(Z_{[1:s+\mu]}, F_{[L\ell s+1:N]}|F_{[1:L\ell s]}\right)\\ &= \frac{\mu}{s+\mu}\left(H\left(F_{[L\ell s+1:N]}|F_{[1:L\ell s]}\right) + \underbrace{H\left(Z_{[1:s+\mu]}|F_{[1:N]}\right)}_{=0}\right) \overset{\text{(a)}}{\leq} \frac{\mu}{s+\mu}(N - L\ell s)^+, \end{aligned} \tag{A.8}$$

where step (a) follows from the fact that the caches are functions of all N files in the library.

Upper Bound on λ : To upper bound λ, we observe from the last step in (A.3) that the transmissions $X_{[1:\ell]}$, along with caches $Z_{[1:s+\mu]}$ can decode the files $F_{[1:L\ell(s+\mu)]}$ within the conditioning, i.e.,

$$\lambda = H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell(s+\mu)]}\right). \tag{A.9}$$

In order to characterize the upper bound on λ, we consider two cases as follows.

• ***Case 1 ($N \leq L\ell(s+\mu)$) :*** All files are decoded by the caches $Z_{[1:s+\mu]}$ and transmissions $X_{[1:\ell]}$ within the conditioning for the term λ in (A.3). We have

$$\lambda = H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:N]}\right) = 0, \tag{A.10}$$

since all transmissions are functions of the file library $F_{[1:N]}$. In the case when, for $N > K$, fewer than K caches suffices to decode all files with the transmissions within the conditioning in λ i.e. $s + \mu \leq K$, we have:

$$KL\ell \geq L\ell(s+\mu) \geq N, \quad \text{i.e.,} \quad \lambda = (N - KL\ell)^+ \ = 0. \tag{A.11}$$

It can also be easily seen that for the case of $K \geq N$, $\lambda = (N - KL\ell)^+ = 0$ since $\ell, L \geq 1$.

• ***Case 2 ($N > L\ell(s+\mu)$) :*** The case when, even with $s + \mu = K$ caches, all files are not decoded by the caches and transmissions within the conditioning for the term λ in (A.3). In this case, $\lambda \neq 0$ and we have:

$$\begin{aligned}\lambda &= H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}\right)\\ &\leq H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}, F_{[KL\ell+1:N]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}\right)\\ &\leq H\left(F_{[KL\ell+1:N]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}\right) + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:N]}\right)\\ &\overset{\text{(a)}}{\leq} H\left(F_{[KL\ell+1:N]}\right) \leq (N - KL\ell), \end{aligned} \tag{A.12}$$

where step (a) follows from the fact that the second entropy term in the previous step goes to zero since transmissions are functions of the N files. Thus from (A.10) and (A.12), we can compactly bound λ as:

$$\lambda \leq (N - KL\ell)^+. \tag{A.13}$$

Substituting (A.8) and (A.13) into (A.3), we have:

$$N \quad \leq sM + \ell R^*_{\mathsf{cen,L}}(M) + \frac{\mu}{s+\mu}(N - L\ell s)^+ + (N - KL\ell)^+ \tag{A.14}$$

Rearranging (A.14), we obtain the following lower bound on the optimal rate $R^*_{\mathsf{cen,L}}(M)$

$$R^*_{\mathsf{cen,L}}(M) \geq \frac{1}{\ell}\left\{N - sM - \frac{\mu}{s+\mu}(N - L\ell s)^+ - (N - KL\ell)^+\right\}. \tag{A.15}$$

Optimizing over all parameter values of s, ℓ, completes the proof of Theorem 4.

A.2 Proof of Theorem 5

From Theorem 4, considering the lower bound on the optimal rate $R^*_{\mathsf{cen,L}}(M)$, we set $\ell = \left\lceil \frac{\beta N}{Ls} \right\rceil \in \left[1 : \left\lceil \frac{N}{Ls} \right\rceil\right]$ with $\beta \in [0, 1]$. Using this, we next derive an upper bound on the term $\left(\frac{\mu}{\mu+s}\right)$ as follows

$$\begin{aligned}\frac{\mu}{\mu+s} = \frac{\min\left\{\left\lceil \frac{N}{L\ell} \right\rceil, K\right\} - s}{\min\left\{\left\lceil \frac{N}{L\ell} \right\rceil, K\right\}} &\leq 1 - \frac{s}{\left\lceil \frac{N}{L\ell} \right\rceil} = 1 - \frac{s}{\left\lceil \frac{N}{L\left\lceil \frac{\beta N}{Ls} \right\rceil} \right\rceil} \leq 1 - \frac{s}{\left\lceil \frac{s}{\beta} \right\rceil} \\ &\leq 1 - \frac{s}{\frac{s}{\beta} + 1} = 1 - \frac{\beta}{1 + \frac{\beta}{s}} \leq 1 - \frac{\beta}{1+\beta} = \frac{1}{1+\beta},\end{aligned} \tag{A.16}$$

where the last inequality follows from the fact that $s \geq 1$. Substituting (A.16) into (3.12), we have:

$$\begin{aligned}R^*_{\mathsf{cen,L}}(M) &\geq \frac{N - sM - \frac{1}{1+\beta}\left(N - L\left\lceil \frac{\beta N}{Ls} \right\rceil s\right)^+ - \left(N - KL\left\lceil \frac{\beta N}{Ls} \right\rceil\right)^+}{\left\lceil \frac{\beta N}{Ls} \right\rceil} \\ &\geq \frac{\left(\frac{2\beta}{1+\beta}\right) N - sM - N\left(1 - K\frac{\beta}{s}\right)^+}{\left\lceil \frac{\beta N}{Ls} \right\rceil}.\end{aligned} \tag{A.17}$$

Next, we consider two cases, namely (i) $\min\left\{\frac{N}{L}, K\right\} \leq 10$; and (ii) $\min\left\{\frac{N}{L}, K\right\} \geq 11$.

• ***Case 1*** $\left(\mathbf{\min\left\{\frac{N}{L}, K\right\} \leq 10}\right)$ **:** For this case, setting $s = 1$ and $\beta = 1$ in (A.17), we have the following form on the lower bound,

$$R^*_{\mathsf{cen,L}}(M) \geq \frac{N\left(1 - \frac{M}{N}\right)}{\left\lceil \frac{N}{L} \right\rceil} \tag{A.18}$$

Consider first, the case when $\frac{N}{L} \leq K$. From (3.10), we have the following upper bound on the achievable rate

$$R_{\mathsf{cen,L}}(M) \leq \min\{N, KL\}\left(1 - \frac{M}{N}\right) \leq N\left(1 - \frac{M}{N}\right). \tag{A.19}$$

Therefore, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq \left\lceil \frac{N}{L} \right\rceil \leq 10. \tag{A.20}$$

Next, consider the case when $K \leq \frac{N}{L}$. Again, from (3.10), we have the following upper bound on the achievable rate

$$R_{\mathsf{cen,L}}(M) \leq \min\{N, KL\} \left(1 - \frac{M}{N}\right) \leq KL \left(1 - \frac{M}{N}\right). \tag{A.21}$$

Again, setting $s = 1$ and $\beta = 1$ in (A.17), we have

$$R^*_{\mathsf{cen,L}}(M) \geq \frac{N\left(1 - \frac{M}{N}\right)}{\frac{N}{L} + 1} = \frac{L\left(1 - \frac{M}{N}\right)}{1 + \frac{L}{N}} \geq \frac{KL\left(1 - \frac{M}{N}\right)}{1 + K} \tag{A.22}$$

Therefore, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq K + 1 \leq 10 + 1 = 11. \tag{A.23}$$

• ***Case 2*** $\left(\mathbf{min}\left\{\frac{N}{L}, \boldsymbol{K}\right\} \geq \boldsymbol{11}\right)$ **:** For this case, we consider three distinct regimes for the cache storage size M: ***Regime 1:*** $0 \leq M \leq 1.275 \max\{L, N/K\}$; ***Regime 2:*** $1.275 \max\{L, N/K\} < M \leq 0.2N$; and ***Regime 3:*** $0.2N < M \leq N$. We consider each of the three regimes separately.

• ***Regime 1*** $(\boldsymbol{0 \leq M \leq 1.275 \max\{L, N/K\}})$ **:**

For this regime, we set $s = \lfloor 0.3049 \min\{N/L, K\} \rfloor \in [1 : \min\{N/L, K\}]$ and $\ell = \left\lceil \frac{0.9649N}{Ls} \right\rceil$, from (A.17), we have

$$\begin{aligned}
R^*_{\mathsf{cen,L}}(M) \;&\geq\; \frac{\left(\frac{2\times 0.9649}{1+0.9649}\right) - s\frac{M}{N} - \left(1 - K\frac{0.9649}{s}\right)^+}{\frac{0.9649}{Ls} + \frac{1}{N}} \\
&= \frac{\left(\frac{2\times 0.9649}{1+0.9649}\right) - \lfloor 0.3049 \min\{N/L, K\} \rfloor \frac{1.275 \max\{L, N/K\}}{N} - \left(1 - K\frac{0.9649}{\lfloor 0.3049 \min\{N/L, K\} \rfloor}\right)^+}{\frac{0.9649}{L\lfloor 0.3049 \min\{N/L, K\} \rfloor} + \frac{1}{N}} \\
&\overset{(a)}{\geq} \frac{\left(\frac{2\times 0.9649}{1+0.9649}\right) - (0.3049 \times 1.275)\frac{\min\{N/L, K\} \max\{L, N/K\}}{N} - \left(1 - K\frac{0.9649}{0.3049 \min\{N/L, K\}}\right)^+}{\frac{0.9649}{L(0.3049 \min\{N/L, K\} - 1)} + \frac{1}{N}} \\
&\geq \frac{L \min\left\{\frac{N}{L}, K\right\} \left(0.3049 - \frac{1}{\min\left\{\frac{N}{L}, K\right\}}\right) \left\{\left(\frac{2\times 0.9649}{1+0.9649}\right) - (0.3049 \times 1.275) - \left(1 - \frac{0.9649}{0.3049}\right)^+\right\}}{0.9649 + \frac{L\left(0.3049 \min\left\{\frac{N}{L}, K\right\} - 1\right)}{N}} \\
&\overset{(b)}{\geq} \frac{\min\{N, KL\} \left(0.3049 - \frac{1}{10+1}\right) \left\{\left(\frac{2\times 0.9649}{1+0.9649}\right) - (0.3049 \times 1.275) - \left(1 - \frac{0.9649}{0.3049}\right)^+\right\}}{0.9649 + 0.3049}
\end{aligned}$$

$$\geq \frac{\min\{N, KL\}}{10}, \tag{A.24}$$

where step (a) follows by using $\lfloor 0.3049 \min\{N/L, K\} \rfloor \leq 0.3049 \min\{N/L, K\}$ in the numerator and $\lfloor 0.3049 \min\{N/L, K\} \rfloor \geq 0.3049 \min\{N/L, K\} - 1$ in the denominator; and step (b) follows by using $\min\{N/L, K\} \leq N/L$ in the second term in the denominator. Again, considering the upper bound in (3.10), we have

$$R_{\mathsf{cen,L}}(M) \leq \min\{N, KL\} \left(1 - \frac{M}{N}\right) \leq \min\{N, KL\}. \tag{A.25}$$

Therefore for *Regime 1*, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq 10. \tag{A.26}$$

- ***Regime 2*** $\boldsymbol{(1.275 \max\{L, N/K\} < M \leq 0.2N)}$ **:**

  For this regime, setting $s = \left\lfloor 0.442 \frac{N}{M} \right\rfloor \in [1 : \min\{N/L, K\}]$[1] and $\ell = \left\lceil \frac{0.984N}{Ls} \right\rceil$, from (A.17), we have

$$\begin{aligned}
R^*_{\mathsf{cen,L}}(M) &\geq \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - s\frac{M}{N} - \left(1 - K\frac{0.984}{s}\right)^+}{\frac{0.984}{Ls} + \frac{1}{N}} \\
&= \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - \left\lfloor 0.442\frac{N}{M} \right\rfloor \frac{M}{N} - \left(1 - K\frac{0.984}{\left\lfloor 0.442\frac{N}{M} \right\rfloor}\right)^+}{\frac{0.984}{L\left\lfloor 0.442\frac{N}{M} \right\rfloor} + \frac{1}{N}} \\
&\overset{\text{(a)}}{\geq} \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - 0.442\frac{N}{M}\frac{M}{N} - \left(1 - \frac{0.984}{0.442}\frac{KM}{N}\right)^+}{\frac{0.984}{L\left(0.442\frac{N}{M} - 1\right)} + \frac{1}{N}} \\
&\overset{\text{(b)}}{\geq} \frac{\frac{LN}{M}\left(0.442 - \frac{M}{N}\right)\left\{\left(\frac{2\times 0.984}{1+0.984}\right) - 0.442 - \left(1 - \frac{0.984}{0.442} \times 1.275\right)^+\right\}}{0.984 + \frac{0.442}{1.275}\frac{L}{M}} \\
&\overset{\text{(c)}}{\geq} \frac{\frac{LN}{M}(0.442 - 0.2)\left\{\left(\frac{2\times 0.984}{1+0.984}\right) - 0.442 - \left(1 - \frac{0.984\times 1.275}{0.442}\right)^+\right\}}{0.984 + \frac{0.442}{1.275}} \geq \frac{LN}{10M},
\end{aligned} \tag{A.27}$$

where step (a) follows again by using $\lfloor 0.442N/M \rfloor \leq 0.442N/M$ in the numerator and $\lfloor 0.442N/M \rfloor \geq 0.442N/M - 1$ in the denominator; step (b) follows from using $KM/N \geq 1.275$; and step (c) follows by using $M/N \leq 0.2$ in the numerator and $M \geq 1.275L$ in the denominator. Again considering the upper bound in (3.10), we have

$$R_{\mathsf{cen,L}}(M) \leq \frac{KL\left(1 - \frac{M}{N}\right)}{1 + \frac{KM}{N}} \leq \frac{LN}{M}\left(1 - \frac{M}{N}\right) \leq \frac{LN}{M}. \tag{A.28}$$

[1]The range of s is validated as follows. Using the upper bound $M \leq 0.2N$, we have $0.442N/M \geq 0.442/0.2 \geq 1$. Again using the lower bound $M \geq 1.275L$, we have $0.442N/M \leq \frac{0.442}{1.275}N/L \leq N/L$. Again using $M \geq 1.275N/K$, we have $0.442N/M \leq K$.

Therefore for *Regime 2*, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq 10. \tag{A.29}$$

- ***Regime 3 ($0.2N < M \leq N$) :***

  In this regime, setting $s = 1$ and $\ell = \left\lceil \frac{N}{L} \right\rceil$ in (A.17), we note that in this case, $\mu = 0$ and $(N - K\ell)^+ = 0$. Thus, we have

$$R^*_{\mathsf{cen,L}}(M) \geq \frac{N\left(1 - \frac{M}{N}\right)}{\frac{N}{L} + 1} \geq \frac{\left(1 - \frac{M}{N}\right)}{\frac{1}{L} + \frac{1}{N}}. \tag{A.30}$$

  From (3.10), we have

$$R_{\mathsf{cen,L}}(M) \leq \frac{KL\left(1 - \frac{M}{N}\right)}{1 + \frac{KM}{N}} \leq \frac{LN}{M}\left(1 - \frac{M}{N}\right). \tag{A.31}$$

  Therefore for *Regime 3*, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq \frac{LN}{M}\left(\frac{1}{L} + \frac{1}{N}\right) \leq \frac{2N}{M} \leq \frac{2}{0.2} \leq 10 \tag{A.32}$$

Combining (A.20),(A.23),(A.26),(A.29) and (A.32), completes the proof of Theorem 5. □

A.3 Proof of Theorem 6

From Corollary 1, considering the lower bound on the optimal rate $R^*_{\mathsf{cen}}(M)$, we set $\ell = \left\lceil \frac{\beta N}{s} \right\rceil \in \{1, 2, \ldots, \left\lceil \frac{N}{s} \right\rceil\}$ with $0 < \beta \leq 1$. Using this we next derive an upper bound on $\left(\frac{\mu}{\mu+s}\right)$.

$$\begin{aligned}
\frac{\mu}{\mu+s} &= \frac{\min\left\{\left\lceil \frac{N-\ell s}{\ell} \right\rceil, K-s\right\}}{\min\left\{\left\lceil \frac{N-\ell s}{\ell} \right\rceil, K-s\right\} + s} = \frac{\min\left\{\left\lceil \frac{N}{\ell} \right\rceil, K\right\} - s}{\min\left\{\left\lceil \frac{N}{\ell} \right\rceil, K\right\}} \\
&= 1 - \frac{s}{\min\left\{\left\lceil \frac{N}{\ell} \right\rceil, K\right\}} = 1 - \frac{s}{\min\left\{\left\lceil \frac{N}{\left\lceil \frac{\beta N}{s} \right\rceil} \right\rceil, K\right\}} \\
&\leq 1 - \frac{s}{\left\lceil \frac{s}{\beta} \right\rceil} \leq 1 - \frac{s}{\frac{s}{\beta} + 1} = 1 - \frac{\beta}{1 + \frac{\beta}{s}} \leq 1 - \frac{\beta}{1+\beta} = \frac{1}{1+\beta},
\end{aligned} \tag{A.33}$$

where the last inequality follows from the fact that $s \geq 1$. Substituting (A.33) into (3.14), we have:

$$R^*_{\mathsf{cen}}(M) \geq \frac{N - sM - \frac{1}{1+\beta}\left(N - \left\lceil \frac{\beta N}{s} \right\rceil s\right)^+ - \left(N - K\left\lceil \frac{\beta N}{s} \right\rceil\right)^+}{\left\lceil \frac{\beta N}{s} \right\rceil}$$

$$\geq \frac{N - sM - N\left(\frac{1-\beta}{1+\beta}\right) - N\left(1 - K\frac{\beta}{s}\right)^{+}}{\left\lceil \frac{\beta N}{s} \right\rceil}. \tag{A.34}$$

Next, we consider two cases, namely (ii) $\min\{N, K\} \leq 8$; and (ii) $\min\{N, K\} \geq 9$. We next consider each case separately.

• ***Case 1*** $\boldsymbol{(\min\{N, K\} \leq 8)}$ **:** For this case, setting $s = 1$ and $\beta = 1$ in (A.34), we have:

$$R^{*}_{\mathsf{cen}}(M) \geq \frac{N - M}{N} = \left(1 - \frac{M}{N}\right). \tag{A.35}$$

Again, from [97, Theorem 1], we have:

$$R_{\mathsf{cen}}(M) \leq \min\{N, K\}\left(1 - \frac{M}{N}\right). \tag{A.36}$$

Thus for this case, the gap between the upper and lower bound is given by:

$$\text{Gap} = \frac{R_{\mathsf{cen}}(M)}{R^{*}_{\mathsf{cen}}(M)} \leq \min\{N, K\} \leq 8. \tag{A.37}$$

• ***Case 2*** $\boldsymbol{(\min\{N, K\} \geq 9)}$ **:** For this case, we consider three distinct regimes for the cache storage size M namely (i) ***Regime 1:*** $0 \leq M \leq 1.01 \max\{1, N/K\}$; (ii) ***Regime 2:*** $1.01 \max\{1, N/K\} < M \leq 0.1250N$; and (ii) ***Regime 3:*** $0.1250N < M \leq N$. We next consider each of the three regimes separately.

• ***Regime 1*** $\boldsymbol{(0 \leq M \leq 1.01 \max\{1, N/K\})}$ **:**

In this regime, setting $s = \lfloor 0.4701 \min\{N, K\} \rfloor \in \{1, 2, \ldots, \min\{N, K\}\}$, $\ell = \left\lceil \frac{0.93N}{s} \right\rceil$, and using the fact that $x \leq \lceil x \rceil \leq x + 1$ and $x - 1 \leq \lfloor x \rfloor \leq x$, we have:

$$\begin{aligned}
R^{*}_{\mathsf{cen}}(M) &\geq \frac{N - sM - N\left(\frac{1-\beta}{1+\beta}\right) - N\left(1 - \frac{K\beta}{s}\right)^{+}}{\frac{\beta N}{s} + 1} \geq \frac{N\left[\frac{2\beta}{1+\beta} - s\frac{M}{N} - \left(1 - \frac{K\beta}{s}\right)^{+}\right]}{\frac{\beta N}{s} + 1} \\
&\geq \frac{\left\{\frac{2\times 0.93}{1+0.93} - \lfloor 0.4701 \min\{N, K\} \rfloor \frac{M}{N} - \left(1 - \frac{0.93K}{\lfloor 0.4701 \min\{N, K\} \rfloor}\right)^{+}\right\}}{\frac{0.93}{\lfloor 0.4701 \min\{N, K\} \rfloor} + \frac{1}{N}} \\
&\geq \frac{\left\{\frac{2\times 0.93}{1+0.93} - 0.4701 \min\{N, K\} \frac{1.01 \max\{1, N/K\}}{N} - \left(1 - \frac{0.93K}{0.4701 \min\{N, K\}}\right)^{+}\right\}}{\frac{0.93}{0.4701 \min\{N, K\} - 1} + \frac{1}{N}} \\
&\geq \frac{\left\{(0.4701 \min\{N, K\} - 1)\left[\frac{2\times 0.93}{1+0.93} - 0.4701 \times 1.01 - \left(1 - \frac{0.93}{0.4701}\right)^{+}\right]\right\}}{0.93 + \frac{0.4701 \min\{N, K\}}{N} - \frac{1}{N}}
\end{aligned}$$

$$\geq \min\{N,K\}\frac{\left(0.4701-\frac{1}{9}\right)}{0.93+0.4701}\left[\frac{2\times 0.93}{1+0.93}-0.4701\times 1.01-\left(1-\frac{0.93}{0.4701}\right)^{+}\right]$$
$$\geq \frac{\min\{N,K\}}{8}. \tag{A.38}$$

Again, from [97, Theorem 1], we have:

$$R_{\mathsf{cen}}(M)\leq \min\{N,K\}\left(1-\frac{M}{N}\right)\leq \min\{N,K\}. \tag{A.39}$$

Thus for this regime, the gap between the upper and lower bound is given by:

$$\text{Gap}=\frac{R_{\mathsf{cen}}(M)}{R^{*}_{\mathsf{cen}}(M)}\leq 8. \tag{A.40}$$

- ***Regime 2*** $\boldsymbol{(1.01\max\{1,N/K\}<M\leq 0.1250N)}$ **:**

  In this regime, we set $s=\lfloor 0.4983\frac{N}{M}\rfloor\in\{1,2,\ldots,\min\{N,K\}\}$, $\ell=\left\lceil\frac{0.991N}{s}\right\rceil$ and using the fact that $x\leq\lceil x\rceil\leq x+1$ and $x-1\leq\lfloor x\rfloor\leq x$, we have:

$$R^{*}_{\mathsf{cen}}(M)\geq\frac{N\left[\frac{2\beta}{1+\beta}-s\frac{M}{N}-\left(1-\frac{K\beta}{s}\right)^{+}\right]}{\frac{\beta N}{s}+1}\geq\frac{N\left[\frac{2\times 0.991}{1+0.991}-0.4983-\left(1-\frac{0.991}{0.4983}\frac{KM}{N}\right)^{+}\right]}{\frac{0.991N}{0.4983\frac{N}{M}-1}+1}$$
$$\geq\frac{N\left[\frac{2\times 0.991}{1+0.991}-0.4983-\left(1-\frac{0.991\times 1.01}{0.4983}\right)^{+}\right]}{\frac{0.991N}{0.4983\frac{N}{M}-1}+1}$$
$$\geq\frac{\left(0.4983\frac{N}{M}-1\right)\left[\frac{2\times 0.991}{1+0.991}-0.4983-\left(1-\frac{0.991\times 1.01}{0.4983}\right)^{+}\right]}{0.991+0.4983\frac{1}{M}-\frac{1}{N}}$$
$$\geq\frac{N}{M}\frac{\left(0.4983-\frac{M}{N}\right)\left[\frac{2\times 0.991}{1+0.991}-0.4983-\left(1-\frac{0.991\times 1.01}{0.4983}\right)^{+}\right]}{0.991+\frac{0.4983}{1.01}}$$
$$\geq\frac{N}{M}\frac{(0.4983-0.1250)}{0.991+\frac{0.4983}{1.01}}\left[\frac{2\times 0.991}{1+0.991}-0.4983-\left(1-\frac{0.991\times 1.01}{0.4983}\right)^{+}\right]\geq\frac{N}{8M}. \tag{A.41}$$

Again, from [97, Theorem 1], we have:

$$R_{\mathsf{cen}}(M)\leq\frac{\min\{N,K\}}{1+\frac{KM}{N}}\left(1-\frac{M}{N}\right)\leq\frac{K}{\frac{KM}{N}}\left(1-\frac{M}{N}\right)$$
$$\leq\frac{N}{M}\left(1-\frac{M}{N}\right)\leq\frac{N}{M}. \tag{A.42}$$

Thus for this regime, the gap between the upper and lower bound is given by:

$$\text{Gap} = \frac{R_{\mathsf{cen}}(M)}{R^*_{\mathsf{cen}}(M)} \leq 8. \tag{A.43}$$

- ***Regime 3 ($0.1250N < M \leq N$) :***

  In this regime, setting $s = 1$ and $\beta = 1$ i.e., $\ell = N$, we have:

$$R^*_{\mathsf{cen}}(M) \geq \frac{N-M}{N} = \left(1 - \frac{M}{N}\right). \tag{A.44}$$

  Again, from [97, Theorem 1], we have:

$$\begin{aligned} R_{\mathsf{cen}}(M) &\leq \frac{\min\{N,K\}}{1+\frac{KM}{N}}\left(1-\frac{M}{N}\right) \leq \frac{K}{\frac{KM}{N}}\left(1-\frac{M}{N}\right) \\ &\leq \frac{N}{M}\left(1-\frac{M}{N}\right) \leq \frac{1}{0.1250}\left(1-\frac{M}{N}\right). \end{aligned} \tag{A.45}$$

  Thus for this regime, the gap between the upper and lower bound is given by:

$$\text{Gap} = \frac{R_{\mathsf{cen}}(M)}{R^*_{\mathsf{cen}}(M)} \leq \frac{1}{0.1250} = 8. \tag{A.46}$$

Thus from (A.37), (A.40), (A.43) and (A.46), we have for all N, K, the gap between the achievability and the proposed converse is upper bounded by 8. This completes proof of Theorem 6. □

A.4 Proof of Theorem 7

Consider the case of D2D-assisted content delivery for cache-aided system with a library of $N \in \mathbb{N}^+$ files $F_{[1:N]}$ each of size B bits, and $K \in \mathbb{N}^+$ users, with cache storage $Z_{[1:K]}$ which satisfies the minimum D2D storage constraint $KM \geq N$. Let s be an integer such that $s \in [1 : \min\{\lceil N/L \rceil, K\}]$. The demand vector is such that each user requests L distinct files at each transmission interval. Consider the first s caches $Z_{[1:s]}$ and a demand vector

$$\mathbf{D}_1 = \left(\mathbf{d}_{[1:s]}, \mathbf{d}_{[s+1:K]}\right) = \left(\underbrace{\{1:L\}, \{L+1:2L\}, \ldots, \{L(s-1)+1:Ls\}}_{=\,\mathbf{d}_{[1:s]}}, \phi\right), \tag{A.47}$$

where the first s user demands are for Ls unique files and last $K-s$ users' demands can be for any arbitrary $L(K-s)$ files. To service this set of demands, consider a composite multicast transmission

$$X_1 = \left\{X^1_{\left(\mathbf{d}_{[1:s]},\phi\right)}, \ldots, X^s_{\left(\mathbf{d}_{[1:s]},\phi\right)}, X^{s+1}_{\left(\mathbf{d}_{[1:s]},\phi\right)}, X^{s+2}_{\left(\mathbf{d}_{[1:s]},\phi\right)}, \ldots, X^K_{\left(\mathbf{d}_{[1:s]},\phi\right)}\right\}, \tag{A.48}$$

composed of K device multicast transmissions, which, along with the s device caches decodes the files $F_{[1:Ls]}$. Similarly consider another demand vector,

$$\mathbf{D}_2 = \Big(\{Ls+1 : L(s+1)\}, \{L(s+1)+1 : L(s+2)\}, \ldots, \{L(2s-1) : 2Ls\}, \phi\Big). \quad \text{(A.49)}$$

A second composite multicast transmission X_2, along with device cache contents $Z_{[1:s]}$, can decode the next Ls files $F_{[Ls+1:2Ls]}$. Thus considering the request vectors $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{\lceil N/(Ls)\rceil}$ and their corresponding composite multicast transmissions $X_1, X_2, \ldots, X_{\lceil N/(Ls)\rceil}$, along with the first s device caches $Z_{[1:s]}$, the whole library of files $F_{[1:N]}$ can be decoded. Note that for an optimal composite transmission rate $R^*_{\mathsf{d2d,L}}(M)$, each device in the D2D cluster multicasts with a rate of $R^*_{\mathsf{d2d,L}}(M)/K$ owing to symmetry and the sum-rate constraint in (3.5) i.e.,

$$H\left(X^k_{\mathbf{D}}\right) \;\leq\; R^*_{\mathsf{d2d,L}}(M)/K, \quad \forall k \in [1:K]. \quad \text{(A.50)}$$

Considering $B = 1$ without loss of generality, we have:

$$\begin{aligned}
N &\leq I\left(F_{1:N}; Z_{[1:s]}, X_{[1:\lceil N/(Ls)\rceil]}\right) \leq H\left(Z_{[1:s]}, X_{[1:\lceil N/(Ls)\rceil]}\right) \leq H\left(Z_{[1:s]}\right) + H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) \\
&\leq sM + H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) \leq sM + H\left(X_{[1:\ell]}|Z_{[1:s]}\right) + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}, X_{[1:\ell]}\right) \\
&\overset{\text{(a)}}{\leq} sM + \frac{\ell(K-s)}{K} R^*_{\mathsf{d2d,L}}(M) + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right) \\
&\overset{\text{(b)}}{\leq} sM + \frac{\ell(K-s)}{K} R^*_{\mathsf{d2d,L}}(M) + H\big(X_{[\ell+1:\lceil N/(Ls)\rceil]}, Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\big) \\
&\leq sM + \frac{\ell(K-s)}{K} R^*_{\mathsf{d2d,L}}(M) + \underbrace{H\left(Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right)}_{\triangleq\, \delta} \\
&\qquad + \underbrace{H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right)}_{\triangleq\, \lambda}, \quad \text{(A.51)}
\end{aligned}$$

where in step (a), the second term follows from (A.50) and the fact that given cache contents $Z_{[1:s]}$, the device transmissions $\left\{X^1_{[1:\ell]}, X^2_{[1:\ell]}, \ldots, X^s_{[1:\ell]}\right\}$ can be obtained and hence the entropy term reduces to $H\left(X^{s+1}_{[1:\ell]}, X^{s+2}_{[1:\ell]}, \ldots, X^K_{[1:\ell]}\right) \leq (\ell(K-s)/K) R^*_{\mathsf{d2d,L}}(M)$; the third term follows from the fact that the device storage contents, $Z_{[1:s]}$, along with the composite transmission vectors $X_{[1:\ell]}$ are capable of decoding the files $F_{[1:L\ell s]}$. In step (b), $\mu = (\min\{\lceil N/(L\ell)\rceil, K\} - s)$ is the number of additional device caches which, along with the transmissions $X_{[1:\ell]}$ can decode all N files. Note that, for $s = K$, we have:

$$H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) = 0, \quad \text{(A.52)}$$

since transmissions are functions of all K caches. As a result, the second step in (A.51) yields the minimum storage constraint for D2D-assisted delivery $KM \geq N$. Next we upper bound the terms δ, λ in (A.51) which finally yields an upper bound on the RHS. We first note that the term δ

is identical to the case of centralized delivery and can be upper bounded using Han's Inequality by following the same steps as in (A.5)-(A.8) in Appendix A.1, yielding the upper bound

$$\delta \leq \frac{\mu}{s+\mu}(N - L\ell s)^{+}. \tag{A.53}$$

Upper Bound on $\boldsymbol{\lambda}$: We next derive an upper bound on the factor λ in (A.51) and consider two distinct cases as follows.

• ***Case 1 ($\boldsymbol{N \leq L\ell(s+\mu)}$) :*** We consider the case that all N files can be decoded with $\mu \leq K - s$ additional device storage contents and transmissions $X_{[1:\ell]}$, within the conditioning in the factor λ in (A.51), i.e., $L\ell(s+\mu) \geq N$. Thus, we have

$$\lambda = H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, F_{[1:N]}\right) = 0, \tag{A.54}$$

which follows from the fact that the transmissions are functions of all N files.

• ***Case 2 ($\boldsymbol{N > L\ell(s+\mu)}$) :*** We consider the complementary case where $\mu = K - s$ additional device storage contents along with the transmissions $X_{[1:\ell]}$, cannot decode all N files. We have:

$$\lambda = H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:K]}, F_{[1:KL\ell]}\right) \leq H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:K]}\right) = 0, \tag{A.55}$$

which follows from the fact that $KM \geq N$ i.e., all files are stored within the collective device caches for D2D-assisted delivery and hence all transmissions are functions of the cache contents. Thus combining (A.54) and (A.55) we have:

$$\lambda = 0. \tag{A.56}$$

Substituting (A.53) and (A.56) into (A.51) and optimizing over all parameter values of s, ℓ, completes the proof of Theorem 7. □

A.5 Proof of Theorem 8

From Theorem 7, considering the lower bound on the optimal rate $R^{*}_{\mathsf{d2d,L}}(M)$, we set $\ell = \left\lceil \frac{\beta N}{Ls} \right\rceil \in \left[1 : \left\lceil \frac{N}{Ls} \right\rceil\right]$ with $\beta \in [0,1]$. We make use of the upper bound on $\left(\frac{\mu}{\mu+s}\right)$ from (A.16) in Appendix A.2. Using this in (3.16) from Theorem 7, we have

$$R^{*}_{\mathsf{d2d,L}}(M) \geq \frac{N - sM - \frac{1}{1+\beta}\left(N - L\left\lceil \frac{\beta N}{Ls} \right\rceil s\right)^{+}}{\left\lceil \frac{\beta N}{Ls} \right\rceil \left(\frac{K-s}{K}\right)} \geq \frac{N\left(\frac{2\beta}{1+\beta} - s\frac{M}{N}\right)}{\left\lceil \frac{\beta N}{Ls} \right\rceil \left(\frac{K-s}{K}\right)} \tag{A.57}$$

In order to facilitate the proof of Theorem 8, we consider two cases namely - (i) *low per-device demand* with $0.5N \geq L$; and (ii) *high per-device demand* with $0.5N \leq L$. We consider the two cases separately.

• ***Case 1*** $(\boldsymbol{0.5N \geq L})$: For the case of low-per device demands, we divide the available cache storage at each device into the following three regimes, namely (i) *Regime 1*: $N/K \leq M \leq L$; (ii) *Regime 2*: $L \leq M \leq 0.2N$; and (iii) *Regime 3*: $0.2N \leq M \leq N$. We consider each regime separately.

• ***Regime 1*** $(\boldsymbol{N/K \leq M \leq L})$ **:**

For this regime of cache storage, we consider two further sub-cases, i.e., (i) $N < K$ and (ii) $N \geq K$. We next treat each of the sub-cases separately.

– ***Sub-case 1*** $(\boldsymbol{N < K})$: For this sub-case, we note that from the minimum storage constraint for D2D-assisted delivery, i.e., $KM \geq N$, the minimum allowable cache storage at each user can be less than unity. Therefore, we divide the available cache storage in this regime into two sub-regimes namely (i) $N/K \leq M \leq 0.5$ and (ii) $0.5 \leq M \leq L$. We these sub-regimes separately as follows. Consider first, the sub-regime i.e., $N/K \leq M \leq 0.5$. For this sub-regime consider the case when $N = 1$. For this case, setting $s = 1$ and $\beta = 1$, from the lower bound in (A.57), we have

$$R^*_{\mathsf{d2d,L}}(M) \geq (1 - M), \tag{A.58}$$

where we have used the fact that $L = 1$ when $N = 1$. Again considering the upper bound in (3.11), we have $R_{\mathsf{d2d,L}} \leq 1$. Using the upper and the lower bounds, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}}{R^*_{\mathsf{d2d,L}}} \leq \frac{1}{1-M} \leq \frac{1}{1-0.5} = 2. \tag{A.59}$$

Next, we consider the case when $N \geq 2$. For this case, setting $s = \lceil \frac{N}{L} \rceil \in \left[1 : \lceil \frac{N}{L} \rceil\right]$ and $\beta = 1$, from (A.57), we have

$$\begin{aligned} R^*_{\mathsf{d2d,L}}(M) &\geq \frac{N\left(1 - \left\lceil \frac{N}{L} \right\rceil \frac{M}{N}\right)}{\left\lceil \frac{N}{L\lceil \frac{N}{L} \rceil} \right\rceil \frac{K - \lceil \frac{N}{L} \rceil}{K}} \geq N\left(1 - \left(\frac{N}{L} + 1\right)\frac{M}{N}\right) \\ &= N\left(1 - \left(\frac{1}{L} + \frac{1}{N}\right)M\right) \overset{\mathsf{(a)}}{\geq} N\left(1 - \frac{3}{2} \times 0.5\right), \end{aligned} \tag{A.60}$$

where step (a) follows from the fact that $N \geq 2$ and $L \geq 1$. Again, from the upper bound in (3.11), we have $R_{\mathsf{d2d,L}} \leq N$. Using this, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}}{R^*_{\mathsf{d2d,L}}} \leq \frac{1}{1 - \frac{3}{2} \times 0.5} = 4. \tag{A.61}$$

We next consider the sub-regime $0.5 \leq M \leq L$. In this regime, setting $s = \left\lfloor 0.5\frac{N}{M} \right\rfloor \in [1 : K]$[2]

[2]The regime of s can be verified as follows. Using the lower bound $0.5 \leq M$, we have $0.5N/M \leq N < K$. Again using the upper bound $M \leq L$, we have $0.5N/M \geq 0.5N/L \geq 1$.

and $\beta = 1$, from the lower bound in (A.57), we have

$$R^*_{\mathsf{d2d,L}} \geq \frac{N\left(1 - \left\lfloor 0.5\frac{N}{M}\right\rfloor \frac{M}{N}\right)}{\left\lceil \frac{N}{L\left\lfloor 0.5\frac{N}{M}\right\rfloor}\right\rceil \frac{K-\left\lfloor 0.5\frac{N}{M}\right\rfloor}{K}} \geq \frac{N\left(1-0.5\right)}{\left\lceil \frac{N/L}{\left\lfloor 0.5\frac{N}{L}\right\rfloor}\right\rceil} \overset{\text{(a)}}{\geq} \frac{N(1-0.5)}{3}, \tag{A.62}$$

where step (a) follows from the fact that for any $N/L \geq 2$, we have $\frac{N/L}{\lfloor 0.5(N/L)\rfloor} \leq 3$. Again from the upper bound in (3.11), we have $R_{\mathsf{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq \frac{3}{1-0.5} = 6. \tag{A.63}$$

– ***Sub-case 2*** $(\boldsymbol{N \geq K})$: For this sub-case, we note that from the minimum storage constraint for D2D-assisted delivery, i.e., $KM \geq N$, we have $M \geq 1$. Therefore, we consider the following regime of available cache storage $0.5 \leq N/K \leq M \leq L$. In this regime, setting $s = \left\lfloor 0.5\frac{N}{M}\right\rfloor \in [1:K]$[3] and $\beta = 1$, from the lower bound in (A.57), we have

$$R^*_{\mathsf{d2d,L}} \geq \frac{N\left(1 - \left\lfloor 0.5\frac{N}{M}\right\rfloor \frac{M}{N}\right)}{\left\lceil \frac{N}{L\left\lfloor 0.5\frac{N}{M}\right\rfloor}\right\rceil \frac{K-\left\lfloor 0.5\frac{N}{M}\right\rfloor}{K}} \geq \frac{N\left(1-0.5\right)}{\left\lceil \frac{N/L}{\left\lfloor 0.5\frac{N}{L}\right\rfloor}\right\rceil} \overset{\text{(a)}}{\geq} \frac{N(1-0.5)}{3}, \tag{A.64}$$

where step (a) again follows from the fact that for any $N/L \geq 2$, we have $\frac{N/L}{\lfloor 0.5(N/L)\rfloor} \leq 3$. Again from the upper bound in (3.11), we have $R_{\mathsf{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq \frac{3}{1-0.5} = 6. \tag{A.65}$$

- ***Regime 2*** $(\boldsymbol{L \leq M \leq 0.2N})$ **:**

For this regime, setting $s = \left\lfloor 0.51\frac{N}{M}\right\rfloor \in [1:K]$[4] and $\ell = \left\lceil \frac{0.984N}{Ls}\right\rceil$, from (A.57), we have

$$\begin{aligned} R^*_{\mathsf{d2d,L}}(M) \;&\geq \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - s\frac{M}{N}}{\left(\frac{0.984}{Ls} + \frac{1}{N}\right)\left[\frac{K-s}{K}\right]} \overset{\text{(a)}}{=} \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - \left\lfloor 0.51\frac{N}{M}\right\rfloor \frac{M}{N}}{\frac{0.984}{L\left\lfloor 0.51\frac{N}{M}\right\rfloor} + \frac{1}{N}} \overset{\text{(b)}}{\geq} \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - 0.51\frac{N}{M}\frac{M}{N}}{\frac{0.984}{L\left(0.51\frac{N}{M}-1\right)} + \frac{1}{N}} \\ &\geq \frac{\frac{LN}{M}\left(0.51 - \frac{M}{N}\right)\left\{\left(\frac{2\times 0.984}{1+0.984}\right) - 0.51\right\}}{0.984 + 0.51\left(\frac{L}{M} - \frac{L}{N}\right)} \overset{\text{(c)}}{\geq} \frac{\frac{LN}{M}\left(0.51 - 0.2\right)\left\{\left(\frac{2\times 0.984}{1+0.984}\right) - 0.51\right\}}{0.984 + 0.51} \geq \frac{LN}{10M}, \end{aligned} \tag{A.66}$$

[3]The regime of s is validated as follows. Using the lower bound $M \geq N/K$, we have $0.5N/M \leq 0.5K \leq K$. Again using the upper bound $M \leq L$, we have $0.5N/M \geq 0.5N/L \geq 1$.

[4]The regime of s can be validated as follows. Consider first, a lower bound on $0.5N/M$. In the given regime, we have $0.5N/M \geq 0.5/0.2 \geq 1$. Next, we consider an upper bound on $0.5N/M$. Consider first, the case when $N/L \leq K$. In this case, its easy to note that $0.5N/M \leq K$. Next consider the case that $N/L \geq K$. In this case, *Regime 2* reduces to $L \leq N/K \leq M \leq 0.2N$ due to the minimum storage constraint and hence we have $0.5N/M \leq 0.5K \leq K$. Therefore we have $\lfloor 0.5N/M\rfloor \in [1:K]$.

where step (a) follows due to the fact that $(K-s)/K \leq 1$; step (b) follows by using $\lfloor 0.51N/M \rfloor \leq 0.51N/M$ in the numerator and $\lfloor 0.51N/M \rfloor \leq 0.51N/M - 1$ in the denominator; and step (c) follows by using $M/N \leq 0.2$ in the numerator and $M \geq L$ in the denominator. Again, considering the upper bound in (3.11), we have

$$R_{\mathsf{d2d,L}}(M) \leq \frac{LN}{M}\left(1-\frac{M}{N}\right) \leq \frac{LN}{M}. \tag{A.67}$$

Therefore for *Regime 2*, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq 10. \tag{A.68}$$

- ***Regime 3*** $(\boldsymbol{0.2N \leq M \leq N})$ **:**

  In this regime, setting $s=1$ and $\beta=1$ in (A.57), we have

$$R^*_{\mathsf{d2d,L}}(M) \geq \frac{N\left(1-\frac{M}{N}\right)}{\frac{N}{L}+1} \geq \frac{\left(1-\frac{M}{N}\right)}{\frac{1}{L}+\frac{1}{N}}. \tag{A.69}$$

  Again, considering the upper bound in (3.11), we have

$$R_{\mathsf{d2d,L}}(M) \leq \frac{LN}{M}\left(1-\frac{M}{N}\right). \tag{A.70}$$

  Therefore for *Regime 3*, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq \frac{LN}{M}\left(\frac{1}{L}+\frac{1}{N}\right) \overset{\text{(a)}}{\leq} \frac{LN}{M} \times \frac{2}{L} \leq \frac{2N}{M} \leq \frac{2}{0.2} \leq 10, \tag{A.71}$$

  where step (a) follows from the fact that $L \leq N$.

• ***Case 2*** $(\boldsymbol{0.5N \leq L})$: For the case of high per-device demands, we divide the available cache storage at each device into the following two regimes, namely (i) *Regime 1*: $N/K \leq M \leq N/3$; and (ii) *Regime 2*: $N/3 \leq M \leq N$. We next consider each regime separately.

- ***Regime 1*** $(\boldsymbol{N/K \leq M \leq N/3})$ **:**

  For this regime, setting $s=1$ and $\ell = \left\lceil \frac{0.5N}{Ls} \right\rceil$, from (A.57), we have

$$R^*_{\mathsf{d2d,L}}(M) \geq \frac{N\left(\left(\frac{2\times 0.5}{1+0.5}\right)-\frac{M}{N}\right)}{\left\lceil \frac{0.5N}{L} \right\rceil \left(\frac{K-1}{K}\right)} \geq \frac{N\left(\left(\frac{2\times 0.5}{1+0.5}\right)-\frac{M}{N}\right)}{\frac{0.5N}{L}+1} \overset{\text{(a)}}{\geq} \frac{N\left(\left(\frac{2\times 0.5}{1+0.5}\right)-\frac{1}{3}\right)}{2}, \tag{A.72}$$

  where step (a) follows by using the lower bound $L \geq 0.5N$. Again, considering the upper bound in (3.11), we have $R_{\mathsf{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq \frac{2}{\left(\frac{2\times 0.5}{1+0.5}\right)-\frac{1}{3}} \leq 6. \tag{A.73}$$

- ***Regime 2*** $(\boldsymbol{N/3 \leq M \leq N})$ **:**

  In this regime, setting $s = 1$ and $\beta = 1$ in (A.57), we have

$$R^*_{\mathsf{d2d,L}}(M) \geq \frac{N\left(1-\frac{M}{N}\right)}{\frac{N}{L}+1} \geq \frac{\left(1-\frac{M}{N}\right)}{\frac{1}{L}+\frac{1}{N}}. \tag{A.74}$$

  From (3.11), we have

$$R_{\mathsf{d2d,L}}(M) \leq \frac{LN}{M}\left(1-\frac{M}{N}\right). \tag{A.75}$$

  Therefore for *Regime 3*, we have

$$\mathrm{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq \frac{LN}{M}\left(\frac{1}{L}+\frac{1}{N}\right) \leq \frac{LN}{M} \times \frac{2}{L} \leq \frac{2N}{M} \leq \frac{2}{1/3} = 6 \tag{A.76}$$

Finally, combining (A.59), (A.61), (A.63), (A.65), (A.68), (A.71), (A.73) and (A.76), completes the proof of Theorem 8. □

A.6 Proof of Theorem 9

From Corollary 2, considering the lower bound on the optimal rate $R^*_{\mathsf{d2d}}(M)$, we set $\ell = \left\lceil \frac{\beta N}{s} \right\rceil \in \{1, 2, \ldots, \left\lceil \frac{N}{s} \right\rceil\}$ with $0 < \beta \leq 1$. Under this setting, we can again use the upper bound on $\left(\frac{\mu}{\mu+s}\right)$ from (A.33) from Appendix A.3. Substituting (A.33) into (3.18), we have the following form on the lower bound

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N - sM - \frac{1}{1+\beta}\left(N - \left\lceil \frac{\beta N}{s} \right\rceil s\right)^+}{\left\lceil \frac{\beta N}{s} \right\rceil \left(\frac{K-s}{K}\right)} \geq \frac{N\left(\frac{2\beta}{1+\beta} - s\frac{M}{N}\right)}{\left\lceil \frac{\beta N}{s} \right\rceil \left(\frac{K-s}{K}\right)}. \tag{A.77}$$

In order to facilitate the proof of Theorem 9, we consider the three cases, namely (i) the gap at $M = N/K$; (ii) the case when $N < K$; and (iii) the case when $N \geq K$. We next consider each of these cases separately.

• ***Case 1*** $(\boldsymbol{Gap\ at\ M = N/K})$: In this case, setting $s = N$ and $\beta = 1$ in (A.77), we have

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N(1-M)}{\left(\frac{K-N}{K}\right)} = \frac{KN\left(1-\frac{N}{K}\right)}{K-N} = N. \tag{A.78}$$

Again, from [105, Theorem 1], we have

$$R_{\mathsf{d2d}}(M) \leq \frac{N}{M}\left(1-\frac{M}{N}\right) \leq N, \tag{A.79}$$

where the last inequality stems from the fact that for $N \geq K$, $M \geq 1$. Thus the gap is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq 1. \tag{A.80}$$

• ***Case 2*** $(\boldsymbol{N < K})$: For $N < K$ we divide the cache storage into 3 distinct regimes: *Regime 1:* $N/K < M \leq 1$; *Regime 2:* $1 < M \leq 0.1250N$; and *Regime 3:* $0.1250N < M \leq N$. We consider each of the regimes separately.

• ***Regime 1*** $(\boldsymbol{N/K < M \leq 1})$ **:**

For this regime, we consider two sub-regimes, namely (i) $N/K < M \leq 2/3$; and (ii) $2/3 < M \leq 1$. We consider each sub-regime separately.

– ***Sub-regime*** $\mathbf{1}$ $\mathbf{(N/K < M \leq 2/3)}$ **:** From (A.77), setting $s = N$ and $\beta = 1$, we have

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N\,(1-M)}{\left(\frac{K-N}{K}\right)} \geq N(1-M) \geq N\left(1 - \frac{2}{3}\right) = \frac{N}{3} \tag{A.81}$$

Again from [105, Theorem 1], we have

$$R_{\mathsf{d2d}} \triangleq \min\left\{\frac{N}{M}\left(1 - \frac{M}{N}\right), N\right\} \leq N \tag{A.82}$$

Thus in this sub-regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq \frac{N}{N/3} \leq 3. \tag{A.83}$$

– ***Sub-regime*** $\mathbf{2}$ $\mathbf{(2/3 < M \leq 1)}$ **:** Consider first, the case when $N = 1$. Setting $s = 1, \beta = 1$ in (A.77), we have

$$R^*_{\mathsf{d2d}}(M) \geq \frac{1 - sM}{\lceil 1/s \rceil} \geq 1 - M. \tag{A.84}$$

From [105, Theorem 1], setting $N = 1$, we have

$$R_{\mathsf{d2d}}(M) \leq \frac{1}{M}(1-M). \tag{A.85}$$

Thus the gap in this case is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq \frac{1}{M} \leq \frac{3}{2}. \tag{A.86}$$

For the case of $N \geq 2$, setting $s = \left\lfloor \frac{2N}{3M} \right\rfloor$ and $\beta = 1$ in (A.77), we have

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N\left(1 - \left\lfloor \frac{2N}{3M} \right\rfloor \frac{M}{N}\right)}{\left\lceil \frac{N}{\left\lfloor \frac{2N}{3M} \right\rfloor} \right\rceil \left(\frac{K-s}{K}\right)} \geq \frac{N\left(1 - \left\lfloor \frac{2N}{3M} \right\rfloor \frac{M}{N}\right)}{\left\lceil \frac{N}{\left\lfloor \frac{2}{3}N \right\rfloor} \right\rceil} \geq \frac{N\left(1 - \frac{2}{3}\right)}{\lceil 2 \rceil} \geq \frac{N}{6}, \tag{A.87}$$

where, we have used the fact that $\frac{N}{\lfloor 2N/3 \rfloor} \leq 2, \ \ \forall N \geq 2$. Again from [105, Theorem 1], we have

$$R_{\mathsf{d2d}} \triangleq \min\left\{ \frac{N}{M}\left(1 - \frac{M}{N}\right), N \right\} \leq N \tag{A.88}$$

Thus in this sub-regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq \frac{N}{N/6} \leq 6. \tag{A.89}$$

Thus, in general, in this regime, we can upper bound the Gap by the constant 6.

- ***Regime 2 ($1 < M \leq 0.1250N$) :***

  For this regime, setting $s = \lfloor 0.5N/M \rfloor$, $\beta = 1$ in (A.77), we have

$$\begin{aligned} R^*_{\mathsf{d2d}}(M) &\geq \frac{N\left(\frac{2\beta}{1+\beta} - s\frac{M}{N}\right)}{\frac{\beta N}{s} + 1} \times \frac{K}{K-s} \geq \frac{N\left(1 - \lfloor 0.5N/M \rfloor \frac{M}{N}\right)}{\frac{N}{\lfloor 0.5N/M \rfloor} + 1} \\ &\geq \frac{N\left(1 - 0.5\right)}{\frac{N}{0.5N/M - 1} + 1} \geq \frac{0.5N\left(0.5N/M - 1\right)}{N + \frac{0.5N}{M} - \frac{1}{N}} \geq \frac{0.5\left(0.5N/M - 1\right)}{1 + \frac{0.5}{M}} \geq \frac{N}{M}\frac{0.5\left(0.5 - \frac{M}{N}\right)}{1 + 0.5} \\ &\geq \frac{N}{M}\frac{0.5\left(0.5 - 0.1250\right)}{1 + 0.5} \geq \frac{N}{8M} \end{aligned} \tag{A.90}$$

  Again, from [105, Theorem 1], we have

$$R_{\mathsf{d2d}}(M) \leq \frac{N}{M} - 1 \leq \frac{N}{M}. \tag{A.91}$$

  Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq 8. \tag{A.92}$$

- ***Regime 3 ($0.1250N < M \leq N$) :***

  For this regime, setting $s = 1$ and $\beta = 1$ in (A.77), we have

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N - M}{N} = \left(1 - \frac{M}{N}\right) \tag{A.93}$$

Again, from [105, Theorem 1], we have

$$R_{\mathsf{d2d}}(M) \leq \frac{N}{M}\left(1-\frac{M}{N}\right) \leq \frac{1}{0.1250}\left(1-\frac{M}{N}\right). \tag{A.94}$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq \frac{1}{0.1250} = 8. \tag{A.95}$$

Thus, for the case of $K > N$, the gap between the upper and lower bounds is 1 at $M = N/K$, at most 6 for $N/K \leq M \leq 1$ and at most 8 for all $1 < M \leq N$.

• ***Case 3 ($N \geq K$)*** **:** For this case, we consider two main sub-cases, namely (i) $K \leq 8$; and (ii) $K \geq 9$. We treat each case separately.

• ***Sub-case 1 ($K \leq 8$)*** **:** For this case, from [105, Theorem 1], we have

$$R_{\mathsf{d2d}}(M) \leq \frac{N}{M}\left(1-\frac{M}{N}\right) \leq K\left(1-\frac{M}{N}\right), \tag{A.96}$$

where the last inequality is derived from the minimum storage constraint $KM > N$. Again, from (A.77), setting $s = 1, \beta = 1$, we have:

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N\left(1-\frac{M}{N}\right)}{\lceil N \rceil} \geq \left(1-\frac{M}{N}\right). \tag{A.97}$$

Thus the gap between the lower and upper bound is given by:

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq K \leq 8. \tag{A.98}$$

• ***Sub-case 2 ($K \geq 8$)*** **:**

For this case, we divide the cache storage into 3 distinct regimes namely, (i) Regime 1: $1 \leq N/K \leq M \leq 1.15N/K$; (ii) Regime 2: $1.15N/K < M \leq 0.1250N$; and (iii) Regime 3: $0.1250N < M \leq N$. We next consider each regime separately.

• ***Regime 1 ($1 \leq N/K \leq M \leq 1.15N/K$)*** **:**

For this regime, setting $s = \lfloor 0.4361K \rfloor$ and $\beta = 0.7398$ in (A.77) and using the fact that $K \geq 9$, we have

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N\left(\frac{2\beta}{1+\beta} - s\frac{M}{N}\right)}{\frac{\beta N}{s}+1} \times \frac{K}{K-s} \geq \frac{N\left(\frac{2\times 0.7398}{1+0.7398} - \lfloor 0.4361K \rfloor \frac{M}{N}\right)}{\frac{0.7398N}{\lfloor 0.4361K \rfloor}+1} \times \frac{1}{1-\frac{\lfloor 0.4361K \rfloor}{K}}$$

$$\geq \frac{N\left(\frac{2\times 0.7398}{1+0.7398} - 0.4361\frac{KM}{N}\right)}{\frac{0.7398N}{0.4361K-1}+1} \times \frac{1}{1-\frac{0.4361K-1}{K}}$$
$$\geq \frac{N\left(0.4361K-1\right)\left[\frac{2\times 0.7398}{1+0.7398} - 0.4361\frac{KM}{N}\right]}{0.7398N+0.4361K-1} \times \frac{1}{1+\frac{1}{K}-0.4361}$$
$$\geq K\frac{\left(0.4361-\frac{1}{9}\right)\left[\frac{2\times 0.7398}{1+0.7398} - 0.4361\frac{KM}{N}\right]}{0.7398+0.4361\frac{K}{N}-\frac{1}{N}} \times \frac{1}{1+\frac{1}{9}-0.4361}$$
$$\geq K\frac{\left(0.4361-\frac{1}{9}\right)\left[\frac{2\times 0.7398}{1+0.7398} - 0.4361\times 1.15\right]}{0.7398+0.4361} \times \frac{1}{1+\frac{1}{9}-0.4361} \geq \frac{K}{7}. \quad \text{(A.99)}$$

Again, from (A.96), we have

$$R_{\mathsf{d2d}}(M) \leq K. \quad \text{(A.100)}$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq \frac{K}{K/7} = 7. \quad \text{(A.101)}$$

- ***Regime 2 ($1.15N/K < M \leq 0.1250N$) :***

For this regime, setting $s = \left\lfloor 0.4470\frac{N}{M} \right\rfloor$ and $\beta = 0.8995$ in (A.77), we have

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N\left(\frac{2\beta}{1+\beta} - s\frac{M}{N}\right)}{\frac{\beta N}{s}+1} \times \frac{K}{K-s} \geq \frac{N\left(\frac{2\times 0.8995}{1+0.8995} - \left\lfloor 0.4470\frac{N}{M}\right\rfloor \frac{M}{N}\right)}{\frac{0.8995N}{\left\lfloor 0.4470\frac{N}{M}\right\rfloor}+1} \geq \frac{N\left(\frac{2\times 0.8995}{1+0.8995} - 0.4470\right)}{\frac{0.8995N}{0.4470\frac{N}{M}-1}+1}$$
$$\geq \frac{\left(0.4470\frac{N}{M}-1\right)\left[\frac{2\times 0.8995}{1+0.8995} - 0.4470\right]}{0.8995+\frac{0.4470}{M}-\frac{1}{N}} \geq \frac{N}{M}\frac{\left(0.4470-\frac{M}{N}\right)\left[\frac{2\times 0.8995}{1+0.8995} - 0.4470\right]}{0.8995+\frac{0.4470K}{1.15N}}$$
$$\geq \frac{N}{M}\frac{\left(0.4470-0.1250\right)\left[\frac{2\times 0.8995}{1+0.8995} - 0.4470\right]}{0.8995+\frac{0.4470}{1.15}} \geq \frac{N}{8M} \quad \text{(A.102)}$$

Again, from [105, Theorem 1], we have

$$R_{\mathsf{d2d}}(M) \leq \frac{N}{M} - 1 \leq \frac{N}{M}. \quad \text{(A.103)}$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq 8. \quad \text{(A.104)}$$

- ***Regime 3 ($0.1250N < M \leq N$) :***

For this regime, setting $s = 1$ and $\beta = 1$ in (A.77), we have

$$R^*_{\mathsf{d2d}}(M) \geq \frac{N-M}{N} = \left(1 - \frac{M}{N}\right) \tag{A.105}$$

Again, from [105, Theorem 1], we have

$$R_{\mathsf{d2d}}(M) \leq \frac{N}{M}\left(1 - \frac{M}{N}\right) \leq \frac{1}{0.1250}\left(1 - \frac{M}{N}\right). \tag{A.106}$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\mathsf{d2d}}(M)}{R^*_{\mathsf{d2d}}(M)} \leq \frac{1}{0.1250} = 8. \tag{A.107}$$

Thus, for the case of $N \geq K$, the gap between the upper and lower bounds is at most 8 for all $N/K \leq M \leq N$. Combining the three cases completes the proof of Theorem 9. □

A.7 Proof of Lemma 3

Each user has a cache storage of MB bits in which they can each store $\frac{M}{N}$ fragments of each of the N files. Now assume each file is of size $B^{'} = \alpha B$ bits. We let each user store $\frac{M}{\alpha N}$ fragment of each file. Now consider the case where $\frac{M}{\alpha} = \frac{Nt}{K}, t \in \{1, \ldots, K\}$. In this case we have $t \triangleq \frac{KM}{\alpha N}$. Following the storage and multicast delivery scheme in [97], each file of size $B^{'}$ bits is then divided into $\binom{K}{t}$ sub-files of size $B^{'}/\binom{K}{t}$ bits. Each user caches a total of $N\binom{K-1}{t-1}$ of these sub-files i.e., each user caches a total of

$$N\binom{K-1}{t-1}\frac{B^{'}}{\binom{K}{t}} = \alpha B\frac{Nt}{K} = MB \text{ bits}, \tag{A.108}$$

which satisfies the cache storage constraint. Further, in the delivery phase, $\binom{K}{t+1}$ transmissions, each of size $B^{'}/\binom{K}{t}$ bits are made. Thus the achievable rate R is given by:

$$\begin{aligned} RB &= \binom{K}{t+1}\frac{B^{'}}{\binom{K}{t}} = \alpha B\frac{K-t}{t+1} \\ \Rightarrow \quad R &= \alpha\frac{K\left(1-\frac{M}{\alpha N}\right)}{1+\frac{KM}{\alpha N}} = \alpha R^m_{\mathsf{hom}}\left(\frac{M}{\alpha}, N, K\right). \end{aligned} \tag{A.109}$$

For any other $M/\alpha \in (0, N]$, cache splitting and time-sharing [97] can achieve the rate $\alpha R^m_{\mathsf{hom}}\left(\frac{M}{\alpha}, N, K\right)$, where R^m_{hom} is given in (3.46) for $t = \left\lfloor \frac{KM}{\alpha N} \right\rfloor$. The following example illustrates Lemma 3.

Example 13. Consider $N = 3$ files A, B, C of unit size, $K = 3$ users and per-user storage $M = 1$. In order to deliver the entire files to the users utilizing M i.e., $\alpha = 1$, the achievable scheme is the following [97]. Divide each file into non-overlapping fragments: $A \to A_1, A_2, A_3$, $B \to B_1, B_2, B_3$ and $C \to C_1, C_2, C_3$. The size of each fragment is $1/3$. The users' caches are stored with the content $Z_1 = \{A_1, B_1, C_1\}$, $Z_2 = \{A_2, B_2, C_2\}$ and $Z_3 = \{A_3, B_3, C_3\}$. Considering the worst-case demand $(d_1, d_2, d_3) = (A, B, C)$ the requested content can be delivered with the multicast transmission - $A_2 \oplus B_1$, $A_3 \oplus C_1$, $B_3 \oplus C_2$. The total rate of this transmission is $R = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$, which can also be obtained from (3.46) by setting $t = \left\lfloor \frac{KM}{N} \right\rfloor = 1$. Now consider the case for $\alpha = 0.5$ i.e., each of A, B, C are of size $1/2$. In this case, with $M = 1$, we can store $\frac{1}{\alpha N} = 2/3$ of each file in the cache. Thus for this case, the following scheme is used. Each file is divided into fragments $A \to A_{12}, A_{13}, A_{23}$, $B \to B_{12}, B_{13}, B_{23}$ and $C \to C_{12}, C_{13}, C_{23}$. The size of each fragment is now $\alpha/3 = 1/6$. The users' caches store the following fragments: $Z_1 = \{A_{12}, A_{13}, B_{12}, B_{13}, C_{12}, C_{13}\}$, $Z_2 = \{A_{12}, A_{23}, B_{12}, B_{23}, C_{12}, C_{23}\}$ and $Z_3 = \{A_{13}, A_{23}, B_{13}, B_{23}, C_{13}, C_{23}\}$. In this case, a request of $(d_1, d_2, d_3) = (A, B, C)$ can be served by the single multicast transmission $A_{23} \oplus B_{13} \oplus C_{12}$ which has a rate of $1/6$. This can be seen in Fig. 3.8(b) and is same as $0.5R^m_{\mathsf{hom}}(1/0.5, 3, 3)$, from (3.46) for $t = \left\lfloor \frac{KM}{0.5N} \right\rfloor = 2$. For any value of M such that $\frac{KM}{0.5N} = 2M \notin \{0, 1, 2, 3\}$, cache sharing between two schemes can achieve $0.5R^m_{\mathsf{hom}}(2M, 3, 3)$. □ □

A.8 Proof of Theorem 11

We prove a lower bound on $R^*_{\mathsf{het}}(\mathcal{M}, N, K)$ for any N, K and ordered set of caches $\mathcal{M}$. Consider the contents of the first $s \in 1, 2, \ldots, \min\{N, K\}$ *smallest* caches $Z_1, \ldots, Z_s$ in $\mathcal{M}$. For a request vector $(d_1, d_2, \ldots, d_s, d_{s+1}, \ldots, d_K) = (1, 2, \ldots, s, \phi, \ldots, \phi)$, the transmission $X_1 = X_{(d_1,\ldots,d_k)}$, along with the contents $Z_1, \ldots, Z_s$ can decode the files $F_1, \ldots, F_s$. Similarly, for another request $(s+1, s+2, \ldots, 2s, \phi, \ldots, \phi)$, the transmission X_2, along with the cache contents $Z_1, \ldots, Z_s$, must be able to decode the files $F_{s+1}, \ldots, F_{2s}$. Thus, considering $\lfloor N/s \rfloor$ different such requests, the transmissions $X_1, \ldots, X_{\lfloor N/s \rfloor}$, along with the s smallest cache contents $Z_1, \ldots, Z_s$, must be able to decode the files $F_1, \ldots, F_{s\lfloor N/s \rfloor}$. The information flow consisting of transmissions $X_1, \ldots, X_{\lfloor N/s \rfloor}$ and the cache contents $Z_1, \ldots, Z_s$ for decoding files $F_1, \ldots, F_{s\lfloor N/s \rfloor}$, has a minimum capacity $s\lfloor N/s \rfloor$. Thus, we have:

$$
\begin{aligned}
s \lfloor N/s \rfloor B &\leq H\left(Z_1, \ldots, Z_s, X_1, \ldots, X_{\lfloor N/s \rfloor}\right) && \text{(A.110)} \\
&\leq H\left(Z_1, \ldots, Z_s\right) + H\left(X_1, \ldots, X_{\lfloor N/s \rfloor}\right) && \text{(A.111)} \\
&\leq \sum_{i=1}^{s} H(Z_i) + \lfloor N/s \rfloor R^*_{\mathsf{het}} B && \text{(A.112)} \\
&\leq \sum_{M_i \in \mathcal{M}, i \in \{1,2,\ldots,s\}} M_i B + \lfloor N/s \rfloor R^*_{\mathsf{het}} B, && \text{(A.113)}
\end{aligned}
$$

where (A.112) results from the fact that each transmission has a rate not exceeding R^*_{het}. Solving for R^*_{het} and optimizing over all possible choices of s, we have:

$$R^*_{\mathsf{het}}(\mathcal{M}, N, K) \geq \max_{\substack{s\in 1,\ldots,\min\{N,K\} \\ M_i\in\mathcal{M}, i\in\{1,2,\ldots,s\}}} \left(s - \frac{\sum_{i=1}^{s} M_i}{\left\lfloor \frac{N}{s} \right\rfloor} \right). \tag{A.114}$$

This completes the proof of Theorem 11. □

A.9 Proof of Theorem 12

We first consider the case of two-level heterogeneity i.e., where the system has only two distinct cache sizes across all K users.

A.9.1 System with Two-Level Heterogeneity

Consider a system with N files and K users having set of heterogeneous caches $\mathcal{M} := \{M_1, M_2, \ldots, M_K\} \in (0, N]$ such that $M_1 = M_2 = \ldots = M_\ell = M_L$ and $M_{\ell+1} = M_{\ell+2} = \ldots = M_K = M_H$ with $M_L < M_H$ as shown in Fig. A.1. To prove the order-optimality of the LHC scheme for this system, we consider two cases namely (i) $\min\{N, K\} \leq C$ and (ii) $\min\{N, K\} \geq C + 1$ for some real constant C. We treat each of the two cases separately.

• ***Case 1*** $\boldsymbol{(\min\{N, K\} \leq C)}$ **:** First we consider an upper bound on the the achievable rate of the LHC scheme. Assume that we group the caches into $G = K$ non-overlapping subsets of caches. In this case, the proposed scheme reduces to a heterogeneous unicast scheme. Also assume that the minimum cache storage in $\mathcal{M}$ i.e., M_L is used for caching thereby reducing the system to a homogeneous unicast scheme with storage M_L. Thus we have:

$$R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M}, N, K) \leq \min\{N, K\} \left(1 - \frac{M_L}{N}\right). \tag{A.115}$$

Now considering the lower bound from Theorem 11, any choice of s further lower bounds the optimal rate. Thus, setting $s = 1$, we have:

$$R^*_{\mathsf{het}}(\mathcal{M}, N, K) \geq \left(1 - \frac{M_L}{N}\right). \tag{A.116}$$

Thus combining (A.115) and (A.116), we have

$$\frac{R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M}, N, K)}{R^*_{\mathsf{het}}(\mathcal{M}, N, K)} \leq \min\{N, K\} \leq C \tag{A.117}$$

• ***Case 2*** $\boldsymbol{(\min\{N, K\} \geq C + 1)}$ **:** For this case, we consider three regimes for the cache

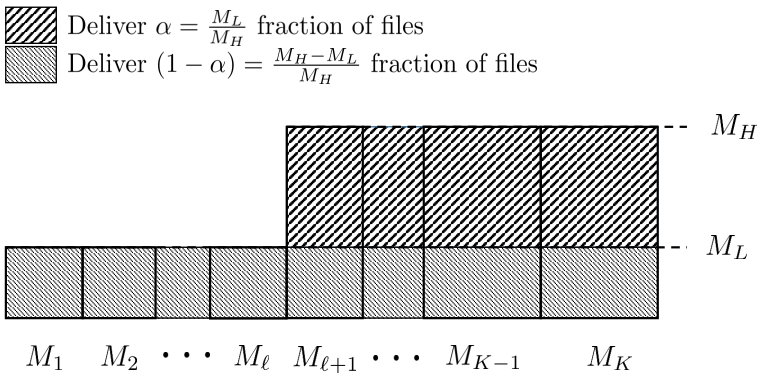


Figure A.1: LHC scheme for two-level heterogeneity.

storage M_H namely (i) ***Regime 1:*** $0 \leq M_H \leq \lambda_1 \max\left(1, \frac{N}{K}\right)$; (ii) ***Regime 2:*** $\lambda_1 \max\left(1, \frac{N}{K}\right) < M_H \leq \lambda_2 \frac{N}{\min\{N,\ell\}}$; and (iii) ***Regime 3:*** $\lambda_2 \frac{N}{\min\{N,\ell\}} < M_H \leq N$. We next consider each of the three regimes separately:

- ***Regime 1 ($0 \leq M_H \leq \lambda_1 \max(1, N/K)$) :***

  For this regime, consider the lower bound from Theorem 11. Setting $s = \lfloor \mu_1 \min\{N, K\} \rfloor$, with $\mu_1 \leq 1$, we have

$$\begin{aligned} R^*_{\mathsf{het}}(\mathcal{M}, N, K) &\geq \frac{N - \sum_{i=1}^{s} M_i}{\frac{N}{s} + 1} \\ &\geq \frac{N - sM_H}{\frac{N}{s} + 1} = \frac{N - \lfloor \mu_1 \min\{N,K\} \rfloor M_H}{\frac{N}{\lfloor \mu_1 \min\{N,K\} \rfloor} + 1} \\ &\geq \frac{N - \mu_1 \min\{N,K\} M_H}{\frac{N}{\mu_1 \min\{N,K\} - 1} + 1} \geq \frac{(\mu_1 \min\{N,K\} - 1)\left[1 - \mu_1 \frac{\min\{N,K\} \max\{1, N/K\}}{N}\right]}{1 + \mu_1 \frac{\min\{N,K\}}{N} - \frac{1}{N}} \\ &\geq \min\{N,K\} \frac{(\mu_1 - \frac{1}{C+1})\left[1 - \mu_1 \lambda_1\right]}{1 + \mu_1} \end{aligned} \tag{A.118}$$

  Next, we consider the upper bound on the LHC rate. Consider a scheme where we have $G = K$ subsets of caches and all $\vec{\alpha}_g = 0,\ \forall g \in \{1, 2, \ldots, G\}$. Thus no storage is used in any cache and the achievable rate can be upper bounded by

$$R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M}, N, K) \leq \min\{N, K\}. \tag{A.119}$$

  Thus combining (A.119) and (A.118), we have

$$\frac{R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M}, N, K)}{R^*_{\mathsf{het}}(\mathcal{M}, N, K)} \leq \frac{1 + \mu_1}{(\mu_1 - \frac{1}{C+1})\left[1 - \mu_1 \lambda_1\right]} \leq C \tag{A.120}$$

- ***Regime 2*** $(\boldsymbol{\lambda_1 \max(1, N/K) < M_H \leq \lambda_2 N/\min\{N,\ell\}})$ **:**

For this regime, consider the lower bound from Theorem 11. Setting $s = \lfloor \mu_2 \frac{N}{M_H} \rfloor$, we have

$$\begin{aligned} R^*_{\mathsf{het}}(\mathcal{M}, N, K) &\geq \frac{N - \sum_{i=1}^{s} M_i}{\frac{N}{s} + 1} \geq \frac{N - sM_H}{\frac{N}{s} + 1} \\ &= \frac{N - \lfloor \mu_2 \frac{N}{M_H} \rfloor M_H}{\frac{N}{\lfloor \mu_2 \frac{N}{M_H} \rfloor} + 1} \geq \frac{N - \mu_2 \frac{N}{M_H} M_H}{\frac{N}{\mu_2 \frac{N}{M_H} - 1} + 1} \\ &\geq \frac{(1-\mu_2)\left[\mu_2 \frac{N}{M_H} - 1\right]}{1 + \frac{\mu_2}{M_H} - \frac{1}{N}} \geq \frac{N}{M_H} \frac{(1-\mu_2)\left[\mu_2 - \frac{M_H}{N}\right]}{1 + \frac{\mu_2}{\lambda_1}} \\ &\geq \frac{N}{M_H} \frac{(1-\mu_2)\left[\mu_2 - \frac{\lambda_2}{\min\{N,\ell\}}\right]}{1 + \frac{\mu_2}{\lambda_1}} \overset{\text{(a)}}{\geq} \frac{N}{M_H} \frac{(1-\mu_2)\left[\mu_2 - \lambda_2\right]}{1 + \frac{\mu_2}{\lambda_1}}, \end{aligned} \tag{A.121}$$

where step (a) follows from the fact that $\min\{N, \ell\} \geq 1$.

Next, we consider the upper bound on the optimal rate. For the two-level heterogeneous system, consider the index ℓ such that the cache storage $M_\ell = M_L$ and $M_{\ell+1} = M_H$ i.e., users $1, 2, \ldots, \ell$ have cache storage M_L and users $\ell+1, \ell+2, \ldots, K$ have cache storage M_H. A fraction α of requested files is delivered to all K users using the storage M_L. For users $\ell+1, \ldots, K$ the remaining $1-\alpha$ fraction of files is delivered using cache storage $M_H - M_L$. For users $1, 2, \ldots, \ell$, the remaining $1-\alpha$ fraction is delivered via unicast with a rate of $\min\{N, \ell\}(1-\alpha)$. Considering $\alpha = \frac{M_L}{M_H}$, the corresponding LHC scheme is shown in Fig. A.1. The rate of this LHC scheme is upper bounded as follows:

$$\begin{aligned} &R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M}, N, K) \\ &\leq \alpha R^m_{\mathsf{hom}}\left(\frac{M_L}{\alpha}, N, K\right) + (1-\alpha) R^m_{\mathsf{hom}}\left(\frac{M_H - M_L}{1-\alpha}, N, K - \ell + 1\right) + \min\{N, \ell\}(1-\alpha) \\ &\overset{\text{(a)}}{\leq} \alpha^2 \frac{N}{M_L} + (1-\alpha)^2 \frac{N}{M_H - M_L} + \min\{N, \ell\}(1-\alpha) \\ &\overset{\text{(b)}}{=} \alpha \frac{N}{M_H} + (1-\alpha)\frac{N}{M_H} + \min\{N, \ell\} \leq \frac{N}{M_H} + \lambda_2 \frac{N}{M_H} = \frac{N}{M_H}(1 + \lambda_2), \end{aligned} \tag{A.122}$$

where step (a) follows by the fact that $R^m_{\mathsf{hom}}(M, N, K) \leq \frac{N}{M}$ for any N, K [97] and step (b) follows from the definition of α. Thus, combining (A.122) and (A.121), we have

$$\frac{R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M}, N, K)}{R^*_{\mathsf{het}}(\mathcal{M}, N, K)} \leq \frac{(1+\lambda_2)\left[1 + \frac{\mu_2}{\lambda_1}\right]}{(\mu_2 - \lambda_2)(1-\mu_2)} \leq C. \tag{A.123}$$

- ***Regime 3*** $(\boldsymbol{\lambda_2 N/\min\{N,\ell\} < M_H \leq N})$ **:**

For this regime, we consider sub-regimes namely (i) $M_L \leq \lambda_2 \frac{N}{\min\{N,\ell\}}$ and (ii) $M_L \geq \lambda_2 \frac{N}{\min\{N,\ell\}}$. We next discuss each of these regimes separately.

– ***Sub-Regime 1*** $(\boldsymbol{M_L \leq \lambda_2 N/\min\{N,\ell\} < M_H \leq N})$ **:**
For this regime, we set $s = \lfloor \mu_3 \min\{N,\ell\} \rfloor$ for the lower bound in Theorem 11 with $\mu_3 \leq 1$. Since we have $s \leq \ell$ the lower bound takes the following form:

$$\begin{aligned} R^*_{\mathsf{het}}(\mathcal{M},N,K) &\geq \frac{N-\sum_{i=1}^{s} M_i}{\frac{N}{s}+1} \geq \frac{N-sM_L}{\frac{N}{s}+1} = \frac{N-\lfloor \mu_3 \min\{N,\ell\}\rfloor M_L}{\frac{N}{\lfloor \mu_3 \min\{N,\ell\}\rfloor}+1} \\ &\geq \frac{N-\mu_3 \min\{N,\ell\} M_L}{\frac{N}{\mu_3 \min\{N,\ell\}-1}+1} \geq \frac{(\mu_3 \min\{N,\ell\}-1)\left[1-\mu_3 \frac{\min\{N,\ell\}}{N}\cdot \frac{\lambda_2 N}{\min\{N,\ell\}}\right]}{1+\mu_3\frac{\min\{N,\ell\}}{N}-\frac{1}{N}} \\ &\geq \min\{N,\ell\}\frac{\left(\mu_3-\frac{1}{\min\{N,\ell\}}\right)[1-\mu_3\lambda_2]}{1+\mu_3} \overset{\text{(a)}}{\geq} \min\{N,\ell\}\frac{\left(\mu_3-\frac{1}{3}\right)[1-\mu_3\lambda_2]}{1+\mu_3}, \end{aligned} \tag{A.124}$$

where in step (a), we assume that $\min\{N,\ell\} \geq 3$. Now, consider the LHC scheme from Fig. A.1. Using $\alpha = \frac{M_L}{M_H}$, the rate of this scheme is upper bounded as follows:

$$\begin{aligned} &R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M},N,K) \\ &\leq \alpha R^m_{\mathsf{hom}}\left(\frac{M_L}{\alpha},N,K\right) + (1-\alpha)R^m_{\mathsf{hom}}\left(\frac{M_H-M_L}{1-\alpha},N,K-\ell+1\right) + \min\{N,\ell\}(1-\alpha) \\ &\overset{\text{(a)}}{\leq} \alpha^2\frac{N}{M_L} + (1-\alpha)^2\frac{N}{M_H-M_L} + \min\{N,\ell\}(1-\alpha) \\ &\overset{\text{(b)}}{\leq} \alpha\frac{N}{M_H} + (1-\alpha)\frac{N}{M_H} + \min\{N,\ell\} = \frac{N}{M_H} + \min\{N,\ell\} \\ &\overset{\text{(c)}}{\leq} \frac{\min\{N,\ell\}}{\lambda_2} + \min\{N,\ell\} = \min\{N,\ell\}\left(\frac{1}{\lambda_2}+1\right) \end{aligned} \tag{A.125}$$

where, step (a) is due to the fact that $R^m_{\mathsf{hom}}(M,N,K) \leq \frac{N}{M}\left(1-\frac{M}{N}\right)$, step (b) follows by using $\alpha = \frac{M_L}{M_H}$ and step (c) follows from the definition of *Regime 3*. Combining (A.125) and (A.124), we have

$$\frac{R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M},N,K)}{R^*_{\mathsf{het}}(\mathcal{M},N,K)} \leq \frac{(1+\mu_3)\left[\frac{1}{\lambda_2}+1\right]}{\left(\mu_3-\frac{1}{3}\right)[1-\mu_3\lambda_2]} \leq C. \tag{A.126}$$

The above holds true when $\min\{N,\ell\} \geq 3$. Next, consider the alternate case when $\min\{N.\ell\} \leq 2$. Setting $s=1$ in the lower bound in Theorem 11, we have $R^*_{\mathsf{het}}(\mathcal{M},N,K) \geq 1 - M_L/N \geq 1-\lambda_2$. Again, the upper bound in (A.125) is given by $R_{\mathsf{het}}(\mathcal{M},N,K) \leq 2\left(1/\lambda_2+1\right)$ yielding

$$\frac{R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M},N,K)}{R^*_{\mathsf{het}}(\mathcal{M},N,K)} \leq \frac{2\left(1/\lambda_2+1\right)}{1-\lambda_2} \leq C. \tag{A.127}$$

– ***Sub-Regime 2*** $(\boldsymbol{\lambda_2 N/\min\{N,\ell\} < M_L \leq M_H \leq N})$ **:**
We further sub-divide this regime into two sub-regimes and treat each of these regimes separately.

∗ ***Sub-Case 1*** $(\boldsymbol{\lambda_2 N/\min\{N,\ell\} < M_L \leq \lambda_3 N})$ **:**
For this regime we set $s = \left\lfloor \mu_4 \frac{N}{M_L} \right\rfloor$. We want to ensure that the largest value that this s can take is less than ℓ. In the regime of interest, this can be ensured by

$$\mu_4 \frac{N}{M_L} \leq \mu_4 \frac{N}{\lambda_2 N} \min\{N,\ell\} \leq \frac{\mu_4}{\lambda_2} \min\{N,\ell\} \leq \ell \;\; \Rightarrow \;\; \mu_4 \leq \lambda_2$$

Since, $s \leq \ell$, the lower bound takes the following form:

$$\begin{aligned} R^*_{\mathsf{het}}(\mathcal{M},N,K) &\geq \frac{N - \sum_{i=1}^{s} M_i}{\frac{N}{s}+1} \geq \frac{N - sM_L}{\frac{N}{s}+1} \\ &= \frac{N - \lfloor \mu_4 \frac{N}{M_L} \rfloor M_L}{\frac{N}{\lfloor \mu_4 \frac{N}{M_L} \rfloor}+1} \geq \frac{N - \mu_4 \frac{N}{M_L} M_L}{\frac{N}{\mu_4 \frac{N}{M_L}-1}+1} \\ &\geq \frac{(1-\mu_4)\left[\mu_4 \frac{N}{M_L} - 1\right]}{1 + \frac{\mu_4}{M_L} - \frac{1}{N}} \geq \frac{N}{M_L} \frac{(1-\mu_4)\left[\mu_4 - \frac{M_L}{N}\right]}{1+\frac{\mu_4}{\lambda_2}} \\ &\geq \frac{N}{M_L} \frac{(1-\mu_4)\left[\mu_4 - \lambda_3\right]}{1+\frac{\mu_4}{\lambda_2}}, \end{aligned} \tag{A.128}$$

Next, consider the upper bound on the LHC scheme. Consider that every user has only a cache storage of M_L. Under this assumption, the following upper bound on the rate of the LHC scheme holds:

$$R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M},N,K) \leq \frac{K\left(1-\frac{M_L}{N}\right)}{1+\frac{KM_L}{N}} \leq \frac{N}{M_L}\left(1-\frac{M_L}{N}\right) \leq \frac{N}{M_L}. \tag{A.129}$$

Combining (A.129) and (A.128), we have

$$\frac{R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M},N,K)}{R^*_{\mathsf{het}}(\mathcal{M},N,K)} \leq \frac{1+\frac{\mu_4}{\lambda_2}}{(1-\mu_4)\left[\mu_4 - \lambda_3\right]} \leq C. \tag{A.130}$$

∗ ***Sub-Case 2*** $(\boldsymbol{\lambda_3 N < M_L \leq N})$ **:**
In this regime, setting $s = 1$ in the lower bound from Theorem 11, we have

$$R^*_{\mathsf{het}}(\mathcal{M},N,K) \geq 1 - \frac{M_L}{N} \tag{A.131}$$

Next consider the upper bound on the LHC scheme with each user having storage M_L,

$$R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M},N,K) \leq \frac{K\left(1-\frac{M_L}{N}\right)}{1+\frac{KM_L}{N}} \leq \frac{N}{M_L}\left(1-\frac{M_L}{N}\right). \tag{A.132}$$

Combining (A.132) and (A.131), we have

$$\frac{R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right)}{R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right)} \leq \frac{N}{M_L} \leq \frac{1}{\lambda_3} \leq C. \tag{A.133}$$

Finding values for the parameters $\lambda_1, \lambda_2, \lambda_3, \mu_1, \mu_2, \mu_3, \mu_4$ which satisfy the inequalities in (A.120),(A.123),(A.126), (A.127), (A.130) and (A.133) for the minimum value of the constant C enables us to find a constant gap of $C = 19$ with $\lambda_1 = 1.1, \lambda_2 = 0.3150, \lambda_3 = 0.0610, \mu_1 = 0.1176, \mu_2 = 0.5460, \mu_3 = 0.9443, \mu_4 = 0.2261$, which further satisfies $\mu_4 \leq \lambda_2$.

A.9.2 System with Three-Level Heterogeneity

Consider a system with N files and K users having set of heterogeneous caches $\mathcal{M} := \{M_1, M_2, \ldots, M_K\} \in (0, N]$ such that

$$\begin{aligned} &M_1 = M_2 = \ldots = M_{\ell_1} = M_L, \\ &M_{\ell_1+1} = M_{\ell_1+2} = \ldots = M_{\ell_2} = M_I, \\ &M_{\ell_2+1} = M_{\ell_2+2} = \ldots = M_K = M_H, \end{aligned} \tag{A.134}$$

with $M_L < M_I < M_H$ as shown in Fig. A.2. To prove the order-optimality of the LHC scheme for such a system, we again consider two cases namely (i) $\min\{N, K\} \leq C$ and (ii) $\min\{N, K\} \geq C + 1$ for some real constant C. We treat each of the two cases separately.

• ***Case 1*** $\boldsymbol{(\min\{N, K\} \leq C)}$ **:** First we consider an upper bound on the the achievable rate of the LHC scheme by partitioning $\mathcal{M}$ into K non-overlapping subsets which reduces LHC to a heterogeneous unicast scheme. Also assume that the minimum cache storage in $\mathcal{M}$ i.e., M_L is used for caching thereby reducing the system to a homogeneous unicast scheme with storage M_L. Thus we have, from (3.47):

$$R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right) \leq \min\{N,K\}\left(1-\frac{M_L}{N}\right). \tag{A.135}$$

Now considering the lower bound from Theorem 11, any choice of s further lower bounds the optimal rate. Thus, setting $s = 1$, we have:

$$R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right) \geq \left(1-\frac{M_L}{N}\right). \tag{A.136}$$

Thus combining (A.135) and (A.136), we have

$$\frac{R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right)}{R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right)} \leq \min\{N,K\} \leq C \tag{A.137}$$

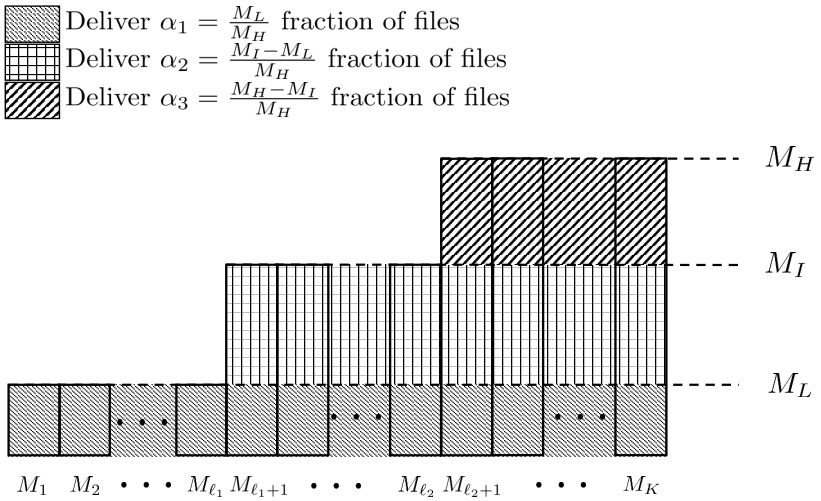


Figure A.2: LHC scheme for three-level heterogeneity.

• ***Case 2*** $\mathbf{(\min\{N, K\} \geq C + 1)}$ **:** For this case, we consider three *main regimes* for the highest cache storage M_H namely (i) ***Regime 1:*** $0 \leq M_H \leq \lambda_1 \max(1, N/K)$; (ii) ***Regime 2:*** $\lambda_1 \max(1, N/K) < M_H \leq \lambda_2 N/\min\{N, \ell_2\}$; and (iii) ***Regime 3:*** $\lambda_2 N/\min\{N, \ell_2\} < M_H \leq N$. We next consider each of the three regimes separately:

• ***Regime 1*** $\mathbf{(0 \leq M_H \leq \lambda_1 \max(1, N/K))}$ **:**

For this regime, consider the lower bound from Theorem 11. Setting $s = \lfloor \mu_1 \min\{N, K\} \rfloor$, with $\mu_1 \leq 1$, we have

$$\begin{aligned} R^*_{\mathsf{het}}(\mathcal{M}, N, K) &\geq \frac{N - \sum_{i=1}^{s} M_i}{\frac{N}{s} + 1} \\ &\geq \frac{N - sM_H}{\frac{N}{s} + 1} = \frac{N - \lfloor \mu_1 \min\{N, K\} \rfloor M_H}{\frac{N}{\lfloor \mu_1 \min\{N,K\} \rfloor} + 1} \\ &\geq \frac{N - \mu_1 \min\{N, K\} M_H}{\frac{N}{\mu_1 \min\{N,K\} - 1} + 1} \geq \frac{(\mu_1 \min\{N, K\} - 1)\left[1 - \mu_1 \frac{\min\{N,K\} \max\{1,N/K\}}{N}\right]}{1 + \mu_1 \frac{\min\{N,K\}}{N} - \frac{1}{N}} \\ &\geq \min\{N, K\} \frac{(\mu_1 - \frac{1}{C+1})[1 - \mu_1 \lambda_1]}{1 + \mu_1} \end{aligned} \tag{A.138}$$

Next, we consider the upper bound on the LHC rate. Consider a scheme where we have $G = K$ subsets of caches and all $\vec{\alpha}_g = 0, \ \forall g \in \{1, 2, \ldots, G\}$. Thus no storage is used in any cache and the achievable rate can be upper bounded by

$$R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M}, N, K) \leq \min\{N, K\}. \tag{A.139}$$

Thus combining (A.139) and (A.138), we have

$$\frac{R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right)}{R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right)} \leq \frac{1+\mu_1}{\left(\mu_1-\frac{1}{C+1}\right)\left[1-\mu_1\lambda_1\right]} \leq C \tag{A.140}$$

- ***Regime 2*** $\boldsymbol{(\lambda_1 \max(1, N/K) < M_H \leq \lambda_2 N/\min\{N,\ell_2\})}$ **:**

  For this regime, consider the lower bound from Theorem 11. Setting $s = \lfloor \mu_2 \frac{N}{M_H} \rfloor$, we have

$$\begin{aligned}
R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right) &\geq \frac{N-\sum_{i=1}^{s} M_i}{\frac{N}{s}+1} \geq \frac{N-sM_H}{\frac{N}{s}+1} \\
&= \frac{N-\lfloor \mu_2 \frac{N}{M_H} \rfloor M_H}{\frac{N}{\lfloor \mu_2 \frac{N}{M_H} \rfloor}+1} \geq \frac{N-\mu_2 \frac{N}{M_H} M_H}{\frac{N}{\mu_2 \frac{N}{M_H}-1}+1} \\
&\geq \frac{(1-\mu_2)\left[\mu_2 \frac{N}{M_H}-1\right]}{1+\frac{\mu_2}{M_H}-\frac{1}{N}} \geq \frac{N}{M_H}\frac{(1-\mu_2)\left[\mu_2-\frac{M_H}{N}\right]}{1+\frac{\mu_2}{\lambda_1}} \\
&\geq \frac{N}{M_H}\frac{(1-\mu_2)\left[\mu_2-\frac{\lambda_2}{\min\{N,\ell_2\}}\right]}{1+\frac{\mu_2}{\lambda_1}} \overset{\text{(a)}}{\geq} \frac{N}{M_H}\frac{(1-\mu_2)\left[\mu_2-\lambda_2\right]}{1+\frac{\mu_2}{\lambda_1}},
\end{aligned} \tag{A.141}$$

  where step (a) follows from the fact that $\min\{N,\ell_2\} \geq 1$. Next, we consider the LHC scheme shown in Fig. A.2for the three-level heterogeneous system. The different multicast and unicast tranmissions for this scheme are as follows:

  – *Multicast:* A fraction $\alpha_1 = M_L/M_H$ of requested files is delivered to all K users using the storage M_L. Another fraction $\alpha_2 = (M_I - M_L)/M_H$ of requested files are delivered to users $\ell_1+1, \ell_1+2, \ldots, K$. Finally a fraction $\alpha_3 = (M_H - M_I)/M_H$ is delivered to users $\ell_2+1, \ell_2+2, \ldots, K$. Note that $\alpha_1+\alpha_2+\alpha_3 = 1$.

  – *Unicast:* The remaining fraction $(1-\alpha_1)$ of files is unicast to users $1, 2, \ldots, \ell_1$ with a rate of $\min\{N,\ell_1\}(1-\alpha_1)$ while the remaining $(1-\alpha_1-\alpha_2)$ fraction of files for users $\ell_1+1, \ldots, \ell_2$ are unicast with a rate $\min\{N, \ell_2-\ell_1\}(1-\alpha_1-\alpha_2)$.

  The total achievable rate for this LHC scheme is given by:

$$\begin{aligned}
&R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right) \leq \alpha_1 R_{\mathsf{hom}}^{m}\left(\frac{M_L}{\alpha_1},N,K\right) + \alpha_2 R_{\mathsf{hom}}^{m}\left(\frac{M_I-M_L}{\alpha_2},N,K-\ell_1+1\right) \\
&\quad + \alpha_3 R_{\mathsf{hom}}^{m}\left(\frac{M_H-M_I}{\alpha_3},N,K-\ell_2+1\right) + \min\{N,\ell_1\}(1-\alpha_1) + \min\{N,\ell_2-\ell_1\}(1-\alpha_1-\alpha_2) \\
&\overset{\text{(a)}}{\leq} \alpha_1^2\frac{N}{M_L} + \alpha_2^2\frac{N}{M_I-M_L} + \alpha_3^2\frac{N}{M_H-M_I} + \min\{N,\ell_2\} \\
&\overset{\text{(b)}}{=} \frac{N}{M_H} + \min\{N,\ell_2\} \leq \frac{N}{M_H} + \lambda_2\frac{N}{M_H} = \frac{N}{M_H}\left(1+\lambda_2\right),
\end{aligned} \tag{A.142}$$

where step (a) follows by the fact that $R^m_{\mathsf{hom}}(M,N,K) \leq \frac{N}{M}$ for any N,K [97] and that the unicast rate can be upper bounded by unicasting $\min\{N,\ell_2\}$ files; step (b) follows from the definition of α. Thus, combining (A.142) and (A.141), we have

$$\frac{R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M},N,K)}{R^*_{\mathsf{het}}(\mathcal{M},N,K)} \leq \frac{(1+\lambda_2)\left[1+\frac{\mu_2}{\lambda_1}\right]}{(\mu_2-\lambda_2)(1-\mu_2)} \leq C. \tag{A.143}$$

- ***Regime 3*** $(\boldsymbol{\lambda_2 N/\min\{N,\ell_2\} < M_H \leq N})$ **:**

  For this regime, we consider the following sub-regimes which cover the entire valid regime of cache storage namely (i) $M_I \leq \lambda_2 N/\min\{N,\ell_2\} < M_H \leq N$, (ii) $\lambda_2 N/\min\{N,\ell_2\} \leq M_I \leq \lambda_3 N/\min\{N,\ell_1\}$, (iii) $M_L \leq \lambda_3 N/\min\{N,\ell_1\} < M_I \leq N$, (iv) $\lambda_3 N/\min\{N,\ell_1\} < M_L \leq \lambda_4 N$ and (v) $\lambda_4 N \leq M_L \leq N$. We next discuss each of these regimes separately.

  – ***Sub-Regime 1*** $(\boldsymbol{M_I \leq \lambda_2 N/\min\{N,\ell_2\} < M_H \leq N})$ **:**

  For this regime, we set $s = \lfloor \mu_3 \min\{N,\ell_2\} \rfloor$ for the lower bound in Theorem 11 with $\mu_3 \leq 1$. Since we have $s \leq \ell_2$ the lower bound takes the following form:

$$\begin{aligned} R^*_{\mathsf{het}}(\mathcal{M},N,K) &\geq \frac{N-\sum_{i=1}^{s} M_i}{\frac{N}{s}+1} \geq \frac{N-sM_I}{\frac{N}{s}+1} = \frac{N-\lfloor \mu_3 \min\{N,\ell_2\}\rfloor M_I}{\frac{N}{\lfloor \mu_3 \min\{N,\ell_2\}\rfloor}+1} \\ &\geq \frac{N-\mu_3 \min\{N,\ell_2\} M_I}{\frac{N}{\mu_3 \min\{N,\ell_2\}-1}+1} \geq \frac{(\mu_3 \min\{N,\ell_2\}-1)\left[1-\mu_3 \frac{\min\{N,\ell_2\}}{N} \cdot \frac{\lambda_2 N}{\min\{N,\ell_2\}}\right]}{1+\mu_3\frac{\min\{N,\ell_2\}}{N}-\frac{1}{N}} \\ &\geq \min\{N,\ell_2\}\frac{\left(\mu_3-\frac{1}{\min\{N,\ell_2\}}\right)[1-\mu_3\lambda_2]}{1+\mu_3} \overset{(a)}{\geq} \min\{N,\ell_2\}\frac{\left(\mu_3-\frac{1}{2}\right)[1-\mu_3\lambda_2]}{1+\mu_3}, \end{aligned} \tag{A.144}$$

  where in step (a), we assume that $\min\{N,\ell_2\} \geq 2$ which holds true for the three-level system[5]. Next, again consider the LHC scheme from Fig. A.2 and following the same steps as in (A.142), we have

$$\begin{aligned} R^{\mathsf{LHC}}_{\mathsf{het}}(\mathcal{M},N,K) &\leq \frac{N}{M_H} + \min\{N,\ell_2\} \overset{(a)}{\leq} \frac{\min\{N,\ell_2\}}{\lambda_2} + \min\{N,\ell_2\} \\ &= \min\{N,\ell_2\}\left(\frac{1}{\lambda_2}+1\right) \end{aligned} \tag{A.145}$$

[5]Note that for $N=1$, we have from Theorem 11, $R^*_{\mathsf{het}(\mathcal{M},1,K)} \geq 1-M_L$. Again, since there is only one file in the system, the users can request only that file and thus the LHC scheme is reduced to a heterogeneous unicast with the minimum cache storage i.e., $R^{\mathsf{LHC}}_{\mathsf{het}(\mathcal{M},1,K)} = 1-M_L$. Thus for $N=1$, the LHC scheme (which reduces to the heterogeneous unicast scheme) is optimal.

where, step (a) follows from the bounds due to the regime under consideration. Combining the bounds in (A.145) and (A.144), we have

$$\frac{R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right)}{R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right)}\leq\frac{\left(1+\mu_3\right)\left[\frac{1}{\lambda_2}+1\right]}{\left(\mu_3-\frac{1}{2}\right)\left[1-\mu_3\lambda_2\right]}\leq C. \tag{A.146}$$

– ***Sub-Regime 2 ($\boldsymbol{\lambda_2 N/\min\{N,\ell_2\}\leq M_I\leq\lambda_3 N/\min\{N,\ell_1\}}$) :*** For this regime, consider the lower bound from Theorem 11. Setting $s=\lfloor\mu_4\frac{N}{M_I}\rfloor$ with $\mu_4\leq\lambda_2$ ensures we have $s\leq\min\{N,\ell_2\}\leq\ell_2$. Therefore we have

$$\begin{aligned}
R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right)&\geq\frac{N-\sum_{i=1}^{s}M_i}{\frac{N}{s}+1}\geq\frac{N-sM_I}{\frac{N}{s}+1}\\
&=\frac{N-\lfloor\mu_4\frac{N}{M_I}\rfloor M_I}{\frac{N}{\lfloor\mu_4\frac{N}{M_I}\rfloor}+1}\geq\frac{N-\mu_4\frac{N}{M_I}M_I}{\frac{N}{\mu_4\frac{N}{M_I}-1}+1}\\
&\geq\frac{\left(1-\mu_4\right)\left[\mu_4\frac{N}{M_I}-1\right]}{1+\frac{\mu_4}{M_I}-\frac{1}{N}}\geq\frac{N}{M_I}\frac{\left(1-\mu_4\right)\left[\mu_4-\frac{M_I}{N}\right]}{1+\frac{\mu_4\min\{N,\ell_2\}}{\lambda_2 N}}\\
&\geq\frac{N}{M_I}\frac{\left(1-\mu_4\right)\left[\mu_4-\frac{\lambda_3}{\min\{N,\ell_1\}}\right]}{1+\frac{\mu_4}{\lambda_2}}\overset{\text{(a)}}{\geq}\frac{N}{M_I}\frac{\left(1-\mu_4\right)\left[\mu_4-\lambda_3\right]}{1+\frac{\mu_4}{\lambda_2}},
\end{aligned} \tag{A.147}$$

where step (a) follows from the fact that $\min\{N,\ell_1\}\geq 1$. Next, consider the system from Fig. A.2 and the following LHC scheme. We use cache storage M_L in all K users to multicast a fraction $\alpha=M_L/M_I$ of the requested files. Again for users $\ell_1+1,\ell_1+2,\ldots,K$, we use cache storage M_I-M_L to multicast the remaining $(1-\alpha)=(M_I-M_L)/M_I$ fraction of the requested files. The $(1-\alpha)$ fraction of requested content of the first ℓ_1 users are unicast with a rate $\min\{N,\ell_1\}(1-\alpha)$. Thus the rate of this LHC scheme is given by

$$\begin{aligned}
&R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right)\\
&\leq\alpha R_{\mathsf{hom}}^{m}\left(\frac{M_L}{\alpha},N,K\right)+(1-\alpha)R_{\mathsf{hom}}^{m}\left(\frac{M_I-M_L}{1-\alpha},N,K-\ell+1\right)+\min\{N,\ell_1\}(1-\alpha)\\
&\overset{\text{(a)}}{\leq}\alpha^2\frac{N}{M_L}+(1-\alpha)^2\frac{N}{M_I-M_L}+\min\{N,\ell_1\}\\
&\overset{\text{(b)}}{\leq}\alpha\frac{N}{M_I}+(1-\alpha)\frac{N}{M_I}+\min\{N,\ell_1\}=\frac{N}{M_I}+\min\{N,\ell_1\}\\
&\overset{\text{(c)}}{\leq}\frac{N}{M_I}+\lambda_3\frac{N}{M_I}=\frac{N}{M_I}\left(1+\lambda_3\right)
\end{aligned} \tag{A.148}$$

where step (a) follows by the fact that $R_{\mathsf{hom}}^{m}(M,N,K)\leq\frac{N}{M}$ for any N,K [97]; step (b) follows from the definition of α and step (c) follows from the choice of the regime of interest.

Combining (A.147) and (A.148), we have

$$\frac{R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right)}{R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right)} \leq \frac{(1+\lambda_3)\left[1+\frac{\mu_4}{\lambda_2}\right]}{(\mu_4-\lambda_3)(1-\mu_4)} \leq C. \tag{A.149}$$

– ***Sub-Regime 3*** $\boldsymbol{(M_L \leq \lambda_3 N/\min\{N,\ell_1\} < M_I \leq N)}$ **:**
For this regime, we set $s = \lfloor \mu_5 \min\{N,\ell_1\} \rfloor$ for the lower bound in Theorem 11 with $\mu_5 \leq 1$. Since we have $s \leq \ell_1$ the lower bound takes the following form:

$$\begin{aligned} R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right) &\geq \frac{N-\sum_{i=1}^{s} M_i}{\frac{N}{s}+1} \geq \frac{N-sM_L}{\frac{N}{s}+1} = \frac{N-\lfloor \mu_5 \min\{N,\ell_1\}\rfloor M_L}{\frac{N}{\lfloor \mu_5 \min\{N,\ell_1\}\rfloor}+1} \\ &\geq \frac{N-\mu_5 \min\{N,\ell_1\} M_L}{\frac{N}{\mu_5 \min\{N,\ell_1\}-1}+1} \geq \frac{(\mu_5 \min\{N,\ell_1\}-1)\left[1-\mu_5 \frac{\min\{N,\ell_1\}}{N}\cdot\frac{\lambda_3 N}{\min\{N,\ell_1\}}\right]}{1+\mu_5\frac{\min\{N,\ell_1\}}{N}-\frac{1}{N}} \\ &\geq \min\{N,\ell_1\}\frac{\left(\mu_5-\frac{1}{\min\{N,\ell_1\}}\right)\left[1-\mu_5\lambda_3\right]}{1+\mu_5} \overset{\text{(a)}}{\geq} \min\{N,\ell_1\}\frac{\left(\mu_5-\frac{1}{2}\right)\left[1-\mu_5\lambda_3\right]}{1+\mu_5}, \end{aligned} \tag{A.150}$$

where in step (a), we assume that $\min\{N,\ell_1\} \geq 2$ which holds true for the three-level system. Next, again consider the LHC scheme *Sub-Regime 2* and following the same steps as in (A.148), we have

$$\begin{aligned} R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right) &\leq \frac{N}{M_I} + \min\{N,\ell_1\} \overset{\text{(a)}}{\leq} \frac{\min\{N,\ell_1\}}{\lambda_3} + \min\{N,\ell_1\} \\ &= \min\{N,\ell_1\}\left(\frac{1}{\lambda_3}+1\right) \end{aligned} \tag{A.151}$$

where, step (a) follows from the bounds due to the regime under consideration. Combining the bounds in (A.151) and (A.150), we have

$$\frac{R_{\mathsf{het}}^{\mathsf{LHC}}\left(\mathcal{M},N,K\right)}{R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right)} \leq \frac{(1+\mu_5)\left[\frac{1}{\lambda_3}+1\right]}{\left(\mu_5-\frac{1}{2}\right)\left[1-\mu_5\lambda_3\right]} \leq C. \tag{A.152}$$

– ***Sub-Regime 4*** $\boldsymbol{(\lambda_3 N/\min\{N,\ell_1\} < M_L \leq \lambda_4 N)}$ **:**
For this regime we set $s = \left\lfloor \mu_6 \frac{N}{M_L} \right\rfloor$. We want to ensure that the largest value that this s can take is less than ℓ_1. In the regime of interest, this can be ensured by

$$\mu_6 \frac{N}{M_L} \leq \mu_6 \frac{N}{\lambda_3 N}\min\{N,\ell_1\} \leq \frac{\mu_6}{\lambda_3}\min\{N,\ell_1\} \leq \ell_1 \;\; \Rightarrow \;\; \mu_6 \leq \lambda_3$$

Since, $s \leq \ell_1$, the lower bound takes the following form:

$$R_{\mathsf{het}}^{*}\left(\mathcal{M},N,K\right) \geq \frac{N-\sum_{i=1}^{s} M_i}{\frac{N}{s}+1} \geq \frac{N-sM_L}{\frac{N}{s}+1}$$

$$= \frac{N - \lfloor \mu_6 \frac{N}{M_L} \rfloor M_L}{\frac{N}{\lfloor \mu_6 \frac{N}{M_L} \rfloor} + 1} \geq \frac{N - \mu_6 \frac{N}{M_L} M_L}{\frac{N}{\mu_6 \frac{N}{M_L} - 1} + 1}$$
$$\geq \frac{(1-\mu_6)\left[\mu_6 \frac{N}{M_L} - 1\right]}{1 + \frac{\mu_6}{M_L} - \frac{1}{N}} \geq \frac{N}{M_L} \frac{(1-\mu_6)\left[\mu_6 - \frac{M_L}{N}\right]}{1 + \frac{\mu_6 \min\{N, \ell_1\}}{\lambda_3 N}}$$
$$\geq \frac{N}{M_L} \frac{(1-\mu_6)\left[\mu_6 - \lambda_4\right]}{1 + \frac{\mu_6}{\lambda_3}}, \tag{A.153}$$

Next, consider the upper bound on the LHC scheme. Consider that every user has only a cache storage of M_L. Under this assumption, the following upper bound on the rate of the LHC scheme holds:

$$R_{\mathsf{het}}^{\mathsf{LHC}}(\mathcal{M}, N, K) \leq \frac{K\left(1 - \frac{M_L}{N}\right)}{1 + \frac{KM_L}{N}} \leq \frac{N}{M_L}\left(1 - \frac{M_L}{N}\right) \leq \frac{N}{M_L}. \tag{A.154}$$

Combining (A.154) and (A.153), we have

$$\frac{R_{\mathsf{het}}^{\mathsf{LHC}}(\mathcal{M}, N, K)}{R_{\mathsf{het}}^{*}(\mathcal{M}, N, K)} \leq \frac{1 + \frac{\mu_6}{\lambda_2}}{(1-\mu_6)\left[\mu_6 - \lambda_3\right]} \leq C. \tag{A.155}$$

– ***Sub-Regime 5*** $(\boldsymbol{\lambda_4 N < M_L \leq N})$ **:**
In this regime, setting $s = 1$ in the lower bound from Theorem 11, we have

$$R_{\mathsf{het}}^{*}(\mathcal{M}, N, K) \geq 1 - \frac{M_L}{N} \tag{A.156}$$

Next consider the upper bound on the LHC scheme with each user having storage M_L,

$$R_{\mathsf{het}}^{\mathsf{LHC}}(\mathcal{M}, N, K) \leq \frac{K\left(1 - \frac{M_L}{N}\right)}{1 + \frac{KM_L}{N}} \leq \frac{N}{M_L}\left(1 - \frac{M_L}{N}\right). \tag{A.157}$$

Combining (A.157) and (A.156), we have

$$\frac{R_{\mathsf{het}}^{\mathsf{LHC}}(\mathcal{M}, N, K)}{R_{\mathsf{het}}^{*}(\mathcal{M}, N, K)} \leq \frac{N}{M_L} \leq \frac{1}{\lambda_4} \leq C. \tag{A.158}$$

Finding values for the parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ which satisfy the inequalities in (A.140), (A.143), (A.146), (A.149), (A.152), (A.155) and (A.158) for the minimum possible value of the constant C leads to a constant multiplicative gap of $C = 28$ with $\lambda_1 = 1.2, \lambda_2 = 0.3950, \lambda_3 = 0.2410, \lambda_4 = 0.0410, \mu_1 = 0.0769, \mu_2 = 0.5660, \mu_3 = 0.8600, \mu_4 = 0.3920, \mu_5 = 0.9810, \mu_6 = 0.1020$, which further satisfies $\mu_4 \leq \lambda_2$ and $\mu_6 \leq \lambda_3$. This completes the proof of Theorem 12.

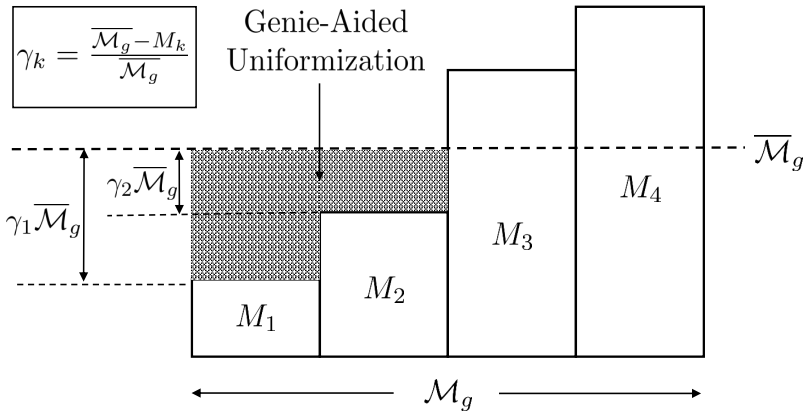


Figure A.3: Genie aided uniformization of the cache sets for a bicriteria approximation.

A.10 Proof of Theorem 13

For the proof we use a *genie-aided uniformization* argument. Consider that at the beginning of each placement phase, a central server determines the optimal partitioning of the cache set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_G\}$. For each set of caches $\mathcal{M}_g,\ \forall g \in \{1, 2, \ldots, G\}$, let the mean cache storage i.e., $\mathbb{E}_{K_g}[\mathcal{M}_g] = \overline{\mathcal{M}_g}$. Now we consider that in every set of caches $\mathcal{M}_g$, a genie provides every user k, having cache storage $M_k \leq \overline{\mathcal{M}_g}$, with additional cache storage of $\overline{\mathcal{M}_g} - M_k$. This is illustrated in Fig. A.3 for an arbitrary set of 4 caches. For any cache set $\mathcal{M}_g$, let $\mathfrak{M}_g$ denote the new genie-aided uniformized cache set. Also, let the mean cache storage of the entire set of caches $\mathcal{M}$ be $\overline{\mathcal{M}} = \mathbb{E}_K[\mathcal{M}]$.

Proof for Equation (3.61): We start by considering the minimum rate of the proposed LHC scheme given in Theorem 10, evaluated over the uniformized cache sets $\mathfrak{M}_g$ and the optimal file splitting strategy $\vec{\alpha}^*$. For any arbitrary partitioning and sub-optimal choice of $\vec{\alpha}$, the sum-rate of the genie-aided scheme will only increase. Using this notion, the following series of inequalities proves the theorem.

$$\begin{aligned}
\sum_{g=1}^{G} R_{\mathsf{het}}\left(\mathfrak{M}_g, N, K_g\right) &\overset{(a)}{=} \sum_{g=1}^{G} R_{\mathsf{het}}\left(\mathfrak{M}_g, N, K_g, \vec{\alpha}^*\right), \\
&\overset{(b)}{\leq} \sum_{g=1}^{G} R_{\mathsf{het}}\left(\mathfrak{M}_g, N, K_g, \vec{\alpha}\right), \quad \text{s.t. } \alpha_1 = 1, \alpha_i = 0 \ \ \forall i \neq 1 \\
&\overset{(c)}{=} \sum_{g=1}^{G} R_{\mathsf{hom}}\left(\left[\min_{M \in \mathfrak{M}_g} \mathfrak{M}_g\right], N, K_g\right), \\
&\overset{(d)}{=} \sum_{g=1}^{G} R_{\mathsf{hom}}\Big(\overline{\mathcal{M}_g}, N, K_g\Big), \quad \text{where } \overline{\mathcal{M}_g} = \mathbb{E}_{K_g}\big[\mathcal{M}_g\big],
\end{aligned}$$

$$\begin{aligned}
&\overset{(e)}{\leq} \sum_{g=1}^{G} 12R^{\mathsf{lb}}_{\mathsf{hom}}\left(\overline{\mathcal{M}}_g, N, K_g\right) \\
&\overset{(f)}{\leq} 12GR^{\mathsf{lb}}_{\mathsf{hom}}\left(\overline{\mathcal{M}}, N, K\right) \overset{(g)}{\leq} 12GR^{\mathsf{lb}}_{\mathsf{het}}\left(\mathcal{M}, N, K\right) \\
&\overset{(h)}{\leq} 12GR^{*}_{\mathsf{het}}\left(\mathcal{M}, N, K\right),
\end{aligned} \tag{A.159}$$

which proves the theorem. We next elaborate on each inequality in (A.159). The equality (a) follows from the fact that the achievable rate of the LHC scheme over the partition of G groups is given by the optimal file splitting strategy $\vec{\alpha}^*$. The inequality (b) follows from the fact that choosing a sub-optimal file splitting strategy $\vec{\alpha}$ can only decrease the rate of the LHC scheme. The equality (c) follows from the observation that the LHC scheme with $\vec{\alpha} : \{\alpha_1 = 1, \alpha_i = 0 \ \ \forall i \neq 1\}$ reduces to the homogeneous scheme of [97] with cache storage equal to the minimum cache storage in each group. The equality (d) follows from the genie-aided uniformization argument which ensures that the minimum cache storage in each group becomes $\overline{\mathcal{M}}_g$ which is the mean cache size for the original ordered group $\mathcal{M}_g$. The inequality (e) follows from the sub-optimality gap of the homogeneous caching schemes in [97, Theorem 3]. The inequality (f) can be proved as follows. Consider the lower bound on the homogeneous caching scheme employed in each group $g \in G$. We have

$$\begin{aligned}
\sum_{g=1}^{G} R^{\mathsf{lb}}_{\mathsf{hom}}\left(\overline{\mathcal{M}}_g, N, K_g\right) &= \sum_{g=1}^{G} \max_{s\in\{1,\ldots,\min N, K_g\}} \left(s - \frac{s\overline{\mathcal{M}}_g}{\lfloor N/s \rfloor}\right) \overset{(i)}{=} \sum_{g=1}^{G} \left(s^* - \frac{s^*\overline{\mathcal{M}}_g}{\lfloor N/s^* \rfloor}\right) \\
&= Gs^* - \frac{s^* \sum_{g=1}^{G} \overline{\mathcal{M}}_g}{\lfloor N/s^* \rfloor} \overset{(ii)}{=} Gs^* - \frac{s^* G\overline{\mathcal{M}}}{\lfloor N/s^* \rfloor} \\
&= G\left(s^* - \frac{s^*\overline{\mathcal{M}}}{\lfloor N/s^* \rfloor}\right) \overset{(iii)}{\leq} G \max_{s\in\{1,\ldots,\min N, K\}} \left(s - \frac{s\overline{\mathcal{M}}}{\lfloor N/s \rfloor}\right) \\
&= GR^{\mathsf{lb}}_{\mathsf{hom}}\left(\overline{\mathcal{M}}, N, K\right).
\end{aligned} \tag{A.160}$$

In (i), the optimal choice of s i.e., s^* is selected. The equality (ii) follows from the fact that $\sum_{g=1}^{G} \overline{\mathcal{M}}_g = G\overline{\mathcal{M}} = M_1 + M_2 + \ldots + M_K$ i.e., both are equal to the total cache storage in the system. The inequality (iii) follows from the fact that s^* is a valid choice of s for the domain $s \in \{1, \ldots, \min\{N, K\}\}$ since $K_g \leq K$. As a result, maximizing over all possible choices of s, upper-bounds the LHS in (iii). Again returning to (A.159), the inequality (g) can be proved as follows. Consider the fact that $s\overline{\mathcal{M}} \geq \sum_{i=1}^{s} M_i$ such that $M_i \in \mathcal{M}$ i.e., the sum of the smallest s caches is upper bounded by s times the mean cache, then we have

$$\max_{s\in 1,\ldots,\min\{N,K\}} \left(s - \frac{s\overline{\mathcal{M}}}{\lfloor N/s \rfloor}\right) \leq \max_{\substack{s\in 1,\ldots,\min\{N,K\} \\ M_i\in\mathcal{M},\forall i\in[s]}} \left(s - \frac{\sum_{i=1}^{s} M_i}{\lfloor N/s \rfloor}\right) \tag{A.161}$$

$$\Rightarrow \quad R^{\mathsf{lb}}_{\mathsf{hom}}\left(\overline{\mathcal{M}}, N, K\right) \leq R^{\mathsf{lb}}_{\mathsf{het}}\left(\mathcal{M}, N, K\right). \tag{A.162}$$

The final inequality follows from the definition of the lower bound on the optimal rate. This completes the proof of (A.159).

Violation of Cache Storage Constraint: The approximation provided in (3.61) is a bicriteria approximation. We next analyse the cache storage violation incurred by the genie-aided uniformization argument presented in the proof of (3.61). In order to characterize the approximation ratio, for any user k with cache storage $M_k \in \mathcal{M}_g$, we define the ratio of the additional cache storage supplied by the genie to the mean cache in each group as

$$\gamma_k = \left(\frac{\overline{\mathcal{M}}_g - M_k}{\overline{\mathcal{M}}_g} \right)^+, \tag{A.163}$$

where the $(x)^+ = \max\{0, x\}$ ensures that the ratio is non-zero for only the caches which have storage lower than the mean. Assume further, that in each group a total of $\overline{K}_g$ out of K_g users have cache storage less than the mean i.e., the genie helps these $\overline{K}_g$ users with additional storage as shown in Fig. A.3. Then the total cache storage across the genie-aided cache set $\mathfrak{M}_g$ can be written as

$$\underbrace{K \cdot \overline{\mathcal{M}}}_{\text{Original Cache Storage}} + \underbrace{\sum_{g=1}^{G} \sum_{k=1}^{\overline{K}_g} \overline{\mathcal{M}}_g \gamma_k}_{\text{Additional Storage provided by Genie}} = K\overline{\mathcal{M}} + \sum_{g=1}^{G} \overline{\mathcal{M}}_g \sum_{k=1}^{\overline{K}_g} \gamma_k. \tag{A.164}$$

Thus the ratio of the new sum cache storage to the original sum storage is given by

$$\frac{K\overline{\mathcal{M}} + \sum_{g=1}^{G} \overline{\mathcal{M}}_g \sum_{k=1}^{\overline{K}_g} \gamma_k}{K\overline{\mathcal{M}}} = 1 + \frac{\sum_{g=1}^{G} \overline{\mathcal{M}}_g \sum_{k=1}^{\overline{K}_g} \gamma_k}{\frac{K}{G} \sum_{g=1}^{G} \overline{\mathcal{M}}_g} = 1 + \frac{G}{K} \sum_{g=1}^{G} \sum_{k=1}^{\overline{K}_g} \gamma_k \tag{A.165}$$

From (A.165), it should be noted that when every user has equal cache storage i.e., the homogeneous case, then all the $\gamma_k = 0$ such that the second term vanishes and the ratio is 1 and there is no cache violation. In this case, there is no need to use $G > 1$. Thus for $\gamma_i = 0,\ \forall i$ and $G = 1$, Theorem 13 reduces to [97, Theorem 3]. Thus the approximation ratio in Theorem 13 is a $\left(12G, \left[1 + \frac{G}{K} \sum_{g=1}^{G} \sum_{k=1}^{\overline{K}_g} \gamma_k\right]\right)-$ bi-criteria approximation.

Remark 23 (*Performance of LHC*)**.** Consider the fact that the bi-criteria approximation was derived using a file splitting strategy such that $\vec{\alpha} : \{\alpha_1 = 1, \alpha_i = 0\ \forall i \neq 1\}$. This is a naive choice of $\vec{\alpha}$ and minimization over all possible $\vec{\alpha}$ always improves the rate when cache sizes are not uniform. Thus in practice, the proposed LHC algorithm provides an achievable rate which is much tighter than the approximation without violation of any cache storage constraint. □

Appendix B

Proofs From Chapter 4

Fundamental Limits of Caching with Secure Delivery

B.1 Proof of Theorem 14

In this section, we discuss the secure centralized caching strategy which achieves the upper bound $R_{s,\mathsf{cen}}(M)$ as stated in Theorem 14. The algorithm achieving the rate in Theorem 14 is presented in Algorithm 2. These are two phases in the caching strategy: the storage phase and the delivery phase. We consider a cache size $M \leq N$ and $M \in \frac{N-1}{K} \cdot \{0, 1. \ldots, K\} + 1$. Let $t \in \{0, 1, \ldots, K\}$ be an integer between 0 and K. The cache storage size can then be parametrized by t as:

$$M = \frac{N-1}{K}t + 1 = \frac{Nt}{K} + 1 - \frac{t}{K}. \tag{B.1}$$

From (B.1), we have $t = \frac{K(M-1)}{N-1}$. Next, we break up the total cache storage into data storage and key storage, $M = M_D + M_K$, as follows:

$$M_K = 1 - \frac{t}{K}; \;\; M_D = M - M_K = \frac{Nt}{K}. \tag{B.2}$$

From the discussion in Section 4.3, we know that the *conventional secure scheme* achieves the $(M, R_{s,\mathsf{cen}})$ pair $(1, K)$ and $(N, 0)$. Thus $R_s^*(1) \leq K$ and $R_s^*(N) = 0$. We therefore consider the case in which $1 < M < N$. In this case, $t \in \{1, 2, \ldots, K-1\}$.

Storage Phase: In the placement phase, each file F_n for $n = 1, \ldots, N$ is split into $\binom{K}{t}$ non-overlapping sub-files of equal size $B/\binom{K}{t}$:

$$F_n = (F_{n,\tau} : \tau \subseteq \{1, \ldots, K\}, |\tau| = t). \tag{B.3}$$

For each n, the sub-file $F_{n,\tau}$ is placed the cache of user k if $k \in \tau$. Since $|\tau| = t$, for each user $k \in \tau$, there are $t-1$ out of $K-1$ possible users with whom it shares a sub-file of a given file

F_n. Thus each user caches $N\binom{K-1}{t-1}$ sub-files. Next we generate a set of keys, each of the size of a sub-file i.e. of size $B/\binom{K}{t}$:

$$(\mathcal{K}_{\tau_k} : \tau_k \subseteq \{1,\ldots,K\}, |\tau_k| = t+1). \tag{B.4}$$

The key $\mathcal{K}_{\tau_k}$ is placed in the cache of user k if $k \in \tau_k$. The keys are generated such that all the keys are orthogonal to each other and each key is distributed according to $\mathcal{K}_{\tau_k} \sim \text{unif}\left\{1,2,\ldots,2^{B/\binom{K}{t}}\right\}$. Again, since $|\tau_k| = t+1$, each user $k \in \tau_k$ shares key $\mathcal{K}_{\tau_k}$ with t out of $K-1$ possible users. Thus there are $\binom{K-1}{t}$ keys in the cache of each user. Given each key and sub-file has size $B/\binom{K}{t}$, number of bits required for storage at each user is:

$$N\binom{N-1}{t-1}\cdot\frac{B}{\binom{K}{t}} + \binom{K-1}{t}\cdot\frac{B}{\binom{K}{t}}$$

$$= \frac{BNt}{K} + B\left(1-\frac{t}{K}\right) = B\left(\frac{Nt}{k}+1-\frac{t}{K}\right) = BM \tag{B.5}$$

which satisfies the storage constraint.

Delivery Phase: We now elaborate on the delivery phase. Consider a request vector $(d_1,\ldots,d_k) \in \{1,\ldots,N^K\}$ where user k requests the file F_{d_k}. Let $\mathcal{S} \subseteq \{1,\ldots,K\}$ be a subset of $|\mathcal{S}| = t+1$ users. Every t users in $\mathcal{S}$ share a sub-file in their cache which is requested by the $t+1$-th user. Given a user $k \in \mathcal{S}$ and $|\mathcal{S}\setminus\{k\}| = t$, the sub-file $F_{d_k,\mathcal{S}\setminus\{k\}}$ is requested by user k as it is a sub-file of F_{d_k} which is missing at user k since $k \notin \mathcal{S}\setminus\{k\}$. The file is present in the cache of the t users $s \in \mathcal{S}\setminus\{k\}$. For each such subset $\mathcal{S} \subseteq \{1,\ldots,K\}$, the server sends the following transmission: $X_{(d_1,\ldots,d_k)} = \left\{\mathcal{K}_{\mathcal{S}} \oplus_{s\in\mathcal{S}} F_{d_s,\mathcal{S}\setminus\{s\}}\right\}$ such that $\{\mathcal{S} \subseteq \{1,2,\ldots,K\}, |\mathcal{S}| = t+1\}$. The number of subsets $\mathcal{S}$ is $\binom{K}{t+1}$. Thus there are $\binom{K}{t+1}$ transmissions and an unique key associated with each transmission i.e., there are $\binom{K}{t+1}$ keys in the system. Each transmission has the size of a subfile and thus the total number of bits sent over the rate-limited link is:

$$R_{s,\mathsf{cen}}(M)B = \binom{K}{t+1}\cdot\frac{B}{\binom{K}{t}} = \frac{K\left(1-\frac{M-1}{N-1}\right)}{1+\frac{K(M-1)}{N-1}}\cdot B$$

$$\Rightarrow R_s^*(M) \leq R_{s,\mathsf{cen}}(M) \triangleq \frac{K\left(1-\frac{M-1}{N-1}\right)}{1+\frac{K(M-1)}{N-1}}. \tag{B.6}$$

Next, we show that the delivery phase does not reveal any information to the wiretapper i.e., we show that

$$I\left(X_{(d_1,\ldots,d_k)}; F_1,\ldots,F_N\right) = 0 \tag{B.7}$$

We have,

$$I\left(X_{(d_1,\ldots,d_K)}; F_1,\ldots,F_N\right) = H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(X_{(d_1,\ldots,d_K)}|F_1,\ldots,F_N\right)$$

$$
\begin{aligned}
&= H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(\left\{\mathcal{K}_{\mathcal{S}} \oplus_{s\in\mathcal{S}} F_{d_s,\mathcal{S}\setminus\{s\}} : |\mathcal{S}| = t+1\right\} | F_1,\ldots,F_N\right) \\
&= H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(\{\mathcal{K}_{\mathcal{S}} : |\mathcal{S}| = t+1\} | F_1,\ldots,F_N\right) \\
&= H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(\{\mathcal{K}_{\mathcal{S}} : |\mathcal{S}| = t+1\}\right), \qquad \text{(B.8)}
\end{aligned}
$$

where, the last equality follows from the fact that the keys are uniformly distributed and are independent of the files $(F_1,\ldots,F_N)$. Using the fact that $H(A,B) \leq H(A) + H(B)$, we have:

$$
\begin{aligned}
H\left(X_{(d_1,\ldots,d_K)}\right) &= H\left(\left\{\mathcal{K}_{\mathcal{S}} \oplus_{s\in\mathcal{S}} F_{d_s,\mathcal{S}\setminus\{s\}} : |\mathcal{S}| = t+1\right\}\right) \\
&\leq \sum_{i=1}^{\binom{K}{t+1}} H\left(\mathcal{K}_{\mathcal{S}_i} \oplus_{s\in\mathcal{S}_i} F_{d_s,\mathcal{S}_i\setminus\{s\}} : |\mathcal{S}_i| = t+1\right) \\
&\leq \sum_{i=1}^{\binom{K}{t+1}} \log_2\left(\frac{B}{\binom{K}{t}}\right) = \binom{K}{t+1}\log_2\left(\frac{B}{\binom{K}{t}}\right). \qquad \text{(B.9)}
\end{aligned}
$$

On the other hand, we have:

$$
\begin{aligned}
H\left(\{\mathcal{K}_{\mathcal{S}} : |\mathcal{S}| = t+1\}\right) &= \sum_{i=1}^{\binom{K}{t+1}} H\left(\mathcal{K}_{\mathcal{S}_i} : |\mathcal{S}_i| = t+1\right) \\
&= \sum_{i=1}^{\binom{K}{t+1}} \log_2\left(\frac{B}{\binom{K}{t}}\right) = \binom{K}{t+1}\log_2\left(\frac{B}{\binom{K}{t}}\right), \qquad \text{(B.10)}
\end{aligned}
$$

where the equality in (B.10) follows from the fact that the keys $\mathcal{K}_{\mathcal{S}_i}$, for all i are mutually independent and distributed as unif $\left\{1,2,\ldots,2^{B/\binom{K}{t}}\right\}$. Substituting (B.9) and (B.10) into (B.8), we have:

$$
I\left(X_{(d_1,\ldots,d_K)}; F_1,\ldots,F_N\right) \leq 0. \qquad \text{(B.11)}
$$

Using the fact that for any X, Y, $I(X;Y) \geq 0$, we have:

$$
I\left(X_{(d_1,\ldots,d_K)}; F_1,\ldots,F_N\right) = 0, \qquad \text{(B.12)}
$$

which proves that the rate $R_{s,\mathsf{cen}}(M)$ is *securely* achievable. This completes the proof of Theorem 14. □

B.2 Proof of Theorem 15

In this section, we prove the information-theoretic lower bound on $R_s^*(M)$ for any $N, K \in \mathbb{N}$. Let s be an integer such that $s \in \{1,\ldots,\min\{N,K\}\}$. Consider the first s caches $Z_1,\ldots,Z_s$. For a request vector $(d_1,d_2,\ldots,d_s,d_{s+1},\ldots,d_K) = (1,2,\ldots,s,\phi,\ldots,\phi)$, the transmission $X_1 = X_{(d_1,\ldots,d_k)}$, along with the caches $Z_1,\ldots,Z_s$ must be able to decode the files $F_1,\ldots,F_s$. Similarly

there for another request $(d_1, d_2, \ldots, d_s, d_{s+1}, \ldots, d_K) = (s+1, s+2, \ldots, 2s, \phi, \ldots, \phi)$, the transmission X_2, which along with caches $Z_1, \ldots, Z_s$, must be able to decode the files $F_{s+1}, \ldots, F_{2s}$. Thus considering $\lfloor N/s \rfloor$ different requests, the transmissions from the central server denoted by $X_1, \ldots, X_{\lfloor N/s \rfloor}$, along with the caches $Z_1, \ldots, Z_s$, must be able to decode the files $F_1, \ldots, F_{s\lfloor N/s \rfloor}$. Let

$$\begin{aligned}
\widetilde{W} &= \left\{F_1, \ldots, F_{s\lfloor N/s \rfloor}\right\} \\
\widetilde{X} &= \left\{X_1, \ldots, X_{\lfloor N/s \rfloor}\right\} \\
\widetilde{X}_{\setminus\{l\}} &= \left\{X_1, \ldots, X_{l-1}, X_{l+1}, \ldots, X_{\lfloor N/s \rfloor}\right\} \\
\widetilde{Z} &= \left\{Z_1, \ldots, Z_s\right\}.
\end{aligned}$$

In addition, we also have constraints based on file retrieval and security. The file retrieval constraint is based on the fact that given all possible transmissions and caches, all files can can be retrieved. The security constraint is that a wiretapper should not be able to retrieve any information about the files from any transmission by the server. Using Definition 6, we have:

$$H(\widetilde{W}|\widetilde{X}, \widetilde{Z}) \leq \epsilon \tag{B.13}$$

$$I(\widetilde{W}; X_l) \leq \epsilon; \quad l = 1, \ldots, \lfloor N/s \rfloor \tag{B.14}$$

We present a novel extension to the cut-set bound argument [174] to include the security and file retrieval constraints. Consider the information flow consisting of transmissions $X_1, \ldots, X_{\lfloor N/s \rfloor}$ and caches $Z_1, \ldots, Z_s$ for decoding files $F_1, \ldots, F_{s\lfloor N/s \rfloor}$. This flow has minimum capacity $s\lfloor N/s \rfloor$. Thus, we have:

$$\begin{aligned}
s\lfloor N/s \rfloor B &\leq H(\widetilde{W}) = I(\widetilde{W}; \widetilde{X}, \widetilde{Z}) + H(\widetilde{W}|\widetilde{X}, \widetilde{Z}) \\
&\overset{(B.13)}{\leq} I(\widetilde{W}; \widetilde{X}, \widetilde{Z}) + \epsilon \\
&= I\left(\widetilde{W}; \{X_1, \ldots, X_{\lfloor N/s \rfloor}\}, \{Z_1, \ldots, Z_s\}\right) + \epsilon \\
&= I(\widetilde{W}; X_l) + I\left(\widetilde{W}; \widetilde{X}_{\setminus\{l\}}, \widetilde{Z}|X_l\right) + \epsilon \\
&\overset{(B.14)}{\leq} I\left(\widetilde{W}; \widetilde{X}_{\setminus\{l\}}, \widetilde{Z}|X_l\right) + 2\epsilon \\
&\leq H\left(\widetilde{X}_{\setminus\{l\}}, \widetilde{Z}\right) + 2\epsilon \\
&\leq \sum_{i=1, i\neq l}^{\lfloor N/s \rfloor} H(X_i) + \sum_{j=1}^{s} H(Z_j) + 2\epsilon \\
&\leq (\lfloor N/s \rfloor - 1)\, R_s^*(M) B + sMB + 2\epsilon \\
\Rightarrow s\lfloor N/s \rfloor &\leq (\lfloor N/s \rfloor - 1)\, R_s^*(M) + sM + \frac{2\epsilon}{B}.
\end{aligned} \tag{B.15}$$

Solving for R_s^* and optimizing over all possible s, we have:

$$R_s^*(M) \;\geq\; \max_{s \in \{1, \ldots, \min\{N,K\}\}} \lim_{\epsilon \to 0} \frac{s\lfloor N/s \rfloor - sM - \frac{2\epsilon}{B}}{\lfloor N/s \rfloor - 1}$$

$$= \max_{s\in\{1,\ldots,\min\{N,K\}\}} \left(s - \frac{s(M-1)}{\left(\lfloor \frac{N}{s} \rfloor - 1\right)} \right), \tag{B.16}$$

which concludes the proof of Theorem 15. □

B.3 Proof of Theorem 16

In this section, we prove that a constant multiplicative gap exists between the securely achievable rate $R_{s,\mathsf{cen}}(M)$ given in Theorem 14 and the optimal secure rate $R_s^*(M)$, for the regime

$$\max\left\{\frac{(K-N)(N-1)}{KN} + 1, 1\right\} \leq M \leq N. \tag{B.17}$$

We consider two cases for the value of K. Firstly, for $K \leq N$, we have from Theorem 14:

$$R_{s,\mathsf{cen}}(M) \leq K\left(1 - \frac{M-1}{N-1}\right) = \min\{N,K\}\left(1 - \frac{M-1}{N-1}\right). \tag{B.18}$$

For the case of $K > N$, (B.17) reduces to $(K-N)(N-1)/KN + 1 \leq M \leq N$. Thus we have:

$$\begin{aligned}
&\frac{(K-N)(N-1)}{KN} + 1 \leq M \\
\Rightarrow\ &\frac{1}{N} - \frac{1}{K} \leq \frac{M-1}{N-1} \ \Rightarrow K \cdot \frac{1}{1 + K\frac{M-1}{N-1}} \leq N \\
\Rightarrow\ &K\left(1 - \frac{M-1}{N-1}\right)\frac{1}{1 + K\frac{M-1}{N-1}} \leq N\left(1 - \frac{M-1}{N-1}\right) \\
\Rightarrow\ &R_{s,\mathsf{cen}}(M) \leq \min\{N,K\}\left(1 - \frac{M-1}{N-1}\right).
\end{aligned} \tag{B.19}$$

To prove the constant gap result, we focus on two cases namely (i) $\min\{N,K\} \leq 17$ and (ii) $\min\{N,K\} \geq 18$. We consider the two cases separately:

• ***Case 1*** $\boldsymbol{(\min\{N,K\} \leq 17)}$ **:** Setting $s = 1$ in Theorem 15 gives the following lower bound on the optimal secure rate:

$$R_s^*(M) \geq \left(1 - \frac{M-1}{N-1}\right). \tag{B.20}$$

Hence from (B.19) and (B.20), we have

$$\frac{R_{s,\mathsf{cen}}(M)}{R_s^*(M)} \leq \min\{N,K\} \leq 17. \tag{B.21}$$

• ***Case 2*** $\boldsymbol{(\min\{N,K\} \geq 18)}$ **:**

For this case, the rate in Theorem 14 has 3 distinct regimes namely (i) ***Regime 1:*** $\max\left\{\frac{(K-N)(N-1)}{KN}, 0\right\} \leq M-1 \leq 1.2\max\left(1, \frac{N-1}{K}\right)$; (ii) ***Regime 2:*** $1.2\max\left(1, \frac{N-1}{K}\right) < M-1 \leq 0.0628(N-1)$; and (iii) ***Regime 3:*** $0.0628(N-1) < M-1 \leq N-1$. We consider each of these regimes separately.

- ***Regime 1*** $\left(\max\left\{\frac{(K-N)(N-1)}{KN}, 0\right\} \leq M-1 \leq 1.2\max\left(1, \frac{N-1}{K}\right)\right)$**:**

  By Theorem 14, we have:

  $$R_{s,\mathsf{cen}}(M) \leq R_{s,\mathsf{cen}}(1) \leq \min\{N, K\}. \tag{B.22}$$

  By Theorem 15 and using the fact that $\lfloor N/s \rfloor \geq N/s - 1$, we have:

  $$R_s^*(M) \geq s - \frac{s^2(M-1)}{N-2s}. \tag{B.23}$$

  Setting $s = \lfloor 0.1586\min\{N, K\}\rfloor \in \{1, \ldots, \min\{N, K\}\}$ we get, for $M-1 \leq 1.2\max\left(1, \frac{N-1}{K}\right)$:

  $$\begin{aligned}
  R_s^*(M) &\geq R_s^*\left(1.2\max\left(1, \frac{N-1}{K}\right) + 1\right) \\
  &\geq 0.1586\min\{N, K\} - 1 - \frac{(0.1586\min\{N, K\})^2 \cdot 1.2\max\left(1, \frac{N-1}{K}\right)}{N - 2\cdot 0.1586\min\{N, K\}} \\
  &\geq \min\{N, K\}\left\{0.1586 - \frac{1}{\min\{N, K\}} - \frac{(0.1586)^2 \cdot 1.2}{1 - 2\cdot(0.1586)\min\{1, K/N\}}\right\} \\
  &\geq \min\{N, K\}\left\{0.1586 - \frac{1}{18} - \frac{1.2\cdot(0.1586)^2}{1 - 2\cdot 0.1586}\right\} \\
  &\geq \frac{1}{17}\min\{N, K\}.
  \end{aligned} \tag{B.24}$$

  Combining (B.22) and (B.24), we have:

  $$\frac{R_{s,\mathsf{cen}}(M)}{R_s^*(M)} \leq 17. \tag{B.25}$$

- ***Regime 2*** $\left(1.2\max\left(1, \frac{N-1}{K}\right) < M-1 \leq 0.0628(N-1)\right)$**:**

  Let $\bar{M}$ be the largest multiple of $\frac{N-1}{K}$ less than equal to M such that

  $$0 \leq M - \frac{N-1}{K} \leq \bar{M} \leq M. \tag{B.26}$$

  Choosing $\bar{M} = M - (N-1)/K$, and using the fact that $R_{s,\mathsf{cen}}(M)$ is monotonically decreasing in M, we have:

  $$R_{s,\mathsf{cen}}(M) \leq R_{s,\mathsf{cen}}(\bar{M}) \leq K\cdot\left\{1 - \frac{M-1}{N-1} + \frac{1}{K}\right\}\cdot\frac{1}{1 + \frac{K(M-1)}{N-1} - 1} \leq \left(\frac{N-1}{M-1}\right), \tag{B.27}$$

where we have used $\frac{M-1}{N-1} > \frac{1}{K}$ in the last inequality. Now setting $s = \lfloor 0.1530\frac{N-1}{M-1} \rfloor \in \{1, \ldots, \min\{N, K\}\}$ in Theorem 15, we have:

$$\begin{aligned} R_s^*(M) &\geq 0.1530\frac{N-1}{M-1} - 1 - \frac{0.1530^2 \cdot \frac{N-1}{M-1}^2 \cdot (M-1)}{N - 2 \cdot 0.1530 \cdot \frac{N-1}{M-1}} \\ &\geq \frac{N-1}{M-1}\left\{0.1530 - 0.0628 - \frac{0.1530^2}{1 - \frac{2 \cdot 0.1530}{1.2}}\right\} \\ &\geq \frac{1}{17}\left(\frac{N-1}{M-1}\right). \end{aligned} \tag{B.28}$$

Combining (B.27) and (B.28), we get:

$$\frac{R_{s,\mathsf{cen}}(M)}{R_s^*(M)} \leq 17. \tag{B.29}$$

- ***Regime 3 $(0.0628(N-1) < M-1 \leq N-1)$:***

Let $\bar{M} - 1$ be a multiple of $(N-1)/K$ less than equal to $0.0628(N-1)$, such that

$$0 \leq 0.0628(N-1) - \frac{N-1}{K} \leq \bar{M} - 1 \leq 0.0628(N-1). \tag{B.30}$$

Then using Theorem 14 and the fact that $\bar{M} \leq M$, we have:

$$\begin{aligned} R_{s,\mathsf{cen}}(M) \cdot \frac{1}{1 - \frac{M-1}{N-1}} &\leq R_{s,\mathsf{cen}}(\bar{M}) \cdot \frac{1}{1 - \frac{\bar{M}-1}{N-1}} \\ \Rightarrow R_{s,\mathsf{cen}}(M) &\leq R_{s,\mathsf{cen}}(\bar{M}) \cdot \frac{1}{1 - \frac{\bar{M}-1}{N-1}} \cdot \left(1 - \frac{M-1}{N-1}\right) \\ &\leq R_{s,\mathsf{cen}}(\bar{M}) \cdot \frac{1}{1 - 0.0628} \cdot \left(1 - \frac{M-1}{N-1}\right). \end{aligned} \tag{B.31}$$

Now by Theorem 14 and using (B.30), we have:

$$R_{s,\mathsf{cen}}(\bar{M}) \leq \frac{1}{\frac{\bar{M}-1}{N-1} + \frac{1}{K}} \leq \frac{1}{0.0628 - \frac{1}{K} + \frac{1}{K}} = \frac{1}{0.0628}. \tag{B.32}$$

Thus we have, from (B.31) and (B.32):

$$R_{s,\mathsf{cen}}(M) \leq \frac{1}{0.0628(1 - 0.0628)}\left(1 - \frac{M-1}{N-1}\right). \tag{B.33}$$

Setting $s = 1$ in Theorem 15, we have the following lower bound:

$$R_s^*(M) \geq \left(1 - \frac{M-1}{N-1}\right). \tag{B.34}$$

Thus combining (B.33) and (B.34), we get:

$$\frac{R_{s,\mathsf{cen}}(M)}{R_s^*(M)} \leq \frac{1}{0.0628(1 - 0.0628)} \leq 17. \tag{B.35}$$

Thus we have proved that for any $N, K \in \mathbb{N}$ and all $\frac{(K-N)(N-1)}{KN} + 1 \leq M \leq N$, there is a constant multiplicative gap of 17 between the achievable rate and the information theoretic optimal. This concludes the proof of Theorem 16.

Remark 24. For $K \leq N$ the gap is bounded for the entire feasible regime of $1 \leq M \leq N$. However, for $K > N$, the gap is unbounded in the regime:

$$1 \leq M < \frac{(K-N)(N-1)}{KN} + 1,$$

and scales with the number of users K. However, $\frac{(K-N)(N-1)}{KN} \leq 1$ for any $K > N$ and thus the regime is a fraction of the value of M and is in general negligible when N is large. Also, the regime is always below the values of M for which the data storage dominates key storage i.e., $M > 2N/(N+1) \geq 1$, thereby making it a regime of lesser practical interest. □ □

B.4 Proof of Theorem 17

The decentralized algorithm which achieves the rate in Theorem 17 is given in Algorithm 3. Given N files and K users, each with a cache size of MF bits, we first show that the storage constraint $M \in \frac{N-1}{N}t + 1$ for $t \in (0, N]$ is valid. We then evaluate the rate of Algorithm 3 and show that the multicast delivery is information theoretically secure.

Considering the proposed decentralized scheme in Algorithm 3, each user is allowed to cache any random subset of $\frac{M-1}{N-1}F$ bits of any file W_n. Since the choice of these subsets is uniform, given a particular bit in file W_n, the probability of the bit being cached at a given user is:

$$q \triangleq \frac{M-1}{N-1} \in (0, 1]. \tag{B.36}$$

Considering a fixed subset of s out of K users, the probability that this bit is cached exactly at these s users and not cached at the remaining $(K-s)$ users is $q^s(1-q)^{K-s}$. The expected number of bits of W_n that are cached at exactly those s users is given by:

$$E\left[\#\text{ of bits of } W_n \text{ at } s \text{ users}\right] = Fq^s(1-q)^{K-s}. \tag{B.37}$$

The actual realization of the random number of bits of a file W_n cached at s users is within the range:

$$Fq^s(1-q)^{K-s} \pm o(F). \tag{B.38}$$

For ease of exposition, we consider all the fragments of files shared by s users have the same size. Hence the factor $o(F)$ can be ignored for large enough F.

B.4.1 Storage Constraint

Next, the server maps the contents of the users' caches to non-overlapping fragments in files such that each fragment reflects which users have cached the bits contained in the fragment. Referring to Algorithm 3, Line 4, the variable i signifies the number of users which share a given file fragment. For $i = 0$, the file fragments are $W_{n,\phi}$ which is not stored at any user. When $i = 1$, the file fragments are $W_{n,k}$ for $k = 1, \ldots, K$ which are stored only at one user and hence shared by none. In general for any i, the fragments $W_{n,\mathcal{S}}$ such that $|\mathcal{S}| = i$ are stored at i users and shared by any given user with $i-1$ other users. Thus, for a given a user k, the number of fragments it shares with $i-1$ out of the remaining $K-1$ users for each i is given by $\binom{K-1}{i-1}$. From (B.37), we have the size of fragments which are stored at exactly i users is $Fq^i(1-q)^{K-i}$. Thus, the total storage at each user for storing data is given by:

$$\begin{aligned} M_D F &= N \cdot \sum_{i=1}^{K} \binom{K-1}{i-1} F q^i (1-q)^{K-i} \\ M_D &= Nq \sum_{i-1=0}^{K-1} \binom{K-1}{i-1} q^{i-1} (1-q)^{(K-1)-(i-1)} = Nq = N\frac{M-1}{N-1}. \end{aligned} \tag{B.39}$$

Next, we describe the centralized key placement. For each sub-set $\mathcal{S} \subseteq \{1, \ldots, K\}$ of size s, i.e., $|\mathcal{S}| = s$, where $s = 1, 2, \ldots, K$, a key $\mathcal{K}_{\mathcal{S}}$ is generated as follows:

$$\mathcal{K}_{\mathcal{S}} \sim \text{unif}\left\{1, 2, \ldots, 2^{Fq^{s-1}(1-q)^{K-s+1}}\right\}. \tag{B.40}$$

Subsequently, the key $\mathcal{K}_{\mathcal{S}}$ is placed in the cache of user k if $k \in \mathcal{S}$. The centralized key generation and placement phase is inherently related to the delivery phase of the decentralized algorithm since the size of a key is related to the size of file fragment which is encoded with the key during coded delivery. Consider the coded delivery phase in Algorithm 3, Line $15 - 19$. Given a request $(d_1, \ldots, d_K)$, the composite transmission $X_{(d_1,\ldots,d_K)}$ is sent by the server. The composite transmission can be written as:

$$X_{(d_1,\ldots,d_K)} = \left\{X^s_{(d_1,\ldots,d_K)}\right\}_{s=1}^{K}, \tag{B.41}$$

where $X^s_{(d_1,\ldots,d_K)}$ consists of $\binom{K}{s}$ transmissions, one for each possible sub-set $\mathcal{S}$ of size s i.e.,

$$X^s_{(d_1,\ldots,d_K)} = \left\{\mathcal{K}_{\mathcal{S}} \oplus_{k\in\mathcal{S}} W_{d_k,\mathcal{S}\setminus\{k\}} : |\mathcal{S}| = s\right\}. \tag{B.42}$$

$W_{d_k,\mathcal{S}\setminus\{k\}}$ denotes the part of the file W_{d_k} requested by user k which is present in the caches all the users in set $\mathcal{S}$ except in the cache of user k. The key $\mathcal{K}_{\mathcal{S}}$ is associated with the transmission $\oplus_{k\in\mathcal{S}} W_{d_k,\mathcal{S}\setminus\{k\}}$. Furthermore, from the design of the key placement, the key $\mathcal{K}_{\mathcal{S}}$ is available in the cache of all the s users in the sub-set $\mathcal{S}$. Since $|\mathcal{S}\setminus\{k\}| = s-1$, from (B.37) we have, the expected

size of the fragment $W_{d_k,\mathcal{S}\setminus\{k\}}$ is given by $Fq^{s-1}(1-q)^{K-s+1}$. For a fixed value of s, the size of each transmission in $X^s_{(d_1,\ldots,d_K)}$ is given by:

$$\max_{k\in\mathcal{S}} |W_{d_k,\mathcal{S}\setminus\{k\}}| = Fq^{s-1}(1-q)^{K-s+1}. \tag{B.43}$$

Thus, each key $\mathcal{K}_\mathcal{S}$ must be chosen with the size:

$$|\mathcal{K}_\mathcal{S}| = \max_{k\in\mathcal{S}} |W_{d_k,\mathcal{S}\setminus\{k\}}| = Fq^{s-1}(1-q)^{K-s+1}, \tag{B.44}$$

which is precisely how each key is generated according to (B.40). Now, for a given value of s, a user k needs file fragments contained in $\mathcal{S}\setminus\{k\}$ i.e., $s-1$ other users in the set $\mathcal{S}$. This set of $s-1$ users need to be chosen out of the remaining $K-1$ users. Thus for each s, there are $\binom{K-1}{s-1}$ keys associated with each user. Thus the total number of keys at each user is given by $\sum_{s=1}^{K}\binom{K-1}{s-1} = 2^{K-1}$. The total storage occupied by keys at each users' cache is given by:

$$\begin{aligned} M_K F &= \sum_{s=1}^{K}\binom{K-1}{s-1} Fq^{s-1}(1-q)^{K-s+1} \\ M_K &= (1-q)\sum_{s=1}^{K}\binom{K-1}{s-1} Fq^{s-1}(1-q)^{(K-1)-(s-1)} \\ &= (1-q) = 1-\frac{M-1}{N-1}. \end{aligned} \tag{B.45}$$

From (B.45) and (B.39), we have:

$$M_D + M_K = N\frac{M-1}{N-1} + 1 - \frac{M-1}{N-1} = M, \tag{B.46}$$

which proves the storage constraint. Putting $M_D = t$, the storage break up can be parametrized as:

$$M = t + (1-\frac{t}{N}) = \frac{N-1}{N}t + 1. \tag{B.47}$$

Now, when $t=0$, $M=1$, which is the condition for storing just keys in caches and sending entire files over the shared link. On the other hand, when $t=N$, $M=N$ i.e., the entire files are stored in the caches and there is no need for a transmission. Thus $t\in(0,N]$ is the region of interest. Hence $M \in \frac{N-1}{N}\cdot(0,N]+1$ is valid. Note that the constraint on M is due to the centralized key placement and is thus the cost for security.

Remark 25. Considering the range for file fragment size in (B.38), if we consider that the fragments are not indeed of equal size, then in turn the key size is also within the range $M_K \pm o(F)$. If this is the case, then the cache storage constraint will be within the range $M \pm o(F)$. Since $o(F)$ can generally be ignored in comparison to M, the cache storage constraint is satisfied on an average. □

B.4.2 Calculation of $R_{s,\mathsf{dec}}(M)$

B.4.2.1 Analysis of Conventional Secure Scheme

In conventional secure delivery scheme, for $N \leq K$, the worst case request corresponds to at least one user requesting every file. Considering all users request file W_n, they all have $F(M-1)/(N-1)$ of its bits already in their cache. Thus at most $F\left(1-\frac{M-1}{N-1}\right)+o(F)$ random linear combinations need to be sent to the users requesting the file n. For ease of exposition, $o(F)$ can be ignored. In the conventional scheme, each user k stores an unique key $\mathcal{K}_k$ of size $\left(1-\frac{M-1}{N-1}\right)F$ bits which is XOR-ed with the data before transmission. Although there are N files, each users' request needs to be secured with a key. Thus, in contrast to the non-secure case in [98], the unicast delivery is done for K users and the normalized delivery rate is $K\left(1-\frac{M-1}{N-1}\right)$.

If $N > K$, then at most K different files can be requested. The transmission thus has a normalized rate of $K\left(1-\frac{M-1}{N-1}\right)$. Thus, for all N and $M \in (1, N]$, the conventional scheme has a normalized rate of:

$$R_s^{\text{conv}}(M) = K\left(1-\frac{M-1}{N-1}\right) \tag{B.48}$$

B.4.2.2 Analysis of the proposed scheme

Considering the secure delivery procedure for the coded caching scheme in Algorithm 3, we can see that there are $\binom{K}{s}$ subsets $\mathcal{S}$ of cardinality s. Thus there are $\binom{K}{s}$ transmissions for each $s = K, K-1, \ldots, 1$. Now, for the coded secure transmission, the unique key $\mathcal{K}_{\mathcal{S}}$ is associated with each subset $\mathcal{S}$. The total number of unique keys in the system is given by $\sum_{s=1}^{K}\binom{K}{s} = 2^K - 1$.

Now, considering the fragment size of $W_{d_k,\mathcal{S}\setminus\{k\}}$ in (B.43) and the transmission $X^s_{(d_1,\ldots,d_K)}$ in (B.42), for each value of s, the size of each transmission is given by:

$$|X^s_{(d_1,\ldots,d_K)}| = \binom{K}{s} F q^{s-1}(1-q)^{K-s+1}. \tag{B.49}$$

Summing over all values of s, the rate $R_s^{\text{dec}}(M)$, of the composite transmission $X_{(d_1,\ldots,d_K)}$ is:

$$\begin{aligned} R_s^{\text{dec}}(M)F &= \sum_{s=1}^{K}\binom{K}{s}Fq^{s-1}(1-q)^{K-s+1} \\ R_s^{\text{dec}}(M) &= \frac{1-q}{q}\cdot\sum_{s=1}^{K}\binom{K}{s}q^s(1-q)^{K-s} \\ &= \frac{1-q}{q}\cdot\left(1-(1-q)^K\right) \end{aligned}$$

$$\overset{(B.36)}{=} \frac{1-\frac{M-1}{N-1}}{\frac{M-1}{N-1}} \cdot \left(1-\left(1-\frac{M-1}{N-1}\right)^K\right)$$
$$= K\left(1-\frac{M-1}{N-1}\right) \cdot \frac{N-1}{K(M-1)} \cdot \left(1-\left(1-\frac{M-1}{N-1}\right)^K\right). \tag{B.50}$$

The server can use either the proposed scheme or the conventional secure scheme, whichever uses the minimal rate. Thus combining (B.48) and (B.50), Algorithm 3 achieves a rate of:

$$R_{s,\mathsf{dec}}(M) = \min\left\{R_s^{\text{conv}}(M), R_s^{\text{dec}}(M)\right\}$$
$$= K\left(1-\frac{M-1}{N-1}\right) \cdot \min\left\{\frac{N-1}{K(M-1)} \cdot \left(1-\left(1-\frac{M-1}{N-1}\right)^K\right), 1\right\}, \tag{B.51}$$

which is the result (4.40) presented in Theorem 17.

B.4.3 Proof of Secure Achievability

Next, we show that the delivery phase does not reveal any information to the wiretapper i.e., we show that:

$$I\left(X_{(d_1,\ldots,d_K)}; W_1, \ldots, W_N\right) = 0 \tag{B.52}$$

In the decentralized scheme, the central server transmits $X_{(d_1,\ldots,d_K)}$ to satisfy the requests $(d_1,\ldots,d_k)$ of the K users. The composite transmission $X_{(d_1,\ldots,d_K)}$, given in (B.41), consists of $\binom{K}{s}$ transmissions for each $s = K, K-1, \ldots, 1$. We have:

$$\begin{aligned}
&I\left(X_{(d_1,\ldots,d_K)}; W_1, \ldots, W_N\right) \\
&\quad = H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(X_{(d_1,\ldots,d_K)} | W_1, \ldots, W_N\right) \\
&\quad = H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(\left\{X^s_{(d_1,\ldots,d_K)}\right\}_{s=1}^K | W_1, \ldots, W_N\right) \\
&\quad = H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(\left\{\left\{\mathcal{K}_{\mathcal{S}} \oplus_{k\in\mathcal{S}} W_{d_k,\mathcal{S}\setminus\{k\}} : |\mathcal{S}| = s\right\}\right\}_{s=1}^K | W_1, \ldots, W_N\right) \\
&\quad = H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(\left\{\left\{\mathcal{K}_{\mathcal{S}} : |\mathcal{S}| = s\right\}\right\}_{s=1}^K | W_1, \ldots, W_N\right) \\
&\quad = H\left(X_{(d_1,\ldots,d_K)}\right) - H\left(\left\{\left\{\mathcal{K}_{\mathcal{S}} : |\mathcal{S}| = s\right\}\right\}_{s=1}^K\right),
\end{aligned} \tag{B.53}$$

where, the last equality follows from the fact that the keys are uniformly distributed and are independent of the files $W_1, \ldots, W_N$. Using the fact that $H(A,B) \le H(A) + H(B)$, we have:

$$H\left(X_{(d_1,\ldots,d_K)}\right) = H\left(\left\{X^s_{(d_1,\ldots,d_K)}\right\}_{s=1}^K\right) \le \sum_{s=1}^K H\left(X^s_{(d_1,\ldots,d_K)}\right)$$

$$\leq \sum_{s=1}^{K} \sum_{i=1}^{\binom{K}{s}} H\left(\mathcal{K}_{\mathcal{S}_i} \oplus_{k\in\mathcal{S}_i} W_{d_k,\mathcal{S}_i\setminus\{k\}} : |\mathcal{S}_i| = s\right)$$
$$\leq \sum_{s=1}^{K} \sum_{i=1}^{\binom{K}{s}} \log_2\left(Fq^{s-1}(1-q)^{K-s+1}\right)$$
$$= \sum_{s=1}^{K} \binom{K}{s} \log_2\left(Fq^{s-1}(1-q)^{K-s+1}\right). \tag{B.54}$$

On the other hand, we have:

$$H\left(\{\{\mathcal{K}_{\mathcal{S}} : |\mathcal{S}| = s\}\}_{s=1}^{K}\right) = \sum_{s=1}^{K} H\left(\{\mathcal{K}_{\mathcal{S}} : |\mathcal{S}| = s\}\right) = \sum_{s=1}^{K} \sum_{i=1}^{\binom{K}{s}} H\left(\mathcal{K}_{\mathcal{S}_i} : |\mathcal{S}_i| = s\right)$$
$$= \sum_{s=1}^{K} \sum_{i=1}^{\binom{K}{s}} \log_2\left(Fq^{s-1}(1-q)^{K-s+1}\right)$$
$$= \sum_{s=1}^{K} \binom{K}{s} \log_2\left(Fq^{s-1}(1-q)^{K-s+1}\right), \tag{B.55}$$

where the equality in (4.36) follows from the fact that the keys are orthogonal to each other and they are uniformly distributed as in (B.40). Substituting (B.54) and (B.55) into (B.53), we have:

$$I\left(X_{(d_1,\ldots,d_K)}; W_1, \ldots, W_N\right) \leq 0 \tag{B.56}$$

Using the fact that for any X, Y, $I(X;Y) \geq 0$, we have:

$$I\left(X_{(d_1,\ldots,d_K)}; W_1, \ldots, W_N\right) = 0 \tag{B.57}$$

which proves that the rate $R_{s,\mathsf{dec}}(M)$ is *securely* achievable. This completes the proof of Theorem 17. □

B.5 Proof of Theorem 18

The proof for Theorem 18 is similar to the proof of Theorem 16 in Appendix B.3. We prove that a constant multiplicative gap exists between the achievable decentralized secure rate in Theorem 17 and the information theoretic optimal for the regime:

$$\frac{N-1}{N} + 1 \leq M \leq N \tag{B.58}$$

For the case of $K < N$, from Theorem 17, we have, for $1 < M \leq N$,

$$R_{s,\mathsf{dec}}(M) \leq K\left(1 - \frac{M-1}{N-1}\right) = \min\{N, K\}\left(1 - \frac{M-1}{N-1}\right). \tag{B.59}$$

Again in the case of $K > N$, we have

$$M \geq \frac{N-1}{N} + 1 \Rightarrow \frac{N-1}{M-1} < N \tag{B.60}$$

Now, setting $r = 1 - \frac{M-1}{N-1}$ and substituting in (B.60), we have:

$$\frac{1}{1-r} < N \tag{B.61}$$

Since $0 \leq r < 1$, we have

$$\frac{1}{1-r} \approx \sum_{i=0}^{K-1} r^i \leq N, \tag{B.62}$$

which becomes tighter as $K \to \infty$. Noting that (B.62) is a geometric series, we get:

$$\sum_{i=0}^{K-1} r^i \leq N \;\; \Rightarrow \;\; \frac{1-r^K}{1-r} \leq N \tag{B.63}$$

Substituting the value of r, we have:

$$\begin{aligned} &\frac{N-1}{M-1}\left(1 - \left(1 - \frac{M-1}{N-1}\right)^K\right) \leq N \\ \Rightarrow \;\; &R_{s,\mathsf{dec}}(M) \leq \min\{N, K\}\left(1 - \frac{M-1}{N-1}\right) \end{aligned} \tag{B.64}$$

Thus in general, $R_{s,\mathsf{dec}}(M) \leq \min\{N, K\}\left(1 - \frac{M-1}{N-1}\right)$ for the regime:

$$\frac{N-1}{N} + 1 \leq M \leq N. \tag{B.65}$$

Next, we consider two cases namely (i) $\min\{N, K\} \leq 17$ and (ii) $\min\{N, K\} \geq 18$.

• ***Case 1*** $(\boldsymbol{\min\{N, K\} \leq 17})$ **:** From (B.64), we have:

$$R_{s,\mathsf{dec}}(M) \leq \min\{N, K\}\left(1 - \frac{M-1}{N-1}\right). \tag{B.66}$$

Also, setting $s = 1$ in Theorem 15 gives:

$$R_s^*(M) \geq \left(1 - \frac{M-1}{N-1}\right). \tag{B.67}$$

Thus we have:

$$\frac{R_{s,\mathsf{dec}}(M)}{R_s^*(M)} \leq \min\{N, K\} \leq 17. \tag{B.68}$$

• ***Case 2*** $\boldsymbol{(\min\{N, K\} \geq 18)}$ **:** For this case, we consider 3 distinct regimes namely (i) *Regime 1:* $\frac{N-1}{N} + 1 \leq M - 1 \leq 1.2 \max\left(1, \frac{N-1}{K}\right)$; (ii) *Regime 2:* $1.2 \max\left(1, \frac{N-1}{K}\right) < M - 1 \leq \frac{(N-1)}{17}$; and (iii) *Regime 3:* $\frac{(N-1)}{17} < M - 1 \leq N - 1$. We consider each of the three regimes separately.

• ***Regime 1*** $\boldsymbol{\left(\frac{N-1}{N} + 1 \leq M - 1 \leq 1.2 \max\left(1, \frac{N-1}{K}\right)\right)}$ **:**

By (B.64), we have:

$$R_{s,\mathsf{dec}}(M) \leq R_{s,\mathsf{dec}}(1) \leq \min\{N, K\}. \tag{B.69}$$

By Theorem 15 and using the fact that $\lfloor N/s \rfloor \geq N/s - 1$, we have:

$$R_s^*(M) \geq s - \frac{s^2(M-1)}{N - 2s}. \tag{B.70}$$

Setting $s = \lfloor 0.1586 \min\{N, K\} \rfloor$ we get, for $M - 1 \leq 1.2 \max\left(1, \frac{N-1}{K}\right)$:

$$\begin{aligned}
R_s^*(M) &\geq R_s^*\left(1.2 \max\left(1, \frac{N-1}{K}\right) + 1\right) \\
&\geq 0.1586 \min\{N, K\} - 1 - \frac{(0.1586 \min\{N, K\})^2 \cdot 1.2 \max\left(1, \frac{N-1}{K}\right)}{N - 2 \cdot 0.1586 \min\{N, K\}} \\
&\geq \min\{N, K\} \left\{0.1586 - \frac{1}{\min\{N, K\}} - \frac{(0.1586)^2 \cdot 1.2}{1 - 2 \cdot (0.1586) \min\{1, K/N\}}\right\} \\
&\geq \min\{N, K\} \left\{0.1586 - \frac{1}{18} - \frac{1.2 \cdot (0.1586)^2}{1 - 2 \cdot 0.1586}\right\} \\
&\geq \frac{1}{17} \min\{N, K\}.
\end{aligned} \tag{B.71}$$

Combining (B.69) and (B.71), we get:

$$\frac{R_{s,\mathsf{dec}}(M)}{R_s^*(M)} \leq 17. \tag{B.72}$$

• ***Regime 2*** $\boldsymbol{\left(1.2 \max\left(1, \frac{N-1}{K}\right) < M - 1 \leq \frac{(N-1)}{17}\right)}$ **:**

Using (B.64), we have:

$$R_{s,\mathsf{dec}}(M) \leq \frac{N-1}{M-1} - 1 \leq \frac{N-1}{M-1}. \tag{B.73}$$

Now setting $s = \lfloor 0.1460\frac{N-1}{M-1} \rfloor$ in Theorem 15, we have:

$$\begin{aligned} R_s^*(M) &\geq 0.1460\frac{N-1}{M-1} - 1 - \frac{0.1460^2 \cdot \frac{N-1}{M-1}^2 \cdot (M-1)}{N - 2 \cdot 0.1460 \cdot \frac{N-1}{M-1}} \\ &\geq \frac{N-1}{M-1}\left\{0.0.1460 - \frac{1}{17} - \frac{0.1460^2}{1 - \frac{2 \cdot 0.1460}{1.2}}\right\} \\ &\geq \frac{1}{17}\left(\frac{N-1}{M-1}\right). \end{aligned} \tag{B.74}$$

Combining (B.73) and (B.74), we get:

$$\frac{R_{s,\mathsf{dec}}(M)}{R_s^*(M)} \leq 17. \tag{B.75}$$

- ***Regime 3*** $\left(\frac{(N-1)}{17} < M - 1 \leq N - 1\right)$**:**

  From (B.64), we have:

$$R_{s,\mathsf{dec}}(M) \leq \frac{N-1}{M-1} - 1. \tag{B.76}$$

  Setting $s = 1$ in Theorem 15, we have again:

$$R_s^*(M) \geq \left(1 - \frac{M-1}{N-1}\right). \tag{B.77}$$

  Thus combining (B.76) and (B.77), we get:

$$\begin{aligned} \frac{R_{s,\mathsf{dec}}(M)}{R_s^*(M)} &\leq \frac{\frac{N-1}{M-1} - 1}{1 - \frac{M-1}{N-1}} \\ &= \frac{N-1}{M-1} \leq 17. \end{aligned} \tag{B.78}$$

Thus we have proved that for any $N, K \in \mathbb{N}$ and all $\frac{N-1}{N} + 1 \leq M \leq N$, there is a constant multiplicative gap of 17 between the achievable secure decentralized rate and the information theoretic optimal for any secure scheme. It is to be noted that for $K > N$, the gap is unbounded in the regime

$$1 < M < \frac{N-1}{N} + 1, \tag{B.79}$$

and scales with the number of users K. But $\frac{N-1}{N} < 1$ for any N and thus the regime of M in which the gap is unbounded is in general negligible, especially when N, K are large. This concludes the proof of Theorem 18. □

B.6 Proof of Theorem 19

The rate $R_{s,\mathsf{dec}}(M,N,K)$ achieved by Algorithm 3 is the worst-case peak rate achievable for every possible user request. For the case of equal file popularity, the expected rate (over all requests) is the same as the peak rate for any specific request. Considering a specific request $(d_1,d_2,\ldots,d_K)$, the users can be partitioned into L sets $\mathbb{K}_1,\mathbb{K}_2,\ldots,\mathbb{K}_L$ with cardinality $\mathsf{K}_!,\mathsf{K}_2,\ldots,\mathsf{K}_L$. The delivery algorithm treats each group independently thereby achieving the rate in (4.30) for a request $(d_1,d_2,\ldots,d_K)$. The only randomness in the rate is due to the random size K_ℓ of the random group $\mathbb{K}_\ell$. Taking an expectation over all K_ℓ yields an upper bound on the expected rate of the optimal secure caching scheme:

$$R_s^*(M,\mathbb{N},K) \leq \sum_{\ell=1}^{L} \mathbb{E}\left[R_{s,\mathsf{dec}}(M_\ell,N_\ell,\mathsf{K}_\ell)\right]. \tag{B.80}$$

This upper bound can be further minimized by optimizing over the choice of cache storage allocation M_ℓ at each user. This yields the expression in (4.39):

$$R_s^*(M,\mathbb{N},K) \leq \min_{\{M_\ell:\sum_\ell M_\ell=M\}} \sum_{\ell=1}^{L} \mathbb{E}\left[R_{s,\mathsf{dec}}(M_\ell,N_\ell,\mathsf{K}_\ell)\right]. \tag{B.81}$$

This completes the proof of Theorem 19.

B.7 Proof of Theorem 20

In this section we will prove equivalently that

$$\sum_{\ell=1}^{L} \mathbb{E}\left[R_{s,\mathsf{dec}}(M,N_\ell,\mathsf{K}_\ell)\right] \leq cLR_s^*(M,\mathbb{N},K) \tag{B.82}$$

The proof outline follows closely the proof outline of Theorem 2 in [99]. Similar to [99], the following three claims aid in proving the theorem.

Claim 1

$$R_{s,\mathsf{dec}}(M,N_\ell,\mathsf{K}_\ell) \leq c_1\bar{R}_s(M,N_\ell,K_\ell) \tag{B.83}$$

where $\bar{R}_s(M,N_\ell,K_\ell)$ denotes the expected rate of the optimal scheme for a system with K_ℓ users and N_ℓ files with *uniform* popularity. This claim upper bounds the peak rate of Algorithm 3 by the optimal expected rate for the caching problem with equal file popularities.

Claim 2

$$\bar{R}_s(M, N_\ell, K_\ell) \leq c_2 R_s^*(M, \mathbb{N}_\ell, K_\ell) \tag{B.84}$$

This claim upper bounds the optimal rate for a system with *uniform* file popularities by the optimal expected rate of a system with almost equal file popularities (i.e., file popularities differing by at most a factor p).

Claim 3

$$\mathbb{E}\left[R_s^*(M, \mathbb{N}_\ell, \mathsf{K}_\ell)\right] \leq R_s^*(M, \mathbb{N}, K) \tag{B.85}$$

This claim states that if the server is only asked to handle the demands of users in $\mathbb{K}_\ell$, ignoring the demands of the remaining users, the rate if the optimal system decreases. The expectation is with respect to the random number of users K_ℓ.

Combining Claim 1 and 2, we get

$$\sum_{\ell=1}^{L} \mathbb{E}\left[R_{s,\mathsf{dec}}(M, N_\ell, \mathsf{K}_\ell)\right] \leq c_1 c_2 \sum_{\ell=1}^{L} \mathbb{E}\left[R_s^*(M, \mathbb{N}_\ell, \mathsf{K}_\ell)\right] \tag{B.86}$$

Combining this with Claim 3, we get

$$\sum_{\ell=1}^{L} \mathbb{E}\left[R_{s,\mathsf{dec}}(M, N_\ell, \mathsf{K}_\ell)\right] \leq c_1 c_2 L R_s^*(M, \mathbb{N}, K) \tag{B.87}$$

which proves (B.82) with $c_1 c_2 \triangleq c$. This completes the proof of Theorem 20. It remains to prove the three claims individually.

B.7.1 Proof of Claim 1

To prove Claim 1, we will show equivalently that

$$\bar{R}_s^D(M, N, K) \geq \frac{1}{102} R_{s,\mathsf{dec}}(M, N, K) \tag{B.88}$$

The left hand side is the expected rate of the optimal secure scheme in the case of uniform file popularity for the N files. The right side is equal to the achievable rate of Algorithm 3. First, we show that

$$\bar{R}_s^D(M, N, K) \geq \frac{1}{6} \max_{s \in \{1, \ldots, \min\{N,K\}\}} s \left(1 - \frac{M-1}{\lfloor N/s - 1 \rfloor}\right). \tag{B.89}$$

Consider a demand vector $\underline{d} \in \{1, 2, \ldots, N\}$ and denote by $w(\underline{d})$, the number of distinct entries in $\underline{d}$. We can write the LHS of (B.88) as

$$\begin{aligned}\bar{R}_s^D(M, N, K) &= \sum_{\underline{d} \in N^K} N^{-K} \bar{R}_s^D(M, N, K, \underline{d}) \\ &= \sum_{j=1}^{K} N^{-K} \sum_{\underline{d} \in N^K; w(\underline{d})=j} \bar{R}_s^D(M, N, K, \underline{d}) \end{aligned} \tag{B.90}$$

where $\bar{R}_s^D(M.N.K.\underline{d})$ denotes the rate of the optimal caching sche,e designed for uniform file popularities when the specific demand vector is $\underline{d}$. Clearly, reducing the number of users can only decrease the rate over the shared link. Hence,

$$\begin{aligned}&\bar{R}_s^D(M, N, K) \geq \\ &\sum_{j=1}^{K} N^{-K} \frac{|\{\underline{d} \in N^K : w(\underline{d}) = j\}|}{|\{\underline{d} \in N^j : w(\underline{d}) = j\}|} \sum_{\underline{d} \in N^j : w(\underline{d})=j} \bar{R}_s^D(M, N, j, \underline{d}). \end{aligned} \tag{B.91}$$

The RHS of (B.91) deals with a system in which j users request distinct files from the set of N files *uniformly*. We next evaluate a cut around some number s of users and derive a lower bound on the expected rate using a cut-set argument. Fixing $s \in \{1, 2, \ldots, \min\{N, K\}/4\}$, we can lower bound the RHS of (B.91) as:

$$\begin{aligned}&\bar{R}_s^D(M, N, K) \geq \\ &\sum_{j=s}^{K} N^{-K} \frac{|\{\underline{d} \in N^K : w(\underline{d}) = j\}|}{|\{\underline{d} \in N^s : w(\underline{d}) = s\}|} \sum_{\underline{d} \in N^s : w(\underline{d})=s} \bar{R}_s^D(M, N, s, \underline{d}). \end{aligned} \tag{B.92}$$

The RHS of (B.92) consists of two factors. Appendix A of [99] derives the following result on the first factor using a coupon collector argument:

$$\sum_{j=s}^{K} N^{-K} |\{\underline{d} \in N^K : w(\underline{d}) = j\}| \geq 2/3 \tag{B.93}$$

To evaluate the second factor, a symmetrization argument is used. Consider $I \triangleq \lfloor N/s \rfloor$ and consider $I-$tuples $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_I$ such that $\mathcal{S}_i \subseteq \{1, 2, \ldots, N\}$ has cardinality s and all distinct subsets are disjoint. Denote by $\mathcal{P}$ the collection of all such ordered $I-$tuples. Noting that, by symmetry, every possible subset $\mathcal{S}$ of cardinality s is contained the same number of times in $I-$tuples in $\mathcal{P}$. Let B be the number. Then we have

$$\sum_{\underline{d} \in N^s : w(\underline{d})=s} \bar{R}_s^D(M, N, s, \underline{d}) = \frac{1}{B} \sum_{\mathcal{S}_1, \ldots, \mathcal{S}_I \in \mathcal{P}} \sum_{i=1}^{I} \sum_{\underline{d} \in \mathcal{S}_i^s; w(\underline{d})=s} \bar{R}_s^D(M, N, s, \underline{d}) \tag{B.94}$$

Fix $(\mathcal{S}_1, \ldots, \mathcal{S}_I)$ in $\mathcal{P}$ and consider corresponding demand vectors $(\underline{d}_1, \ldots, \underline{d}_I)$, where $\underline{d}_i \in \mathcal{S}_i^s$ with $w(\underline{d}_i) = s$. We use the cut set argument to lower bound the sum

$$\sum_{i=1}^{I} \bar{R}_s^D(M, N, s, \underline{d}_i)$$

Note that $\bar{R}_s^D(M, N, s, \underline{d}_i)$ is the rate of a system with s users. Considering the I different demands, the transmissions from the central server denoted by $X_1, \ldots, X_I$, along with the caches of the s users, $Z_1, \ldots, Z_s$ must be able to decode the files $W_1, \ldots, W_I$. Let

$$\begin{aligned} \widetilde{W} &= \{W_1, \ldots, W_I\} \\ \widetilde{X} &= \{X_1, \ldots, X_I\} \\ \widetilde{X}_{\setminus\{l\}} &= \left\{X_1, \ldots, X_{l-1}, X_{l+1}, \ldots, X_I\right\} \\ \widetilde{Z} &= \{Z_1, \ldots, Z_s\}. \end{aligned}$$

For the secure caching problem, we also have constraints based on file retrieval and security. The file retrieval constraint is based on the fact that given all possible transmissions and caches, all files can can be retrieved. The security constraint is that a wiretapper should not be able to retrieve any information about the files from any transmission by the server. The constraints are thus given by

$$H(\widetilde{W}|\widetilde{X}, \widetilde{Z}) \leq \epsilon \tag{B.95}$$

$$I(\widetilde{W}; X_l) \leq \epsilon; \quad l = 1, \ldots, \lfloor N/s \rfloor \tag{B.96}$$

Using the cut-set bound, and following the proof of Theorem 15 (Appendix B.2), we have for the secure caching problem,

$$\begin{aligned} sIF \leq H(\widetilde{W}) &= I(\widetilde{W}; \widetilde{X}, \widetilde{Z}) + H(\widetilde{W}|\widetilde{X}, \widetilde{Z}) \\ &\overset{\text{(B.95)}}{\leq} I(\widetilde{W}; \widetilde{X}, \widetilde{Z}) + \epsilon \\ &= I(\widetilde{W}; X_l) + I\left(\widetilde{W}; \widetilde{X}_{\setminus\{l\}}, \widetilde{Z}|X_l\right) + \epsilon \\ &\overset{\text{(B.96)}}{\leq} I\left(\widetilde{W}; \widetilde{X}_{\setminus\{l\}}, \widetilde{Z}|X_l\right) + 2\epsilon \leq H\left(\widetilde{X}_{\setminus\{l\}}, \widetilde{Z}\right) + 2\epsilon \\ &\leq \sum_{i=1, i\neq l}^{I} H(X_i) + \sum_{j=1}^{s} H(Z_j) + 2\epsilon \\ &\leq \sum_{i=1, i\neq l}^{I} \bar{R}_s^D(M, N, s, \underline{d}_i)F + sMF + 2\epsilon \\ \Rightarrow sI \leq &\sum_{i=1, i\neq l}^{I} \bar{R}_s^D(M, N, s, \underline{d}_i) + sM + \frac{2\epsilon}{F}. \end{aligned} \tag{B.97}$$

Simplifying and taking $\lim \epsilon \to 0$, we have

$$\sum_{i=1, i\neq l}^{I} \bar{R}_s^D(M, N, s, \underline{d}_i) \geq s(I - M) \tag{B.98}$$

Using the above, we have for any l,

$$\sum_{i=1}^{I} \bar{R}_s^D(M, N, s, \underline{d}_i) \geq \sum_{i=1, i\neq l}^{I} \bar{R}_s^D(M, N, s, \underline{d}_i) \geq s(I - M)^+ \tag{B.99}$$

where the first inequality simply follows from the fact that the sum of I terms is greater than the sum of $I - 1$ terms and the second inequality follows from the fact that the LHS is strictly non-negative. $(x)^+$ denotes $\max\{x, 0\}$. Combining (B.99) with (B.94), we lower bound the second factor of (B.92) as

$$\begin{aligned} \frac{1}{|\{\underline{d} \in N^s : w(\underline{d}) = s\}|} \sum_{\underline{d}\in N^s : w(\underline{d})=s} & \bar{R}_s^D(M, N, s, \underline{d}) \\ & \geq \frac{s(I - M)^+}{I - 1}, \end{aligned} \tag{B.100}$$

where the normalization $1/(I - 1)$ arises because we have lower bounded the sum of $I - 1$ terms at a time. Substituting (B.93) and (B.100) into (B.92), yields

$$\bar{R}_s^D(M, N, K) \geq \frac{2}{3} \cdot s \left(1 - \frac{M - 1}{\lfloor N/s \rfloor - 1}\right)^+. \tag{B.101}$$

Since this is true for any $s \in \{1, 2, \ldots, \min\{N, K\}/4\}$,

$$\bar{R}_s^D(M, N, K) \geq \frac{2}{3} \max_{s\in\{1,2,\ldots,\min\{N,K\}/4\}} s \left(1 - \frac{M - 1}{\lfloor N/s \rfloor - 1}\right)^+. \tag{B.102}$$

Using a result from [99], we have

$$\begin{aligned} \max_{s\in\{1,2,\ldots,\min\{N,K\}/4\}} & s \left(1 - \frac{M - 1}{\lfloor N/s \rfloor - 1}\right)^+ \\ & \geq \frac{1}{4} \max_{s\in\{1,2,\ldots,\min\{N,K\}\}} s \left(1 - \frac{M - 1}{\lfloor N/s \rfloor - 1}\right) \end{aligned} \tag{B.103}$$

Substituting (B.103) into (B.102), we have the desired result in (B.89). On the other hand from Theorems 15 and 18, we have,

$$\max_{s\in\{1,\ldots,\min\{N,K\}\}} s \left(1 - \frac{(M - 1)}{\left(\lfloor \frac{N}{s} \rfloor - 1\right)}\right) \geq \frac{1}{17} R_{s,\mathsf{dec}}(M, N, K). \tag{B.104}$$

Using the above we obtain

$$\bar{R}_{s,\mathsf{dec}}(M, N, K) \geq \frac{1}{102} R_{s,\mathsf{dec}}(M, N, K), \tag{B.105}$$

which completes the proof of Claim 1.

B.7.2 Proof of Claim 2

We will prove equivalently, that if $p_N/p_n \geq 1/\mathsf{p}$, for all $n \in \mathbb{N}$, then

$$R_s^*(M, \mathbb{N}, K) \geq \frac{1}{\mathsf{c}} \bar{R}_s^D(M, N, K) \tag{B.106}$$

for any constant c. The LHS is the rate of the optimal secure scheme for which K users request files N files in the database $\mathbb{N}$ with popularity $p_1, p_2, \ldots, p_N$. The RHS is the expected rate of the optimal scheme with uniform file popularities. Similar to the corresponding proof in [99], assume that at the beginning of the delivery phase, a *genie* arrives to aid the transmission of files in the following manner. Consider a user requesting a file n. The genie flips a biased coin yielding head with probability $p_N/p_n \geq 1/\mathsf{p}$. If the coin shows a tail, the genie provides the file to the user for free. If the coin shows head, he does not help the user. Thus the probability that a user requests a file n and is not helped by the genie is $p_n \cdot (p_N/p_n) = p_N$, which is the same for each file n. The genie repeats the procedure independently for each user.

The users that have their file delivered by the genie can be ignored in the subsequent delivery phase. Thus we have converted a system with K users requesting files with non-uniform popularities to a system with a random number of users $\tilde{\mathsf{K}}$ requesting files which have uniform popularity. The rate of the optimal scheme for this new system is given by

$$\sum_{i=1}^{K} \mathbb{P}(\tilde{\mathsf{K}} = i) \sum_{\underline{d} \in \mathbb{N}^{\tilde{\mathsf{K}}}} N^{-\tilde{K}} R_s^*(M, \mathbb{N}, i, \underline{d}) \geq \sum_{i=1}^{K} \mathbb{P}(\tilde{\mathsf{K}} = i) \bar{R}_s^D(M, N, i) \tag{B.107}$$

where the inequality follows since $\bar{R}_s^D(M, N, i)$ is the optimal rate expected rate under uniform file popularities. Consider the number of users $K - \tilde{\mathsf{K}}$ that are helped by the genie. Since the probability $1 - p_N/p_n$ is upper bounded by $(\mathsf{p} - 1)/\mathsf{p}$, by assumption on $p_1, p_2, \ldots, p_N$, we have

$$\mathbb{E}\left[K - \tilde{\mathsf{K}}\right] \leq \frac{K(\mathsf{p} - 1)}{\mathsf{p}}. \tag{B.108}$$

By Markov's inequality, we have

$$\mathbb{P}(K - \tilde{\mathsf{K}} \geq K(c-1)/c) \leq \frac{c(\mathsf{p} - 1)}{\mathsf{p}(c-1)} \tag{B.109}$$

for any constant c. Simplifying, we have

$$\mathbb{P}(\tilde{\mathsf{K}} \geq K/c) \geq \frac{c - \mathsf{p}}{\mathsf{p}(c-1)} \tag{B.110}$$

Using this inequality (B.107) can be upper bounded as:

$$\sum_{i \geq K/c} \mathbb{P}(\tilde{\mathsf{K}} = i) \bar{R}_s^D(M, N, \tilde{\mathsf{K}}) \geq \mathbb{P}(\tilde{\mathsf{K}} \geq K/c) \bar{R}_s^D(M, N, K/c) \geq \frac{c - \mathsf{p}}{\mathsf{p}(c-1)} \bar{R}_s^D(M, N, K/c). \tag{B.111}$$

Next we aim to relate the RHS of the above inequality to a system with K users, divided into c groups with each group operating in parallel. The sum rate the system will then be the sum of the delivery rates of the c parallel systems. Since the optimal scheme can be no worse than this, we have

$$\frac{c-\mathsf{p}}{\mathsf{p}(c-1)}\bar{R}_s^D(M,N,K/c) \geq \frac{c-\mathsf{p}}{\mathsf{p}c(c-1)}\bar{R}_s^D(M,N,K), \tag{B.112}$$

which is a lower bound on expected rate of the optimal scheme in the genie aided system. Since the aid of the genie can only reduce the rate, the optimal expected rate of the actual system should be higher than the genie-aided system i.e.,

$$R_s^*(M,\mathbb{N},\mathsf{K}) \geq \frac{c-\mathsf{p}}{\mathsf{p}c(c-1)}\bar{R}_s^D(M,N,K) \tag{B.113}$$

which proves Claim 2 with $\mathsf{c} \triangleq \frac{c-\mathsf{p}}{\mathsf{p}c(c-1)}$.

B.7.3 Proof of Claim 3

We will show that

$$R_s^*(M,\mathbb{N},K) \geq \mathbb{E}\left[R_s^*(M,\mathbb{N}_\ell,\mathsf{K}_\ell)\right] \tag{B.114}$$

The LHS is the expected rate of the optimal scheme of the original problem with K users requesting files in $\mathbb{N}$. Now, we assume that at the beginning of the delivery phase, a genie provides for free, the requested files to each user who requests a file outside of $\mathbb{N}_\ell$. This can only reduce the rate of the system over the shared link. The RHS is the expected rate of the optimal scheme for this genie-aided system. This concludes the proof of Claim 3 and in turn Theorem 20

Appendix C

Proofs From Chapter 5

Fundamental Limits of Cloud and Cache-Aided Wireless Networks

C.1 Proof of Theorem 21

In this section, we present a detailed proof of Theorem 21. To obtain a lower bound on the NDT, we fix a specific request vector $\mathbf{D}$, namely one for which all requested files $(F_1, ..., F_K) = F_{[1:K]}$ are different, and a given channel realization $\mathbf{H}$. Note that this is possible given the assumption $N \geq K$. We denote as T_F and T_E the fronthaul and edge transmission latencies, as per Definition 7 for any given feasible policy $\pi = (\pi_c, \pi_f, \pi_e, \pi_d)$ which guarantees a vanishing probability of error P_e as $L \to \infty$ for the given request $\mathbf{D}$, channel $\mathbf{H}$ and fronthaul rate $C_F = r \log(P)$. Our goal is to obtain a lower bound on the minimum NDT $\delta^*(\mu, r)$ for any $r \geq 0$. To this end, consider the fronthaul messages $\mathbf{U}_m^{T_F}$ which are $1 \times T_F$ row vectors and the corresponding channel outputs in (5.5), where $\mathbf{Y}_k^{T_E}, \mathbf{X}_m^{T_E}$ and $\mathbf{n}_k^{T_E}$ are $1 \times T_E$ row vectors.

For ease of exposition, we next introduce the following notation which we use throughout the appendix. For any integer pair (a, b) with $a \leq b \leq K$, let $\mathbf{Y}_{[a:b]}^{T_E}$ be the $(b - a + 1) \times T$ matrix of channel outputs of a subset $[a : b]$, of receivers. The notation is also used for the channel inputs $\mathbf{X}^{T_E}$ and noise $\mathbf{n}^{T_E}$. Furthermore, for any integers $1 \leq a \leq b \leq K$ and $1 \leq c \leq d \leq M$, we define the following sub-matrix of the channel matrix $\mathbf{H}$:

$$\mathbf{H}_{[a:b]}^{[c:d]} = \begin{bmatrix} h_{a,c} & h_{a,c+1} & \cdots & h_{a,d} \\ h_{a+1,c} & h_{a+1,c+1} & \cdots & h_{a+1,d} \\ \vdots & \vdots & \ddots & \vdots \\ h_{b,c} & h_{a,c+1} & \cdots & h_{b,d} \end{bmatrix}.$$

Using this notation, we can represent the channel outputs at all K receivers as

$$\mathbf{Y}_{[1:K]}^{T_E} = \mathbf{H}_{[1:K]}^{[1:M]}\, \mathbf{X}_{[1:M]}^{T_E} + \mathbf{n}_{[1:K]}^{T_E}, \tag{C.1}$$

To obtain the constraint (5.16), we make the following key observation. Given any set of $\ell \leq \min\{M, K\}$ output signals $\mathbf{Y}_k^{T_E}$, say $\mathbf{Y}_{[1:\ell]}^{T_E}$, and the content of any $(M-\ell)^+$ caches, say $S_{[1:(M-\ell)^+]}$ and their corresponding fronthaul messages $\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}$, all transmitted signals $\mathbf{X}_{[1:M]}^{T_E}$, and hence also all the files $F_{[1:K]}$, can be resolved in the high-SNR regime. This is because: (i) from the cache contents $S_{[1:(M-\ell)^+]}$ and fronthaul messages $\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}$, one can reconstruct the corresponding channel inputs $\mathbf{X}_{[1:(M-\ell)^+]}^{T_E}$; (ii) neglecting the noise in the high-SNR regime, the relationship between the variables $\mathbf{Y}_{[1:\ell]}^{T_E}$ and the remaining inputs $\mathbf{X}_{[(M-\ell)^+:M]}^{T_E}$ is given almost surely by an invertible linear system as in (5.5). We use this argument in the following:

$$\begin{aligned}
KL &= H\left(F_{[1:K]}\right) \\
&\overset{\text{(a)}}{=} H\left(F_{[1:K]}|F_{[K+1:N]}\right) \\
&= I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^{T_E}, \mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}|F_{[K+1:N]}\right) + H\left(F_{[1:K]}|\mathbf{Y}_{[1:\ell]}^{T_E}, \mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right)
\end{aligned} \tag{C.2}$$

where step (a) follows from the fact that all files are independent of each other. The first term in (C.2) can be upper bounded as follows:

$$\begin{aligned}
&I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^{T_E}, \mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}|F_{[K+1:N]}\right) \\
&= I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^{T_E}|F_{[K+1:N]}\right) + I\left(F_{[1:K]}; \mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}|\mathbf{Y}_{[1:\ell]}^{T_E}, F_{[K+1:N]}\right) \\
&\leq I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^{T_E}|F_{[K+1:N]}\right) + I\left(F_{[1:K]}; \mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}, F_{[1:\ell]}|\mathbf{Y}_{[1:\ell]}^{T_E}, F_{[K+1:N]}\right) \\
&= I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^{T_E}|F_{[K+1:N]}\right) + I\left(F_{[1:K]}; F_{[1:\ell]}|\mathbf{Y}_{[1:\ell]}^{T_E}, F_{[K+1:N]}\right) \\
&\quad + I\left(F_{[1:K]}; \mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}|\mathbf{Y}_{[1:\ell]}^{T_E}, F_{[1:\ell]\cup[K+1:N]}\right) \\
&\overset{\text{(a)}}{\leq} I\left(F_{[1:K]}; \mathbf{Y}_{[1:\ell]}^{T_E}|F_{[K+1:N]}\right) + H\left(F_{[1:\ell]}|\mathbf{Y}_{[1:\ell]}^{T_E}\right) \\
&\quad + H\left(\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}|\mathbf{Y}_{[1:\ell]}^{T_E}, F_{[1:\ell]\cup[K+1:N]}\right) - H\left(\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}|\mathbf{Y}_{[1:\ell]}^{T_E}, F_{[1:N]}\right) \\
&\overset{\text{(b)}}{\leq} h\left(\mathbf{Y}_{[1:\ell]}^{T_E}\right) - h\left(\mathbf{Y}_{[1:\ell]}^{T_E}|F_{[1:N]}\right) + L\epsilon_L \\
&\quad + H\left(\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}|F_{[1:\ell]\cup[K+1:N]}\right) - H\left(\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}|\mathbf{Y}_{[1:\ell]}^{T_E}, F_{[1:N]}\right) \\
&\overset{\text{(c)}}{\leq} \ell T_E \log\Big(2\pi e\big(\Lambda P + 1\big)\Big) - h\left(\mathbf{n}_{[1:\ell]}^{T_E}|F_{[1:N]}\right) + L\epsilon_L + H\left(\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}\right) + \sum_{i=1}^{(M-\ell)^+} H\Big(S_{i,[1:N]}|F_{[1:\ell]}, F_{[K+1:N]}\Big)
\end{aligned}$$

$$\overset{\text{(d)}}{\leq} \ell T_E \log\left(\Lambda P + 1\right) + L\epsilon_L + (M-\ell)^+(K-\ell)^+\mu L + (M-\ell)^+ r T_F \log(P), \tag{C.3}$$

where, the steps in (C.3) are explained as follows:

- Step (a) follows from careful expansion of the second term in the previous step and that conditioning reduces entropy.
- Step (b) follows from the fact that $\mathbf{Y}_{[1:\ell]}^{T_E}$ are continuous random variables and that dropping the conditioning in the first term increases entropy. We apply Fano's inequality to the second term where ϵ_L is a function, independent of P, which vanishes as $L \to \infty$.
- Step (c) can be explained as follows. The first term is upper bounded by the use of Lemma 12 detailed in Appendix C.8. The parameter Λ is a constant dependent only on the channel parameters. The last term is zero since the cache contents $S_{[1:(M-\ell)^+]}$ and fronthaul messages $\mathbf{U}_{[1:(M-\ell)^+]}^{T_F}$ are functions of the library of files $F_{[1:N]}$. Moreover, given all the files, the channel outputs are a function of the channel noise at each receiver.
- Step (d) follows from the fact that the channel noise is i.i.d. across time and distributed as $\mathcal{N}(0,1)$.

Next, the second term in (C.2) can be upper bounded by use of Lemma 13 as follows:

$$H\left(F_{[1:K]}|\mathbf{Y}_{[1:\ell]}^{T_E}, \mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right) \leq \; L\epsilon_L + T_E \log\det\left(\mathbf{I}_{[K-\ell]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right), \tag{C.4}$$

where ϵ_L is a function, independent of P and vanishes as $L \to \infty$. Furthermore, the term $\log\det\left(\mathbf{I}_{[K-\ell]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right)$ is independent of signal power P and file size L and is dependent only on the noise variance and the channel coefficients. The proof of (C.4) follows from Lemma 13 which is detailed in Appendix C.8. Substituting (C.3) and (C.4) into (C.2), we have

$$\begin{aligned} KL \leq \ell T_E \log\left(\Lambda P + 1\right) + (M-\ell)^+(K-\ell)^+\mu L + (M-\ell)^+ r T_F \log(P) \\ + L\epsilon_L + T_E \log\det\left(\mathbf{I}_{[K-\ell]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right). \end{aligned} \tag{C.5}$$

Rearranging (C.5), we get the following

$$\ell\delta_E\left[1 + \frac{\ell\log\left(\Lambda + \frac{1}{P}\right) + \log\det\left(\mathbf{I}_{[K-\ell]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right)}{\ell\log(P)}\right] + (M-\ell)^+ r\delta_F \geq K - (M-\ell)^+(K-\ell)^+\mu - \epsilon_L. \tag{C.6}$$

Now, using (C.6), we first take the limit of $L \to \infty$ such that $\epsilon_L \to 0$ as $P_e \to 0$. Further, taking the limit $P \to \infty$, for the high-SNR regime, we arrive at (5.16):

$$\ell\delta_E + (M-\ell)^+ r\delta_F \geq K - (M-\ell)^+(K-\ell)^+\mu, \tag{C.7}$$

where the multiplier of $\ell\delta_E$ converges to 1 under the limit $P \to \infty$.

Next, we prove that the constraint $\delta_E \geq 1$ in (5.17) holds under the decodability constraint for file delivery to all users irrespective of the value of the fronthaul gain r. To this end, without loss of generality, we consider that the users $[1:K]$ demand the first K distinct files $F_{[1:K]}$ i.e., $\mathbf{D} = (d_1, d_2, \ldots, d_K) = (1, 2, \ldots, K)$. Next, consider the following set of inequalities:

$$\begin{aligned} KL &= H\left(F_{[1:K]}\right) \\ &\leq I\left(F_{[1:K]}; \mathbf{Y}_{[1:K]}^{T_E}\right) + H\left(F_{[1:K]}|\mathbf{Y}_{[1:K]}^{T_E}\right) \\ &\overset{\text{(a)}}{\leq} h\left(\mathbf{Y}_{[1:K]}^{T_E}\right) - h\left(\mathbf{Y}_{[1:K]}^{T_E}|F_{[1:K]}\right) + L\epsilon_L \\ &= h\left(\mathbf{Y}_{[1:K]}^{T_E}\right) - h\left(\mathbf{n}_{[1:K]}^{T_E}\right) + L\epsilon_L \\ &\overset{\text{(b)}}{\leq} KT_E \log\left(\Lambda P + 1\right) + L\epsilon_L, \end{aligned} \tag{C.8}$$

where step (a) follows from a Fano's Inequality and the fact that all requested files should be decoded by the received signals. Step (b) follows from the use of Lemma 12 (see Appendix C.8). Again taking the limits $P \to \infty$ and $L \to \infty$, and rearranging, we arrive at the constraint $\delta_E \geq 1$. Note that, by substituting $\ell = M$ in (5.16), we get the following lower bound on the edge latency:

$$\delta_E \geq K/M, \quad \forall\, K, M.$$

This bound is tighter for $M \leq K$, while the constraint $\delta_E \geq 1$, proved here, supersedes the bound for the case when $M \geq K$. Using constraints (5.16)-(5.17) to minimize the sum-latency, i.e., using linear combinations of the family of constraints in (5.16) and (5.17) over all possible choices of $\ell \in [0 : \min\{M, K\}]$, gives the family of lower bounds for the $M \times K$ cache-aided F-RAN.

We conclude this section by addressing the scenario discussed in Section 5.8.1 in which the relaxed cache placement constraints (5.65)-(5.66) are imposed. To prove (5.16), under the relaxed constraints, we consider all possible *sets* of $(M-\ell)^+$ ENs and follow steps similar to (C.2)-(C.3). Considering the step (c) in (C.3) and using the different sets of $(M-\ell)^+$ ENs to decode the files, we will obtain $\binom{M}{(M-\ell)^+}$ different inequalities of this form. Summing and symmetrizing over all the $\binom{M}{(M-\ell)^+}$ inequalities and using the constraint in (5.65) to upper bound the overall number of bits required to store $(K-\ell)^+$ files across the M ENs yields a bound which is identical to (5.16). This shows that the strategy of allocating an equal number of bits to each file at every EN as in Definition 7 is in fact information-theoretically optimal under the assumption of uncoded cache placement.

C.2 Proof of Theorem 23

In order to prove Theorem 23, we first discuss the NDT performance of a scheme that uses fronthaul and wireless channels in the standard fashion that is adopted, for instance, in the CPRI fronthaul interface in C-RANs [172, 226]. In this scheme, the cloud quantizes the encoded baseband

samples, and all the ENs *simultaneously* transmit the quantized baseband signals. We then generalize this policy by allowing for a more general transmission schedule in which different *clusters* of ENs can transmit on the wireless channel at distinct time intervals as introduced in Section 5.4.2 (cf. Fig. 5.5). The proof is divided into two parts as follows.

C.2.1 Standard Soft-Transfer Fronthauling

Here, we prove that an NDT equal to

$$\delta(\mu, r) = \frac{K}{\min\{M, K\}} \left(1 + \frac{1}{r}\right), \tag{C.9}$$

is achievable by means of standard serial soft-transfer fronthauling for any fractional cache size $\mu \geq 0$ and for any fronthaul gain $r \geq 0$. To interpret (C.9), we note that the NDT $\delta(\mu, r) = K/\min\{M, K\}$ can be achieved by means of ZF-beamforming in an ideal system in which there is either full caching, i.e., $\mu = 1$, or no fronthaul capacity limitations, i.e., $r \to \infty$. In fact, in such systems, full cooperation is possible at the ENs for any users' demand vector, including the worst case in which users request distinct files, and hence transmission at the maximum per-user multiplexing gain $\min\{M, K\}/K$ can be attained (see Example 8). The achievable NDT (C.9) hence shows a multiplicative penalty term equal to $1 + 1/r$ due to fronthaul capacity limitations.

The proof of (C.9) relies on the use of the fronthaul and transmission policies introduced in Example 11. Note that caching is not used, in accordance with the assumption that μ may be zero. The cloud encodes the signals using ZF beamforming under a power constraints smaller than P that will be specified below. The resulting baseband signals are quantized and sent to the ENs on the fronthaul links. The ENs transmit simultaneously the respective received quantized samples on the wireless channel. Reception at the users is affected by the fronthaul quantization noise, as well as by the channel noise. If the quantization rate is properly chosen, it can be proved that the achievable NDT is (C.9), where the term $K/\min\{M, K\}$ is the edge-NDT in (5.19), which is the same as for the ideal ZF scheme, and the term $K/(r\min\{M, K\})$ is the fronthaul-NDT (5.18). A more detailed discussion is provided next.

In the cloud-based scheme under study, the cloud performs ZF precoding, producing signal $\bar{X}_i$ for each EN_i with power constraint $\bar{P} = E[|\bar{X}_i|^2]$. The signal $\bar{X}_i$ is quantized to obtain the signal X_i that is to be transmitted by EN_i as

$$X_i = \bar{X}_i + Z_i, \tag{C.10}$$

where $Z_i \sim \mathcal{CN}(0, \sigma^2)$ represents the quantization noise with zero mean and variance σ^2. In order to satisfy the power constraint P, we enforce the condition

$$P = \bar{P} + \sigma^2. \tag{C.11}$$

Furthermore, let B denote the number of bits used for each baseband signal sample on the fronthaul link. From rate-distortion arguments [174] and using (C.10), we obtain the condition

$$\begin{aligned} I\left(X_i; \bar{X}_i\right) &= \log_2\left(1+\frac{\bar{P}}{\sigma^2}\right) = B \\ \text{i.e.,} \qquad \sigma^2 &= \frac{\bar{P}}{2^B-1}. \end{aligned} \tag{C.12}$$

Therefore, from (C.11) and (C.12), we obtain the power constraint on the precoded signal as

$$\bar{P} = P\left(1-2^{-B}\right), \tag{C.13}$$

and the quantization noise power as

$$\sigma^2 = 2^{-B}P. \tag{C.14}$$

The quantization noise terms Z_i for all ENs $i \in [1:M]$, contribute to raising the noise level at each user. In particular, for any user $k \in [1:K]$, the power of the effective noise on the received signals in (5.5) is given by

$$1+\sigma^2\sum_{m=1}^{M}|h_{km}|^2 = 1+\sigma^2 G, \tag{C.15}$$

where $G = \sum_{m=1}^{M}|h_{km}|^2$. Normalizing the received signal so that the variance of the effective noise is 1, using (C.13) and (C.14), we obtain an equivalent signal model in which the effective power constraint is

$$\frac{\bar{P}}{1+\sigma^2 G} = \frac{P(1-2^{-B})}{1+2^{-B}PG}. \tag{C.16}$$

Now, setting $B = \log(P)$, the effective power becomes $(P-1)/(1+G)$, which scales linearly with P. Using the proposed soft-transfer fronthaul scheme, it follows that the fronthaul latency is given by

$$T_F = T_E\frac{B}{C_F}, \tag{C.17}$$

since BT_E bits need to be sent on each fronthaul link at a rate of $C_F = r\log(P)$ to represent the quantized signals. It follows that the total latency of this scheme is

$$T_E + T_F = T_E\left(1+\frac{B}{C_F}\right) \overset{(\mathsf{a})}{=} T_E\left(1+\frac{1}{r}\right), \tag{C.18}$$

where (a) follows from the choice of $B = \log(P)$. Furthermore, in the high-SNR regime we have the following limit:

$$\lim_{P\to\infty}\lim_{L\to\infty}\frac{T_E\log\left((P-1)(1+G)\right)}{L} = \frac{K}{\min\{M,K\}}, \tag{C.19}$$

due to achievability of the NDT $K/\min\{M,K\}$ in the ideal ZF system mentioned above and due to the effective noise power $(P-1)/(1+G)$ for the scheme at hand. We can thus conclude our proof by computing the NDT

$$\lim_{P\to\infty}\lim_{L\to\infty}\frac{(T_E+T_F)\log(P)}{L}=\left(1+\frac{1}{r}\right)\lim_{P\to\infty}\lim_{L\to\infty}\frac{T_E\log(P)}{L}$$
$$=\left(1+\frac{1}{r}\right)\frac{K}{\min\{M,K\}}, \tag{C.20}$$

where the second equality follows due to (C.19).

C.2.2 Soft-Transfer Fronthauling with Clustering

Here, we prove that the following NDT is achievable by means of a generalized soft-transfer fronthaul scheme based on sequential scheduling of distinct clusters of ENs on the wireless channel, when the number M of ENs is larger than the number K of users. In particular we show that for any $M \geq K$, an NDT of

$$\delta(\mu,r)=1+\frac{K}{Mr} \tag{C.21}$$

is achievable by means of soft-transfer fronthauling in conjunction with EN clustering with sequential scheduling for any fractional cache size $\mu \geq 0$ and any fronthaul gain $r \geq 0$.

We start by observing that, if $M \geq K$, the NDT for the ideal system with full caching or unlimited fronthaul is given by $\delta = 1$, which is achieved by ZF-beamforming. Comparing the NDT (C.21) with (C.9), and recalling the discussion in the previous subsection, we can conclude that clustering and sequential scheduling of ENs allows one to reduce the normalized latency associated with the fronthaul transmission from $1/r$ to $K/(Mr)$. We also emphasize that, unlike the NDT in (C.9), which is based on standard C-RAN fronthauling, the improved NDT (C.21) tends to the ideal NDT $\delta = 1$ when the number of transmit antennas grows large. As detailed next, this is due to a novel use of the fronthaul in soft-transfer mode, whereby quantized baseband signals received at the same time on the fronthaul by different ENs can be scheduled at different times on the wireless channel.

We first present the proposed scheme for the case in which M is a multiple of K, so that M/K is an integer number, and then we generalize the strategy for any M. As explained in Section 5.4.2, the main idea is to partition the ENs into M/K disjoint clusters of K ENs and to schedule each cluster for a time equal to $T_E K/M$, that is, on one of M/K equal time intervals dividing T_E. Note that the fact that K ENs are active at any given time enables the use of ZF-beamforming for all time intervals on the wireless channel. In particular, we can use the same scheme based on fronthaul quantization presented in the previous subsection with a key caveat: each cluster needs to receive only $T_E K/M$ baseband samples, and hence the fronthaul latency is

$$T_F = T_E\frac{BK}{MC_F}, \tag{C.22}$$

i.e., the fronthaul latency is M/K times smaller than the latency in (C.17) for the scheme discussed in the previous section. Following the same reasoning as in (C.19)-(C.20) concludes the proof of (C.21) for the case of M/K being an integer number.

We consider now, the more general case in which $M/K \geq 1$ may not be an integer. Here, we proceed by clustering the ENs into all possible $\binom{M}{K}$ subsets of K ENs, and then scheduling each cluster into distinct time intervals of duration $T_E/\binom{M}{K}$. Note that, unlike the case with integer M/K, here the clusters of ENs overlap. The number of samples that each EN needs to receive on its fronthaul is equal to

$$T_E\binom{M-1}{K-1}/\binom{M}{K} = T_E\frac{K}{M}, \tag{C.23}$$

since each EN participates in $\binom{M-1}{K-1}$ clusters and the fronthaul latency is again given by (C.22). Following the same arguments above leads to the NDT in (C.21). Finally, combining the fronthaul latency expressions in (C.17) and (C.22), we have

$$T_F = T_E\frac{B\min\{M,K\}}{MC_F}. \tag{C.24}$$

Using this and following the same arguments as in the previous cases leads to the NDT in (5.32) which completes the proof of Theorem 23.

C.3 Proof of Theorem 25

To prove Theorem 25, we expound on the minimum NDT for the two extremal values of fractional cache size $\mu \in \{1/M, 1\}$. For $\mu = 1/M$, we substitute $\ell = 1$ in (5.20) to get

$$\delta^*(1/M, 0) \geq K - \frac{(M-1)(K-1)}{M} = \frac{M+K-1}{M}. \tag{C.25}$$

To obtain an upper bound on NDT, consider the cache-aided EN coordination scheme achieving the NDT $\delta_{\mathsf{Ca-IA}}$ given in (5.27) as discussed in Lemma 6. Thus, we have the upper bound

$$\delta^*(1/M, 0) \leq \delta_{\mathsf{Ca-IA}} = \frac{M+K-1}{M}. \tag{C.26}$$

Combining (C.25) and (C.26) shows that the lower bound in Corollary 4 is tight at $\mu = 1/M$. Next, considering the NDT at $\mu = 1$, substituting $\ell = \min\{M, K\}$ into (5.20), we get

$$\delta^*(1, 0) \geq \frac{K}{\min\{M, K\}}, \quad \text{for } r = 0. \tag{C.27}$$

Again, when $\mu = 1$, consider the cache-aided EN cooperation scheme leveraging ZF-beamforming achieving the NDT $\delta_{\mathsf{Ca-ZF}}$ given in (5.25) as discussed in Lemma 5. Using this, we have the upper bound

$$\delta^*(1,0) \leq \delta_{\mathsf{Ca-ZF}} = \frac{K}{\min\{M,K\}}. \tag{C.28}$$

Combining (C.27) and (C.28), shows that the lower bound in Corollary 4 is tight at $\mu = 1$. This concludes the proof of Theorem 25.

C.4 Proof of Theorem 27

In this section, we present the proof of the approximate optimality of the achievable schemes presented in Section 5.4. To this end, we consider two regimes for the fractional cache size μ namely low-cache regime with $\mu \in [0, 1/M]$ and high-cache regime with $\mu \in [1/M, 1]$. Next, we consider each of the two regimes separately.

Low-Cache Regime **($\mu \in [0, 1/M]$):** For the low cache size regime, we consider two different cases where (i) the number of users exceeds the number of ENs, i.e., $M \leq K$; and (ii) the number of ENs exceeds the number of users, i.e., $M \geq K$. Next, we treat each of the two cases separately.

• ***Case 1*** **($M \leq K$):** For the case when the number of users exceed the number of ENs, we consider two different subcases: (i) a high fronthaul regime with $r \geq 1$; and (ii) a low fronthaul regime with $r \in (0, 1]$. We consider each of these regimes separately.

High Fronthaul Regime $r \geq 1$: In this regime, consider the achievable NDT in (5.37). We have

$$\delta_{\mathsf{Ach}}(\mu, r) \leq \frac{K}{M} + (1-\mu)\frac{K}{Mr} \leq \frac{K}{M}\left(1 + \frac{1}{r}\right). \tag{C.29}$$

Consider the LP in Theorem 21 and the fact that any lower bound on the solution of this LP is also a valid lower bound on the minimum NDT. Thus, substituting $\ell = M$ in constraint (5.16) and using the fact that $\delta_F \geq 0$, we have

$$\delta^*(\mu, r) \geq \delta_E + \delta_F \geq \frac{K}{M}. \tag{C.30}$$

Thus, we have

$$\frac{\delta_{\mathsf{Ach}}(\mu, r)}{\delta^*(\mu, r)} \leq \left(1 + \frac{1}{r}\right) \leq 2, \tag{C.31}$$

for any fronthaul gain $r \geq 1$. Thus the proposed schemes are approximately optimal to within a factor of 2 for any parameter values of M, K in the high fronthaul regime.

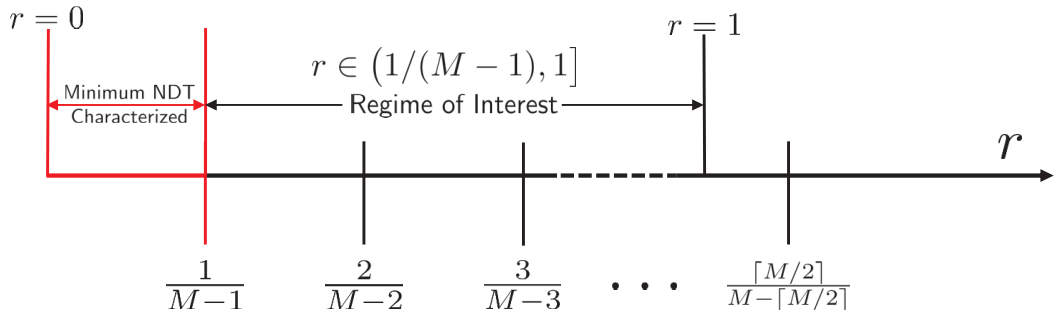


Figure C.1: Division of fronthaul gain $r \in (0, 1]$ into parametrized regimes.

Low Fronthaul Regime $r \in (0, 1]$: For the low fronthaul regime of $r \in (0, 1]$, we first divide the fronthaul gain into multiple non-overlapping regimes based on the number of ENs M in the F-RAN as shown in Fig. C.1, which indicates that for $r \in (0, 1/(M-1)]$, the result in Theorem 28 characterizes the minimum NDT for any $M \leq K$. Therefore, we focus our attention to the remaining regimes of interest for which $r \in (1/(M-1), 1]$. As illustrated in Fig. C.1, we further sub-divide these intervals into the sub-intervals

$$r \in \left[\frac{\ell - 1}{(M - \ell + 1)^+} \, , \, \frac{\ell}{(M - \ell)^+} \right], \tag{C.32}$$

indexed by $\ell \in [1 : \min\{M, K\}]$.

For each sub-interval indexed by $\ell - 1$ and ℓ, we obtain a lower bound on the minimum NDT by considering the constraints (5.16), which are rewritten here as

$$\begin{aligned} \textsf{Ineq 1}: \quad & (\ell-1)\delta_E + (M-\ell+1)r\delta_F \;\geq\; K - (M-\ell+1)(K-\ell+1)\mu, \\ \textsf{Ineq 2}: \quad & \ell\delta_E + (M-\ell)r\delta_F \;\geq\; K - (M-\ell)(K-\ell)\mu. \end{aligned}$$

Specifically, we take a a linear combination of the two inequalities

$$\alpha \times \textsf{Ineq 1} \;+\; \beta \times \textsf{Ineq 2},$$

which with $\alpha, \beta \geq 0$ to yield the following lower bound on the minimum NDT

$$\delta^*(\mu, r) \geq \alpha\big[K - (M-\ell+1)(K-\ell+1)\mu\big] + \beta\big[K - (M-\ell)(K-\ell)\mu\big]. \tag{C.33}$$

Choosing the weights α and β as

$$\alpha \;=\; \frac{\ell}{M}\left(1 + \frac{1}{r}\right) - 1, \quad \beta \;=\; 1 - \frac{\ell - 1}{M}\left(1 + \frac{1}{r}\right), \tag{C.34}$$

we have the following set of inequalities:

$$\delta^*(\mu, r) \geq \big[K - (M-\ell+1)(K-\ell+1)\mu\big] \left[\frac{\ell}{M}\left(1 + \frac{1}{r}\right) - 1\right]$$

$$
\begin{aligned}
&\quad + \big[K-(M-\ell)(K-\ell)\mu\big]\left[1-\frac{\ell-1}{M}\left(1+\frac{1}{r}\right)\right] \\
&= \big[K-(M-\ell)(K-\ell)\mu\big]\left[\frac{1}{M}\left(1+\frac{1}{r}\right)\right] - \mu\big(M+K+1-2\ell\big)\left[\frac{\ell}{M}\left(1+\frac{1}{r}\right)-1\right] \\
&= \left[\frac{K-(M-\ell)(K-\ell)\mu}{M} - \frac{(M+K+1-2\ell)\ell\mu}{M}\right]\left(1+\frac{1}{r}\right) + \mu\big(M+K+1-2\ell\big) \\
&= \left[\frac{K}{M} - K\mu + \frac{(\ell^2-\ell)\,\mu}{M}\right]\left(1+\frac{1}{r}\right) + (M+K-1)\mu + 2\mu(1-\ell) \\
&= (M+K-1)\mu + \frac{K(1-\mu M)}{M}\left(1+\frac{1}{r}\right) + \frac{\ell^2-\ell}{M}\left(1+\frac{1}{r}\right)\mu + 2\mu(1-\ell). \quad \text{(C.35)}
\end{aligned}
$$

An achievable NDT is obtained by considering (5.35), that is, the first term inside the $\min(\cdot)$ function in (5.33) and substituting $\min\{M,K\}=M$, yielding

$$
\begin{aligned}
\delta_{\mathsf{Ach}}(\mu,r) &\le (M+K-1)\mu + \frac{K(1-\mu M)}{M}\left(1+\frac{1}{r}\right) \\
&\overset{\text{(a)}}{\le} (M+K-1)\mu + \frac{K(1-\mu M)}{M}\left(1+\frac{1}{r}\right) + \frac{\mu}{M}, \quad \text{(C.36)}
\end{aligned}
$$

where step (a) follows by adding a positive factor μ/M to achievable NDT from the previous step. Now, from (C.35) and (C.36) we have

$$
\begin{aligned}
\frac{\delta_{\mathsf{Ach}}(\mu,r)}{\delta^*(\mu,r)} &\le \frac{(M+K-1)\mu + \frac{K(1-\mu M)}{M}\left(1+\frac{1}{r}\right)}{(M+K-1)\mu + \frac{K(1-\mu M)}{M}\left(1+\frac{1}{r}\right) + \frac{\ell^2-\ell}{M}\left(1+\frac{1}{r}\right)\mu + 2\mu(1-\ell)} \\
&\overset{\text{(a)}}{=} 1 + \frac{-\frac{\ell^2-\ell}{M}\left(1+\frac{1}{r}\right)\mu - 2\mu(1-\ell)}{(M+K-1)\mu + \frac{K(1-\mu M)}{M}\left(1+\frac{1}{r}\right) + \frac{\ell^2-\ell}{M}\left(1+\frac{1}{r}\right)\mu + 2\mu(1-\ell)} \\
&\overset{\text{(b)}}{\le} 1 + \frac{2\mu(\ell-1)}{(M+K-1)\mu + \frac{K(1-\mu M)}{M}\left(1+\frac{1}{r}\right) + \frac{\ell^2-\ell}{M}\left(1+\frac{1}{r}\right)\mu + 2\mu(1-\ell)} \\
&\overset{\text{(c)}}{\le} 1 + \frac{2\mu(\ell-1)}{(M+K-1)\mu + \frac{\ell^2-\ell}{M}\left(1+\frac{1}{r}\right)\mu + 2\mu(1-\ell)} \\
&\overset{\text{(d)}}{\le} 1 + \frac{2(\ell-1)}{(M+K-1) + 2\left(\frac{\ell^2-\ell}{M}\right) + 2(1-\ell)} \\
&= 1 + \frac{2}{\frac{M+K-1}{\ell-1} + 2\left(\frac{\ell}{M}-1\right)} \\
&\overset{\text{(e)}}{\le} 1 + \frac{2}{\frac{2M-1}{\ell-1} + 2\left(\frac{\ell}{M}-1\right)} \\
&\overset{\text{(f)}}{\le} 1 + \frac{2}{\frac{4M-2}{M-2} - 1} \le 1 + \frac{2(M-2)}{3M} \le 1 + \frac{2}{3} < 2, \quad \text{(C.37)}
\end{aligned}
$$

where step (a) follows by addition and subtraction of the term $\frac{\ell^2-\ell}{M}\left(1+\frac{1}{r}\right)\mu + 2\mu(1-\ell)$ from the numerator; step (b) follows from the fact that the term $\frac{\ell^2-\ell}{M}\left(1+\frac{1}{r}\right)\mu$ in the numerator of the second term is positive in the regime of interest; step (c) follows from the fact that the term $\frac{K(1-\mu M)}{M}\left(1+\frac{1}{r}\right)$ in the denominator of the second term is positive and omitting it leads to an upper bound; step (d) follows by setting by setting $r=1$ which, in turn, follows from the fact that we are interested in the low fronthaul regime with $r \leq 1$; step (e) follows from the fact that $M \leq K$ in the regime of interest and step (f) follows from the fact that for $r \in (0,1]$, we have $\ell \leq M/2$ and putting $\ell = M/2$ minimizes the denominator of the second term which is a decreasing function of ℓ. We conclude that the maximum multiplicative gap between the achievable NDT and the minimum NDT is at most 2 for all the sub-intervals in Fig. C.1.

• ***Case 2 ($M \geq K$):*** For this case, we obtain a lower bound by considering the sum of the constraints in (5.16) with $\ell = 0$ and (5.17), yielding

$$\delta^*(\mu, r) \geq \delta_E + \delta_F = 1 + \frac{K\left(1-\mu M\right)}{Mr}. \tag{C.38}$$

For an achievable NDT, we consider again, the first term inside the $\min(\cdot)$ function in (5.33), which for $K = \min\{M, K\}$ gives the following upper bound

$$\delta_{\mathsf{Ach}}(\mu, r) \leq (M+K-1)\mu + (1-\mu M)\left(1+\frac{K}{Mr}\right) = 1 + \frac{K\left(1-\mu M\right)}{Mr} + (K-1)\mu. \tag{C.39}$$

Thus, from (C.38) and (C.39), we have

$$\frac{\delta_{\mathsf{Ach}}(\mu, r)}{\delta^*(\mu, r)} \leq 1 + \frac{(K-1)\mu}{1+\frac{K(1-\mu M)}{Mr}} \overset{\text{(a)}}{\leq} 1 + \frac{(K-1)}{M} \leq 1 + \frac{K}{M} \overset{\text{(b)}}{\leq} 2, \tag{C.40}$$

where step (a) follows from setting the fractional cache size $\mu = 1/M$ which is the maximum value it can assume in the low-cache memory regime at hand; and step (b) follows from the fact that $M \geq K$. Next, we consider the regime of high cache i.e., $\mu \in [1/M, 1]$.

High-Cache Regime ($\mu \in [1/M, 1]$): In this regime, we consider the achievable NDT in (5.36), i.e., the first term in (5.34), which, using $\mu = 1/M$ yields the upper bound

$$\delta_{\mathsf{Ach}}(\mu, r) \leq \frac{M+K-1}{M}. \tag{C.41}$$

For the lower bounds, we first consider the case of $M \leq K$. From constraint (5.16) in Theorem 21, using $\ell = M$, we have $\delta_E \geq K/M$ which yields the lower bound on the minimum NDT

$$\delta^*(\mu, r) \geq \frac{K}{M}, \tag{C.42}$$

where we have used the fact that $\delta_F \geq 0$. Thus we have the desired gap

$$\frac{\delta_{\mathsf{Ach}}(\mu, r)}{\delta^*(\mu, r)} \leq \frac{(M+K-1)}{M}\frac{M}{K} \leq 1 + \frac{M}{K} \leq 2. \tag{C.43}$$

Next consider the case of $M \geq K$, From constraint (5.17) in Theorem 21, we have $\delta_E \geq 1$. Again, using the fact that that $\delta_F \geq 0$, we have

$$\delta^*(\mu, r) \geq 1. \tag{C.44}$$

Thus we have the desired gap

$$\frac{\delta_{\mathsf{Ach}}(\mu, r)}{\delta^*(\mu, r)} \leq \frac{(M+K-1)}{M} \leq 1 + \frac{K}{M} \leq 2. \tag{C.45}$$

This concludes the proof of Theorem 27.

C.5 Proof of Theorem 28

The minimum NDT is first proved to be upper bounded by the right-hand side of (5.41) by substituting $M = \min\{M, K\}$ into the achievable rate in Theorem 24 for the regime $\mu \in [0, 1/M]$. For the matching lower bound, in the LP of Theorem 21, we substitute $\ell = 1$ and $\ell = 0$ in (5.16), yielding respectively:

$$\mathsf{Ineq\ 1}: \ \delta_E + (M-1)r\delta_F \geq (M+K-1)\mu + K(1-\mu M), \tag{C.46}$$

$$\mathsf{Ineq\ 2}: \ \delta_F \geq K(1-\mu M)/Mr. \tag{C.47}$$

Since $r \in (0, 1/(M-1)]$, we obtain a lower bound by considering the linear combination $\mathsf{Ineq\ 1} + (1-(M-1)r) \times \mathsf{Ineq\ 2}$, leading to the expression on the right-hand side of (5.41). This concludes the proof.

C.6 Converse for Corollary 6

We characterize the lower bounds for the 2×2 F-RAN in order to show the optimality of the achievable schemes discussed in Section 5.4. We again consider each of the fronthaul regimes separately.

Cache-Only F-RAN ($r = 0$): For the cache-only F-RAN, considering the lower bound from Corollary 4 and using $\ell = 1$, we get

$$\delta^*(\mu, r) \geq 2 - \mu, \tag{C.48}$$

which is identical to the achievable NDT in (5.45). Next, we consider the more general case when fronthaul is available i.e., $r > 0$. To this end, we consider the LP in Theorem 21. The constraints of the LP can be rewritten as:

$$\mathsf{Ineq\ 1}: \quad (\delta_E + r\delta_F) \geq (2 - \mu) \tag{C.49}$$

$$\mathsf{Ineq\ 2}: \quad \delta_F \geq (1 - 2\mu)/r \tag{C.50}$$

$$\mathsf{Ineq\ 3}: \quad \delta_E \geq 1. \tag{C.51}$$

Ineq 1 and Ineq 2 are obtained from (5.16) by substituting $\ell = 1$ and $\ell = 0$ respectively, while Ineq 3 follow directly from (5.17). We next utilize these inequalities to prove the converse for different regimes of r.

Low Fronthaul ($r \in (0, 1]$): In this regime, using Ineq 1 $+ (1 - r) \times$ Ineq 2 gives the lower bound:

$$\delta^*(\mu, r) \geq 1 + \mu + \frac{1 - 2\mu}{r}. \tag{C.52}$$

Substituting $r = 1$ in Ineq 1 gives the lower bound

$$\delta^*(\mu, r) \geq 2 - \mu. \tag{C.53}$$

Combining this with the upper bounds in the previous section gives the minimum NDT for the low fronthaul regime as shown in Fig. 5.6(a).

High Fronthaul ($r \geq 1$): For this regime, using Ineq 1 $+ (r - 1) \times$ Ineq 3, we have

$$\delta^*(\mu, r) \geq 1 + \frac{1 - \mu}{r}. \tag{C.54}$$

Combining with the upper bound in the previous section gives the minimum NDT for the high fronthaul regime as shown in Fig. 5.6(b). Thus, the lower bound on the NDT in Theorem 21 and the achievable scheme presented in Theorem 24 completely characterizes the minimum NDT for the 2×2 F-RAN.

C.7 Proof of Corollary 7

To prove the result in Corollary 7 for the 3×3 F-RAN with serial fronthaul-edge transmission, we next expound on the achievable schemes and the converse.

C.7.1 Achievability

The schemes achieving the NDT in Corollary 7 are discussed next.

Cache-Only F-RAN ($r = 0$): For the cache-only F-RAN ($r = 0$), we adapt the results in [146, Theorem 1] to obtain the following achievable NDT:

$$\delta_{\mathsf{MN}}(\mu, 0) = \begin{cases} 13/6 - 3\mu/2 & \text{for } \mu \in [1/3, 2/3], \\ 3/2 - \mu/2 & \text{for } \mu \in [2/3, 1]. \end{cases} \tag{C.55}$$

The two corner points for $\mu = 1/3$ and $\mu = 1$ of the achievable NDT in (C.55) are achieved as in Theorem 25. The inner point at $\mu = 2/3$, instead uses an interference alignment and ZF-beamforming based hybrid scheme to achieve an NDT of $7/6$ as described in [146, 192].

Low Fronthaul ($r \in (0, 1/2]$): In this regime, the minimum NDT for $\mu \in [0, 1/3]$ is obtained by file-splitting between the cloud-aided soft-transfer fronthaul scheme that yields $\delta_{\mathsf{Cl-Sf}}$ in (5.32) requiring $\mu = 0$, and the cache-aided EN coordination strategy, yielding $\delta_{\mathsf{Ca-IA}}$ in (5.27) and requiring $\mu = 1/3$. Therefore, from (5.35), we have

$$\delta^*(\mu, r) = \delta_{\mathsf{Ach}}(\mu, r) = 1 + 2\mu + \frac{1 - 3\mu}{r} \quad \text{for } \mu \in [0, 1/3]. \tag{C.56}$$

Instead, for the high cache memory regime of $\mu \geq 1/3$, it can be seen that transmitting a part of the files by means of the cloud-aided soft-transfer fronthauling does not improve the NDT, and the achievable NDT is therefore given by (C.55). Specifically, for the cache memory regime of $\mu \in [1/3, 2/3]$, the achievable NDT is given by a strategy that performs file-splitting and cache-sharing between the cache-aided EN coordination via interference alignment yielding $\delta_{\mathsf{Ca-IA}}$ and requiring $\mu = 1/3$, and the strategy of [146] as described above, requiring $\mu = 2/3$. In the cache memory regime of $\mu \in [2/3, 1]$, file-splitting and cache-sharing between the strategy of [146] at $\mu = 2/3$ and the cache-aided EN cooperation strategy yielding $\delta_{\mathsf{Ca-ZF}}$ requiring $\mu = 1$, is achieves the minimum NDT, i.e.,

$$\delta^*(\mu, r) = \delta_{\mathsf{MN}}(\mu, 0) = 3/2 - \mu/2 \qquad \text{for } \mu \in [2/3, 1]. \tag{C.57}$$

Intermediate Fronthaul 1 ($r \in [1/2, 6/7]$): In this regime, the achievable schemes are identical to those in the low fronthaul regime. However, in the low cache memory regime when $\mu \in [0, 1/3]$, the NDT obtained by file-splitting between the cloud-aided soft-transfer fronthaul scheme that yields $\delta_{\mathsf{Cl-Sf}}$ in (5.32) and requiring $\mu = 0$, and the cache-aided EN coordination strategy yielding $\delta_{\mathsf{Ca-IA}}$ in (5.27), requiring $\mu = 1/3$, is no longer optimal and yields an upper bound on the minimum NDT. For the high cache memory regime where $\mu \in [1/3, 1]$, the results from the previous case, including the optimal strategy for $\mu \in [2/3, 1]$ achieving the NDT of (C.57), still holds.

Intermediate Fronthaul 2 ($r \in [6/7, 2]$): In this regime, file-splitting between the cloud-aided soft transfer fronthauling scheme yielding $\delta_{\mathsf{Cl-Sf}}$ in (5.32) requiring $\mu = 0$, and the cache-aided EN coordination and cooperation based hybrid scheme from [146], yielding an NDT of $\delta_{\mathsf{MN}}(\mu = 2/3, 0) = 7/6$ from (C.55), achieves the NDT:

$$\delta^*(\mu, r) \leq \delta_{\mathsf{Ach}}(\mu, r) = 1 + \frac{\mu}{4} + \frac{2 - 3\mu}{2r}, \quad \text{for } \mu \in [0, 2/3]. \tag{C.58}$$

For $\mu \in [2/3, 1]$, the minimum NDT is given by (C.57), which is achieved by the optimal strategy as discussed above.

High Fronthaul ($r \geq 2$): In this regime, the minimum NDT is achieved by file-splitting between the cloud-aided soft-transfer fronthauling scheme yielding $\delta_{\mathsf{Cl-Sf}}$ in (5.32) requiring $\mu = 0$, and the cache-aided EN cooperation strategy that yields $\delta_{\mathsf{Ca-ZF}}$ in (5.25) requiring $\mu = 1$. Therefore, from (5.37), we have

$$\delta^*(\mu, r) = \delta_{\mathsf{Ach}}(\mu, r) = 1 + \frac{1-\mu}{r} \quad \text{for } \mu \in [0, 1]. \tag{C.59}$$

C.7.2 Converse

The converse is obtained from Theorem 21 (and Corollary 4) by considering specific weighted sums of the constraints in the LP therein to obtain lower bounds on the optimal value of the LP, and hence on the minimum NDT. We next look at different regimes of r to characterize the lower bound on the NDT in each regime.

Cache-Only F-RAN ($r = 0$): For the case of cache-only F-RAN, we consider the result in Corollary 4 by setting $M = K = 3$, yielding

$$\delta^*(\mu, 0) \geq 3 - 4\mu \qquad \text{for} \quad \ell = 1, \tag{C.60}$$
$$\delta^*(\mu, 0) \geq 3/2 - \mu/2 \quad \text{for} \quad \ell = 2, \tag{C.61}$$
$$\delta^*(\mu, 0) \geq 1 \qquad \text{for} \quad \ell = 3. \tag{C.62}$$

It can be seen that the lower bound coincides with the upper bound at $\mu = 1/3$ and for the regime $\mu \in [2/3, 1]$. Hence, the proposed lower bound in conjunction with the recent result from [146], partially characterizes the minimum NDT for the $M = K = 3$ system. For the regime $\mu \in [1/3, 2/3]$, characterizing the minimum NDT remains an open problem.

Next, we consider the system for the general F-RAN with fronthaul and edge-caching i.e., when $r > 0$ and $\mu \in [0, 1]$. To this end, we consider the LP in Theorem 21. The constraints of the LP can be rewritten as:

$$\mathsf{Ineq\ 1}: \quad (\delta_E + 2r\delta_F) \geq (3 - 4\mu) \tag{C.63}$$
$$\mathsf{Ineq\ 2}: \quad (2\delta_E + r\delta_F) \geq (3 - \mu) \tag{C.64}$$
$$\mathsf{Ineq\ 3}: \quad \delta_F \geq (1 - 3\mu)/r \tag{C.65}$$
$$\mathsf{Ineq\ 4}: \quad \delta_E \geq 1. \tag{C.66}$$

Ineq 1, Ineq 2 and Ineq 3 are obtained from (5.16) by substituting $\ell = 1$, $\ell = 2$ and $\ell = 0$ respectively, while Ineq 4 follows directly from (5.17). We next utilize these inequalities to prove the converse for different regimes of r.

Low Fronthaul ($r \in (0, 1/2]$): In this regime, Ineq 1 $+ (1 - 2r) \times$ Ineq 3 gives the lower bound:

$$\delta^*(\mu, r) \geq 1 + 2\mu + \frac{1 - 3\mu}{r}. \tag{C.67}$$

Also, considering $\mathsf{lneq}\ 1$ and using $r \leq 1/2$, we have

$$\delta^*(\mu, r) \geq 3 - 4\mu. \tag{C.68}$$

Low and Intermediate Fronthaul ($r \leq 2$): Considering $\mathsf{lneq}\ 2$, and using $r \leq 2$, we have the desired lower bound:

$$\delta^*(\mu, r) \geq \frac{3-\mu}{2}. \tag{C.69}$$

Intermediate Fronthaul ($r \in [1/2, 2]$): In this regime, $\left(\frac{2-r}{3r}\right) \times \mathsf{lneq}\ 1 + \left(\frac{2r-1}{3r}\right) \times \mathsf{lneq}\ 2$ yields the lower bound:

$$\delta^*(\mu, r) \geq 1 + \frac{2}{3}\mu + \frac{3-7\mu}{r}. \tag{C.70}$$

High Fronthaul ($r \geq 2$): In this regime, $\mathsf{lneq}\ 2 + (r-2) \times \mathsf{lneq}\ 4$ gives us the lower bound:

$$\delta^*(\mu, r) \geq 1 + \frac{1-\mu}{r}. \tag{C.71}$$

Combining the upper bounds on the minimum NDT from Section 5.6.2 and the lower bounds on the NDT presented above, characterizes the NDT trade-off for the 3×3 F-RAN presented in Corollary 7.

C.8 Lemmas used in Appendix C.1

In this section, we state and prove the lemmas used in the proof of Theorem 21. First, we state and prove Lemma 12 which was used in (C.3) in Appendix C.1.

Lemma 12. *For the cloud and cache-aided wireless network under consideration, the differential entropy of any ℓ channel outputs $\mathbf{Y}_{[1:\ell]}^{T_E}$ can be upper bounded as*

$$h\left(\mathbf{Y}_{[1:\ell]}^{T_E}\right) \leq \ell T_E \log\Big(2\pi e\left(\Lambda P + 1\right)\Big), \tag{C.72}$$

where the parameter Λ is a function of the channel coefficients in $\mathbf{H}$ and is defined as

$$\Lambda = \left(\max_{k\in[1:\ell]}\left[\sum_{m=1}^{M} h_{km}^2 + \sum_{m\neq\tilde{m}} h_{km}h_{k\tilde{m}}\right]\right).$$

Proof. The entropy of the received signals $\mathbf{Y}_{[1:\ell]}^{T_E}$ can be upper bounded as follows:

$$h\left(\mathbf{Y}_{[1:\ell]}^{T_E}\right) \leq \sum_{k=1}^{\ell}\sum_{t=1}^{T_E} h\Big(Y_k[t]\Big). \tag{C.73}$$

Now, we upper bound the inner sum as follows:

$$\begin{aligned}
\sum_{t=1}^{T_E} h\Big(Y_k[t]\Big) &= \sum_{t=1}^{T_E} h\left(\sum_{m=1}^{M} h_{km}X_m[t] + n_k[t]\right) \\
&\leq \sum_{t=1}^{T_E} \log\left(2\pi e \,\text{Var}\left[\sum_{m=1}^{M} h_{km}X_m[t] + n_k[t]\right]\right) \\
&\overset{\text{(a)}}{=} \sum_{t=1}^{T_E} \log\left(2\pi e\left(\text{Var}\left[\sum_{m=1}^{M} h_{km}X_m[t]\right] + \text{Var}\left[n_k[t]\right]\right)\right) \\
&\overset{\text{(b)}}{=} \sum_{t=1}^{T_E} \log\left(2\pi e\left(\sum_{m=1}^{M} h_{km}^2\text{Var}\left[X_m[t]\right] + \sum_{m\neq\tilde{m}} h_{km}h_{k\tilde{m}}\text{Cov}(X_m[t], X_{\tilde{m}}[t]) + 1\right)\right) \\
&\overset{\text{(c)}}{\leq} \sum_{t=1}^{T_E} \log\left(2\pi e\left(\sum_{m=1}^{M} h_{km}^2\text{Var}\left[X_m[t]\right] + \sum_{m\neq\tilde{m}} h_{km}h_{k\tilde{m}}\sqrt{\text{Var}[X_m[t]]\text{Var}[X_{\tilde{m}}[t]]} + 1\right)\right) \\
&\overset{\text{(d)}}{\leq} \sum_{t=1}^{T_E} \log\left(2\pi e\left(\sum_{m=1}^{M} h_{km}^2 P + \sum_{m\neq\tilde{m}} h_{km}h_{k\tilde{m}}P + 1\right)\right) \\
&= \sum_{t=1}^{T_E} \log\Big(2\pi e\big(\Lambda P + 1\big)\Big) = T_E \log\Big(2\pi e\big(\Lambda P + 1\big)\Big) \qquad \text{(C.74)}
\end{aligned}$$

where $\Lambda = \max_{k\in[1:\ell]}\left[\sum_{m=1}^{M} h_{km}^2 + \sum_{m\neq\tilde{m}} h_{km}h_{k\tilde{m}}\right]$. The steps in (C.74) as explained as follows:

- Step (a) follows from the fact that noise is i.i.d. and uncorrelated with the input symbols.
- Step (b) follows from the fact that $\text{Var}\left[n_k[t]\right] = 1$.
- Step (c) follows from the Cauchy-Schwartz Inequality.
- Step (d) follows from the average power constraint P on the input symbols.

Substituting (C.74) into (C.73), we have

$$h\left(\mathbf{Y}_{[1:\ell]}^{T_E}\right) \;\leq\; \sum_{k=1}^{\ell} T_E \log\Big(2\pi e\big(\Lambda P + 1\big)\Big) \;=\; \ell T_E \log\Big(2\pi e\big(\Lambda P + 1\big)\Big), \qquad \text{(C.75)}$$

which completes the proof of the Lemma 12. □

Next, we state and prove Lemma 13 which was used to bound the second term in (C.2) in Appendix C.1.

Lemma 13. *For the cloud and cache-aided wireless network under consideration, for any feasible policy $\pi = (\pi_f, \pi_c, \pi_e, \pi_d)$, the entropy of the K requested files $F_{[1:K]}$, conditioned on the channel outputs $\mathbf{Y}^{T_E}_{[1:\ell]}$, on any $(M-\ell)^+$ fronthaul transmissions $\mathbf{U}^{T_F}_{[1:(M-\ell)^+]}$ with corresponding cache contents $S_{[1:(M-\ell)^+]}$ and on the remaining files $F_{[K+1:M]}$, can be upper bounded as*

$$H\left(F_{[1:K]}|\mathbf{Y}^{T_E}_{[1:\ell]}, \mathbf{U}^{T_F}_{[1:(M-\ell)^+]}, S_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right) \leq \; L\epsilon_L + T_E \log\det\left(\mathbf{I}_{[K-\ell]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right), \tag{C.76}$$

where ϵ_L is a function of the probability of error P_e that vanishes as $L \to \infty$, the matrix $\tilde{\mathbf{H}}$ is a function solely of the channel matrix $\mathbf{H}$ and $\mathbf{I}_{[K-\ell]}$ is a $(K-\ell)\times(K-\ell)$ identity matrix.

Proof. In order to prove this lemma, we first consider the following set of inequalities:

$$\begin{aligned}
&H\left(F_{[1:K]}|\mathbf{Y}^{T_E}_{[1:\ell]}, \mathbf{U}^{T_F}_{[1:(M-\ell)^+]}, S_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right)\\
&\overset{\text{(a)}}{=} H\left(F_{[1:K]}|\mathbf{Y}^{T_E}_{[1:\ell]}, \mathbf{U}^{T_F}_{[1:(M-\ell)^+]}, S_{[1:(M-\ell)^+]}, \mathbf{X}^{T_E}_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right)\\
&\overset{\text{(b)}}{\leq} H\left(F_{[1:K]}|\mathbf{Y}^{T_E}_{[1:\ell]}, \mathbf{X}^{T_E}_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right)\\
&\overset{\text{(c)}}{\leq} H\left(F_{[1:\ell]}|\mathbf{Y}^{T_E}_{[1:\ell]}\right) + H\left(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[1:\ell]}, \mathbf{X}^{T_E}_{[1:(M-\ell)^+]}, F_{[1:\ell]}, F_{[K+1:N]}\right)\\
&\overset{\text{(d)}}{\leq} L\epsilon_L + H\left(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[1:\ell]}, \mathbf{X}^{T_E}_{[1:(M-\ell)^+]}, F_{[1:\ell]\cup[K+1:N]}\right),
\end{aligned} \tag{C.77}$$

where the steps in (C.77) are explained as follows:

- Step (a) follows from the fact that the channel inputs $\mathbf{X}^{T_E}_{[1:(M-\ell)^+]}$ are functions of the fronthaul transmissions $\mathbf{U}^{T_F}_{[1:(M-\ell)^+]}$ and the corresponding cache contents $S_{[1:(M-\ell)^+]}$.
- Step (b) follows from the fact that conditioning reduces entropy.
- Step (c) follows from the chain rule of entropy and from the fact that conditioning reduces entropy.
- In step (d), we use Fano's inequality on the first term where ϵ_L is a function, independent of P, that vanishes as $L \to \infty$.

Next, we consider the second term in (C.77). We have

$$\begin{aligned}
&H\left(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[1:\ell]}, \mathbf{X}^{T_E}_{[1:(M-\ell)^+]}, F_{[1:\ell]\cup[K+1:N]}\right)\\
&\overset{\text{(a)}}{=} H\left(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[1:\ell]}, \mathbf{X}^{T_E}_{[1:(M-\ell)^+]}, \mathbf{n}^{T_E}_{[\ell+1:K]}, F_{[1:\ell]\cup[K+1:N]}\right)\\
&\overset{\text{(b)}}{\leq} H\Big(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[\ell+1:K]} + \tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}, \mathbf{Y}^{T_E}_{[1:\ell]}, F_{[1:\ell]\cup[K+1:N]}\Big)
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{(c)}}{\leq} H\Big(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[\ell+1:K]}+\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]},F_{[1:\ell]\cup[K+1:N]}\Big)-H\Big(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[\ell+1:K]},F_{[1:\ell]\cup[K+1:N]}\Big)\\
&\qquad\qquad+H\Big(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[\ell+1:K]},F_{[1:\ell]\cup[K+1:N]}\Big)\\
&\overset{\text{(d)}}{\leq} H\Big(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[\ell+1:K]}+\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]},F_{[1:\ell]\cup[K+1:N]}\Big)-H\Big(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[\ell+1:K]},F_{[1:\ell]\cup[K+1:N]}\Big)+L\epsilon_L
\end{aligned} \tag{C.78}$$

where the steps in (C.78) are explained as follows:

- Step (a) follows from the fact that the noise term $\mathbf{n}^{T_E}_{[\ell+1:K]}$ is independent of all the other random variables in the entropy term and can be introduced into the conditioning.
- In Step (b), we use Lemma 14 stated in Appendix C.8 and the fact that conditioning reduces entropy. We observe that $\mathbf{n}^{T_E}_{[\ell+1:K]}\rightarrow(\mathbf{Y}^{T_E}_{[1:\ell]},\mathbf{X}^{T_E}_{[1:(M-\ell)^+]},F_{[1:\ell]\cup[K+1:N]})\rightarrow F_{[\ell+1:K]}$ forms a Markov chain and as a result, the data-processing inequality [174] applies. The additive noise term $\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}$ is defined as

  $$\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}=\left(\mathbf{H}_2\cdot{\mathbf{H}_1}^{\dagger}\right)\mathbf{n}^{T_E}_{[1:\ell]},$$

  which is a $[K-\ell]\times T_E$ matrix, where each column is an independent Gaussian random vector distributed as $\mathcal{N}\left(0,\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right)$ with $\tilde{\mathbf{H}}=\left(\mathbf{H}_2\cdot{\mathbf{H}_1}^{\dagger}\right)$, where the matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ are sub-matrices of the channel matrix $\mathbf{H}$ and are defined in Lemma 14 (see (C.82)), and ${\mathbf{H_1}}^{\dagger}$ is the Moore-Penrose pseudo-inverse. We note here that the noise term $\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}$ is independent of channel inputs $\mathbf{X}^{T_E}_{[1:M]}$ and noise terms $\mathbf{n}^{T_E}_{[\ell+1:K]}$.
- Step (c) follows from the fact that conditioning reduces entropy.
- Step (d) follows from applying Fano's inequality to the last entropy term in the previous step, where ϵ_L is again, a function independent of P that vanishes as $L\rightarrow\infty$.

Now, from (C.78), considering the first and second entropy terms together we have:

$$\begin{aligned}
&H\Big(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[\ell+1:K]}+\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]},F_{[1:\ell]\cup[K+1:N]}\Big)-H\Big(F_{[\ell+1:K]}|\mathbf{Y}^{T_E}_{[\ell+1:K]},F_{[1:\ell]\cup[K+1:N]}\Big)\\
&=I\left(F_{[\ell+1:K]};\mathbf{Y}^{T_E}_{[\ell+1:K]}|F_{[1:\ell]\cup[K+1:N]}\right)-I\left(F_{[\ell+1:K]};\mathbf{Y}^{T_E}_{[\ell+1:K]}+\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}|F_{[1:\ell]\cup[K+1:N]}\right)\\
&=h\left(\mathbf{Y}^{T_E}_{[\ell+1:K]}+\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}|F_{[1:N]}\right)-h\left(\mathbf{Y}^{T_E}_{[\ell+1:K]}+\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}|F_{[1:\ell]\cup[K+1:N]}\right)\\
&\qquad+h\left(\mathbf{Y}^{T_E}_{[\ell+1:K]}|F_{[1:\ell]\cup[K+1:N]}\right)-h\left(\mathbf{Y}^{T_E}_{[\ell+1:K]}|F_{[1:N]}\right)\\
&\overset{\text{(a)}}{\leq}h\left(\mathbf{Y}^{T_E}_{[\ell+1:K]}+\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}|F_{[1:N]}\right)-h\left(\mathbf{Y}^{T_E}_{[\ell+1:K]}+\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]}|\tilde{\mathbf{n}}^{T_E}_{[\ell+1:K]},F_{[1:\ell]\cup[K+1:N]}\right)\\
&\qquad+h\left(\mathbf{Y}^{T_E}_{[\ell+1:K]}|F_{[1:\ell]\cup[K+1:N]}\right)-h\left(\mathbf{Y}^{T_E}_{[\ell+1:K]}|F_{[1:N]}\right)
\end{aligned}$$

$$
\begin{aligned}
&= h\left(\mathbf{Y}_{[\ell+1:K]}^{T_E} + \tilde{\mathbf{n}}_{[\ell+1:K]}^{T_E} | F_{[1:N]}\right) - h\left(\mathbf{Y}_{[\ell+1:K]}^{T_E} | F_{[1:N]}\right) \\
&\overset{\text{(b)}}{=} h\Big(\mathbf{n}_{[\ell+1:K]}^{T_E} + \tilde{\mathbf{n}}_{[\ell+1:K]}^{T_E}\Big) - h\Big(\mathbf{n}_{[\ell+1:K]}^{T_E}\Big) \\
&\overset{\text{(c)}}{=} T_E \log\left((2\pi e)^{K-\ell}\left|\mathbf{I}_{[K-\ell]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right|\right) - T_E \log\left((2\pi e)^{K-\ell}\right) \\
&= T_E \log\det\left(\mathbf{I}_{[K-\ell]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right). \qquad \text{(C.79)}
\end{aligned}
$$

The steps in (C.79) are explained as follows:

- Step (a) follows from the fact that conditioning reduces entropy.
- Step (b) follows from the fact that, given all the files $F_{[1:N]}$, the channel outputs are functions of the channel noise.
- Step (c) follows from the fact that the noise terms are jointly Gaussian and are i.i.d. across time T_E. The function $|\cdot|$ is the determinant.

Thus, using (C.78) and (C.79) in (C.77), we have

$$
H\left(F_{[1:K]} | \mathbf{Y}_{[1:\ell]}^{T_E}, \mathbf{U}_{[1:(M-\ell)^+]}^{T_F}, S_{[1:(M-\ell)^+]}, F_{[K+1:N]}\right) \;\leq\; L\epsilon_L + T_E \log\det\left(\mathbf{I}_{[K-\ell]} + \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right), \qquad \text{(C.80)}
$$

which completes the proof of the Lemma 13. □

Finally, we state and prove Lemma 14 which was used in (C.78) for the proof of Lemma 13.

Lemma 14. *Given any $\ell \in [1 : \min\{N, K\}]$, there exists a (deterministic) function of the channel outputs $\mathbf{Y}_{[1:\ell]}^{T_E}$, input symbols $\mathbf{X}_{[1:(M-\ell)]^+}^{T_E}$ and channel noise $\mathbf{n}_{[\ell+1:K]}^{T_E}$, that yields*

$$
\mathbf{Y}_{[\ell+1:K]}^{T_E} + \tilde{\mathbf{n}}_{[\ell+1:K]}^{T_E}, \qquad \text{(C.81)}
$$

where we have defined $\tilde{\mathbf{n}}_{[\ell+1:K]}^{T_E} = \left(\mathbf{H}_2 \cdot \mathbf{H}_1{}^{\dagger}\right)\mathbf{n}_{[1:\ell]}^{T_E}$ and $\mathbf{H}_1{}^{\dagger}$ is the Moore-Penrose pseudo-inverse. The matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ are sub-matrices of the channel matrix $\mathbf{H}$ and are defined as

$$
\mathbf{H}_1 = \mathbf{H}_{[(M-\ell)^+ + 1:M]}^{[1:\ell]}; \quad \mathbf{H}_2 = \mathbf{H}_{[(M-\ell)^+ + 1:M]}^{[\ell+1:K]}. \qquad \text{(C.82)}
$$

Proof. Given any $\ell \in [1 : \min\{M, K\}]$, from (C.1), the channel outputs $\mathbf{Y}_{[1:\ell]}^{T_E}$ are a function of the M input symbols $\mathbf{X}_{[1:M]}^{T_E}$ and of the noise $\mathbf{n}_{[1:\ell]}^{T_E}$. Given the input symbols $\mathbf{X}_{[1:(M-\ell)^+]}^{T_E}$, we can cancel the contribution of these input symbols from the channel outputs $\mathbf{Y}_{[1:\ell]}^{T_E}$ to obtain

$$
\tilde{\mathbf{Y}}_{[1:\ell]}^{T_E} = \mathbf{H}_{[1:\ell]}^{[1:M]}\mathbf{X}_{[1:M]}^{T_E} + \mathbf{n}_{[1:\ell]}^{T_E} - \mathbf{H}_{[1:\ell]}^{[1:M]}\begin{bmatrix}\mathbf{X}_{[1:(M-\ell)^+]}^{T_E} \\ \mathbf{0}_{[(M-\ell)^+ + 1:M]}^{T_E}\end{bmatrix}
$$

$$= \mathbf{H}_1 \left[\mathbf{X}_{[(M-\ell)^+ +1:M]}^{T_E}\right] + \left[\mathbf{n}_{[1:\ell]}^{T_E}\right], \tag{C.83}$$

where $\mathbf{0}_{[(M-\ell)^+ +1:M]}^{T_E}$ is an $\ell \times T_E$ matrix of zeros. As a result, multiplying both sides of (C.83) by $\mathbf{H_1}^\dagger$, we get

$$\mathbf{H_1}^\dagger \tilde{\mathbf{Y}}_{[1:\ell]}^{T_E} = \mathbf{X}_{[(M-\ell)^+ +1:M]}^{T_E} + \mathbf{H_1}^\dagger \mathbf{n}_{[1:\ell]}^{T_E}. \tag{C.84}$$

Now let

$$\mathbf{H}_3 = \mathbf{H}_{[\ell+1:K]}^{[1:M]}. \tag{C.85}$$

Using this definition, we have

$$\begin{aligned}
\mathbf{Y}_{[\ell+1:K]}^{T_E} &= \mathbf{H}_3 \mathbf{X}_{[1:M]}^{T_E} + \mathbf{n}_{[\ell+1:K]}^{T_E} \\
&= \mathbf{H}_3 \begin{bmatrix} \mathbf{X}_{[1:(M-\ell)^+]}^{T_E} \\ \mathbf{H_1}^\dagger \tilde{\mathbf{Y}}_{[1:\ell]}^{T_E} - \mathbf{H_1}^\dagger \mathbf{n}_{[1:\ell]}^{T_E} \end{bmatrix} + \mathbf{n}_{[\ell+1:K]}^{T_E} \\
&\overset{\text{(a)}}{=} \mathbf{H}_3 \begin{bmatrix} \mathbf{X}_{[1:(M-\ell)^+]}^{T_E} \\ \mathbf{H_1}^\dagger \tilde{\mathbf{Y}}_{[1:\ell]}^{T_E} \end{bmatrix} - \mathbf{H}_3 \begin{bmatrix} \mathbf{0}_{[1:(M-\ell)^+]}^{T_E} \\ \mathbf{H_1}^\dagger \mathbf{n}_{[1:\ell]}^{T_E} \end{bmatrix} + \mathbf{n}_{[\ell+1:K]}^{T_E} \\
&= \mathbf{H}_3 \begin{bmatrix} \mathbf{X}_{[1:(M-\ell)^+]}^{T_E} \\ \mathbf{H_1}^\dagger \tilde{\mathbf{Y}}_{[1:\ell]}^{T_E} \end{bmatrix} - \mathbf{H}_2 \left[\mathbf{H_1}^\dagger \mathbf{n}_{[1:\ell]}^{T_E}\right] + \mathbf{n}_{[\ell+1:K]}^{T_E},
\end{aligned} \tag{C.86}$$

where, in (a), $\mathbf{0}_{[1:(M-\ell)^+]}^{T_E}$ is a $[(M-\ell)^+] \times T_E$ matrix of zeros. Rearranging (C.86), we obtain

$$\mathbf{Y}_{[\ell+1:K]}^{T_E} + \tilde{\mathbf{n}}_{[\ell+1:K]}^{T_E} = \mathbf{H}_3 \begin{bmatrix} \mathbf{X}_{[1:(M-\ell)^+]}^{T_E} \\ \mathbf{H_1}^\dagger \tilde{\mathbf{Y}}_{[1:\ell]}^{T_E} \end{bmatrix} + \mathbf{n}_{[\ell+1:K]}^{T_E}, \tag{C.87}$$

where the RHS is a function of the ℓ channel outputs $\mathbf{Y}_{[1:\ell]}^{T_E}$, input symbols $\mathbf{X}_{[1:(M-\ell)]^+}^{T_E}$ and channel noise $\mathbf{n}_{[\ell+1:K]}^{T_E}$. This completes the proof Lemma 14. Note that we assumed in (C.84) that the sub-matrix $\mathbf{H}_1$ is invertible, which is true for almost all channel realizations, i.e., it is true with probability 1. □

C.9 Pipelined Fronthaul-Edge Transmission

C.9.1 Proof of Theorem 29

We adopt block-Markov coding as explained in Section 5.7.2, whereby, in each block, the constituent policies for fronthaul and edge transmission are obtained by file-splitting between two

policies. To elaborate, for some $\alpha \in [0,1]$ fraction of each file, a (serial) policy requiring fronthaul and edge NDTs $\delta_F^{(1)}$ and $\delta_E^{(1)}$ is used, and for the remaining $(1-\alpha)$ fraction of each file, a (serial) policy requiring NDTs $\delta_F^{(2)}$ and $\delta_E^{(2)}$ is used. From (5.56), the achievable NDT with the resulting block-Markov policy is given by

$$\delta_{\mathsf{P,Ach}} = \max\left(\alpha\delta_F^{(1)} + (1-\alpha)\delta_F^{(2)}, \alpha\delta_E^{(1)} + (1-\alpha)\delta_E^{(2)}\right). \tag{C.88}$$

We next identify the constituent policies used to prove (5.58) for the three regimes namely (i) *low cache regime* with $\mu \in [0,\mu_1]$; (ii) *intermediate cache regime* with $\mu \in [\mu_1,\mu_2]$; and *high cache regime* with $\mu \in [\mu_2,1]$.

C.9.1.1 Low Cache Regime $(\mu \in [0,\mu_1])$

For the regime $\mu \in [0,\mu_1]$, we set $\alpha = \mu M$, and perform file-splitting between cache-aided EN coordination, as described in Lemma 6, which is characterized by the NDTs

$$\delta_F^{(1)} = 0; \quad \delta_E^{(1)} = \delta_{\mathsf{Ca-IA}} = \frac{M+K-1}{M}; \tag{C.89}$$

and soft-transfer fronthauling in conjunction with EN cooperation via ZF-beamforming, as described in Theorem 23, which yields the NDTs

$$\delta_F^{(2)} = \frac{K}{Mr}; \quad \delta_E^{(2)} = \frac{K}{\min\{M,K\}}. \tag{C.90}$$

Note that we have $\mu_1 \leq 1/M$ and hence $\alpha \leq 1$ for $\mu \leq \mu_1$. Substituting (C.89)-(C.90) in (C.88), we obtain

$$\begin{aligned}\delta_{\mathsf{P-IA}} &= \max\left\{\frac{(1-\mu M)K}{Mr}, \frac{(1-\mu M)K}{\min\{M,K\}} + \mu(M+K-1)\right\}, \\ &= \frac{(1-\mu M)K}{Mr}, \qquad \text{for} \quad \mu \leq \mu_1.\end{aligned} \tag{C.91}$$

C.9.1.2 High Cache Regime $(\mu \in [\mu_2,1])$

For the regime $\mu \in [\mu_2,1]$, we set $\alpha = \mu$, and perform file-splitting between cache-aided EN cooperation in the form of ZF-beamforming, as described in Lemma 5, which yields the NDTs

$$\delta_F^{(1)} = 0; \quad \delta_E^{(1)} = \delta_{\mathsf{Ca-ZF}} = \frac{K}{\min\{M,K\}}; \tag{C.92}$$

and soft-transfer fronthauling with ZF-beamforming on the edge, as described in Theorem 23 leading to

$$\delta_F^{(2)} = \frac{K}{Mr}; \quad \delta_E^{(2)} = \frac{K}{\min\{M,K\}}. \tag{C.93}$$

Substituting (C.92)-(C.93) in (C.88), we obtain

$$\begin{aligned}\delta_{\mathsf{P-ZF}} &= \max\left\{\frac{(1-\mu)K}{Mr}, \frac{K}{\min\{M,K\}}\right\}, \\ &= \frac{K}{\min\{M,K\}}, \qquad \text{for} \quad \mu \geq \mu_2. \end{aligned} \tag{C.94}$$

C.9.1.3 Intermediate Cache Regime ($\mu \in [\mu_1, \mu_2]$)

For the intermediate cache regime of $\mu \in [\mu_1, \mu_2]$, we consider a strategy which performs file-splitting between the schemes achieving $\delta_{\mathsf{P-IA}}$ at $\mu = \mu_1$ and $\delta_{\mathsf{P-ZF}}$ at $\mu = \mu_2$ discussed above. Specifically, using first scheme which yields an NDT

$$\delta_{\mathsf{P-IA}} = \frac{(1-\mu_1 M)K}{Mr} \tag{C.95}$$

for a fraction $\left(\frac{\mu_2-\mu}{\mu_2-\mu_1}\right)^+$ of the files, and the second scheme for the remaining fraction, which yields an NDT

$$\delta_{\mathsf{P-ZF}} = \frac{(1-\mu_2)K}{Mr}, \tag{C.96}$$

we obtain the achievable NDT

$$\begin{aligned}\delta_{\mathsf{P-FS}} &= \left(\frac{\mu_2-\mu}{\mu_2-\mu_1}\right)^+ \frac{(1-\mu_1 M)K}{Mr} + \left(1-\left(\frac{\mu_2-\mu}{\mu_2-\mu_1}\right)^+\right)\frac{(1-\mu_2)K}{Mr} \\ &= \frac{K}{Mr}\left[1-\mu_2-[\mu_1 M-\mu_2]\left(\frac{\mu_2-\mu}{\mu_2-\mu_1}\right)^+\right]. \end{aligned} \tag{C.97}$$

This concludes the proof of Theorem 29.

C.9.2 Proof of Theorem 30

In the regime of low cache size $\mu \in [0, \mu_1]$, the upper bound (5.58), rewritten here as

$$\delta_{\mathsf{P,Ach}}(\mu, r) \leq \delta_{\mathsf{P-IA}} = \frac{(1-\mu M)K}{Mr}, \tag{C.98}$$

matches the lower bound in Corollary 8 by setting $\ell = 0$, thereby characterizing the minimum NDT for the low cache regime with $\mu \in [0, \mu_1]$.

For the regime of high cache size $\mu \in [\mu_2, 1]$, from (5.58) we have the upper bound

$$\delta_{\mathsf{P,Ach}}(\mu, r) \leq \delta_{\mathsf{P-ZF}} = \frac{K}{\min\{M,K\}}. \tag{C.99}$$

For a matching lower bound, when $M \geq K$, from (5.54), we have $\delta^*_{\mathsf{P}}(\mu, r) \geq 1$, while, for $M \leq K$, using $\ell = M$ in the first term inside the $\max(\cdot)$ function in (5.54) yields $\delta^*_{\mathsf{P}}(\mu, r) \geq K/M$. Combining the two bounds yields the following lower bound on the minimum NDT:

$$\delta^*_{\mathsf{P}}(\mu, r) \geq \frac{K}{\min\{M, K\}}, \tag{C.100}$$

which matches the upper bound (C.99), thereby characterizing the minimum NDT for the high cache regime with $\mu \in [\mu_2, 1]$.

Finally, we consider the high fronthaul regime i.e., $r \geq ((1-\mu)\min\{M, K\})/M$. In this regime, considering the NDT in (C.94), which is achieved by file-splitting between cloud-aided soft-transfer fronthauling and cache-aided ZF beamforming, it can be seen that the second term inside the $\max(\cdot)$ dominates and we have

$$\delta_{\mathsf{P},\mathsf{Ach}}(\mu, r) \leq \delta_{\mathsf{P}-\mathsf{ZF}} = \frac{K}{\min\{M, K\}}, \tag{C.101}$$

which matches the lower bound in (C.100). This completes the proof of Theorem 30.

C.9.3 Proof of Theorem 31

In this section, we present the proof of the approximate optimality of the achievable schemes presented in Section 5.7.2 in the regime of intermediate fractional cache sizes with $\mu \in [\mu_1, \mu_2]$. To this end, we consider two sub-regimes for the fractional cache size μ namely (i) the *intermediate cache regime 1* with $\mu \in [\mu_1, 1/M]$; and (ii) the *intermediate cache regime 2* with $\mu \in [1/M, \mu_2]$. We consider each of the two regimes separately.

Intermediate Cache Regime 1 ($\mu \in [\mu_1, 1/M]$): For this regime, considering the achievable NDT presented in (C.91), we have the upper bound

$$\begin{aligned}\delta_{\mathsf{P},\mathsf{Ach}}(\mu, r) &\leq \max\left\{\frac{(1-\mu M)K}{Mr}\,,\ \frac{(1-\mu M)K}{\min\{M, K\}} + \mu(M + K - 1)\right\} \\ &= \frac{(1-\mu M)K}{\min\{M, K\}} + \mu(M + K - 1),\end{aligned} \tag{C.102}$$

since the edge latency i.e., the second term inside the $\max(\cdot)$ function, dominates when $\mu \geq \mu_1$ (see Appendix C.9.1 for details). A lower bound is given by (C.100). Using the mentioned upper and lower bounds on the minimum NDT, we have

$$\begin{aligned}\frac{\delta_{\mathsf{P},\mathsf{Ach}}(\mu, r)}{\delta^*_{\mathsf{P}}(\mu, r)} &\leq \left[\frac{(1-\mu M)K}{\min\{M, K\}} + \mu(M + K - 1)\right] \times \frac{\min\{M, K\}}{K} \\ &= (1-\mu M) + \mu M\left[\frac{\min\{M, K\}(M + K - 1)}{KM}\right]\end{aligned}$$

$$\leq (1-\mu M) + \mu M \left[\frac{M+K}{\max\{M,K\}}\right] = (1-\mu M) + \mu M \left[1 + \frac{\min\{M,K\}}{\max\{M,K\}}\right]$$
$$\leq (1-\mu M) + 2\mu M \leq 1 + \mu M$$
$$\overset{(a)}{\leq} 2, \tag{C.103}$$

where step (a) follows by using $\mu \leq 1/M$.

Intermediate Cache Regime 2 ($\mu \in [1/M, \mu_2]$): For this regime, considering the achievable NDT presented in Theorem 29, we have the upper bound

$$\delta_{\mathsf{P,Ach}}(\mu, r) \leq \delta_{\mathsf{P,Ach}}(\mu_1, r) = \frac{(1-\mu_1 M)K}{Mr} \leq \frac{M+K-1}{M}, \tag{C.104}$$

for any $M, K \geq 1$ and $r > 0$, and where the first inequality follows since the NDT is a non-decreasing function of the cache size μ. Again, for this regime, considering the lower bound in (C.100), we have

$$\frac{\delta_{\mathsf{P,Ach}}(\mu, r)}{\delta^*_{\mathsf{P}}(\mu, r)} \leq \frac{M+K-1}{M} \times \frac{\min\{M,K\}}{K}$$
$$\leq \frac{M+K}{\max\{M,K\}} = 1 + \frac{\min\{M,K\}}{\max\{M,K\}} \leq 2. \tag{C.105}$$

Finally combining (C.103) and (C.105) concludes the proof of Theorem 31.

C.9.4 Proof of Corollary 9

Using $M = K = 2$ in (5.60), we obtain the minimum NDT

$$\delta^*_{\mathsf{P}}(\mu, r) = \begin{cases} \dfrac{1-2\mu}{r}, & \text{for} \quad \mu \in [0, \mu_1 = (1-r)/(2+r)] \\ 1, & \text{for} \quad \mu \in [\mu_2 = (1-r), 1]. \end{cases} \tag{C.106}$$

For the remaining intermediate cache regime with $\mu \in [\mu_1, \mu_2]$, we adopt the achievable NDT $\delta_{\mathsf{P-FS}}$ given in (5.58), which yields the upper bound

$$\delta_{\mathsf{P,Ach}}(\mu, r) \leq \frac{2-\mu}{1+r}, \quad \text{for} \quad \mu \in [\mu_1, \mu_2], \tag{C.107}$$

and the lower bound in Corollary 8 with $\ell = 1$ which can be seen to match (C.107). In the high fronthaul regime, i.e., $r \geq 1$, using $M = K = 2$ in (5.61) yields the minimum NDT. This concludes the proof.

Appendix D

Proofs From Chapter 6

Learning-Aided Collaborative Caching

D.1 Proof of Lemma 8

The proof is based on the proof of the upper bound on CMAB regret presented in [213]. First, we present the Chernoff-Hoeffding Lemma which is integral to the proof.

Lemma 15 (Chernoff-Hoeffding Bound)**.** *Let $X_1, X_2, \ldots, X_n$ be random variables with common support $[0,1]$ and $\mathbb{E}[X_i] = \theta$. Let $S_n = \sum_{i=1}^{n} X_i$. Then for all $t \geq 0$, we have*

$$\Pr\left[S_n \geq n\theta + t\right] \leq e^{-2t^2/n} \text{and } \Pr\left[S_n \leq n\theta + t\right] \leq e^{-2t^2/n}.$$

D.1.1 Popularity Estimation Process

Let $T_{f,n}^t$ be the value of the variable $T_{f,n}$ after t rounds i.e., the number of times file f is cached at $\mathcal{S}_n$ after t rounds is $T_{f,n}^t$. Also let $\widehat{\theta}_{f,n,s}$ be the value of the variable $\widehat{\theta}_{f,n}$ after the file $f \in \mathbb{F}$ has been cached s times. The estimation of popularity profile is assumed to be *well behaved* if, at time t, the empirical popularity estimate $\widehat{\theta}_{f,n}$ is close to the actual expectation $\theta_{f,n}$ i.e., if

$$|\widehat{\theta}_{f,n,T_{f,n}^{t-1}} - \theta_{f,n}| < \sqrt{\frac{\Psi_{f,n} \log U_n t}{2T_{f,n}^{t-1}}}, \ \ \forall f \in \mathbb{F} \tag{D.1}$$

where $U_n = |\mathcal{U}(\mathcal{S}_n)|$. Let $\mathcal{P}_t$ be the event such that it is true when the estimation process satisfies (D.1). Using Chernoff-Hoeffding Bound, for any $f \in \mathbb{F}$, we have:

$$\Pr\left[|\widehat{\theta}_{f,n,T_{f,n}^{t-1}} - \theta_{f,n}| \geq \sqrt{\frac{\Psi_{f,n} \log U_n t}{2T_{f,n}^{t-1}}}\right] = \sum_{s=1}^{t-1} \Pr\left[\left\{|\widehat{\theta}_{f,n,s} - \theta_{f,n}| \geq \sqrt{\frac{\Psi_{f,n} \log U_n t}{2s}}, T_{f,n}^{t-1} = s\right\}\right]$$

$$\leq \sum_{s=1}^{t-1} \Pr\left[\left\{|\widehat{\theta}_{f,n,s} - \theta_{f,n}| \geq \sqrt{\frac{\Psi_{f,n}\log U_n t}{2s}}\right\}\right]$$
$$\leq 2te^{-\Psi_{f,n}\log U_n t} = 2U_n^{-\Psi_{f,n}} t^{1-\Psi_{f,n}}. \tag{D.2}$$

Taking an union bound on f, we have

$$\Pr\left[\mathcal{P}_t\right] = \Pr\left[\forall f \in \mathbb{F}, |\widehat{\theta}_{f,n,T_{f,n}^{t-1}} - \theta_{f,n}| < \sqrt{\frac{\Psi_{f,n}\log U_n t}{2T_{f,n}^{t-1}}}\right] \leq 1 - 2FU_n^{-\Psi_{f,n}} t^{1-\Psi_{f,n}}. \tag{D.3}$$

The value of $\widehat{\theta}_{f,n}$ at the end of time step t is $\widehat{\theta}_{f,n,T_{f,n}^t}$ since by definition file f is cached $T_{f,n}^t$ times after t time steps. Also, for $\overline{\theta}_{f,n}$, let $\overline{\theta}_{f,n,t}$ be its value after t rounds and $\overline{\Theta}_{n,t} = [\overline{\theta}_{1,n,t}, \overline{\theta}_{2,n,t}, \ldots, \overline{\theta}_{F,n,t}]$ be the input to the CCP solver at round t. Then, from Algorithm 4, we have:

$$\overline{\theta}_{f,n,t} = \widehat{\theta}_{f,n,T_{f,n}^{t-1}} + \sqrt{\frac{\Psi_{f,n}\log\left(U_n t\right)}{2T_{f,n}^{t-1}}}. \tag{D.4}$$

D.1.2 An α-sub-optimal Caching Strategy

A caching strategy $\mathbf{c}_n^\pi(t)$ is defined to be α-*sub-optimal* if the reward obtained by the strategy is less than α fraction of the optimal reward:

$$R_{\Theta_n}(\mathbf{c}_n^\pi(t)) < \alpha \cdot R_{\Theta_n}^{\text{opt}}.$$

The set of all α-sub-optimal caching strategies for $\mathcal{S}_n$ is denoted by:

$$\mathbb{C}_n^B = \{\mathbf{c}_n^\pi(t) | R_{\Theta_n}(\mathbf{c}_n^\pi(t)) < \alpha \cdot R_{\Theta_n}^{\text{opt}}\}.$$

Let the set of all K_f α-sub-optimal caching strategies at $\mathcal{S}_n$ in which the file $f \in \mathbb{F}$ is cached be denoted by

$$\mathbb{C}_{f,n}^B = \{C_{f,n}^{B,i}, \forall i = 1, \ldots, K_f\}$$

Without loss of generality, the strategies $C_{f,n}^{B,i}$ are reordered in increasing order of expected rewards $C_{f,n}^{B,1}, C_{f,n}^{B,2}, \ldots, C_{f,n}^{B,K_f}$, such that $C_{f,n}^{B,K_f}$ is strategy which yields the best expected reward. We define:

$$\Delta_n^{f,j} = \alpha \cdot R_{\Theta_n}^{\text{opt}} - R_{\Theta_n}(C_{f,n}^{B,j}) \tag{D.5}$$

Based on this, we further define:

$$\Delta_{n,\max}^f = \Delta_n^{f,1} \text{ and } \Delta_{n,\min}^f = \Delta_n^{f,K_f}.$$

For any underlying arm (file) $f \in \mathbb{F}$, similar to [213], we define:

$$\Delta_{\text{max}} = \max_{f \in \mathbb{F}} \left[\alpha \cdot \mathsf{R}^{\text{opt}}_{\Theta_n} - \min \left\{ \mathsf{R}_{\Theta_n}(\mathbf{c}_n) | \mathbf{c}_n \in \mathbb{C}_B, f \in \mathbf{c}_n \right\} \right] \tag{D.6}$$

$$\Delta_{\text{min}} = \max_{f \in \mathbb{F}} \left[\alpha \cdot \mathsf{R}^{\text{opt}}_{\Theta_n} - \max \left\{ \mathsf{R}_{\Theta_n}(\mathbf{c}_n) | \mathbf{c}_n \in \mathbb{C}_B, f \in \mathbf{c}_n \right\} \right]. \tag{D.7}$$

Intuitively, Δ_{max} is the difference in expected reward between the optimal and the reward for playing the worst α-*sub-optimal* super-arm, while Δ_{min} is difference with the optimal for the case of playing the best possible α-*sub-optimal* super-arm. If all arms (files) are sampled sufficiently w.r.t $\Delta_{\min}$, then the sample means $\widehat{\theta}_{f,n}$ are close to their true means and the probability of the algorithm playing a α-*sub-optimal* super-arm $\mathbf{c}_n \in \mathbb{C}_B$ is low. However, if the sampling is insufficient, then we incur a regret proportional to $\Delta_{\max}$.

D.1.3 Sampling of Cached Files

For the proof, at each sBS $\mathcal{S}_n$, a counter N_f is maintained for each file $f \in \mathbb{F}$ after the F initial rounds (when each file is sequentially placed in the cache to initialize the algorithm). Let $N_{f,t}$ be the value of the counter after t time instants. Thus $N_{f,F} = 1$ and $\sum_{f \in \mathbb{F}} N_{f,F} = F$. The counters are updated as follows: For any instant $t > F$, if $\mathbf{c}^{\pi}_n(t) \in \mathbb{C}^B_n$, then $f^* = \arg\min_{j \in \mathbf{c}^{\pi}_n(t)} N_{j,t-1}$ and we increment the counter N_{f^*} by 1 i.e., $N_{f^*,t} = N_{f^*,t-1} + 1$. If f^* is not unique, any random file f with the smallest counter in $\mathbf{c}^{\pi}_n(t)$ is picked and its counter is incremented. In every α-sub-optimal caching round, exactly one counter is incremented. The counter N_f is further sub-divided into counters $\{N^l_f\}_{l=1}^{K_f}$, whose value at a round $t = T$ is defined as

$$N^l_{f,T} = \sum_{t=F+1}^{T} \mathbb{I}\{\mathbf{c}^{\pi}_n(t) = C^{B,l}_{f,n}, N_{f,t} > N_{f,t-1}\}, \ \forall l \in [K_f], \tag{D.8}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function such that $\mathbb{I}\{x\} = 1$ iff x is true. We define

$$\ell_T(\Delta) = \frac{2\Psi_{f,n} \log U_n T}{(g^{-1}(\Delta))^2}, \tag{D.9}$$

which is the number of samplings that is considered *sufficient* for a caching strategy which yields a reward Δ away from the $\alpha-$approximation w.r.t time horizon T. If $N_{f,t-1} > \ell_T(\Delta^{f,l})$, the α-sub-optimal strategy $C^{B,l}_{f,n}$ is *sufficiently sampled*. Otherwise it is *under-sampled*. Thus, we have the number of sufficiently sampled and under-sampled periods

$$N^{l,\text{suf}}_{f,T} = \sum_{t=F+1}^{T} \mathbb{I}\{\mathbf{c}^{\pi}_n(t) = C^{B,l}_{f,n}, N_{f,t} > N_{f,t-1} > \ell_T(\Delta^{f,l}_n)\} \tag{D.10}$$

$$N^{l,\text{und}}_{f,T} = \sum_{t=F+1}^{T} \mathbb{I}\{\mathbf{c}^{\pi}_n(t) = C^{B,l}_{f,n}, N_{f,t} > N_{f,t-1}, N_{f,t-1} < \ell_T(\Delta^{f,l}_n)\} \tag{D.11}$$

Thus we have,

$$N_{f,T} = 1 + \sum_{l \in [K_f]} (N_{f,T}^{l,\text{suf}} + N_{f,T}^{l,\text{und}}). \tag{D.12}$$

The total reward at round T is at least

$$T\alpha R_{\Theta_n}^{\text{opt}} - \mathbb{E}\left[\sum_{f \in \mathbb{F}} \left\{ \Delta_n^{f,1} + \sum_{l \in [K_f]} (N_{f,T}^{l,\text{suf}} + N_{f,T}^{l,\text{und}}) \cdot \Delta_n^{f,l} \right\}\right] \tag{D.13}$$

D.1.4 Upper Bound on $N_{f,T}^{l,\text{suf}}$

Define $\Lambda_{f,t} = \sqrt{\frac{\Psi_{f,n}\log(U_n t)}{2T_{f,n}^{t-1}}}$ which is a random variable since $T_{f,n}^{t-1}$ is a random variable. We have $\Lambda_t = \max\{\Lambda_{f,t} | f \in \mathbf{c}_n^\pi(t)\}$. Further, we define $\Lambda^{f,l} = \sqrt{\frac{\Psi_{f,n}\log(U_n t)}{2\ell_T(\Delta_n^{f,l})}}$. Let $\mathcal{B}_t$ be the event that the $(\alpha,\beta)-$approximation solver running in the CCP fails to produce an $\alpha-$approximate solution w.r.t $\overline{\Theta}_{n,t}$ in round t. Also, let $\neg\mathcal{B}_t$ be the event when an the solver does produce an $\alpha-$approximate solution. Thus, by the definition of the $(\alpha,\beta)-$approximation solver, we have:

$$\text{Pr}(\mathcal{B}_t) = \mathbb{E}[\mathbb{I}\{\mathcal{B}_t\}] \leq 1 - \beta.$$

Based on (D.4), the following holds true:

$$\mathcal{P}_t \Rightarrow \overline{\theta}_{f,n,t} - \theta_{f,n} > 0,\ \forall f \in \mathbb{F} \tag{D.48}$$

$$\mathcal{P}_t \Rightarrow \overline{\theta}_{f,n,t} - \theta_{f,n} < 2\Lambda_{f,t},\ \forall f \in \mathbf{c}_n^\pi(t) \tag{D.49}$$

Again, we have $\forall f \in \mathbb{F},\ \forall l \in [K_f]$, and $\forall f^{'} \in \mathbf{c}_n^\pi(t)$:

$$\left\{\mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}, N_{f,t} > N_{f,t-1}, N_{f^{'},t-1} > \ell_T(\Delta_n^{f,l})\right\} \Rightarrow \Lambda^{f,l} > \Lambda_{f,t}. \tag{D.50}$$

For any $f \in \mathbb{F}, l \in [K_f]$, if

$$\left\{\mathcal{P}_t, \neg\mathcal{B}_t, \mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}, N_{f,t} > N_{f,t-1}, N_{f^{'},t-1} > \ell_T(\Delta_n^{f,l})\right\}$$

holds at time t, the following holds true:

$$\begin{aligned} R_{\Theta_n}(\mathbf{c}_n^\pi(t)) + g(2\Lambda^{f,l}) &> R_{\Theta_n}(\mathbf{c}_n^\pi(t)) + g(2\Lambda_{f,t}) \\ &\geq R_{\overline{\Theta}_{n,t}}(\mathbf{c}_n^\pi(t)) \geq \alpha R_{\overline{\Theta}_{n,t}}^{\text{opt}} \geq \alpha R_{\Theta_n}^{\text{opt}}, \end{aligned} \tag{D.51}$$

where the first inequality follows from the strict monotonicity of $g(\cdot)$ and (D.50), the second follows from the bounded smoothness property and (D.49). The third inequality is true since $\neg\mathcal{B}_t \Rightarrow \mathbf{c}_n^\pi(t)$ is an $\alpha-$approximation w.r.t $\overline{\Theta}_{n,t}$. Thus we have

$$R_{\Theta_n}(C_{f,n}^{B,l}) + g(2\Lambda^{f,l}) > \alpha R_{\Theta_n}^{\text{opt}} \tag{D.52}$$

Now,

$$\ell_T(\Delta_n^{f,l}) = \frac{2\Psi_{f,n}\log U_n T}{(g^{-1}(\Delta_n^{f,l}))^2}$$

and we have

$$2\Lambda^{f,l} = g^{-1}(\Delta_n^{f,l})\sqrt{\frac{\log t}{\log T}} \Rightarrow g(2\Lambda^{f,l}) \leq \Delta_n^{f,l}.$$

Thus, (D.51) contradicts (D.5) i.e., $\forall f \in \mathbb{F}, l \in [K_f]$ we have

$$\Pr\Big\{\mathcal{P}_t, \neg\mathcal{B}_t, \mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}, N_{f,t} > N_{f,t-1}, \forall f^{'} \in \mathbf{c}_n^\pi(t), N_{f^{'},t-1} > \ell_T(\Delta_n^{f,l})\Big\} = 0, \quad \text{(D.53)}$$

$$\Rightarrow \quad \Pr\left\{\begin{matrix}\mathcal{P}_t, \neg\mathcal{B}_t, \exists f \in \mathbb{F}, \exists l \in [K_f], \mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}, \\ N_{f,t} > N_{f,t-1}, \forall f^{'} \in \mathbf{c}_n^\pi(t), N_{f^{'},t-1} > \ell_T(\Delta_n^{f,l})\end{matrix}\right\} = 0 \quad \text{(D.54)}$$

$$\Rightarrow \quad \sum_{f\in\mathbb{F}, l\in[K_f]} \Pr\left\{\begin{matrix}\mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}, N_{f,t} > N_{f,t-1}, \\ \forall f^{'} \in \mathbf{c}_n^\pi(t), N_{f^{'},t-1} > \ell_T(\Delta_n^{f,l})\end{matrix}\right\} \leq \Pr\{\neg\mathcal{P}_t \vee \mathcal{B}_t\}$$

$$\leq (1-\beta) + 2FU_n^{-\Psi_{f,n}} t^{1-\Psi_{f,n}} \quad \text{(D.55)}$$

Thus by the definition of $N_{f,T}^{l,\text{suf}}$, we have

$$\mathbb{E}\left[\sum_{f\in\mathbb{F}, l\in[K_f]} N_{f,T}^{l,\text{suf}}\right] \leq \sum_{t=1}^{T}\left[(1-\beta) + 2FU_n^{-\Psi_{f,n}} t^{1-\Psi_{f,n}}\right]$$

$$= (1-\beta)T + 2F\frac{\zeta\left(\Psi_{f,n}-1\right)}{U_n^{\Psi_{f,n}}}, \quad \text{(D.56)}$$

where we have used $\zeta\left(\Psi_{f,n}-1\right) = \sum_{t=1}^{T} t^{1-\Psi_{f,n}}$ and $\zeta(\cdot)$ is Riemann Zeta function.

D.1.5 Upper Bound on $N_{f,T}^{l,\textbf{und}}$

Consider α-sub-optimal caching strategies $\mathbf{c}_n^\pi(t) \in \mathbb{C}_n^B$ which are under-sampled when played (i.e., $f \in \mathbf{c}_n^\pi(t)$ are cached). The counter N_f for a file f increases from 1 to $\ell_T(\Delta_n^{f,K_f})$. The range of counters N_f can be broken into segments $(\ell_T(\Delta_n^{f,j-1}), \ell_T(\Delta_n^{f,j})]$ for $j \in [K_f]$. Assume that for a file f, $N_{f,t-1} \in (\ell_T(\Delta_n^{f,j-1}), \ell_T(\Delta_n^{f,j})]$ for some j. In a *α-sub-optimal* round t, for $\mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}$ for some $l > j$, the regret suffered $\Delta_n^{f,l} < \Delta_n^{f,j}$. Thus for the counter $N_{f,t}$ in the range $(\ell_T(\Delta_n^{f,j-1}), \ell_T(\Delta_n^{f,j})]$, the total regret for the under-sampled files is at most $(\ell_T(\Delta_n^{f,j}) - \ell_T(\Delta_n^{f,j-1}))\cdot\Delta_n^{f,j})$ in the rounds the counter $N_{f,t}$ is incremented. These are used in the following. For any file $\{f \in \mathbb{F} | \Delta_{n,\min}^f > 0\}$, we have

$$\sum_{l\in[K_f]} N_{f,T}^{l,\text{und}} \cdot \Delta_n^{f,l} = \sum_{t=F+1}^{T} \sum_{l\in[K_f]} \mathbb{I}\{\mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}, N_{f,t} > N_{f,t-1}, N_{f,t-1} < \ell_T(\Delta_n^{f,l})\} \cdot \Delta_n^{f,l} \tag{D.57}$$

$$= \sum_{t=F+1}^{T} \sum_{l\in[K_f]} \sum_{j=1}^{l} \mathbb{I}\{\mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}, N_{f,t} > N_{f,t-1}, N_{f,t-1} \in (\ell_T(\Delta_n^{f,j-1}), \ell_T(\Delta_n^{f,j})]\} \cdot \Delta_n^{f,l} \tag{D.58}$$

$$\leq \sum_{t=F+1}^{T} \sum_{l\in[K_f]} \sum_{j\in[K_f]} \mathbb{I}\{\mathbf{c}_n^\pi(t) = C_{f,n}^{B,l}, N_{f,t} > N_{f,t-1}, N_{f,t-1} \in (\ell_T(\Delta_n^{f,j-1}), \ell_T(\Delta_n^{f,j})]\} \cdot \Delta_n^{f,j} \tag{D.59}$$

$$\leq \sum_{j\in[K_f]} \sum_{t=F+1}^{T} \mathbb{I}\{\mathbf{c}_n^\pi(t) \in \mathbb{C}_{f,n}^{B}, N_{f,t} > N_{f,t-1}, N_{f,t-1} \in (\ell_T(\Delta_n^{f,j-1}), \ell_T(\Delta_n^{f,j})]\} \cdot \Delta_n^{f,j} \tag{D.60}$$

$$\leq \sum_{j\in[K_f]} \left[\ell_T(\Delta_n^{f,j}) - \ell_T(\Delta_n^{f,j-1})\right] \cdot \Delta_n^{f,j} = \ell_T(\Delta_n^{f,K_f})\Delta_n^{f,K_f} + \sum_{j=1}^{K_f-1} \ell_T\left[\Delta_n^{f,j})(\Delta_n^{f,j} - \Delta_n^{f,j+1}\right] \tag{D.61}$$

$$\leq \ell_T(\Delta_{n,\min}^f) \cdot \Delta_{n,\min}^f + \ell_T(\Delta_{n,\min}^f) \left[\Delta_{n,\max}^{f,K_f} - \Delta_{n,\min}^{f,K_f}\right] = \ell_T(\Delta_{n,\min}^f) \cdot \Delta_{n,\max}^f. \tag{D.62}$$

The last inequality (D.62) follows from the definitions of $\Delta_{n,\min}^f$ and $\Delta_{n,\max}^f$ and the fact that $\ell_T(\Delta)$ is a decreasing function of Δ.

Combining (D.13), (D.56) and (D.62), and substituting in (6.10), we have

$$\begin{aligned}
\text{Reg}_{\Theta_n,\alpha,\beta}^{\pi}(T) &= T\alpha\beta R_{\Theta_n}^{\text{opt}} - \left(T\alpha R_{\Theta_n}^{\text{opt}} - \mathbb{E}\left[\sum_{f\in\mathbb{F}} \left(\Delta_n^{f,l} + \sum_{l\in K_f} \left(N_{f,T}^{l,\text{suf}} + N_{f,T}^{l,\text{und}}\right) \cdot \Delta_n^{f,l}\right)\right]\right) \\
&\leq \left(F + \mathbb{E}\left[\sum_{f\in\mathbb{F}, l\in[K_f]} N_{f,T}^{l,\text{suf}}\right]\right) \cdot \Delta_{n,\max} + \sum_{\substack{f\in\mathbb{F}, \\ \Delta_{n,\min}^f > 0}} \left(\ell_T(\Delta_{n,\min}^f) \cdot \Delta_{n,\max}^f\right) - (1-\beta) \cdot T\alpha R_{\Theta_n}^{\text{opt}} \\
&\leq \left(\frac{2\zeta(\Psi_{f,n}-1)}{U_n^{\Psi_{f,n}}} + 1\right) \cdot F \cdot \Delta_{n,\max} + \sum_{f\in\mathbb{F}, \Delta_{n,\min}^f > 0} \frac{2\Psi_{f,n} \log U_n t}{(g^{-1}(\Delta_{n,\min}^f))^2} \cdot \Delta_{n,\max}^f
\end{aligned} \tag{D.63}$$

which completes the proof of the Lemma. □

Bibliography

[1] "Quality of Service (QoS) Concept and Architecture," *3GPP 23.107 V7.4.0*, June 2006. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/23107.htm

[2] "E-UTRA and E-UTRAN Overall description; Stage 2 (Release 8)," *3GPP Technical Specification TS 36.300 V8.7.0*, Dec. 2008.

[3] "Requirements for Further Advancements for E-UTRA (LTE-Advanced) (Release 8)," *3GPP Technical Specification TR 36.913 V8.0.0*, Jun 2008.

[4] "Radio resource control (RRC); Protocol Specification, (Release 8)," *3GPP, Tech. Rep. Ts 25.331*, Apr 2008.

[5] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[6] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[7] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfrojd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive mimo in 5g," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 44–51, May 2014.

[8] S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From Theory to Practice*, ser. Wiley InterScience online books. Wiley, 2009.

[9] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005, cambridge Books Online. [Online]. Available: http://dx.doi.org/10.1017/CBO9780511841224

[10] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[11] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta, "Are we approaching the fundamental limits of wireless network densification?" *arXiv:1512.00413*, 2015. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1512/1512.00413.pdf

[12] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. Andrews, P. Xia, H. Jo, H. Dhillon, and T. Novlan, "Heterogeneous cellular networks: From theory to practice," *Communications Magazine, IEEE*, vol. 50, no. 6, pp. 54–64, June 2012.

[13] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, April 2012.

[14] H. Dhillon, M. Kountouris, and J. Andrews, "Downlink coverage probability in mimo hetnets," in *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Nov 2012, pp. 683–687.

[15] N. Saquib, E. Hossain, and D. I. Kim, "Fractional frequency reuse for interference management in lte-advanced hetnets," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 113–122, 2013.

[16] T. Novlan, J. Andrews, I. Sohn, R. Ganti, and A. Ghosh, "Comparison of fractional frequency reuse approaches in the ofdma cellular downlink," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec. 2010, pp. 1 –5.

[17] A. Kumar, A. Sengupta, R. Tandon, and T. C. Clancy, "Dynamic Resource Allocation for Cooperative Spectrum Sharing in LTE Networks," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2014.

[18] 3GPP, "Technical Specification Group Services and System Aspects - Feasibility study for Proximity Services (ProSe) (Release 12)," *3GPP TR 22.803 V12.1.0*, March 2013.

[19] Qualcomm Research, "LTE Direct - The Case for Device-to-Device Proximate Discovery," Tech. Rep., Feb 2013.

[20] C.-H. Yu, K. Doppler, C. Ribeiro, and O. Tirkkonen, "Performance impact of fading interference to device-to-device communication underlaying cellular networks," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sept 2009, pp. 858–862.

[21] P. Janis, V. Koivunen, C. Ribeiro, J. Korhonen, K. Doppler, and K. Hugl, "Interference-aware resource allocation for device-to-device radio underlaying cellular networks," in *Proc. IEEE Vehicular Technology Conference (VTC-Spring 2009)*, April 2009, pp. 1–5.

[22] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, and B. Jiao, "Interference-aware resource allocation for device-to-device communications as an underlay using sequential second price auction," in *Proc. IEEE International Conference on Communications (ICC)*, June 2012, pp. 445–449.

[23] S. Xu, H. Wang, T. Chen, Q. Huang, and T. Peng, "Effective interference cancellation scheme for device-to-device communication underlaying cellular networks," in *Proc. IEEE Vehicular Technology Conference Fall (VTC 2010-Fall)*, Sept 2010, pp. 1–5.

[24] W. Xu, L. Liang, H. Zhang, S. Jin, J. Li, and M. Lei, "Performance enhanced transmission in device-to-device communications: Beamforming or interference cancellation?" in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec 2012, pp. 4296–4301.

[25] N. Golrezaei, A. Molisch, and A. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," in *Proc. IEEE International Conference on Communications (ICC)*, June 2012, pp. 7077–7081.

[26] N. Golrezaei, M. Ji, A. Molisch, A. Dimakis, and G. Caire, "Device-to-device communications for wireless video delivery," in *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Nov 2012, pp. 930–933.

[27] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3665–3676, 2014.

[28] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Communications*, vol. 19, no. 3, pp. 96–104, 2012.

[29] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Communications Surveys Tutorials*, 2014.

[30] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: a new architecture for wireless communications," in *Proc. IEEE International Conference on Computer Communications(INFOCOM)*, vol. 3, Mar 2000, pp. 1273–1282 vol.3.

[31] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "Flashlinq: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Transactions on Networking*, vol. 21, no. 4, pp. 1215–1228, Aug 2013.

[32] "Spectrum policy task force report," *FCC*, no. 02-155, 2002.

[33] D. Arianandа, M. K. Lakshmanan, and H. Nikookar, "A survey on spectrum sensing techniques for cognitive radio," in *Proc. International Workshop on Cognitive Radio and Advanced Spectrum Management*, May 2009, pp. 74–79.

[34] J. Mitola III and G. Maguire Jr., "Cognitive radio: making software radios more personal," *Personal Communications, IEEE*, vol. 6, no. 4, pp. 13 –18, Aug. 1999.

[35] J. Mitola III, "Software radios-survey, critical evaluation and future directions," in *Proc. National Telesystems Conference*, May 1992, pp. 13/15 –13/23.

[36] T. Clancy, "Achievable capacity under the interference temperature model," in *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, May 2007, pp. 794 –802.

[37] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Communications Surveys and Tutorials*, vol. 11, no. 1, pp. 116–130, 2009.

[38] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Physical Communications*, vol. 4, no. 1, pp. 40–62, Mar 2011.

[39] R. Tandra and A. Sahai, "Snr walls for signal detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 4 –17, Feb. 2008.

[40] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *Signal Processing Magazine, IEEE*, vol. 24, no. 3, pp. 79–89, May.

[41] Y. Zhao, D. Raymond, C. da Silva, J. Reed, and S. Midkiff, "Performance evaluation of radio environment map-enabled cognitive spectrum-sharing networks," in *Proc. IEEE Military Communications Conference (MILCOM)*, Oct 2007, pp. 1–7.

[42] V. Osa, C. Herranz, J. Monserrat, and X. Gelabert, "Implementing opportunistic spectrum access in lte-advanced," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012:99, no. 1, 2012.

[43] C. Santivanez, R. Ramanathan, C. Partridge, R. Krishnan, M. Condell, and S. Polit, "Opportunistic spectrum access: Challenges, architecture, protocols," in *Proc. International Workshop on Wireless Internet*, no. 13. New York, NY, USA: ACM, 2006.

[44] L. M. Lopez-Ramos, A. G. Marqués, and J. Ramos, "Jointly optimal sensing and resource allocation for multiuser overlay cognitive radios," *IEEE Journal on Selected Areas in Communications (submitted)*, 2012.

[45] T. Doumi, M. Dolan, S. Tatesh, A. Casati, G. Tsirtsis, K. Anchan, and D. Flore, "LTE for public safety networks," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 106–112, 2013.

[46] A. Padaki, A. Sengupta, M. Adbelbar, J. H. Reed, and W. H. Tranter, "An Orthogonal Spectrum Sharing Scheme for Cognitive LTE Networks," in *Proc. Wireless Innovation Forum Conference on Wireless Communications Technologies and Software Defined Radio (SDR-WInnComm)*. Wireless Innovation Forum, 2014, pp. 50–59.

[47] J. Peha, "Sharing spectrum through spectrum policy reform and cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 708–719, April 2009.

[48] R. Menon, R. Buehrer, and J. Reed, "Outage probability based comparison of underlay and overlay spectrum sharing techniques," in *Proc. IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Nov 2005, pp. 101–109.

[49] ——, "On the impact of dynamic spectrum sharing techniques on legacy radio systems," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 11, pp. 4198–4207, November 2008.

[50] H. Shajaiah, A. Abdel-Hadi, and C. Clancy, "Spectrum sharing between public safety and commercial users in 4g-lte," in *Proc. IEEE International Conference on Computing, Networking and Communications (ICNC)*, Feb 2014, pp. 674–679.

[51] A. Khawar, A. Abdel-Hadi, and T. Clancy, "Spectrum sharing between s-band radar and lte cellular system: A spatial approach," in *Proc. International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, April 2014, pp. 7–14.

[52] F. Fund, S. Shahsavari, S. S. Panwar, E. Erkip, and S. Rangan, "Spectrum and infrastructure sharing in millimeter wave cellular networks: An economic perspective," *arXiv: 1605.04602*, 2016. [Online]. Available: http://arxiv.org/abs/1605.04602

[53] L. Anchora, L. Badia, E. Karipidis, and M. Zorzi, "Capacity gains due to orthogonal spectrum sharing in multi-operator lte cellular networks," in *Proc. IEEE International Symposium on Wireless Communication Systems (ISWCS)*, 2012, pp. 286–290.

[54] B. Aazhang, J. Lilleberg, and G. Middleton, "Spectrum sharing in a cellular system," in *Proc. IEEE International Symposium on Spread Spectrum Techniques and Applications*, 2004, pp. 355–359.

[55] M. Pereirasamy, J. Luo, M. Dillinger, and C. Hartmann, "Dynamic inter-operator spectrum sharing for umts fdd with displaced cellular networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 3, 2005, pp. 1720–1725 Vol. 3.

[56] G. Middleton, K. Hooli, A. Tolli, and J. Lilleberg, "Inter-operator spectrum sharing in a broadband cellular network," in *Proc. IEEE International Symposium on Spread Spectrum Techniques and Applications*, 2006, pp. 376–380.

[57] M. Bennis and J. Lilleberg, "Inter base station resource sharing and improving the overall efficiency of b3g systems," in *Proc. IEEE Vehicular Technology Conference (VTC 2007-Fall)*, 2007, pp. 1494–1498.

[58] E. Karipidis, D. Gesbert, M. Haardt, K.-M. Ho, E. Jorswieck, E. G. Larsson, J. Li, J. Lindblom, C. Scheunert, M. Schubert *et al.*, "Transmit beamforming for inter-operator spectrum sharing," in *Proc. Future Network & Mobile Summit (FutureNetw)*. IEEE, 2011, pp. 1–8.

[59] F. Mazzenga, M. Petracca, R. Pomposini, F. Vatalaro, and R. Giuliano, "Performance evaluation of spectrum sharing algorithms in single and multi operator scenarios," in *Proc. IEEE Vehicular Technology Conference (VTC 2011-Spring)*, 2011, pp. 1–5.

[60] B. Mehdi, L. Samson, and D. Merouane, "Inter-operator spectrum sharing from a game theoretical perspective," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[61] Y. Wu, B. Wang, K. Liu, and T. Clancy, "Repeated open spectrum sharing game with cheat-proof strategies," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 4, pp. 1922–1933, April 2009.

[62] F. C. Tom Wheeler, "Innovation in the 3.5 ghz band: Creating a new citizens broadband radio service," 2015. [Online]. Available: http://www.fcc.gov/blog/innovation-35-ghz-band-creating-new-citizens-broadband-radio-service

[63] Qualcom, "R13-laa (license-assisted-access)," 2015. [Online]. Available: https://www.qualcomm.com/invention/research/projects/lte-unlicensed/r13-laa-licensed-assisted-access

[64] Ericsson, "Lte license assisted access," 2015. [Online]. Available: http://www.ericsson.com/res/thecompany/docs/press/media_kits/ericsson-license-assisted-access-laa-january-2015.pdf

[65] Q. Research, "Lte in unlicensed spectrum," Qualcomm Technologies Inc., Tech. Rep., 2014. [Online]. Available: https://www.qualcomm.com/media/documents/files/lte-unlicensed-coexistence-whitepaper.pdf

[66] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101–107, June 2011.

[67] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. H. Jr., "Modeling and analyzing millimeter wave cellular systems," *arXiv: 1605.04283*, 2016. [Online]. Available: http://arxiv.org/abs/1605.04283

[68] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5g cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[69] R. G. Fellers, "Millimeter waves and their applications," *Electrical Engineering*, vol. 75, no. 10, pp. 914–917, Oct 1956.

[70] Qualcomm, "Exploring the potential of mmwave 5g mobile access," Tech. Rep., June 2016. [Online]. Available: https://www.qualcomm.com/documents/heavyreading-whitepaper-exploring-potential-mmwave-5g-mobile-access

[71] M. Al-Ayyoub, M. Buddhikot, and H. Gupta, "Self-regulating spectrum management: A case of fractional frequency reuse patterns in lte networks," in *Proc. IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, April 2010, pp. 1–12.

[72] A. Subramanian, M. Al-Ayyoub, H. Gupta, S. Das, and M. Buddhikot, "Near-optimal dynamic spectrum allocation in cellular networks," in *Proc. IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Oct 2008, pp. 1–11.

[73] R. Kwan, C. Leung, and J. Zhang, "Resource allocation in an lte cellular communication system," in *IEEE International Conference on Communications (ICC)*, 2009, pp. 3915–3919.

[74] S. Ali and V. Leung, "Dynamic frequency allocation in fractional frequency reused ofdma networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4286 –4295, Aug. 2009.

[75] A. G. Marquis, L. M. Lopez-Ramos, G. B. Giannakis, and J. Ramos, "Resource allocation for interweave and underlay CRs under probability-of-interference constraints." *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 10, pp. 1922–1933, 2012.

[76] C.-H. Chang, H.-L. Chao, and C.-L. Liu, "Sum throughput-improved resource allocation for lte uplink transmission," in *Proc. IEEE Vehicular Technology Conference (VTC)*, Sept 2011, pp. 1–5.

[77] H. Shajaiah, A. Khawar, A. Abdel-Hadi, and T. Clancy, "Resource allocation with carrier aggregation in lte advanced cellular system sharing spectrum with s-band radar," in *Proc. IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, April 2014, pp. 34–37.

[78] X. Xiao, X. Tao, and J. Lu, "Energy-efficient resource allocation in lte-based mimo-ofdma systems with user rate constraints," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2014.

[79] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the K-user interference channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, 2008.

[80] ——, "Interference alignment and the degrees of freedom of wireless X-networks," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 3893–3908, Sept 2009.

[81] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani, "Communication over mimo x channels: Interference alignment, decomposition, and performance analysis," *Information Theory, IEEE Transactions on*, vol. 54, no. 8, pp. 3457–3470, 2008.

[82] H. Maleki, V. Cadambe, and S. Jafar, "Index coding: An interference alignment perspective," *Information Theory, IEEE Transactions on*, vol. 60, no. 9, pp. 5402–5432, Sept 2014.

[83] V. Ntranos, M. Maddah-Ali, and G. Caire, "Cellular interference alignment," *Information Theory, IEEE Transactions on*, vol. 61, no. 3, pp. 1194–1217, March 2015.

[84] R. Tandon, S. Jafar, S. Shamai Shitz, and H. Poor, "On the synergistic benefits of alternating csit for the miso broadcast channel," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4106–4128, July 2013.

[85] M. A. Maddah-Ali and D. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4418–4431, July 2012.

[86] A. Barbieri, P. Gaal, S. Geirhofer, T. Ji, D. Malladi, Y. Wei, and F. Xue, "Coordinated downlink multi-point communications in heterogeneous cellular networks," in *Proc. Information Theory and Applications Workshop (ITA)*, Feb 2012, pp. 7–16.

[87] A. Lozano, R. Heath, and J. Andrews, "Fundamental limits of cooperation," *Information Theory, IEEE Transactions on*, vol. 59, no. 9, pp. 5213–5226, Sept 2013.

[88] F. Pantisano, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "On the impact of heterogeneous backhauls on coordinated multipoint transmission in femtocell networks," in *Proc. IEEE International Conference on Communications (ICC)*, June 2012, pp. 5064–5069.

[89] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard, and J. Crowcroft, "XORs in the Air: Practical Wireless Network Coding," *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 497–510, June 2008.

[90] Y. Wu, "Network coding for wireless networks," Microsoft Research, Tech. Rep. MSR-TR-2007-89, July 2007. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=70463

[91] C. Fragouli, D. Katabi, A. Markopoulou, M. Medard, and H. Rahul, "Wireless network coding: Opportunities and challenges," in *Proc. IEEE Military Communications Conference(MILCOM)*, Oct 2007, pp. 1–8.

[92] S. H. Dau, V. Skachek, and Y. M. Chee, "Error correction for index coding with side information," *Information Theory, IEEE Transactions on*, vol. 59, no. 3, pp. 1517–1531, March 2013.

[93] Z. Bar-Yossef, Y. Birk, T. Jayram, and T. Kol, "Index coding with side information," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1479–1494, March 2011.

[94] N. Naderializadeh and A. Avestimehr, "ITLinQ: A New Approach for Spectrum Sharing in Device-to-Device Communication Systems," *Selected Areas in Communications, IEEE Journal on*, vol. 32, no. 6, pp. 1139–1151, June 2014.

[95] C. Geng, N. Naderializadeh, A. Avestimehr, and S. Jafar, "On the optimality of treating interference as noise," *Information Theory, IEEE Transactions on*, vol. 61, no. 4, pp. 1753–1767, April 2015.

[96] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6524–6540, Oct 2012.

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

[98] ——, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Transactions on Networking*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015. [Online]. Available: http://dx.doi.org/10.1109/TNET.2014.2317316

[99] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," in *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2014, pp. 221–226.

[100] R. Pedarsani, M. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, April 2016.

[101] U. Niesen and M. A. Maddah-Ali, "Coded caching for delay-sensitive content," *arXiv:1407.4489*, July 2014. [Online]. Available: http://arxiv.org/abs/1407.4489

[102] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, July 2013, pp. 1461–1465.

[103] ——, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 176–189, Jan 2016.

[104] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "On the average performance of caching and coded multicasting with random demands," in *Proc. International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2014, pp. 922–926.

[105] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless d2d networks," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 849–869, Feb 2016.

[106] N. Karamchandani, U. Niesen, M. Maddah-Ali, and S. Diggavi, "Hierarchical coded caching," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 2142–2146.

[107] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-Optimal Rate of Caching and Coded Multicasting with Random Demands," *arXiv:1502.03124*, 2015.

[108] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, vol. 1, 1999, pp. 126–134 vol.1.

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

[110] I. Baev, R. Rajaraman, and C. Swamy, "Approximation Algorithms for Data Placement Problems," *SIAM J. Comput.*, vol. 38, no. 4, pp. 1411–1429, Aug. 2008.

[111] A. Meyerson, K. Munagala, and S. Plotkin, "Web caching using access statistics," in *Proc. Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '01. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2001, pp. 354–363. [Online]. Available: http://dl.acm.org/citation.cfm?id=365411.365479

[112] M. Ji, "Fundamental limits of caching networks: Turning memory into bandwidth," Ph.D. dissertation, University of Southern California, 2015.

[113] J. Roberts and N. Sbihi, "Exploring the memory-bandwidth tradeoff in an information-centric network," in *Proc. International Teletraffic Congress (ITC)*, Sept 2013, pp. 1–9.

[114] S. K. Fayazbakhsh, Y. Lin, A. Tootoonchian, A. Ghodsi, T. Koponen, B. Maggs, K. Ng, V. Sekar, and S. Shenker, "Less pain, most of the gain: Incrementally deployable icn," in *Proc. ACM SIGCOMM Conference*, ser. SIGCOMM '13. New York, NY, USA: ACM, 2013, pp. 147–158. [Online]. Available: http://doi.acm.org/10.1145/2486001.2486023

[115] L. W. Dowdy and D. V. Foster, "Comparative Models of the File Assignment Problem," *ACM Comput. Surv.*, vol. 14, no. 2, pp. 287–313, Jun. 1982.

[116] K. C. Almeroth and M. H. Ammar, "The Use of Multicast Delivery to Provide a Scalable and Interactive Video-on-Demand Service," *Selected Areas in Communications, IEEE Journal on*, vol. 14, no. 6, pp. 1110–1122, 1996.

[117] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," in *Proc. Information Theory and Applications Workshop (ITA)*, Feb 2015, pp. 98–107.

[118] M. Ji, A. Tulino, J. Llorca, and G. Caire, "Caching-aided coded multicasting with multiple random requests," in *Proc. IEEE Information Theory Workshop (ITW)*, April 2015, pp. 1–5.

[119] P. Hassanzadeh, A. Tulino, J. Llorca, and E. Erkip, "Cache-aided coded multicast for correlated sources," in *Proc. International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*. IEEE, 2016, pp. 360–364.

[120] J. Hachem, N. Karamchandani, and S. N. Diggavi, "Coded caching for heterogeneous wireless networks with multi-level access," *arXiv: 1404.6560*, 2014. [Online]. Available: http://arxiv.org/abs/1404.6560

[121] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 56–60.

[122] ——, "Effect of number of users in multi-level coded caching," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1701–1705.

[123] S. P. Shariatpanahi, A. S. Motahari, and B. H. Khalaj, "Multi-server coded caching," *arXiv: 1503.00265*, 2015. [Online]. Available: http://arxiv.org/abs/1503.00265

[124] S. Sahraei and M. Gastpar, "Multi-library coded caching," *arXiv:1601.06016*, 2016. [Online]. Available: http://arxiv.org/abs/1601.06016

[125] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: improved bounds for users with small buffers," *IET Communications*, July 2016. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-com.2015.1205

[126] M. M. Amiri and D. Gunduz, "Fundamental limits of caching: Improved delivery rate-cache capacity trade-off," *arXiv:1604.03888*, 2016. [Online]. Available: http://arxiv.org/pdf/1604.03888v1.pdf

[127] M. M. Amiri, Q. Yang, and D. Gündüz, "Coded caching for a large number of users," *arXiv: 1605.01993*, 2016. [Online]. Available: http://arxiv.org/abs/1605.01993

[128] K. Wan, D. Tuninetti, and P. Piantanida, "On caching with more users than files," *arXiv:1601.06383*, 2016. [Online]. Available: http://arxiv.org/abs/1601.06383

[129] C. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: Sequential coding for computing," *arXiv: 1504.00553*, 2015. [Online]. Available: http://arxiv.org/abs/1504.00553

[130] S. H. Lim, C. Wang, and M. Gastpar, "Information theoretic caching: The multi-user case," *arXiv: 1604.02333*, 2016. [Online]. Available: http://arxiv.org/abs/1604.02333

[131] S. Sahraei and M. Gastpar, "k users caching two files: An improved achievable rate," *arXiv:1512.06682*, 2015. [Online]. Available: http://arxiv.org/abs/1512.06682

[132] C.-Y. Wang, S. H. Lim, and M. Gastpar, "A new converse bound for coded caching," *arXiv:1601.05690*, 2016. [Online]. Available: http://arxiv.org/abs/1601.05690

[133] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1696–1700.

[134] A. Sengupta, R. Tandon, and T. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1691–1695.

[135] A. Sengupta and R. Tandon, "Improved approximation of storage-rate tradeoff for caching with multiple demands," Oct. 2016. [Online]. Available: https://filebox.ece.vt.edu/~aviksg/fundconv.pdf

[136] N. Ajaykrishnan, N. S. Prem, V. M. Prabhakaran, and R. Vaze, "Critical database size for effective caching," *arXiv:1501.02549*, 2015.

[137] C. Tian, "A note on the fundamental limits of coded caching," *arXiv:1503.00010*, 2015.

[138] ——, "Symmetry, demand types and outer bounds in caching systems," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 825–829.

[139] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *arXiv:1609.07817*, September 2016.

[140] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," *arXiv preprint arXiv:1511.02256*, 2015.

[141] C. Tian and J. Chen, "Caching and delivery via interference elimination," *arXiv:1604.08600*, 2016.

[142] S. S. Bidokhti, M. Wigger, and R. Timo, "An upper bound on the capacity-memory tradeoff of degraded broadcast channels," in *Proc. International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, Sept 2016, pp. 350–354.

[143] ——, "Erasure broadcast networks with receiver caching," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1819–1823.

[144] S. S. Bidokhti, M. A. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *arXiv: 1605.02317*, 2016. [Online]. Available: http://arxiv.org/abs/1605.02317

[145] R. Timo and M. Wigger, "Joint cache-channel coding over erasure broadcast channels," in *Proc. International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2015, pp. 201–205.

[146] M. A. Maddah-Ali and U. Niesen, "Cache aided interference channels," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 809–813.

[147] N. Naderializadeh, M. A. Maddah-Ali, and A. Salman Avestimehr, "Fundamental limits of cache-aided interference management," *arXiv:1602.04207*, Feb. 2016. [Online]. Available: http://arxiv.org/pdf/1602.04207v1

[148] J. Hachem, U. Niesen, and S. N. Diggavi, "A layered caching architecture for the interference channel," *arXiv: 1605.01668*, 2016. [Online]. Available: http://arxiv.org/abs/1605.01668

[149] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 2029–2033.

[150] ——, "Harnessing cloud and edge synergies: Towards an information theory of fog radio access networks," *to appear in IEEE Communications Magazine, Special Issue on Communications, Caching, and Computing for Content-Centric Mobile Networks*, Aug 2016.

[151] S. Yang, K. H. Ngo, and M. Kobayashi, "Content delivery with coded caching and massive mimo in 5g," in *Proc. International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, Sept 2016, pp. 370–374.

[152] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[153] P. Blasco and D. Gündüz, "Learning-based optimization of cache content in a small cell base station," in *Proc. IEEE International Conference on Communications (ICC)*, June 2014, pp. 1897–1903.

[154] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, Aug 2014.

[155] F. Shen, K. Hamidouche, E. Bastug, and M. Debbah, "A stackelberg game for incentive proactive caching mechanisms in wireless networks," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, December 2016.

[156] E. Bastug, M. Bennis, and M. Debbah, "Think Before Reacting: Proactive Caching in 5G Small Cell Networks," *Towards 5G: Applications, Requirements and Candidate Technologies, Wiley, 2015 (Submitted)*, September 2014.

[157] E. Batu, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," in *Proc. International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2015, pp. 161–166.

[158] B. B. Nagaraja and K. G. Nagananda, "Caching with unknown popularity profiles in small cell networks," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[159] B. N. Bharath, K. G. Nagananda, and H. V. Poor, "A learning-based approach to caching in heterogenous small cell networks," *arXiv: 1508.03517*, 2015. [Online]. Available: http://arxiv.org/abs/1508.03517

[160] S. Mller, O. Atan, M. van der Schaar, and A. Klein, "Smart caching in wireless small cell networks via contextual multi-armed bandits," in *Proc. IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–7.

[161] S. Li, J. Xu, M. van der Schaar, and W. Li, "Trend-aware video caching through online learning," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, July 2016.

[162] P. Blasco and D. Gündüz, "Multi-armed bandit optimization of cache content in wireless infostation networks," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 51–55.

[163] ——, "Content-level selective offloading in heterogeneous networks: Multi-armed bandit optimization and regret bounds," *arXiv:1407.6154*, 2014.

[164] D. Malak, M. Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Transactions on Communications*, vol. 64, no. 10, p. 4365, 2016.

[165] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing the spatial content caching distribution for device-to-device communications," in *Proc. IEEE International Symposium on Information Theory (ISIT)*. IEEE, July 2016, pp. 280–284.

[166] D. Malak, M. Al-Shalash, and J. G. Andrews, "Spatially correlated content caching for device-to-device communications," *arXiv: 1609.00419*, 2016. [Online]. Available: http://arxiv.org/abs/1609.00419

[167] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Fundamentals of cluster-centric content placement in cache-enabled device-to-device networks," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2511–2526, June 2016.

[168] M. Afshang and H. S. Dhillon, "Optimal geographic caching in finite wireless networks," in *Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–5.

[169] S. Krishnan and H. S. Dhillon, "Distributed caching in device-to-device networks: A stochastic geometry perspective," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Nov 2015, pp. 1280–1284.

[170] S. Krishnan, M. Afshang, and H. S. Dhillon, "Effect of retransmissions on optimal caching in cache-enabled small cell networks," *arXiv: 1606.03971*, 2016. [Online]. Available: http://arxiv.org/abs/1606.03971

[171] M. Ji, A. Tulino, J. Llorca, and G. Caire, "Caching and coded multicasting: Multiple Groupcast Index Coding," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 881–885.

[172] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *arXiv:1512.07743*, Dec 2015. [Online]. Available: http://arxiv.org/abs/1512.07743

[173] A. S. Motahari, S. Oveis-Gharan, M. A. Maddah-Ali, and A. Khandani, "Real interference alignment: Exploiting the potential of single antenna systems," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4799–4810, Aug 2014.

[174] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*. Hoboken, NJ, USA: Wiley-Interscience, John Wiley and Sons. Inc., 2006.

[175] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.

[176] A. Sengupta, R. Tandon, and T. C. Clancy, "Fundamental limits of caching with secure delivery," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 355–370, Feb 2015.

[177] ——, "Secure caching with non-uniform demands," in *Proc. IEEE Global Wireless Summit (GWS) - International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace Electronic Systems (VITAE)*, May 2014, pp. 1–5.

[178] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning distributed caching strategies in small cell networks," in *Proc. International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2014, pp. 917–921.

[179] A. Sengupta and R. Tandon, "Beyond cut-set bounds-the approximate capacity of D2D networks," in *Proc. Information Theory and Applications (ITA)*, February 2015, pp. 78–83.

[180] S. Wang, W. Li, X. Tian, and H. Liu, "Coded caching with heterogenous cache sizes," *arXiv: 1504.01123*, 2015. [Online]. Available: http://arxiv.org/abs/1504.01123

[181] M. M. Amiri, Q. Yang, and D. Gündüz, "Decentralized coded caching with distinct cache capacities," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Nov 2016. [Online]. Available: https://arxiv.org/pdf/1610.03792v1.pdf

[182] H. S. Wilf, *Generatingfunctionology*. Natick, MA, USA: A. K. Peters, Ltd., 2006.

[183] C. E. Shannon, "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, Sep 1949.

[184] M. Hefeeda and O. Saleh, "Traffic Modeling and Proportional Partial Caching for Peer-to-Peer Systems," *IEEE/ACM Transactions on Networking*, vol. 16, no. 6, pp. 1447–1460, Dec. 2008.

[185] V. Ravindrakumar, P. Panda, N. Karamchandani, and V. Prabhakaran, "Fundamental limits of secretive coded caching," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 425–429.

[186] Z. H. Awan and A. Sezgin, "Fundamental limits of caching in d2d networks with secure delivery," in *Proc. IEEE International Conference on Communication (ICC) Workshops*, June 2015, pp. 464–469.

[187] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud RAN for mobile network - A technology overview," *IEEE Communications Surveys Tutorial*, vol. 17, no. 1, pp. 405–426, March 2015.

[188] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog computing based radio access networks: Issues and challenges," *arXiv:1506.04233*, 2015. [Online]. Available: http://arxiv.org/abs/1506.04233

[189] M. Leconte, G. S. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," *arXiv: 1601.03926*, 2016. [Online]. Available: http://arxiv.org/abs/1601.03926

[190] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite length analysis of caching-aided coded multicasting," in *Proc. Allerton Conference on Communication, Control, and Computing*, Sept 2014, pp. 914–920.

[191] V. Bioglio, F. Gabry, and I. Land, "Optimizing mds codes for caching at the edge," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[192] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," *arXiv: 1510.06121*, Oct 2015. [Online]. Available: http://arxiv.org/abs/1510.06121

[193] X. Peng, J. C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *Proc. IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, Sept 2014, pp. 1370–1374.

[194] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *arXiv:1512.06938*, 2015. [Online]. Available: http://arxiv.org/abs/1512.06938

[195] B. Azari, O. Simeone, U. Spagnolini, and A. M. Tulino, "Hypergraph-based analysis of clustered cooperative beamforming with application to edge caching," *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 84–87, Feb 2016.

[196] S. Park, O. Simeone, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," *arXiv:1601.02460*, Jan. 2016.

[197] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching," *arXiv: 1512.02385*, 2015. [Online]. Available: http://arxiv.org/abs/1512.02385

[198] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. 50th Annual Conference on Information Sciences and Systems (CISS)*, March 2016.

[199] ——, "Cache aided wireless networks: Tradeoffs between storage and latency," *arXiv:1512.07856*, Dec 2015. [Online]. Available: http://arxiv.org/pdf/1512.07856v1.pdf

[200] ——, "Cloud ran and edge caching: Fundamental performance trade-offs," in *Proc. IEEE International workshop on Signal Processing advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–5.

[201] F. Xu, M. Tao, and K. Liu, "Fundamental Tradeoff between Storage and Latency in Cache-Aided Wireless Interference Networks," *arXiv: 1605.00203*, May 2016. [Online]. Available: http://arxiv.org/pdf/1605.00203v1.pdf

[202] Y. Liu and E. Erkip, "Completion time in multi-access channel: An information theoretic perspective," in *Proc. IEEE Information Theory Workshop (ITW)*, Oct 2011, pp. 708–712.

[203] ——, "Completion time in broadcast channel and interference channel," in *Proc. Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2011, pp. 1694–1701.

[204] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, Sept 2006.

[205] S. A. Jafar, "Interference alignment a new look at signal dimensions in a communication network," *Foundations and Trends in Communications and Information Theory*, vol. 7, no. 1, pp. 1–134, 2010. [Online]. Available: http://dx.doi.org/10.1561/0100000047

[206] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 3:1–3:10, Feb 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/840814

[207] A. Ghasemi, A. S. Motahari, and A. K. Khandani, "On the degrees of freedom of X-channel with delayed CSIT," in *Proc. IEEE International Symposium on Information Theory*, July 2011, pp. 767–770.

[208] C. S. Vaze and M. K. Varanasi, "The degree-of-freedom regions of MIMO broadcast, interference, and cognitive radio channels with no CSIT," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5354–5374, Aug 2012.

[209] A. Sengupta, R. Tandon, and O. Simeone, "Cloud and cache-aided wireless networks: Fundamental latency trade-offs," *arXiv: 1605.01690*, May 2016. [Online]. Available: http://arxiv.org/abs/1605.01690

[210] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Wireless Video Content Delivery through Coded Dstributed Caching," in *Proc. IEEE International Conference on Communications (ICC)*. IEEE, 2012, pp. 2467–2472.

[211] S. Bubeck and N. Cesa-Bianchi, "Regret Analysis of Stochastic and Non-Stochastic Multi-armed Bandit Problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–22, 2012.

[212] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002. [Online]. Available: http://dx.doi.org/10.1023/A%3A1013689704352

[213] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. International Conference on Machine Learning (ICML)*, 2013, pp. 151–159.

[214] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1746–1778, Jan. 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=2946645.2946695

[215] Y. Zhou and X. Li, "Multi-armed bandits with combinatorial strategies under stochastic bandits," *arXiv:1307.5438*, 2013. [Online]. Available: http://arxiv.org/abs/1307.5438

[216] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network measurements, models, and implications," *Computer Networks*, vol. 53, no. 4, pp. 501 – 514, 2009, content Distribution Infrastructures for Community Networks. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128608003423

[217] C. Szepesvári, "The asymptotic convergence-rate of q-learning," in *Proc. Conference on Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1998, pp. 1064–1070. [Online]. Available: http://dl.acm.org/citation.cfm?id=302528.302898

[218] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, June 2006.

[219] D. Pisinger, "Algorithms for knapsack problems," Ph.D. dissertation, University of Copenhagen, 1995.

[220] D. B. Shmoys and E. Tardos, "An approximation algorithm for the generalized assignment problem," *Mathematical Programming*, vol. 62, no. 1-3, pp. 461–474, 1993. [Online]. Available: http://dx.doi.org/10.1007/BF01585178

[221] L. Fleischer, M. X. Goemans, V. S. Mirrokni, and M. Sviridenko, "Tight approximation algorithms for maximum general assignment problems," in *Proc. ACM-SIAM Symposium on Discrete Algorithm*. Society for Industrial and Applied Mathematics, 2006, pp. 611–620.

[222] D. Brélaz, "New Methods to Color the Vertices of a Graph," *Communications of the ACM*, vol. 22, no. 4, pp. 251–256, Apr. 1979. [Online]. Available: http://doi.acm.org/10.1145/359094.359101

[223] A. Korbut and I. Sigal, "Exact and greedy solutions of the knapsack problem: the ratio of values of objective functions," *International Journal of Computer and Systems Sciences*, vol. 49, no. 5, pp. 757–764, 2010. [Online]. Available: http://dx.doi.org/10.1134/S1064230710050102

[224] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[225] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.

[226] Ericsson, Huawei Technologies, NEC Corporation, Alcatel Lucent, and Nokia Siemens Networks, "Common public radio interface (CPRI); Interface specification," *CPRI specification version 5.0*, Sep 2011.