Coded Caching for Heterogeneous Systems: An Optimization Perspective

Abdelrahman M. Ibrahim, *Student Member, IEEE*, Ahmed A. Zewail, *Member, IEEE*, and Aylin Yener, *Fellow, IEEE*

***Abstract*— In cache-aided networks, the server populates the cache memories at the users during low-traffic periods in order to reduce the delivery load during peak-traffic hours. In turn, there exists a fundamental tradeoff between the delivery load on the server and the cache sizes at the users. In this paper, we study this tradeoff in a multicast network, where the server is connected to users with unequal cache sizes and the number of users is less than or equal to the number of library files. We propose centralized uncoded placement and linear delivery schemes which are optimized by solving a linear program. Additionally, we derive a lower bound on the delivery memory tradeoff with uncoded placement that accounts for the heterogeneity in cache sizes. We explicitly characterize this tradeoff for the case of three end-users, as well as an arbitrary number of end-users when the total memory size at the users is small, and when it is large. Next, we consider a system where the server is connected to the users via rate-limited links of different capacities and the server assigns the users' cache sizes subject to a total cache budget. We characterize the optimal cache sizes that minimize the delivery completion time with uncoded placement and linear delivery. In particular, the optimal memory allocation balances between assigning larger cache sizes to users with low capacity links and uniform memory allocation.**

***Index Terms*— Coded caching, uncoded placement, cache size optimization, multicast networks.**

I. Introduction

THE immense growth in wireless data traffic is driven by video-on-demand services, which are expected to account for 82% of all consumer Internet traffic by 2020 [1]. The high temporal variation in video traffic leads to under-utilization of network resources during off-peak hours and congestion in peak hours [2]. Caching improves uniformization of network utilization, by pushing data into the cache memories at the network edge during off-peak hours, which in turn reduces congestion during peak hours. The seminal work [3] has proposed a novel caching technique for a downlink setting, in which a server jointly designs the content to be placed during off-peak hours and the delivered during peak hours, in order to ensure that multiple end-users can benefit from delivery transmissions simultaneously. These multicast coding opportunities are shown to provide gains beyond local caching gains, which result from the availability of a fraction of the requested file at the user's local cache. They are termed global caching gains since they scale with the network size. Reference [3] has shown that there exists a fundamental trade-off between the delivery load on the server and the users' cache sizes.

The characterization of this trade-off has been the focus of extensive recent efforts [4]–[14]. In particular, references [4]–[6] have characterized the delivery load memory trade-off with the uncoded placement assumption, i.e., assuming that the users cache only uncoded pieces of the files. The delivery load memory trade-off with general (coded) placement has been studied in [7]–[14]. Coded caching schemes were developed for various cache-aided network architectures, such as multi-hop [15]–[17], device-to-device (D2D) [18], [19], multi-server [20], lossy broadcast [21]–[24], and interference networks [25], [26]. In addition to network topology, several practical considerations have been studied, such as the time-varying nature of the number of users [27], distortion requirements at the users [28]–[30], non-uniform content distribution [31]–[35], delay-sensitive content [36], and systems with security requirements [37]–[39].

End-users in practical caching networks have varying storage capabilities. In this work, we address this system constraint by allowing the users to have distinct cache sizes. In particular, we study the impact of heterogeneity in cache sizes on the delivery load memory trade-off with uncoded placement. Models with similar traits have been studied in references [28], [40]–[42]. In particular, references [40], [41] have extended the decentralized caching scheme in [27] to systems with unequal cache sizes. References [28], [42] have proposed a centralized scheme in which the system is decomposed into layers such that the users in each layer have equal cache size. More specifically, the scheme in [3] is applied on each layer and the optimal fraction of the file delivered in each layer is optimized. Additionally, reference [42] has proposed grouping the users before applying the layered scheme which requires solving a combinatorial problem. In a follow-up work to some of our preliminary results presented in [43], reference [44] proposed optimizing over uncoded placement schemes assuming the delivery scheme in [27].

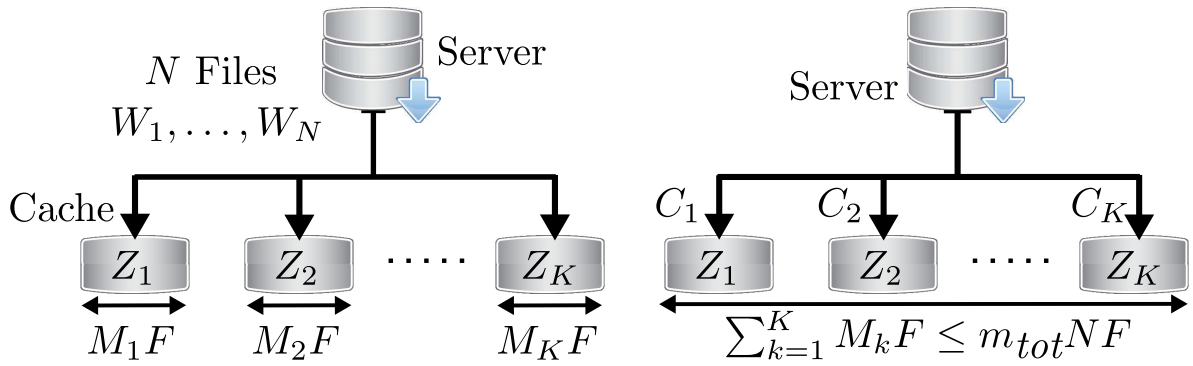
Manuscript received October 5, 2018; revised March 6, 2019; accepted April 23, 2019. Date of publication May 1, 2019; date of current version August 14, 2019. This work was supported in part by NSF Grants 1526165 and 1749665. This paper was presented in part at the IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, USA, in 2017, and in part at the IEEE International Conference on Communications (ICC), Paris, France, in 2017. The associate editor coordinating the review of this paper and approving it for publication was R. Thobaben. *(Corresponding author: Aylin Yener.)*

The authors are with the Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: ami137@psu.edu; aiz103@psu.edu; yener@ee.psu.edu).

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

Digital Object Identifier 10.1109/TCOMM.2019.2914393

0090-6778 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.



(a) Fixed cache sizes and equal download rates.

(b) Cache memory budget and unequal download rates.

Fig. 1. Centralized caching system with unequal cache sizes.

In this work, we focus on uncoded placement and linear delivery, where the server places uncoded pieces of the files at the users' cache memories, and the multicast signals are formed using linear codes. Our proposed caching scheme outperforms the schemes in [28], [42], [44], because it allows flexible utilization of the side-information in the creation of the multicast signals, i.e., the side-information stored exclusively at t users is not restricted to multicast signals of size $t+1$ as in [3], [28], [40]–[42], [44]. We show that the worst-case delivery load is minimized by solving a linear program over the parameters of the proposed caching scheme. In order to evaluate the performance of our caching scheme, we derive a lower bound on the worst-case delivery load with uncoded placement. Using this bound, we explicitly characterize the delivery load memory trade-off for arbitrary number of users with uncoded placement in the small total memory regime, large total memory regime, the definitions of which are made precise in the paper, and for any memory regime for the instance of three-users. Furthermore, we compare the achievable delivery load with the proposed lower bound with uncoded placement, and the lower bounds with general placement in [13], [41]. From the numerical results, we observe that our achievable delivery load coincides with the uncoded placement bound.

Next, inspired by the schemes developed for distinct cache sizes we consider a middle ground between noiseless setups [3], [28], [42] and noisy broadcast channels with cache-aided receivers [21]–[24]. More specifically, we assume that the server is connected to the users via rate limited links of different capacities, and the server assigns the users' cache sizes subject to a cache memory budget. Reference [45] has considered a similar model and proposed jointly designing the caching and modulation schemes. Different from [21]–[24], [45], we consider a separation approach where the caching scheme and the physical layer transmission scheme are designed separately. This is inline in general with the approach of [3] and followup works that consider server to end-users links as bit pipes. We focus on the joint optimization of the users' cache sizes and the caching scheme in order to minimize the worst-case delivery completion time. More specifically, the optimal memory allocation, uncoded placement, and linear delivery schemes are again obtained by solving a linear program. For the case where the cache memory budget is less than or equal to the library size at the server, we derive closed form expressions for the optimal memory allocation and caching scheme. We observe that the optimal solution balances between assigning larger cache memories to users with low capacity links, delivering fewer bits to them, and uniform memory allocation, which maximizes the multicast gain.

II. System Model

Notation: Throughout the paper, vectors are represented by boldface letters, sets of policies are represented by calligraphic letters, e.g., $\mathfrak{A}$, $\oplus$ refers to bitwise XOR operation, $(x)^+ \triangleq \max\{0, x\}$, $|W|$ denotes the size of W, $\mathcal{A} \setminus \mathcal{B}$ denotes the set of elements in $\mathcal{A}$ and not in $\mathcal{B}$, ϕ denotes the empty set, $[K] \triangleq \{1, \ldots, K\}$, $\mathcal{A} \subset \mathcal{B}$ denotes $\mathcal{A}$ being a subset of or equal to $\mathcal{B}$, $\subsetneq_\phi [K]$ denotes non-empty subsets of $[K]$, and $\mathcal{P}_\mathcal{A}$ is the set of all permutations of the elements in the set $\mathcal{A}$, e.g., $\mathcal{P}_{\{1,2\}} = \{[1,2],\ [2,1]\}$.

Consider a centralized system consisting of a server connected to K users via an error-free multicast link [3], see Fig. 1(a). A library $\{W_1, \ldots, W_N\}$ of N files, each with size F bits, is stored at the server. User k is equipped with a cache memory of size M_kF bits. Without loss of generality, we assume that $M_1 \leq M_2 \leq \cdots \leq M_K$. We define $m_k = M_k/N$ to denote the memory size of user k normalized by the library size NF, i.e., $m_k \in [0,1]$ for $M_k \in [0, N]$. The cache size vector is denoted by $\boldsymbol{M} = [M_1, \ldots, M_K]$ and its normalized version by $\boldsymbol{m} = [m_1, \ldots, m_K]$. We focus on the case where the number of files is larger than or equal to the number of users, i.e., $N \geq K$.

In Section VII, we introduce rate limited download links of distinct capacities to the model. In particular, we consider that the link between the server and user k has capacity C_k bits per channel use, which we refer to as the *download rate* at user k, as illustrated in Fig. 1(b). We denote the collection of link capacities by $\boldsymbol{C} = [C_1, \ldots, C_K]$. In this setup, we seek the system configuration with best performance, including the memory sizes $\{M_k\}$, subject to $\sum_{k=1}^{K} M_kF \leq m_{\text{tot}}NF$ bits, where m_{tot} is the cache memory budget normalized by the library size.

The system operates over two phases: placement phase and delivery phase. In the placement phase, the server populates users' cache memories without knowing the users' demands. User k stores Z_k, subject to its cache size constraint, i.e., $|Z_k| \leq M_k F$ bits. Formally, the users' cache contents are defined as follows.

Definition 1 (Cache Placement): A cache placement function $\varphi_k : [2^F]^N \rightarrow [2^{\lfloor M_k F \rfloor}]$ *maps the files in the library to the cache of user k, i.e.,* $Z_k = \varphi_k(W_1, W_2, .., W_N)$.

In the delivery phase, user k requests file W_{d_k} from the server. Users' demand vector $\boldsymbol{d} = [d_1, \ldots, d_K]$ consists of independent uniform random variables over the files as in [3]. In order to deliver the requested files, the server transmits a sequence of unicast/multicast signals, $X_{\mathcal{T},\boldsymbol{d}}$, to the users in the sets $\mathcal{T} \subsetneq_\phi [K]$. $X_{\mathcal{T},\boldsymbol{d}}$ has length $v_\mathcal{T} F$ bits, and is defined as follows.

Definition 2 (Encoding): Given $\mathbf{d}$*, an encoding function* $\psi_{\mathcal{T},\boldsymbol{d}} : [2^F]^K \rightarrow [2^{\lfloor v_\mathcal{T} F \rfloor}]$ *maps requested files to a signal with length* $v_\mathcal{T} F$ *bits, sent to users in* $\mathcal{T}$*, i.e.,* $X_{\mathcal{T},\boldsymbol{d}} = \psi_{\mathcal{T},\boldsymbol{d}}(W_{d_1}, .., W_{d_K})$.

At the end of the delivery phase, user k must be able to reconstruct W_{d_k} from the transmitted signals $X_{\mathcal{T},\boldsymbol{d}}, \mathcal{T} \subsetneq_\phi [K]$ and its cache content Z_k, with negligible probability of error.

Definition 3 (Decoding): A decoding function $\mu_{\boldsymbol{d},k}$: $[2^{\lfloor RF \rfloor}] \times [2^{\lfloor M_k F \rfloor}] \rightarrow [2^F]$*, with* $R \triangleq \sum_{\mathcal{T} \subsetneq_\phi [K]} v_\mathcal{T}$*, maps cache content of user k,* Z_k*, and the signals* $X_{\mathcal{T},\boldsymbol{d}}, \mathcal{T} \subsetneq_\phi [K]$ *to* $\hat{W}_{d_k}$*, i.e.,* $\hat{W}_{d_k} = \mu_{\boldsymbol{d},k}\left(X_{\{1\},\boldsymbol{d}}, X_{\{2\},\boldsymbol{d}}, \ldots, X_{[K],\boldsymbol{d}}, Z_k\right)$.

A caching scheme is defined by $(\varphi_k(.), \psi_{\mathcal{T},\boldsymbol{d}}(.), \mu_{\boldsymbol{d},k}(.))$. The performance is measured in terms of the achievable delivery load, which represents the amount of data transmitted by the server in order to deliver the requested files.

Definition 4: For a given normalized cache size vector $\boldsymbol{m}$*, the delivery load* $R(\boldsymbol{m})$ *is said to be achievable if for every* $\epsilon > 0$ *and large enough F, there exists* $(\varphi_k(.), \psi_{\mathcal{T},\boldsymbol{d}}(.), \mu_{\boldsymbol{d},k}(.))$ *such that* $\max_{\boldsymbol{d},k \in [K]} Pr(\hat{W}_{d_k} \neq W_{d_k}) \leq \epsilon$*, and* $R^*(\boldsymbol{m}) \triangleq \inf\{R : R(\boldsymbol{m})$ *is achievable*$\}$.

The set of cache placement policies $\mathfrak{A}$ considered in this work are the so-called uncoded policies, i.e., only pieces of individual files are placed in the cache memories. Since we have uniform demands, the cache memory at each user k is divided equally over the files, i.e., $m_k F$ bits per file. We consider the set of delivery schemes $\mathfrak{D}$, in which multicast signals are formed using linear codes. The worst-case delivery load achieved by a caching scheme in $(\mathfrak{A}, \mathfrak{D})$ is defined as follows.

Definition 5: With placement and delivery policies in $\mathfrak{A}$ *and* $\mathfrak{D}$*, the worst-case delivery load is defined as* $R_{\mathfrak{A},\mathfrak{D}} \triangleq \max_{\boldsymbol{d}} R_{\boldsymbol{d},\mathfrak{A},\mathfrak{D}} = \sum_{\mathcal{T} \subsetneq_\phi [K]} v_\mathcal{T}$*, and the minimum delivery load overall* $R_{\mathfrak{A},\mathfrak{D}}$ *is denoted by* $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m}) \triangleq \inf\{R_{\mathfrak{A},\mathfrak{D}} : R_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$ *is achievable*$\}$.

Definition 6: The minimum delivery load achievable with a placement policy in $\mathfrak{A}$ *and any delivery scheme, is defined as* $R^*_{\mathfrak{A}}(\boldsymbol{m}) \triangleq \inf\{R_{\mathfrak{A}} : R_{\mathfrak{A}}(\boldsymbol{m})$ *is achievable*$\}$.

Remark 1: Note that $R^*_{\mathfrak{A},\mathfrak{D}} \geq R^*_{\mathfrak{A}} \geq R^*$*, since* R^* *is obtained by taking the infimum over all achievable delivery loads,* $R^*_{\mathfrak{A}}$ *is restricted to uncoded placement policies in* $\mathfrak{A}$*, and* $R^*_{\mathfrak{A},\mathfrak{D}}$ *is restricted to cache placement and delivery policies in* $\mathfrak{A}$ *and* $\mathfrak{D}$*, respectively.*

In Section VII, we consider download links with limited and unequal capacities. Thus, $X_{\mathcal{T},\boldsymbol{d}}$ will need to have a rate $\leq \min_{j \in \mathcal{T}} C_j$ [46]. Additionally, there is no guarantee that the users outside the set $\mathcal{T}$ can decode $X_{\mathcal{T},\boldsymbol{d}}$, as their download rates may be lower than $\min_{j \in \mathcal{T}} C_j$. Consequently, a more relevant system-wide metric is the total time needed by the server to deliver all the requested files to all the users, defined as follows, assuming uncoded placement and linear delivery.

Definition 7: With a placement policy in $\mathfrak{A}$*, and a delivery policy in* $\mathfrak{D}$*, the worst-case delivery completion time (DCT) is defined as* $\Theta_{\mathfrak{A},\mathfrak{D}} \triangleq \max_{\boldsymbol{d}} \Theta_{\boldsymbol{d},\mathfrak{A},\mathfrak{D}} = \sum_{\mathcal{T} \subsetneq_\phi [K]} \frac{v_\mathcal{T}}{\min_{j \in \mathcal{T}} C_j}$.

Observe that, for $C_k = 1, \forall k \in [K]$, $\Theta_{\mathfrak{A},\mathfrak{D}} = R_{\mathfrak{A},\mathfrak{D}}$.

III. Motivational Example

In order to motivate our caching scheme which is tailored to capitalize on multicast opportunities to the fullest extent, we consider an example and compare the state-of-the-art caching schemes in [28], [42], [44] with our scheme.

Consider a three-user system with three files, $\{A, B, C\}$, and $\boldsymbol{m} = [0.4, 0.5, 0.7]$. Without loss of generality, we assume that the users request files A, B, and C, respectively. In the placement phase, the files are divided into subfiles, which are labeled by the users exclusively storing them, e.g., subfile $A_{i,j}$ is stored at users i and j.

1) *The layered scheme [28], [42]:* In the placement phase, the files are partitioned over three layers, we denote the files in layer l by the superscript (l). By optimizing the file partitioning over the layers, we get the following scheme. In layer 1, users have equal caches with size $M_1 F$ bits and files $A^{(1)}, B^{(1)}, C^{(1)}$ with size $0.9F$ bits, each of which is split into six disjoint subfiles, e.g., $A^{(1)}$ is divided into $A^{(1)}_1$, $A^{(1)}_2$, $A^{(1)}_3$, $A^{(1)}_{1,2}$, $A^{(1)}_{1,3}$, $A^{(1)}_{2,3}$, where $|A^{(1)}_i| = 0.2F$, and $|A^{(1)}_{i,j}| = 0.1F$. In delivery phase, the server sends the multicast signals $B^{(1)}_1 \oplus A^{(1)}_2$, $C^{(1)}_1 \oplus A^{(1)}_3$, $C^{(1)}_2 \oplus B^{(1)}_3$, and $C^{(1)}_{1,2} \oplus B^{(1)}_{1,3} \oplus A^{(1)}_{2,3}$. In layer 2, we have a single user with no cache and a two-user system with file size $0.1F$ bits and equal cache size $(M_2 - M_1)F = 0.1NF$ bits. The server only needs to send a unicast signal of size $0.1F$ bits to user 1. In layer 3, the $(M_3 - M_2)F$ bits of the cache at user 3 are not utilized.
2) *The caching scheme in [44]:* Each file is split into six disjoint subfiles, e.g., A is divided into A_1, A_2, A_3, $A_{1,2}$, $A_{1,3}$, $A_{2,3}$, where $|A_i| = 0.4F/3$, $|A_{1,2}| = 0.1F/3$, $|A_{1,3}| = 0.7F/3$, and $|A_{2,3}| = F/3$. In delivery phase, the server sends $B_1 \oplus A_2$, $C_1 \oplus A_3$, $C_2 \oplus B_3$, and $C_{1,2} \bar{\oplus} B_{1,3} \bar{\oplus} A_{2,3}$, where $\bar{\oplus}$ denotes an XOR operation that allows zero padding. Note that $C_{1,2} \bar{\oplus} B_{1,3} \bar{\oplus} A_{2,3}$ can be decomposed into $C_{1,2} \oplus B'_{1,3} \oplus A'_{2,3}$, $B''_{1,3} \oplus A''_{2,3}$, and the unicast signal $A'''_{2,3}$, where $|B'_{1,3}| = |A'_{2,3}| = |C_{1,2}|$, $|B''_{1,3}| = |A''_{2,3}| = |B_{1,3}| - |C_{1,2}|$, and $|A'''_{2,3}| = |A_{2,3}| - |B_{1,3}|$.

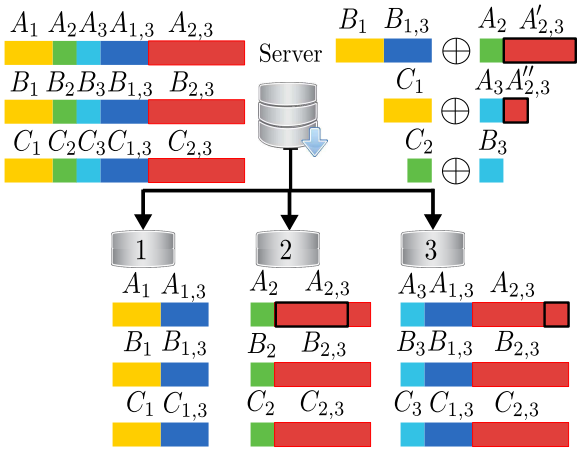


Fig. 2. Optimal scheme with uncoded placement for $K = N = 3$ and $\boldsymbol{M} = [1.2,\ 1.5,\ 2.1]$.

3) *Our proposed scheme:* In the placement phase, each file is split into five disjoint subfiles, e.g., A is divided into A_1, A_2, A_3, $A_{1,3}$, $A_{2,3}$, where $|A_1| = |A_{1,3}| = 0.2F$, $|A_2| = |A_3| = 0.1F$, and $|A_{2,3}| = 0.4F$. First, the server partitions $A_{2,3}$ into $A'_{2,3}$, $A''_{2,3}$ such that $|A'_{2,3}| = 0.3F$ and $|A''_{2,3}| = 0.1F$. Then, the server sends the multicast signals $\left(B_1 \cup B_{1,3}\right) \oplus \left(A_2 \cup A'_{2,3}\right)$, $C_1 \oplus \left(A_3 \cup A''_{2,3}\right)$, and $C_2 \oplus B_3$. One can easily verify that these multicast signals enable the users to decode the requested files. The caching scheme is illustrated in Fig. 2.

Our caching scheme achieves a delivery load equal to 0.7, compared to 0.8 by the layered scheme [28], [42], and 0.7333 by the scheme in [44]. The schemes in [28], [42], [44] need an additional unicast transmission compared with our scheme, as we have better utilization of side-information, e.g., $A'_{2,3}$ is used in the multicast signal to users $\{1, 2\}$. Additionally, in this example, the layered scheme does not utilize $(M_3 - M_2)F$ bits of the cache at user 3. In Theorem 4, we show that our proposed scheme is optimal with uncoded placement.

IV. Cache Placement Phase

Each file W_l is partitioned into 2^K subfiles. A subfile $\tilde{W}_{l,\mathcal{S}}$ is labeled by the set of users $\mathcal{S}$ exclusively storing it. The set of uncoded placement schemes for a given $\boldsymbol{m}$ is defined as

$$\mathfrak{A}(\boldsymbol{m}) = \Bigg\{ \boldsymbol{a} \in [0,1]^{2^K} \Bigg| \sum_{\mathcal{S}\subset[K]} a_{\mathcal{S}} = 1, \sum_{\mathcal{S}\subset[K]:\, k\in\mathcal{S}} a_{\mathcal{S}} \le m_k, \ \forall k \in [K] \Bigg\}, \tag{1}$$

where $\boldsymbol{a}$ is the vector of allocation variables $a_{\mathcal{S}}$, $\mathcal{S} \subset [K]$ and $|\tilde{W}_{l,\mathcal{S}}| = a_{\mathcal{S}} F$ bits, $\forall l \in [N]$. For example, for $K = 3$, we have

$$a_\phi + a_{\{1\}} + a_{\{2\}} + a_{\{3\}} + a_{\{1,2\}} + a_{\{1,3\}} + a_{\{2,3\}} + a_{\{1,2,3\}} = 1, \tag{2}$$

$$a_{\{i\}} + a_{\{i,j\}} + a_{\{i,k\}} + a_{\{i,j,k\}} \le m_i, \quad i,j,k \in \{1,2,3\}, i \ne j \ne k. \tag{3}$$

V. Delivery Phase

A. Multicast Signals $X_{\mathcal{T},\boldsymbol{d}}$

A multicast signal $X_{\mathcal{T},\boldsymbol{d}}$ delivers a piece of the file W_{d_j}, $W^{\mathcal{T}}_{d_j}$, to user $j \in \mathcal{T}$. The server generates $X_{\mathcal{T},\boldsymbol{d}}$ by XORing $W^{\mathcal{T}}_{d_j}$, $\forall j \in \mathcal{T}$, where $|W^{\mathcal{T}}_{d_j}| = v_{\mathcal{T}} F$ bits, $\forall j \in \mathcal{T}$. Each user in $\mathcal{T} \setminus \{j\}$ must be able to cancel $W^{\mathcal{T}}_{d_j}$ from $X_{\mathcal{T},\boldsymbol{d}}$, in order to decode its requested piece. Consequently, $W^{\mathcal{T}}_{d_j}$ is constructed using the side-information cached by all the users in $\mathcal{T} \setminus \{j\}$ and not available at user j:

$$X_{\mathcal{T},\boldsymbol{d}} = \oplus_{j\in\mathcal{T}}\ W^{\mathcal{T}}_{d_j} = \oplus_{j\in\mathcal{T}} \Bigg(\bigcup_{\mathcal{S}\in\mathcal{B}^{\mathcal{T}}_j} W^{\mathcal{T}}_{d_j,\mathcal{S}} \Bigg), \tag{4}$$

where $W^{\mathcal{T}}_{d_j,\mathcal{S}} \subset W^{\mathcal{T}}_{d_j}$ which is stored exclusively at the users in the set $\mathcal{S}$ and

$$\mathcal{B}^{\mathcal{T}}_j \triangleq \Big\{ \mathcal{S} \subset [K] : \mathcal{T}\setminus\{j\} \subset \mathcal{S}, j \notin \mathcal{S} \Big\}, \forall j \in \mathcal{T}, \tag{5}$$

for example, for $K = 3$ and $i, j, k \in \{1, 2, 3\}$, $i \ne j \ne k$, the multicast signals are defined as

$$\begin{aligned} X_{\{i,j\},\boldsymbol{d}} &= W^{\{i,j\}}_{d_i} \oplus W^{\{i,j\}}_{d_j} \\ &= \Big(W^{\{i,j\}}_{d_i,\{j\}} \bigcup W^{\{i,j\}}_{d_i,\{j,k\}}\Big) \oplus \Big(W^{\{i,j\}}_{d_j,\{i\}} \bigcup W^{\{i,j\}}_{d_j,\{i,k\}}\Big), \end{aligned} \tag{6}$$

$$\begin{aligned} X_{\{1,2,3\},\boldsymbol{d}} &= W^{\{1,2,3\}}_{d_1} \oplus W^{\{1,2,3\}}_{d_2} \oplus W^{\{1,2,3\}}_{d_3} \\ &= W^{\{1,2,3\}}_{d_1,\{2,3\}} \oplus W^{\{1,2,3\}}_{d_2,\{1,3\}} \oplus W^{\{1,2,3\}}_{d_3,\{1,2\}}. \end{aligned} \tag{7}$$

where $|W^{\{i,j\}}_{d_i}| = |W^{\{i,j\}}_{d_j}| = v_{\{i,j\}} F$ bits and $|W^{\{1,2,3\}}_{d_1}| = |W^{\{1,2,3\}}_{d_2}| = |W^{\{1,2,3\}}_{d_3}| = v_{\{1,2,3\}} F$. $|W^{\mathcal{T}}_{d_j,\mathcal{S}}| = u^{\mathcal{T}}_{\mathcal{S}} F$ bits, i.e., the assignment variable $u^{\mathcal{T}}_{\mathcal{S}} \in [0, a_{\mathcal{S}}]$ represents the fraction of $W_{d_j,\mathcal{S}}$ involved in the multicast signal $X_{\mathcal{T},\boldsymbol{d}}$. Note that one subfile can contribute to multiple multicast transmissions, for example in a three-user system $\tilde{W}_{d_k,\{i,j\}}$ is used in $X_{\{i,k\},\boldsymbol{d}}$, $X_{\{j,k\},\boldsymbol{d}}$, $X_{\{i,j,k\},\boldsymbol{d}}$. Therefore, in order to guarantee that no redundant bits are transmitted, each subfile $\tilde{W}_{d_k,\mathcal{S}}$ is partitioned into disjoint pieces, e.g., for $K = 3$, we have

$$\tilde{W}_{d_k,\{i,j\}} = W^{\{i,k\}}_{d_k,\{i,j\}} \bigcup W^{\{j,k\}}_{d_k,\{i,j\}} \bigcup W^{\{i,j,k\}}_{d_k,\{i,j\}} \bigcup W^{\phi}_{d_k,\{i,j\}}, \tag{8}$$

where $W^{\phi}_{d_k,\mathcal{S}}$ denotes the remaining piece which is not involved in any transmission.

Remark 2: By contrast with [3], [28], [42], [44], where multicast signals of size $t+1$ utilize only the side-information stored exclusively at t users, i.e., $X_{\mathcal{T},\boldsymbol{d}} = \oplus_{k\in\mathcal{T}} W^{\mathcal{T}}_{d_k,\mathcal{T}\setminus\{k\}}$, the structure of the multicast signal in (4) represents all feasible utilizations of the side-information. This flexibility is instrumental in achieving the delivery load memory trade-off with uncoded placement $R^*_{\mathfrak{A}}$.

B. Unicast Signals $X_{\{i\}}$

A unicast signal $X_{\{i\}}$ delivers the fraction of the requested file which is not stored at user i and will not be delivered by

the multicast transmissions. For example, for $K = 3$, we have

$$X_{\{i\},\boldsymbol{d}} = W_{d_i} \setminus \Big(\bigcup_{\mathcal{S}: i \in \mathcal{S}} \tilde{W}_{d_i,\mathcal{S}} \bigcup W_{d_i}^{\{i,j\}} \bigcup W_{d_i}^{\{i,k\}} \bigcup W_{d_i}^{\{i,j,k\}} \Big), \quad i,j,k \in \{1,2,3\},\ i \neq j \neq k \quad (9)$$

where $\bigcup_{\mathcal{S}: i \in \mathcal{S}} \tilde{W}_{d_i,\mathcal{S}}$ is stored at user i and $W_{d_i}^{\mathcal{T}}$ is delivered to user i via $X_{\mathcal{T},\boldsymbol{d}}$.

C. Delivery Phase Constraints

Recall that $v_{\mathcal{T}} \in [0,\ 1]$ and $u_{\mathcal{S}}^{\mathcal{T}} \in [0, a_{\mathcal{S}}]$ represent $|X_{\mathcal{T},\boldsymbol{d}}|/F$, and $|W_{d_j,\mathcal{S}}^{\mathcal{T}}|/F$, respectively. Our delivery scheme can be represented by constraints on $v_{\mathcal{T}}$ and $u_{\mathcal{S}}^{\mathcal{T}}$ as follows. First, the structure of the multicast signals in (6), (7) imposes

$$\sum_{\mathcal{S} \in \mathcal{B}_j^{\mathcal{T}}} u_{\mathcal{S}}^{\mathcal{T}} = v_{\mathcal{T}},\ \forall\, \mathcal{T} \subsetneq_{\phi} [K],\ \forall\, j \in \mathcal{T}. \quad (10)$$

For example, for $K = 3$, we have

$$v_{\{i,j\}} = u_{\{j\}}^{\{i,j\}} + u_{\{j,k\}}^{\{i,j\}} = u_{\{i\}}^{\{i,j\}} + u_{\{i,k\}}^{\{i,j\}}, \quad (11)$$

$$v_{\{1,2,3\}} = u_{\{2,3\}}^{\{1,2,3\}} = u_{\{1,3\}}^{\{1,2,3\}} = u_{\{1,2\}}^{\{1,2,3\}}. \quad (12)$$

In order to prevent transmitting redundant bits from the subfile $\tilde{W}_{d_j,\mathcal{S}}$ to user j, we need

$$\sum_{\mathcal{T} \subsetneq_{\phi} [K]:\ j \in \mathcal{T}, \mathcal{T} \cap \mathcal{S} \neq \phi, \mathcal{T} \setminus \{j\} \subset \mathcal{S}} u_{\mathcal{S}}^{\mathcal{T}} \leq a_{\mathcal{S}},\ \forall\, j \notin \mathcal{S}, \quad \forall\, \mathcal{S} \in \Big\{ \tilde{\mathcal{S}} \subset [K] :\ 2 \leq |\tilde{\mathcal{S}}| \leq K-1 \Big\}, \quad (13)$$

where the condition $\mathcal{T} \setminus \{j\} \subset \mathcal{S}$ follows from (5).

For example, for $K = 3$, (13) implies

$$u_{\{i,j\}}^{\{i,k\}} + u_{\{i,j\}}^{\{j,k\}} + u_{\{i,j\}}^{\{i,j,k\}} \leq a_{\{i,j\}}. \quad (14)$$

Finally, the delivery signals sent by the server must complete all the requested files:

$$\sum_{\mathcal{T} \subsetneq_{\phi} [K]:\ k \in \mathcal{T}} v_{\mathcal{T}} \geq 1 - \sum_{\mathcal{S} \subset [K]:\ k \in \mathcal{S}} a_{\mathcal{S}}, \forall\, k \in [K], \quad (15)$$

for example, for $K = 3$, the delivery completion constraint for user i is given by

$$v_{\{i\}} + v_{\{i,j\}} + v_{\{i,k\}} + v_{\{i,j,k\}} \geq 1 - (a_{\{i\}} + a_{\{i,j\}} + a_{\{i,k\}} + a_{\{i,j,k\}}). \quad (16)$$

Therefore, for given $\boldsymbol{a}$, the set of feasible delivery schemes, $\mathfrak{D}(\boldsymbol{a})$, is defined as

$$\mathfrak{D}(\boldsymbol{a}) = \Bigg\{ (\boldsymbol{v}, \boldsymbol{u}) \Bigg| \sum_{\mathcal{T} \subsetneq_{\phi} [K]: k \in \mathcal{T}} v_{\mathcal{T}} \geq 1 - \sum_{\mathcal{S} \subset [K]: k \in \mathcal{S}} a_{\mathcal{S}}, \forall k \in [K],\ \sum_{\mathcal{S} \in \mathcal{B}_j^{\mathcal{T}}} u_{\mathcal{S}}^{\mathcal{T}} = v_{\mathcal{T}},\ \forall \mathcal{T} \subsetneq_{\phi} [K],\ \forall j \in \mathcal{T},$$
$$\sum_{\mathcal{T} \subsetneq_{\phi} [K]:\ j \in \mathcal{T}, \mathcal{T} \cap \mathcal{S} \neq \phi, \mathcal{T} \setminus \{j\} \subset \mathcal{S}} u_{\mathcal{S}}^{\mathcal{T}} \leq a_{\mathcal{S}},\ \forall j \notin \mathcal{S},\ \forall \mathcal{S} \in \Big\{ \tilde{\mathcal{S}} \subset [K] :\ 2 \leq |\tilde{\mathcal{S}}| \leq K-1 \Big\},$$
$$0 \leq u_{\mathcal{S}}^{\mathcal{T}} \leq a_{\mathcal{S}}, \forall \mathcal{T} \subsetneq_{\phi} [K], \forall \mathcal{S} \in \bigcup_{j \in \mathcal{T}} \mathcal{B}_j^{\mathcal{T}} \Bigg\}, \quad (17)$$

where the transmission and assignment variables are represented by $\boldsymbol{v}$ and $\boldsymbol{u}$ respectively.

D. Discussion

The linear constraints in (17) guarantee the delivery of the requested files. Successful delivery is guaranteed by 1) By (10), user $j \in \mathcal{T}$ can retrieve $W_{d_j}^{\mathcal{T}}$ from the signal $X_{\mathcal{T},\boldsymbol{d}}$. 2) By (13) and (15), W_{d_j} can be reconstructed from the pieces decoded at user j. The delivery completion constraints ensure that the number of decoded bits are sufficient for decoding the file, and the redundancy constraints prevent the server from transmitting redundant bits. Formally, we have:

Proposition 1: For $\mathcal{S}' \subset [K]$ such that $1 \leq |\mathcal{S}'| \leq K-2$, and some user $j \notin \mathcal{S}'$, the size of the multicast transmissions $X_{\mathcal{T},\boldsymbol{d}}$, where $\{j\} \cup \mathcal{S}' \subset \mathcal{T}$, is limited by the amount of side-information stored at the users in $\mathcal{S}'$ and not available at user j, i.e.,

$$\sum_{\mathcal{T} \subsetneq_{\phi} [K]:\ \{j\} \cup \mathcal{S}' \subset \mathcal{T}} v_{\mathcal{T}} \leq \sum_{\mathcal{S} \subset [K]:\ \mathcal{S}' \subset \mathcal{S}, j \notin \mathcal{S}} a_{\mathcal{S}}, \quad (18)$$

which is guaranteed by (10) and (13).

The proof of Proposition 1 is provided in Appendix A.

VI. Formulation and Results

In this section, we first show that the optimal uncoded placement and linear delivery schemes can be obtained by solving a linear program. Next, we present a lower bound on the delivery load with uncoded placement. Based on this bound, we show that linear delivery is optimal with uncoded placement for three cases; namely, $\sum_{k=1}^{K} m_k \leq 1$, $\sum_{k=1}^{K} m_k \geq K-1$, and the three-user case. That is, for these cases we explicitly characterize the delivery load memory trade-off with uncoded placement $R_{\mathfrak{A}}^{*}(\boldsymbol{m})$.

A. Caching Scheme Optimization

In Sections IV and V, we have demonstrated that an uncoded placement scheme in $\mathfrak{A}$ is completely characterized by the allocation vector $\boldsymbol{a}$, which represents the fraction of files stored exclusively at each subset of users $\mathcal{S} \subset [K]$. Additionally, the assignment and transmission vectors $(\boldsymbol{u}, \boldsymbol{v})$ completely characterize a delivery scheme in $\mathfrak{D}$, where $\boldsymbol{v}$ represents the size of the transmitted signals, and $\boldsymbol{u}$ determines the structure of the transmitted signals. For a given normalized memory vector $\boldsymbol{m}$, the following optimization problem characterizes the minimum worst-case delivery load $R_{\mathfrak{A},\mathfrak{D}}^{*}(\boldsymbol{m})$ and the optimal caching scheme in $\mathfrak{A}, \mathfrak{D}$, i.e., the optimal values for $\boldsymbol{a}$, $\boldsymbol{v}$, and $\boldsymbol{u}$.

$$\textbf{\textit{O1}}: \quad \min_{\boldsymbol{a},\boldsymbol{u},\boldsymbol{v}} \quad \sum_{\mathcal{T} \subsetneq_{\phi} [K]} v_{\mathcal{T}} \quad (19a)$$

$$\text{s.t. } \boldsymbol{a} \in \mathfrak{A}(\boldsymbol{m}),\ \text{and } (\boldsymbol{u}, \boldsymbol{v}) \in \mathfrak{D}(\boldsymbol{a}), \quad (19b)$$

where $\mathfrak{A}(\boldsymbol{m})$ and $\mathfrak{D}(\boldsymbol{a})$ are defined in (1) and (17), respectively.

Remark 3: For equal cache sizes, $R_{\mathfrak{A},\mathfrak{D}}^{*}(\boldsymbol{m})$ is equal to the worst-case delivery load of [3], which was shown to be optimal for uncoded placement in [4] for $N \geq K$. For $N < K$, the optimal scheme for uncoded placement was proposed in [6]. The solution of (19) is equivalent to the memory sharing solution proposed in [3].

Remark 4: In Section V, $X_{\mathcal{T},\boldsymbol{d}}$ is formed by XORing pieces of equal size. A delivery scheme with $\tilde{X}_{\mathcal{T},\boldsymbol{d}} = \bar{\oplus}_{j\in\mathcal{T}} W^{\mathcal{T}}_{d_j}$, where $\bar{\oplus}$ denotes an XOR operation that allows zero padding so that the pieces are of equal length, is equivalent to a delivery scheme in $\mathfrak{D}$ and both yield the same delivery load. For example, $\tilde{X}_{\{1,2\},\boldsymbol{d}} = W^{\{1,2\}}_{d_1,\{2\}} \bar{\oplus} W^{\{1,2\}}_{d_2,\{1\}}$, with $u^{\{1,2\}}_{\{2\}} > u^{\{1,2\}}_{\{1\}}$, is equivalent to a multicast signal $X_{\{1,2\},\boldsymbol{d}}$ and a unicast signal $X_{\{2\},\boldsymbol{d}}$, with sizes $u^{\{1,2\}}_{\{1\}}F$ bits, and $\left(u^{\{1,2\}}_{\{2\}} - u^{\{1,2\}}_{\{1\}}\right)F$ bits, respectively.

Remark 5: In this work, we assume the file size to be large enough, such that the cache placement and delivery schemes can be tailored to the unequal cache sizes by optimizing over continuous variables. More specifically, for F large enough, $u^{\mathcal{T}}_{\mathcal{S}}F$ can be used instead of $\lceil u^{\mathcal{T}}_{\mathcal{S}}F\rceil$ for $u^{\mathcal{T}}_{\mathcal{S}} \in [0,1]$. The required subpacketization level is the least-common-denominator of the assignment variables $u^{\mathcal{T}}_{\mathcal{S}}$. That is, the minimum packet size is equal to the greatest-common-divisor (gcd) of all $u^{\mathcal{T}}_{\mathcal{S}}F$, assuming $u^{\mathcal{T}}_{\mathcal{S}}F$ are integers.

B. Lower Bounds

Next, using the approach in [4], [6], we show that $R^*_{\mathfrak{A}}(\boldsymbol{m})$ is lower bounded by the linear program in (20).

Theorem 1: (Uncoded placement bound) Given K, $N \geq K$, and $\boldsymbol{m}$, the minimum worst-case delivery load with uncoded placement $R^*_{\mathfrak{A}}(\boldsymbol{m})$ is lower bounded by

$$\textbf{\textit{O2}}: \quad \max_{\lambda_0\in\mathbb{R},\lambda_k\geq 0,\alpha_{\boldsymbol{q}}\geq 0} -\lambda_0 - \sum_{k=1}^{K} m_k\lambda_k \tag{20a}$$

$$\text{s.t. } \lambda_0 + \sum_{k\in\mathcal{S}}\lambda_k + \gamma_{\mathcal{S}} \geq 0,\ \forall \mathcal{S}\subset[K], \tag{20b}$$

$$\sum_{\boldsymbol{q}\in\mathcal{P}_{[K]}} \alpha_{\boldsymbol{q}} = 1, \tag{20c}$$

where

$$\gamma_{\mathcal{S}} \triangleq \begin{cases} K, & \text{for } \mathcal{S}=\phi,\\ 0, & \text{for } \mathcal{S}=[K],\\ \sum\limits_{j=1}^{K-|\mathcal{S}|} \sum\limits_{\substack{\boldsymbol{q}\in\mathcal{P}_{[K]}:\, q_{j+1}\in\mathcal{S},\\ \{q_1,\ldots,q_j\}\cap\mathcal{S}=\phi}} j\,\alpha_{\boldsymbol{q}}, & \text{otherwise.}\end{cases} \tag{21}$$

and $\mathcal{P}_{[K]}$ is the set of all permutations of $[1,2,\ldots,K]$.

The proof of Theorem 1 is provided in Appendix B. We compare the achievable delivery load $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$ with the following lower bounds on the worst-case delivery load R^*. From [41], $R^*(\boldsymbol{m})$ is lower bounded by

$$\max_{s\in[K],\, l\in\left[\lceil\frac{N}{s}\rceil\right]} \left\{ \frac{N-(N-Kl)^+}{l} - \frac{sN\sum\limits_{i=1}^{s+\gamma} m_i + \gamma(N-ls)^+}{l(s+\gamma)} \right\}, \tag{22}$$

where $\gamma \triangleq \min\left\{\left(\lceil\frac{N}{l}\rceil - s\right)^+, K-s\right\}$ and $m_1 \leq \cdots \leq m_K$. The following proposition is a straightforward generalization of the lower bounds in [13] for systems with distinct cache sizes.

Proposition 2: Given K, N, $\boldsymbol{m}$, and $m_1 \leq \cdots \leq m_K$, we have

$$R^*(\boldsymbol{m}) \geq \max\left\{ \max_{s\in[\min\{K,N\}]} \left\{ s - \sum_{k=1}^{s} \frac{N\sum_{i=1}^{k} m_i}{N-k+1} \right\}, \max_{s\in[\min\{K,N\}]} \left\{ s\Big(1-\sum_{i=1}^{s} m_i\Big) \right\} \right\}. \tag{23}$$

C. Special Cases

Next, we consider three special cases, for which we explicitly characterize $R^*_{\mathfrak{A}}(\boldsymbol{m})$ and show that the solution of (19) is the optimal caching scheme with uncoded placement. First, we consider the small cache regime, $\sum_{i=1}^{K} m_i \leq 1$,

Theorem 2: The minimum worst-case delivery load with uncoded placement is given by

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = K - \sum_{j=1}^{K}(K-j+1)m_j, \tag{24}$$

for $m_1 \leq \cdots \leq m_K$, $\sum_{i=1}^{K} m_i \leq 1$, and $N \geq K$.

Proof: **Achievability:** In the placement phase, each file is split into $K+1$ subfiles such that $a_{\{j\}} = m_j$ and $a_{\phi} = 1 - \sum_{k=1}^{K} m_k$. In the delivery phase, we have $v_{\{j\}} = 1 - \sum_{i=1}^{j-1} m_i - (K-j+1)m_j$ and $v_{\{i,j\}} = u^{\{i,j\}}_{\{i\}} = u^{\{i,j\}}_{\{j\}} = \min\{a_{\{i\}}, a_{\{j\}}\}$. In turn, $R_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m}) = K - \sum_{j=1}^{K}(K-j+1)m_j$ is achievable. **Converse:** By substituting $\alpha_{[1,2,\ldots,K]}=1$ in Theorem 1, we get

$$\max_{\lambda_k\geq 0,\lambda_0} \quad -\lambda_0 - \sum_{k=1}^{K} m_k\,\lambda_k \tag{25a}$$

$$\text{s.t. } \lambda_0 + \sum_{k\in\mathcal{S}}\lambda_k + \gamma_{\mathcal{S}} \geq 0,\ \forall \mathcal{S}\subset[K], \tag{25b}$$

where $\gamma_{\mathcal{S}} = j-1$ if $\{j\} \in \mathcal{S}$ and $[j-1]\cap\mathcal{S} = \phi$ for $j\in[K]$. $\lambda_0 = -K$, $\lambda_j = K-j+1$ is a feasible solution to (25), since $\lambda_0 + \lambda_j + (j-1) = 0$. Therefore, $R^*_{\mathfrak{A}}(\boldsymbol{m}) \geq K - \sum_{j=1}^{K}(K-j+1)m_j$. ∎

Next theorem characterizes $R^*_{\mathfrak{A}}(\boldsymbol{m})$ in the large total memory regime where $\sum_{i=1}^{K} m_i \geq K-1$.

Theorem 3: The minimum worst-case delivery load with uncoded placement is given by

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = R^*(\boldsymbol{m}) = 1 - m_1, \tag{26}$$

for $m_1 \leq \cdots \leq m_K$, $\sum_{i=1}^{K} m_i \geq K-1$, and $N \geq K$.

Proof: **Achievability:** In the placement phase, W_j is partitioned into subfiles $\tilde{W}_{j,[K]\setminus\{i\}}, i \in [K]$ and $\tilde{W}_{j,[K]}$, such that $a_{[K]} = \sum_{i=1}^{K} m_i - (K-1)$ and $a_{[K]\setminus\{i\}} = 1-m_i$, $i\in[K]$. In the delivery phase, we have the following cases.

- For $(K-2)m_1 \geq \sum_{i=2}^{K} m_i - 1$, we have the following multicast transmissions

$$X_{[K]\setminus\{i\},\boldsymbol{d}} = \oplus_{k\in[K]\setminus\{i\}} W^{[K]\setminus\{i\}}_{d_k,[K]\setminus\{k\}}, i\in\{2,\ldots,K\}, \tag{27}$$

$$X_{[K],\boldsymbol{d}} = \oplus_{k\in[K]} W^{[K]}_{d_k,[K]\setminus\{k\}}, \tag{28}$$

with

$$v_{[K]\backslash\{i\}} = m_i - m_1,\ i \in \{2,\dots,K\}, \tag{29}$$

$$v_{[K]} = 1 + (K-2)m_1 - \sum_{k=2}^{K} m_k. \tag{30}$$

- For $(K-l-1)m_l < \sum_{i=l+1}^{K} m_i - 1$ and $(K-l-2)m_{l+1} \geq \sum_{i=l+2}^{K} m_i - 1$, where $l \in [K-2]$, we have the following transmissions

$$X_{[i],\boldsymbol{d}} = \oplus_{k\in[i]} W^{[i]}_{d_k,[K]\backslash\{k\}},\ i \in [l], \tag{31}$$

$$X_{[K]\backslash\{i\},\boldsymbol{d}} = \oplus_{k\in[K]\backslash\{i\}} W^{[K]\backslash\{i\}}_{d_k,[K]\backslash\{k\}},\ i\in\{l+1,\dots,K\}. \tag{32}$$

with

$$v_{[i]} = m_{i+1} - m_i,\ i \in [l-1], \tag{33}$$

$$v_{[l]} = \frac{1}{K-l-1}\left(\sum_{j=l+1}^{K} m_j - 1 - (K-l-1)m_l\right), \tag{34}$$

$$v_{[K]\backslash\{i\}} = \frac{1}{K-l-1}\left((K-l-1)m_i + 1 - \sum_{j=l+1}^{K} m_j\right), \quad i \in \{l+1,\dots,K\}. \tag{35}$$

In both cases, the size of the assignment variables satisfies $u^{\mathcal{T}}_{[K]\backslash\{k\}} = v_{\mathcal{T}}, \forall k \in \mathcal{T}$. **Converse:** By substituting $s = 1$ in (23), we get $R^*(\boldsymbol{m}) \geq 1 - m_1$. ∎

In the next theorem, we characterize $R^*_{\mathfrak{A}}(\boldsymbol{m})$ for $K = 3$.

Theorem 4: For $K = 3$, $N \geq 3$, and $m_1 \leq m_2 \leq m_3$, the minimum worst-case delivery load with uncoded placement

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = \max\left\{3 - \sum_{j=1}^{3}(4-j)m_j,\ \tfrac{5}{3} - \sum_{j=1}^{3}\tfrac{(4-j)m_j}{3},\ 2 - 2m_1 - m_2,\ 1 - m_1\right\}. \tag{36}$$

In particular, we have the following regions

1) *For $\sum_{j=1}^{3} m_j \leq 1$, $R^*_{\mathfrak{A}}(\boldsymbol{m}) = 3 - \sum_{j=1}^{3}(4-j)m_j$.*
2) *For $1 \leq \sum_{j=1}^{3} m_j \leq 2$, we have three cases*
   - *If $m_3 < m_2 + 3m_1 - 1$, and $2m_2 + m_3 < 2$, then $R^*_{\mathfrak{A}}(\boldsymbol{m}) = \frac{5}{3} - \sum_{j=1}^{3}\frac{(4-j)m_j}{3}$.*
   - *If $m_3 \geq m_2 + 3m_1 - 1$, and $m_1 + m_2 < 1$, then $R^*_{\mathfrak{A}}(\boldsymbol{m}) = 2 - 2m_1 - m_2$.*
   - *If $2m_2 + m_3 \geq 2$, and $m_1 + m_2 \geq 1$, then $R^*_{\mathfrak{A}}(\boldsymbol{m}) = 1 - m_1$.*
3) *For $\sum_{j=1}^{3} m_j \geq 2$, $R^*_{\mathfrak{A}}(\boldsymbol{m}) = 1 - m_1$.*

Proof of Theorem 4 is provided in Appendix C.

Remark 6: By substituting $m_3 = 1$ in Theorem 4, we obtain the two-user delivery load memory trade-off with uncoded placement, given as $R^*_{\mathfrak{A}}(\boldsymbol{m}) = \max\{2 - 2\ m_1 - m_2,\ 1 - m_1\}$.

Remark 7: From the proposed schemes, we observe that the allocation variables satisfy

$$\sum_{\mathcal{S}\subset[K]:\ |\mathcal{S}|=t} a_{\mathcal{S}} = t + 1 - \sum_{i=1}^{K} m_i, \tag{37}$$

$$\sum_{\mathcal{S}\subset[K]:\ |\mathcal{S}|=t+1} a_{\mathcal{S}} = \sum_{i=1}^{K} m_i - t, \tag{38}$$

for $t < \sum_{i=1}^{K} m_i \leq t+1$ and $a_{\mathcal{S}} = 0$ for $|\mathcal{S}| \notin \{t, t+1\}$. That is, the proposed placement scheme generalizes the memory sharing scheme in [3], where $a_{\mathcal{S}} = a_{\mathcal{S}'}$ for $|\mathcal{S}| = |\mathcal{S}'|$.

D. Comparison With Other Schemes With Heterogeneous Cache Sizes

Previous work includes the layered heterogeneous caching (LHC) scheme [28], [42], where the users are divided into groups and within each group the users' cache memories are divided into layers such that the users in each layer have equal cache sizes. The Maddah-Ali–Niesen caching scheme [3] is applied to the fraction of the file assigned to each layer. Let $R_{\text{LHC}}(\boldsymbol{m})$ denote delivery load of this scheme. We have

Proposition 3: Given $K, N \geq K$ and $\boldsymbol{m}$, we have $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m}) \leq R_{LHC}(\boldsymbol{m})$.

Proof: LHC scheme is a feasible (but not necessarily optimal) solution to (19) shown as follows. *Grouping:* Dividing the users into disjoint groups can be represented in the placement phase by setting $a_{\mathcal{S}} = 0$ for any set $\mathcal{S}$ containing two or more users from different groups. Similarly, in the delivery phase $v_{\mathcal{T}} = 0$ if $\{q_1, q_2\} \subset \mathcal{T}$, q_1 and q_2 belong to distinct groups. *Layering:* Without loss of generality, assume there is one group, i.e. there are K layers. Let α_l be the fraction of the file considered in layer l, and assign $a_{\mathcal{S},l}$, $v_{\mathcal{T},l}$, and $u^{\mathcal{T}}_{\mathcal{S},l}$ to layer l, such that $a_{\mathcal{S}} = \sum_{l=1}^{K} a_{\mathcal{S},l}$, $v_{\mathcal{T}} = \sum_{l=1}^{K} v_{\mathcal{T},l}$, and $u^{\mathcal{T}}_{\mathcal{S}} = \sum_{l=1}^{K} u^{\mathcal{T}}_{\mathcal{S},l}$. Additionally, we have $\sum_{\mathcal{S}\subset\{l,\dots,K\}} a_{\mathcal{S},l} = \alpha_l$, and $\sum_{\mathcal{S}\subset\{l,\dots,K\}:\ \{k\}\in\mathcal{S}} a_{\mathcal{S},l} + \sum_{\mathcal{T}\subset\{l,\dots,K\}:\ \{k\}\in\mathcal{T}} v_{\mathcal{T},l} \geq \alpha_l$ for $k \in \{l,\dots,K\}$. Thus, the LHC scheme is a feasible solution to (19). ∎

The recent reference [44] has proposed optimizing over uncoded placement schemes $\mathfrak{A}$ with the decentralized delivery scheme in [27], i.e., the multicast signals are defined as $X_{\mathcal{T},\boldsymbol{d}} = \bar{\oplus}_{k\in\mathcal{T}} \tilde{W}_{d_k,\mathcal{T}\backslash\{k\}}$ where $v_{\mathcal{T}} = \max_{k\in\mathcal{T}} a_{\mathcal{T}\backslash\{k\}}$, which limits the multicast opportunities [33] as illustrated in Section III.

Remark 8: For fixed cache contents, reference [33] proposed a procedure for redesigning the multicast signals, formed by XORing pieces of unequal size, in order to better utilize the side-information stored at the users. In contrast, our scheme is a centralized scheme, where we jointly optimize the cache contents and the delivery procedure which allows flexible utilization of the side-information at the users.

Different from [28], [42], [44], we propose a more general delivery scheme that allows flexible utilization of the side-information. Both our solution and that of [44] is exponential in the number of users. Notably, for systems with only two distinct cache sizes over all users, reference [44] has a caching

scheme which is obtained by solving an optimization problem with polynomial complexity.

VII. Optimizing Cache Sizes With Total Memory Budget

In this section, we consider a centralized system where the server is connected to the K users via rate limited download links of distinct capacities, as described by Fig. 1(b).

A. Problem Formulation

Next, we consider the joint optimization of both the caching scheme and the users' cache sizes for given download rates $\boldsymbol{C}$ (see Fig. 1(b)) and normalized cache budget m_{tot}. More specifically, the minimum worst-case DCT, $\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}},\boldsymbol{C})$, is characterized by

$$\textbf{\textit{O3}}: \quad \min_{\boldsymbol{a},\boldsymbol{u},\boldsymbol{v},\boldsymbol{m}} \sum_{\mathcal{T}\subsetneq_\phi[K]} \frac{v_{\mathcal{T}}}{\min\limits_{j\in\mathcal{T}} C_j} \tag{39a}$$

$$\text{s.t. } \boldsymbol{a}\in\mathfrak{A}(\boldsymbol{m}),\ (\boldsymbol{v},\boldsymbol{u})\in\mathfrak{D}(\boldsymbol{a}), \tag{39b}$$

$$\sum_{k=1}^{K} m_k \le m_{\text{tot}}, \tag{39c}$$

$$0\le m_k\le 1,\ \forall\, k\in[K], \tag{39d}$$

where $\mathfrak{A}(\boldsymbol{m})$ is defined in (1) and $\mathfrak{D}(\boldsymbol{a})$ is defined in (17).

B. Optimal Cache Sizes

The linear program in (39) characterizes the optimal memory allocation assuming uncoded placement and linear delivery schemes. For the case where $m_{\text{tot}}\le 1$, we are able to derive a closed form expression for the optimal memory allocation, and show that the optimal solution balances between allocating larger cache memories to users with low decoding rates and uniform memory allocation. In particular, the cache memory budget m_{tot} is allocated uniformly between users $\{1,\dots,q\}$, where q is determined by $\boldsymbol{C}$ as illustrated in the following theorem.

Theorem 5: For $C_1\le\cdots\le C_K$ and $m_{tot}\le 1$, the minimum worst-case delivery completion time (DCT) is given by

$$\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{tot},\boldsymbol{C}) = \sum_{j=1}^{K}\frac{1}{C_j} - \max_{i\in[K]}\left\{\sum_{j=1}^{i}\frac{j\,m_{tot}}{i\,C_j}\right\}, \tag{40}$$

and the optimal memory allocation is given by $m_1^*=\cdots=m_q^*=\frac{m_{tot}}{q}$, where the user index $q=\operatorname{argmax}_{i\in[K]}\left\{\sum_{j=1}^{i} j/(i\,C_j)\right\}$.

Proof of Theorem 5 is provided in Appendix D. Note that if the optimal solution is not unique, i.e., $q\in\{q_1,\dots,q_L\}$, for some $L\le K$, then $\boldsymbol{m}^*=\sum_{i=1}^{L}\alpha_i\big[\frac{m_{\text{tot}}}{q_i},\dots,\frac{m_{\text{tot}}}{q_i},0,\dots,0\big]$, where $\sum_{i=1}^{L}\alpha_i=1$ and $\alpha_i\ge 0$. The next proposition shows that uniform memory allocation combined with the Maddah-Ali–Niesen caching scheme yields an upper bound on $\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}},\boldsymbol{C})$.

Proposition 4: For $m_{tot}\in[K]$ and $C_1\le C_2\le\cdots\le C_K$, we have

$$\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{tot},\boldsymbol{C})\le\Theta_{unif}(m_{tot},\boldsymbol{C}) = \frac{1}{\binom{K}{m_{tot}}}\sum_{j=1}^{K-m_{tot}}\frac{\binom{K-j}{m_{tot}}}{C_j}. \tag{41}$$

Proof: Assuming $m_j=m_{\text{tot}}/K,\forall j\in[K]$, the placement phase is described by $a_{\mathcal{S}}=1/\binom{K}{m_{\text{tot}}}$ for $|\mathcal{S}|=m_{\text{tot}}$ and zero otherwise. While, the delivery phase is defined by $v_{\mathcal{T}}=1/\binom{K}{m_{\text{tot}}}$ for $|\mathcal{T}|=m_{\text{tot}}+1$ and $u^{\mathcal{T}}_{\mathcal{S}}=v_{\mathcal{T}}$ for $\mathcal{S}\in\{\mathcal{T}\setminus\{j\}: j\in\mathcal{T}\}$. In turn, we have

$$\begin{aligned}\Theta_{\text{unif}}(m_{\text{tot}},\boldsymbol{C}) &= \sum_{\mathcal{T}\subsetneq_\phi[K]}\frac{v_{\mathcal{T}}}{\min\limits_{j\in\mathcal{T}}C_j},\\ &= \frac{1}{\binom{K}{m_{\text{tot}}}}\sum_{\mathcal{T}\subsetneq_\phi[K]:\ |\mathcal{T}|=m_{\text{tot}}+1}\frac{1}{\min\limits_{j\in\mathcal{T}}C_j},\\ &= \frac{1}{\binom{K}{m_{\text{tot}}}}\sum_{j=1}^{K-m_{\text{tot}}}\frac{\binom{K-j}{m_{\text{tot}}}}{C_j},\end{aligned} \tag{42}$$

since $C_1\le C_2\le\cdots\le C_K$ and there are $\binom{K-j}{m_{\text{tot}}}$ sets of size $m_{\text{tot}}+1$ that include user j and do not include users $\{1,2,\dots,j-1\}$. Finally, $\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}},\boldsymbol{C})\le\Theta_{\text{unif}}(m_{\text{tot}},\boldsymbol{C})$, since uniform memory allocation is a feasible solution (but not necessarily optimal) to (39). ∎

VIII. Numerical Results

First, we provide a numerical example for the optimal caching scheme obtained from (19).

Example 1: Consider a caching system with $K=3$, and $\boldsymbol{m}=[0.4,\ 0.5,\ 0.6]$. The caching scheme obtained from (19), is described as follows.

Placement phase: Every file $W^{(l)}$ is divided into six subfiles, such that $a_{\{1\}}=7/30$, $a_{\{2\}}=4/30$, $a_{\{3\}}=4/30$, $a_{\{1,2\}}=1/30$, $a_{\{1,3\}}=4/30$, and $a_{\{2,3\}}=10/30$.

Delivery phase: We have the following transmissions

$$X_{\{1,2\},\boldsymbol{d}} = \left(W^{\{1,2\}}_{d_1,\{2\}}\bigcup W^{\{1,2\}}_{d_1,\{2,3\}}\right)\oplus\left(W^{\{1,2\}}_{d_2,\{1\}}\bigcup W^{\{1,2\}}_{d_2,\{1,3\}}\right),$$

$$X_{\{2,3\},\boldsymbol{d}} = W^{\{2,3\}}_{d_2,\{3\}}\oplus W^{\{2,3\}}_{d_3,\{2\}},$$

$$X_{\{1,3\},\boldsymbol{d}} = \left(W^{\{1,3\}}_{d_1,\{3\}}\bigcup W^{\{1,3\}}_{d_1,\{2,3\}}\right)\oplus W^{\{1,3\}}_{d_3,\{1\}},$$

$$X_{\{1,2,3\},\boldsymbol{d}} = W^{\{1,2,3\}}_{d_1,\{2,3\}}\oplus W^{\{1,2,3\}}_{d_2,\{1,3\}}\oplus W^{\{1,2,3\}}_{d_3,\{1,2\}},$$

and the subfile sizes are as follows

1) *$v_{\{1,2\}}=10/30$, where $u^{\{1,2\}}_{\{2\}}=a_{\{2\}}$, $u^{\{1,2\}}_{\{2,3\}}=0.2$, $u^{\{1,2\}}_{\{1\}}=a_{\{1\}}$, and $u^{\{1,2\}}_{\{1,3\}}=0.1$.*
2) *$v_{\{1,3\}}=7/30$, where $u^{\{1,3\}}_{\{3\}}=a_{\{3\}}$, $u^{\{1,3\}}_{\{2,3\}}=0.1$, and $u^{\{1,3\}}_{\{1\}}=a_{\{1\}}$.*
3) *$v_{\{2,3\}}=4/30$, where $u^{\{2,3\}}_{\{3\}}=a_{\{3\}}$ and $u^{\{2,3\}}_{\{2\}}=a_{\{2\}}$.*
4) *$v_{\{1,2,3\}}=1/30$, where $u^{\{1,2,3\}}_{\{2,3\}}=u^{\{1,2,3\}}_{\{1,3\}}=u^{\{1,2,3\}}_{\{1,2\}}=a_{\{1,2\}}$.*

The minimum worst-case delivery load $R^*_{\mathfrak{A}}(\boldsymbol{m})=R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})=22/30$. Note that, per Remark 5, the required subpacketization level for the proposed scheme is 30, i.e., the minimum packet size is given by

$$\gcd(\boldsymbol{u})=\gcd\left(\frac{7F}{30},\frac{6F}{30},\frac{4F}{30},\frac{3F}{30},\frac{F}{30}\right)=\frac{F}{30},$$

for $F=30n$, $n=1,2,\dots$.

In Fig. 3, we compare the delivery load $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$ obtained from optimization problem (19), with the lower bounds on

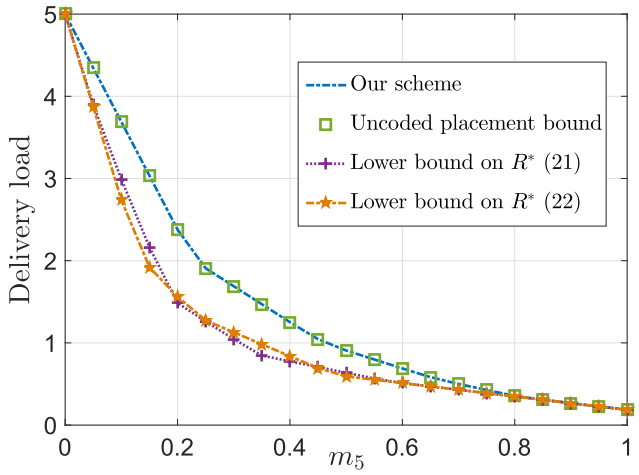


Fig. 3. Comparing $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$ with the lower bounds in (20)-(23), for $K = 5$, and $m_k = 0.95\, m_{k+1}$.

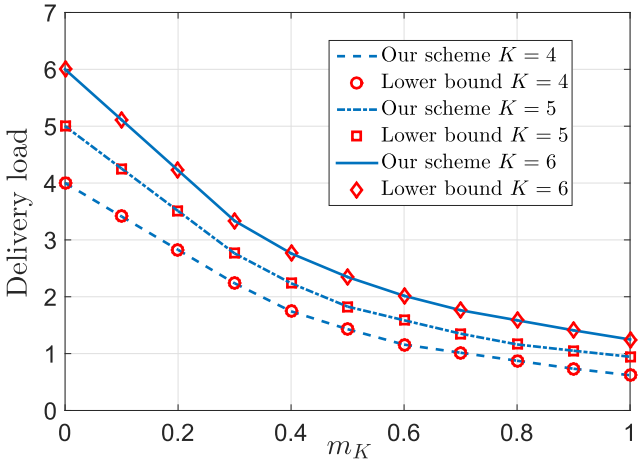


Fig. 4. Comparing $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$ with the lower bounds on $R^*_{\mathfrak{A}}(\boldsymbol{m})$, for $m_k = 0.75\, m_{k+1}$.

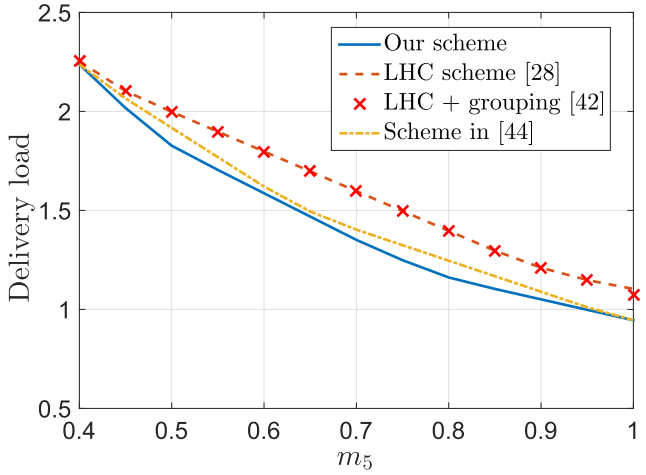


Fig. 5. Comparing $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$ with the achievable delivery loads in [28], [42], [44], for $K = 5$ and $m_k = 0.75\, m_{k+1}$.

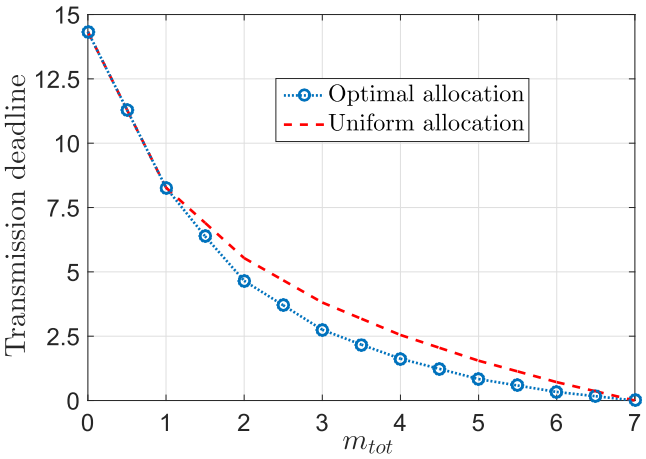


Fig. 6. Comparing $\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}}, \boldsymbol{C})$ and $\Theta_{\text{unif}}(m_{\text{tot}}, \boldsymbol{C})$ for $\boldsymbol{C} = [0.2, 0.4, 0.6, 0.6, 0.8, 0.8, 1]$.

$R^*(\boldsymbol{m})$ in (22), (23), and the lower bound with uncoded placement in (20), for $N = K = 5$ and $m_k = 0.95\, m_{k+1}$.

From Fig. 3, we observe that $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m}) = R^*_{\mathfrak{A}}(\boldsymbol{m})$, which is also demonstrated in Fig. 4, for $K = 4, 5, 6$, and $m_k = 0.75\, m_{k+1}$. In Fig. 5, we compare $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$ with the achievable delivery loads in [28], [42], [44], for $K = 5$ and $m_k = 0.75\, m_{k+1}$.

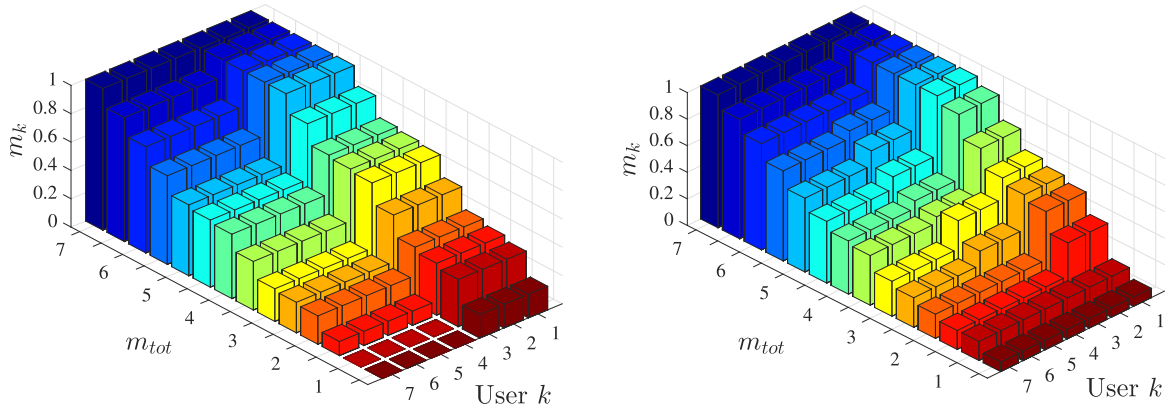
The next example concerns with solving (39) for a system with unequal download rates, and comparing the optimal memory allocation and caching scheme with the Maddah-Ali–Niesen caching scheme with uniform memory allocation.

Example 2: Consider a caching system with $K = 3$, memory budget $m_{tot} = 1$, and $C_1 \leq C_2 \leq C_3$, which implies $\Theta_{unif}(1, \boldsymbol{C}) = \frac{1}{3}\left(\frac{2}{C_1} + \frac{1}{C_2}\right)$, and $q = \text{argmax}_{i\in[3]} \left\{ \sum_{j=1}^{i} \frac{j}{i\, C_j} \right\}$. In particular, we consider the following cases for the download rates:

1) *For $\boldsymbol{C} = [0.2,\ 0.4,\ 0.5]$, we get $q = 3$, hence the optimal solution is the Maddah-Ali–Niesen caching scheme with $\boldsymbol{m}^* = [1/3,\ 1/3,\ 1/3]$, and we have $\Theta^*_{\mathfrak{A},\mathfrak{D}} = 4.1667$.*
2) *For $\boldsymbol{C} = [0.3,\ 0.3,\ 0.6]$, we get $q \in \{2, 3\}$, i.e., the optimal solution is not unique. $\boldsymbol{m}^* = [\frac{\alpha}{2} + \frac{1-\alpha}{3},\ \frac{\alpha}{2} + \frac{1-\alpha}{3},\ \frac{1-\alpha}{3}]$, where $\alpha \in [0, 1]$, and $\Theta^*_{\mathfrak{A},\mathfrak{D}} = \Theta_{unif} = 3.3333$.*
3) *For $\boldsymbol{C} = [0.2,\ 0.3,\ 0.6]$, we get $q = 2$. $\boldsymbol{m}^* = [0.5,\ 0.5,\ 0]$ and the optimal caching scheme is $a^*_{\{1\}} = a^*_{\{2\}} = 0.5$, $v^*_{\{1,2\}} = u^{*\{1,2\}}_{\{1\}} = u^{*\{1,2\}}_{\{2\}} = 0.5$, $v^*_{\{3\}} = 1$, which results in $\Theta^*_{\mathfrak{A},\mathfrak{D}} = 4.1667 < \Theta_{unif} = 4.4444$.*

In Fig. 6, we compare $\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}}, \boldsymbol{C})$ with $\Theta_{\text{unif}}(m_{\text{tot}}, \boldsymbol{C})$ for $K = 7$, and $\boldsymbol{C} = [0.2,\ 0.4,\ 0.6, 0.6,\ 0.8,\ 0.8,\ 1]$. We observe that $\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}}, \boldsymbol{C}) \leq \Theta_{\text{unif}}(m_{\text{tot}}, \boldsymbol{C})$. For $m_{\text{tot}} \leq 1$, we have $\text{argmax}_{i\in[K]} \sum_{j=1}^{i} (j\ m_{\text{tot}})/(i\ C_j) = K$, which implies $\Theta^*_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}}, \boldsymbol{C}) = \Theta_{\text{unif}}(m_{\text{tot}}, \boldsymbol{C})$.

In Fig. 7(a) and 7(b), we show the optimal memory allocation for $\boldsymbol{C} = [0.2,\ 0.2,\ 0.2,\ 0.5,\ 0.6,\ 0.7,\ 0.7]$ and $\boldsymbol{C} = [0.2,\ 0.2,\ 0.4,\ 0.4,\ 0.6,\ 0.7,\ 0.8]$, respectively. A general observation is that the optimal memory allocation balances the gain attained from assigning larger memories for users with weak links and the multicast gain achieved by equating the



(a) $C_1 = C_2 = C_3 = 0.2, C_4 = 0.5, C_5 = 0.6$, and $C_6 = C_7 = 0.7$.

(b) $C_1 = C_2 = 0.2, C_3 = C_4 = 0.4, C_5 = 0.6, C_6 = 0.7$, and $C_7 = 0.8$.

Fig. 7. The optimal memory allocations with different link capacities.

cache sizes. Consequently, in the optimal memory allocation, the users are divided into groups according to their rates, the groups that include users with low rates are assigned larger fractions of the cache memory budget, and users within each group are given equal cache sizes. These characteristics are illustrated in Fig. 7(a), which shows that the users are grouped into $\mathcal{G}_1 = \{1, 2, 3\}$ and $\mathcal{G}_2 = \{4, 5, 6, 7\}$ for all $m_{\text{tot}} \in [0, 7)$. Fig. 7(b) shows that the grouping not only depends on the rates $\boldsymbol{C}$, but also on the cache memory budget m_{tot}. For instance, for $m_{\text{tot}} = 2$, we have two groups $\mathcal{G}_1 = \{1, 2\}$ and $\mathcal{G}_2 = \{3, 4, 5, 6, 7\}$, however, for $m_{\text{tot}} = 3$, we have $\mathcal{G}_1 = \{1, 2\}$, $\mathcal{G}_2 = \{3, 4\}$, and $\mathcal{G}_2 = \{5, 6, 7\}$.

IX. Conclusion

In this paper, we have considered a downlink where the end-users are equipped with cache memories of different sizes. We have shown that the problem of minimizing the worst-case delivery load with uncoded placement and linear delivery can be modeled as a linear program. We have derived a lower bound on the worst-case delivery load with uncoded placement. We have characterized the exact delivery load memory trade-off with uncoded placement for the case where the aggregate cache size is less than or equal to the library size (small memory), the case where the aggregate cache size is greater than or equal to $K - 1$ times the library size (large memory), and the three-user case for arbitrary memory size. The proposed scheme outperforms other works in the same setup [28], [42], [44], and is numerically observed to verify the excellent performance of uncoded placement for parameters of interest.

We have also considered a system where the links between the server and the users have unequal capacities. In this scenario, the server suggests the memory sizes for cached contents along with contents to the users subject to a total memory budget, in order to minimize the delivery completion time. We have observed that the optimal solution balances between allocating larger cache sizes to the users with low link capacities and uniform memory allocation which maximizes the multicast gain. For when the total cache budget is less than the library size, we have shown that the optimal memory allocation distributes the cache memory budget uniformly over some number of users with the lowest link capacities. This number is a function of the users' link capacities.

The optimization perspective in this work provides a principled analysis of optimal caching and delivery schemes for cache-enabled networks, by translating the design elements of cache placement and delivery into structural optimization constraints. Future directions include different network topologies and systems with multiple servers and multiple libraries.

Appendix A
Proof of Proposition 1

$$\sum_{\mathcal{S}\subset[K]:\ \mathcal{S}'\subset\mathcal{S}, j\notin\mathcal{S}} a_{\mathcal{S}}$$

$$= a_{\mathcal{S}'}\mathbb{1}_{(|\mathcal{S}'|=1)} + \sum_{\mathcal{S}\subset[K]:\ \mathcal{S}'\subset\mathcal{S}, j\notin\mathcal{S}, |\mathcal{S}|\geq 2} a_{\mathcal{S}} \tag{43}$$

$$\geq u_{\mathcal{S}'}^{\{j\}\cup\mathcal{S}'}\mathbb{1}_{(|\mathcal{S}'|=1)} + \sum_{\mathcal{S}\subset[K]:\ \mathcal{S}'\subset\mathcal{S}, j\notin\mathcal{S}, |S|\geq 2} \sum_{\mathcal{T}\subsetneq\phi[K]:\ j\in\mathcal{T}, \mathcal{T}\cap\mathcal{S}\neq\phi, \mathcal{T}\setminus\{j\}\subset\mathcal{S}} u_{\mathcal{S}}^{\mathcal{T}}, \tag{44}$$

$$\geq u_{\mathcal{S}'}^{\{j\}\cup\mathcal{S}'}\mathbb{1}_{(|\mathcal{S}'|=1)} + \sum_{\mathcal{S}\subset[K]:\ \mathcal{S}'\subset\mathcal{S}, j\notin\mathcal{S}, |\mathcal{S}|\geq 2} \sum_{\mathcal{T}\subsetneq\phi[K]:\ \{j\}\cup\mathcal{S}'\subset\mathcal{T}, \mathcal{T}\setminus\{j\}\subset\mathcal{S}} u_{\mathcal{S}}^{\mathcal{T}}, \tag{45}$$

$$= u_{\mathcal{S}'}^{\{j\}\cup\mathcal{S}'}\mathbb{1}_{(|\mathcal{S}'|=1)} + \sum_{\mathcal{T}\subsetneq\phi[K]:\ \{j\}\cup\mathcal{S}'\subset\mathcal{T}} \sum_{\mathcal{S}\subset[K]:\ \mathcal{S}'\subset\mathcal{S}, \mathcal{T}\setminus\{j\}\subset\mathcal{S}, j\notin\mathcal{S}, |\mathcal{S}|\geq 2} u_{\mathcal{S}}^{\mathcal{T}}, \tag{46}$$

$$= \sum_{\mathcal{T}\subsetneq\phi[K]:\ \{j\}\cup\mathcal{S}'\subset\mathcal{T}} \sum_{\mathcal{S}\subset[K]:\ \mathcal{T}\setminus\{j\}\subset\mathcal{S}, j\notin\mathcal{S}} u_{\mathcal{S}}^{\mathcal{T}}, \tag{47}$$

$$= \sum_{\mathcal{T}\subsetneq\phi[K]:\ \{j\}\cup\mathcal{S}'\subset\mathcal{T}} v_{\mathcal{T}}, \tag{48}$$

where the indicator function $\mathbb{1}_{(|\mathcal{S}'|=1)} = 1$, if $|\mathcal{S}'| = 1$ and zero otherwise, (44) follows from the redundancy constraints in (13) and $u_{\mathcal{S}'}^{\{j\}\cup\mathcal{S}'} \leq a_{\mathcal{S}'}$, (45) follows from the fact that

TABLE I
OPTIMAL CACHING SCHEME FOR REGION II

Placement scheme	Delivery scheme
$a_{\{1\}} = (2 + m_2 - m_3)/3 - m_1$,	$v_{\{1,2\}} = (2 + 2m_3 - 2m_2)/3 - m_1$, $u_{\{1,3\}}^{\{1,2\}} = m_3 - m_2$,
$a_{\{2\}} = a_{\{3\}} = (2 - 2\,m_2 - m_3)/3$,	$v_{\{1,3\}} = (2 + m_2 - m_3)/3 - m_1$, $u_{\{2,3\}}^{\{1,3\}} = m_2 - m_1$,
$a_{\{1,2\}} = m_1 - (m_3 + 1 - m_2)/3$,	$v_{\{2,3\}} = u_{\{2\}}^{\{2,3\}} = u_{\{3\}}^{\{2,3\}} = (2 - 2m_2 - m_3)/3$,
$a_{\{1,3\}} = m_1 - (2m_2 + 1 - 2m_3)/3$,	$v_{\{1,2,3\}} = m_1 + (m_2 - m_3 - 1)/3$, $u_{\{2,3\}}^{\{1,2\}} = m_3 - m_1$,
$a_{\{2,3\}} = (4m_2 + 2m_3 - 1)/3 - m_1$.	$u_{\{1\}}^{\{1,2\}} = u_{\{1\}}^{\{1,3\}} = a_{\{1\}}$, $u_{\{3\}}^{\{1,3\}} = a_{\{3\}}$, $u_{\{2\}}^{\{1,2\}} = a_{\{2\}}$.

$\mathcal{S}' \subset \mathcal{S}$, and $\mathcal{S}' \subset \mathcal{T}$ implies $\mathcal{T} \cap \mathcal{S} \neq \phi$. By interchanging the order of summations over $\mathcal{S}$ and $\mathcal{T}$ in (45), we get (46), since both represent the set defined by

$$\Big\{(\mathcal{T},\mathcal{S}) \big|\, \mathcal{S}' \subset \mathcal{S}, j \notin \mathcal{S}, |\mathcal{S}| \geq 2, \{j\} \cup \mathcal{S}' \subset \mathcal{T}, \mathcal{T} \setminus \{j\} \subset \mathcal{S}\Big\}. \tag{49}$$

The equality in (47) follows from the fact that $\{j\} \cup \mathcal{S}' \subset \mathcal{T}$ and $\mathcal{T} \setminus \{j\} \subset \mathcal{S}$ implies $\mathcal{S}' \subset \mathcal{S}$, which can be proved by contradiction. More specifically, if $\mathcal{S}' \not\subset \mathcal{S}$, i.e., $\exists\, l \in \mathcal{S}'$ and $l \notin \mathcal{S}$, then $\{j\} \cup \mathcal{S}' \subset \mathcal{T}$ implies $l \in \mathcal{T} \setminus \{j\}$. This contradicts $\mathcal{T} \setminus \{j\} \subset \mathcal{S}$, since $l \in \mathcal{T} \setminus \{j\}$ and $l \notin \mathcal{S}$. The last equality follows from the structural constraints in (10).

APPENDIX B
PROOF OF THEOREM 1: LOWER BOUND WITH UNCODED PLACEMENT

References [4], [5] have shown that the delivery phase is equivalent to an index-coding problem and the delivery load is lower bounded by the acyclic index-coding bound [47, Corollary 1]. Reference [6] has proposed an alternative proof for the uncoded placement bound [4], [5] using a genie-aided approach. For ease of exposition, we will follow the genie-aided approach [6]. We consider a virtual user whose cache memory is populated by a genie. For any permutation of the users $[q_1, \ldots, q_K]$, the virtual user caches the file pieces stored at user q_j excluding files requested by $\{q_1, \ldots, q_{j-1}\}$ for $j \in [K]$, i.e., the virtual users cache content is given by

$$Z_{vir} = \bigcup_{j=1}^{K} \bigcup_{l \in [N] \setminus \{d_{q_1}, \ldots, d_{q_{j-1}}\}} \; \bigcup_{\mathcal{S} \subset [K]:\, \{q_j\} \in \mathcal{S}, \{q_1, \ldots, q_{j-1}\} \cap \mathcal{S} = \phi} \tilde{W}_{l,\mathcal{S}}. \tag{50}$$

Using the virtual user cache content and the transmitted signals, we can decode all the requested files. Additionally, for any uncoded placement $\boldsymbol{a} \in \mathfrak{A}(\boldsymbol{m})$, the worst-case delivery load $R^*_{\mathfrak{A}}(\boldsymbol{m}, \boldsymbol{a})$ satisfies [6]

$$R^*_{\mathfrak{A}}(\boldsymbol{m}, \boldsymbol{a}) \geq \sum_{j=1}^{K} \sum_{\mathcal{S} \subset [K]: \{q_1, \ldots, q_j\} \cap \mathcal{S} = \phi} a_{\mathcal{S}}, \;\; \forall \boldsymbol{q} \in \mathcal{P}_{[K]}, \tag{51}$$

where $\mathcal{P}_{[K]}$ is the set of all permutations of $[K]$. Hence, by taking the convex combination over all possible permutations of the users, we get

$$R^*_{\mathfrak{A}}(\boldsymbol{m}, \boldsymbol{a}) \geq \sum_{\boldsymbol{q} \in \mathcal{P}_{[K]}} \alpha_{\boldsymbol{q}} \left(\sum_{j=1}^{K} \sum_{\mathcal{S} \subset [K]: \{q_1, \ldots, q_j\} \cap \mathcal{S} = \phi} a_{\mathcal{S}} \right), \tag{52}$$

$$= \sum_{\boldsymbol{q} \in \mathcal{P}_{[K]}} \alpha_{\boldsymbol{q}} \left(K a_{\phi} + \sum_{j=1}^{K-1} j \sum_{\mathcal{S} \subset [K]: \{q_1, \ldots, q_j\} \cap \mathcal{S} = \phi,\; q_{j+1} \in \mathcal{S}} a_{\mathcal{S}} \right), \tag{53}$$

where $\sum_{\boldsymbol{q} \in \mathcal{P}_{[K]}} \alpha_{\boldsymbol{q}} = 1$, and $\alpha_{\boldsymbol{q}} \geq 0, \forall \boldsymbol{q} \in \mathcal{P}_{[K]}$. By rearranging the summations, we get

$$R^*_{\mathfrak{A}}(\boldsymbol{m}, \boldsymbol{a}) \geq \sum_{\mathcal{S} \subset [K]} \gamma_{\mathcal{S}}\, a_{\mathcal{S}}, \tag{54}$$

where $\gamma_{\mathcal{S}}$ is given by (21).

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) \geq \min_{\boldsymbol{a} \in \mathfrak{A}(\boldsymbol{m})} \sum_{\mathcal{S} \subset [K]} \gamma_{\mathcal{S}}\, a_{\mathcal{S}}, \tag{55}$$

Furthermore, the dual of the linear program in (55) is given by

$$\max_{\lambda_k \geq 0, \lambda_0} \; -\lambda_0 - \sum_{k=1}^{K} m_k\, \lambda_k \tag{56a}$$

$$\text{subject to } \lambda_0 + \sum_{k \in \mathcal{S}} \lambda_k + \gamma_{\mathcal{S}} \geq 0, \; \forall \mathcal{S} \subset [K], \tag{56b}$$

where λ_0 and λ_k are the dual variables associated with $\sum_{\mathcal{S} \subset [K]} a_{\mathcal{S}} = 1$, and $\sum_{\mathcal{S} \subset [K]: k \in \mathcal{S}} a_{\mathcal{S}} \leq m_k$, respectively. By taking the maximum over all convex combination, we obtain (20).

APPENDIX C
PROOF OF THEOREM 4 : $R^*_{\mathfrak{A}}(\boldsymbol{m})$ FOR $K = 3$

A. Region I: $\sum_{j=1}^{3} m_j \leq 1$ *(Follows from Theorem 2)*

B. Region II: $1 < \sum_{j=1}^{3} m_j \leq 2$, $m_3 < m_2 + 3\; m_1 - 1$, *and* $m_3 < 2(1 - m_2)$

Achievability: The caching scheme defined in Table I achieves $R^*_{\mathfrak{A}}(\boldsymbol{m}) = 5/3 - m_1 - 2\, m_2/3 - m_3/3$.

Converse: For any $\boldsymbol{a} \in \mathfrak{A}(\boldsymbol{m})$ and a permutation $\boldsymbol{q}$, we have

$$R^*_{\mathfrak{A}}(\boldsymbol{m}, \boldsymbol{a}) \geq 3a_{\phi} + 2a_{\{3\}} + a_{\{2\}} + a_{\{2,3\}}, \text{ for } \boldsymbol{q} = [1,\, 2,\, 3], \tag{57}$$

$$R^*_{\mathfrak{A}}(\boldsymbol{m}, \boldsymbol{a}) \geq 3a_{\phi} + 2a_{\{1\}} + a_{\{3\}} + a_{\{1,3\}}, \text{ for } \boldsymbol{q} = [2,\, 3,\, 1], \tag{58}$$

$$R^*_{\mathfrak{A}}(\boldsymbol{m}, \boldsymbol{a}) \geq 3a_{\phi} + 2a_{\{2\}} + a_{\{3\}} + a_{\{2,3\}}, \text{ for } \boldsymbol{q} = [1,\, 3,\, 2]. \tag{59}$$

TABLE II
OPTIMAL CACHING SCHEME FOR REGION III

Conditions	Placement scheme	Delivery scheme
$m_1 \le 1/3$, $m_1+m_3<1$	$a_{\{1\}}=m_1$, $a_{\{2\}}=1-(m_1+m_3)$, $a_{\{3\}}=1-(m_1+m_2)$, $a_{\{2,3\}}=\sum_{j=1}^{3} m_j-1$.	$v_{\{1\}}=1-3m_1$, $v_{\{2\}}=m_3-m_2$, $v_{\{1,2\}}=u_{\{1\}}^{\{1,2\}}=u_{\{2\}}^{\{1,2\}}=m_1$, $v_{\{1,3\}}=u_{\{1\}}^{\{1,3\}}=u_{\{3\}}^{\{1,3\}}=m_1$, $v_{\{2,3\}}=u_{\{2\}}^{\{2,3\}}=u_{\{3\}}^{\{2,3\}}=1-(m_1+m_3)$.
$m_1 > 1/3$, $m_3<2m_1$	$a_{\{1\}}=1-2m_1$, $a_{\{2\}}=2m_1-m_3$, $a_{\{3\}}=1-(m_1+m_2)$, $a_{\{1,3\}}=3m_1-1$, $a_{\{2,3\}}=m_2+m_3-2m_1$.	$v_{\{2\}}=1+m_3-3m_1-m_2$, $v_{\{1,2\}}=u_{\{1\}}^{\{1,2\}}+u_{\{1,3\}}^{\{1,2\}}=m_1$, $u_{\{2\}}^{\{1,2\}}=a_{\{2\}}$, $v_{\{1,3\}}=u_{\{1\}}^{\{1,3\}}=1-2m_1$, $u_{\{3\}}^{\{1,3\}}=a_{\{3\}}$ $v_{\{2,3\}}=u_{\{2\}}^{\{2,3\}}=u_{\{3\}}^{\{2,3\}}=2m_1-m_3$, $u_{\{2,3\}}^{\{1,2\}}=m_3-m_1$, $u_{\{2,3\}}^{\{1,3\}}=m_2-m_1$.
$m_1+m_3\ge 1$, $m_3\ge 2m_1$	$a_{\{1\}}=1-m_3$, $a_{\{3\}}=1-(m_1+m_2)$, $a_{\{1,3\}}=m_1+m_3-1$, $a_{\{2,3\}}=m_2$.	$v_{\{1\}}=m_3-2m_1$, $v_{\{2\}}=1-(m_1+m_2)$, $v_{\{1,2\}}=u_{\{1\}}^{\{1,2\}}+u_{\{1,3\}}^{\{1,2\}}=u_{\{2,3\}}^{\{1,2\}}=m_1$, $v_{\{1,3\}}=u_{\{1\}}^{\{1,3\}}=1-m_3$, $u_{\{3\}}^{\{1,3\}}+u_{\{2,3\}}^{\{1,3\}}=1-m_3$.

Hence, by taking the average of (57)-(59), we get

$$R^*_{\mathfrak{A}}(\boldsymbol{m},\boldsymbol{a}) \ge 3\ a_\phi+\frac{2\ a_{\{1\}}}{3}+a_{\{2\}}+\frac{4\ a_{\{3\}}}{3} +\frac{2\ a_{\{2,3\}}}{3}+\frac{a_{\{1,3\}}}{3}, \tag{60}$$

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) \ge \min_{\boldsymbol{a}\in\mathfrak{A}(\boldsymbol{m})}\left\{\frac{5a_\phi}{3}+\frac{2a_{\{1\}}}{3}+a_{\{2\}}+\frac{4a_{\{3\}}}{3} +\frac{2a_{\{2,3\}}}{3}+\frac{a_{\{1,3\}}}{3}\right\}, \tag{61}$$

$$=\frac{5}{3}-m_1-\frac{2\,m_2}{3}-\frac{m_3}{3}, \tag{62}$$

which is obtained by solving the dual linear program, as in Appendix B.

C. Region III: $1<\sum_{j=1}^{3} m_j\le 2$, $m_1+m_2<1$, and $m_3\ge m_2+3\ m_1-1$

Achievability: There are multiple caching schemes that achieve $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})=2-2\ m_1-m_2$. In particular, we consider caching schemes that satisfy

$$a_{\{1\}}+a_{\{1,3\}}=m_1, \tag{63}$$
$$a_{\{2\}}+a_{\{2,3\}}=m_2, \tag{64}$$
$$a_{\{3\}}=1-m_1-m_2, \tag{65}$$
$$a_{\{1,3\}}+a_{\{2,3\}}=m_1+m_2+m_3-1, \tag{66}$$
$$v_{\{1,2\}}=m_1,\ v_{\{1,3\}}=a_{\{1\}},\ v_{\{2,3\}}=a_{\{2\}}, \tag{67}$$
$$v_{\{1,3\}}+v_{\{2,3\}}=1-m_3, \tag{68}$$
$$v_{\{1\}}+v_{\{1,3\}}=1-2m_1, \tag{69}$$
$$v_{\{2\}}+v_{\{2,3\}}=1-m_1-m_2. \tag{70}$$

In Table II, we provide one feasible solution to (63)-(70).

Converse: For any $\boldsymbol{a}\in\mathfrak{A}(\boldsymbol{m})$ and $\boldsymbol{q}=[1,\ 2,\ 3]$, we have

$$R^*_{\mathfrak{A}}(\boldsymbol{m},\boldsymbol{a}) \ge 3\ a_\phi+2\ a_{\{3\}}+a_{\{2\}}+a_{\{2,3\}}, \tag{71}$$
$$\ge 2\ a_\phi+2\ a_{\{3\}}+a_{\{2\}}+a_{\{2,3\}}, \tag{72}$$
$$R^*_{\mathfrak{A}}(\boldsymbol{m}) \ge \min_{\boldsymbol{a}\in\mathfrak{A}(\boldsymbol{m})}\left\{2a_\phi+2a_{\{3\}}+a_{\{2\}}+a_{\{2,3\}}\right\}, \tag{73}$$
$$=2-2m_1-m_2. \tag{74}$$

D. Region IV: $m_1+m_2>1$, and $m_3\ge 2(1-m_2)$

Achievability: There are multiple caching schemes that achieve $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})=1-m_1$. In particular, we consider caching schemes that satisfy

$$a_{\{2,3\}}=1-m_1, \tag{75}$$
$$a_{\{1\}}-a_{\{1,2,3\}}=2-(m_1+m_2+m_3), \tag{76}$$
$$a_{\{1,2\}}+a_{\{1,2,3\}}=m_1+m_2-1, \tag{77}$$
$$a_{\{1,3\}}+a_{\{1,2,3\}}=m_1+m_3-1, \tag{78}$$
$$v_{\{1\}}+v_{\{1,2\}}+v_{\{1,3\}}+v_{\{1,2,3\}}=1-m_1, \tag{79}$$
$$v_{\{1,2\}}+v_{\{1,2,3\}}=1-m_2, \tag{80}$$
$$v_{\{1,3\}}+v_{\{1,2,3\}}=1-m_3. \tag{81}$$

In Table III, we provide one feasible solution to (75)-(81).

Converse: From the cut-set bound in (23), we have $R^*(\boldsymbol{m})\ge 1-m_1$.

APENDIX D
PROOF OF THEOREM 5: OPTIMAL CACHE SIZES

First, for a given memory allocation with $\sum_{k=1}^{K} m_k\le 1$, we have the following Lemma.

Lemma 1: For $C_1\le\cdots\le C_K$ and memory allocation $\boldsymbol{m}$ satisfying $\sum_{k=1}^{K} m_k\le 1$, the optimal caching scheme for (39) is given by $a^*_{\{j\}}=m_j$, $v^*_{\{i,j\}}=u^{*\{i,j\}}_{\{i\}}=u^{*\{i,j\}}_{\{j\}}=\min\{a^*_{\{i\}},a^*_{\{j\}}\}$, and $v^*_{\{j\}}=1-m_j-\sum_{i=1,i\ne j}^{K}\min\{m_i,m_j\}$.

Proof: By combining (15) with (1), dividing it by C_k, and summing over k, we get

$$\sum_{k=1}^{K}\sum_{\mathcal{T}\subsetneq_\phi[K]:k\in\mathcal{T}}\frac{v_{\mathcal{T}}}{C_k}\ge\sum_{k=1}^{K}\frac{1-m_k}{C_k}, \tag{82}$$

$$\sum_{k=1}^{K}\frac{v_{\{k\}}}{C_k}\ge\sum_{k=1}^{K}\frac{1-m_k}{C_k}-\sum_{\mathcal{T}\subsetneq_\phi[K]:|\mathcal{T}|\ge 2}\sum_{j\in\mathcal{T}}\frac{v_{\mathcal{T}}}{C_j}. \tag{83}$$

TABLE III
OPTIMAL CACHING SCHEME FOR REGION IV

Conditions	Placement scheme	Delivery scheme
$m_1+m_2+m_3\geq 2$, $1+m_1\geq m_2+m_3$	$a_{\{1,2\}}=1-m_3$, $a_{\{1,3\}}=1-m_2$, $a_{\{2,3\}}=1-m_1$, $a_{\{1,2,3\}}=\sum_{j=1}^{3}m_j-2$.	$v_{\{1,2\}}=u_{\{1,3\}}^{\{1,2\}}=u_{\{2,3\}}^{\{1,2\}}=m_3-m_1$, $v_{\{1,3\}}=u_{\{1,2\}}^{\{1,3\}}=u_{\{2,3\}}^{\{1,3\}}=m_2-m_1$, $v_{\{1,2,3\}}=1+m_1-(m_2+m_3)$,
$m_1+m_2+m_3\geq 2$, $1+m_1< m_2+m_3$	$a_{\{1,2\}}=1-m_3$, $a_{\{1,3\}}=1-m_2$, $a_{\{2,3\}}=1-m_1$, $a_{\{1,2,3\}}=\sum_{j=1}^{3}m_j-2$.	$v_{\{1\}}=m_2+m_3-(1+m_1)$, $v_{\{1,2\}}=u_{\{1,3\}}^{\{1,2\}}=u_{\{2,3\}}^{\{1,2\}}=1-m_2$, $v_{\{1,3\}}=u_{\{1,2\}}^{\{1,3\}}=u_{\{2,3\}}^{\{1,3\}}=1-m_3$.
$m_1+m_2+m_3< 2$, $1+m_1\geq m_2+m_3$	$a_{\{1\}}=2-\sum_{j=1}^{3}m_j$, $a_{\{1,2\}}=m_1+m_2-1$, $a_{\{1,3\}}=m_1+m_3-1$, $a_{\{2,3\}}=1-m_1$.	$v_{\{1,2\}}=u_{\{2,3\}}^{\{1,2\}}=m_3-m_1$, $u_{\{1\}}^{\{1,2\}}=a_{\{1\}}$, $v_{\{1,3\}}=u_{\{2,3\}}^{\{1,3\}}=m_2-m_1$, $u_{\{1\}}^{\{1,3\}}=a_{\{1\}}$, $v_{\{1,2,3\}}=1+m_1-(m_2+m_3)$, $u_{\{1,3\}}^{\{1,2\}}=m_2+2m_3-2$, $u_{\{1,2\}}^{\{1,3\}}=2m_2+m_3-2$.
$m_1+m_2+m_3< 2$, $1+m_1< m_2+m_3$	$a_{\{1\}}=2-\sum_{j=1}^{3}m_j$, $a_{\{1,2\}}=m_1+m_2-1$, $a_{\{1,3\}}=m_1+m_3-1$, $a_{\{2,3\}}=1-m_1$.	$v_{\{1\}}=m_2+m_3-(1+m_1)$, $v_{\{1,2\}}=u_{\{2,3\}}^{\{1,2\}}=1-m_2$, $u_{\{1\}}^{\{1,2\}}=a_{\{1\}}$, $v_{\{1,3\}}=u_{\{2,3\}}^{\{1,3\}}=1-m_3$, $u_{\{1\}}^{\{1,3\}}=a_{\{1\}}$, $u_{\{1,3\}}^{\{1,2\}}=m_1+m_3-1$, $u_{\{1,2\}}^{\{1,3\}}=m_1+m_2-1$.

Therefore, we get the lower bound

$$\Theta_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}},\boldsymbol{C})\geq\sum_{k=1}^{K}\frac{1-m_k}{C_k}-\sum_{\mathcal{T}\subsetneq_\phi[K]:|\mathcal{T}|\geq 2}v_{\mathcal{T}}\times\left(\sum_{j\in\mathcal{T}}\frac{1}{C_j}+\frac{1}{\min_{i\in\mathcal{T}}C_i}\right). \quad (84)$$

Additionally, for $C_1\leq\cdots\leq C_K$, we have

$$\Theta_{\mathfrak{A},\mathfrak{D}}(m_{\text{tot}},\boldsymbol{C})\geq\sum_{k=1}^{K}\frac{1-m_k}{C_k}-\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{1}{C_j}\sum_{\mathcal{T}\subsetneq_\phi[K]:\{i,j\}\subset\mathcal{T}}v_{\mathcal{T}}, \quad (85)$$

$$\geq\sum_{k=1}^{K}\frac{1-m_k}{C_k}-\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\min\{m_i,m_j\}}{C_j}, \quad (86)$$

where the last inequality follows from the fact that the multicast transmissions that include users $\{i,j\}$ are limited by the side-information stored at each of them, which is upper bounded by the cache memory size, i.e., $\sum_{\mathcal{T}\subsetneq_\phi[K]:\,\{i,j\}\subset\mathcal{T}}v_{\mathcal{T}}\leq\min\{m_i,m_j\}$.

Finally, for $\sum_{i=1}^{K}m_i\leq 1$, $a_{\{j\}}^*=m_j$, $v_{\{j\}}^*=1-m_j-\sum_{i=1,i\neq j}^{K}\min\{m_i,m_j\}$, and $v_{\{i,j\}}^*=u_{\{i\}}^{*\{i,j\}}=u_{\{j\}}^{*\{i,j\}}=\min\{a_{\{i\}}^*,a_{\{j\}}^*\}$, is a feasible solution to (39) that achieves the lower bound. ∎

Now, using Lemma 1, we can simplify (39) to

$$\min_{\boldsymbol{m}}\ \sum_{k=1}^{K}\frac{1-m_k}{C_k}-\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\min\{m_i,m_j\}}{C_j} \quad (87a)$$

$$\text{s.t.}\ \sum_{k=1}^{K}m_k\leq m_{\text{tot}},\ 0\leq m_k\leq 1,\ \forall\, k\in[K]. \quad (87b)$$

Next, we show that the optimal memory allocation from (87) satisfies $m_1^*\geq m_2^*\geq\cdots\geq m_K^*$.

Lemma 2: For $C_1\leq\cdots\leq C_K$ and $m_{tot}\leq 1$, the objective function of (87) satisfies $\Theta_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})\leq\Theta_{\mathfrak{A},\mathfrak{D}}(\tilde{\boldsymbol{m}})$, where $m_i=\tilde{m}_i$, for $i\in[K]\backslash\{r,r+1\}$, and some $r\in[K-1]$. Additionally, $m_r=\tilde{m}_{r+1}=\alpha+\delta$, $m_{r+1}=\tilde{m}_r=\alpha$, for $\delta,\alpha\geq 0$, and $m_1\geq m_2\geq\cdots\geq m_r$.

Proof: For $\boldsymbol{m}=[m_1,m_2,\ldots,m_{r-1},\alpha+\delta,\alpha,m_{r+2},\ldots,m_K]$ and $\tilde{\boldsymbol{m}}=[m_1,m_2,\ldots,m_{r-1},\alpha,\alpha+\delta,m_{r+2},\ldots,m_K]$, we have $\Theta_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})-\Theta_{\mathfrak{A},\mathfrak{D}}(\tilde{\boldsymbol{m}})=\chi_1+\chi_2$, where

$$\chi_1=\frac{1-m_r}{C_r}+\frac{1-m_{r+1}}{C_{r+1}}-\frac{1-\tilde{m}_r}{C_r}-\frac{1-\tilde{m}_{r+1}}{C_{r+1}}, \quad (88)$$

$$=\delta\left(\frac{1}{C_{r+1}}-\frac{1}{C_r}\right), \quad (89)$$

$$\chi_2=\sum_{i=1}^{r-1}\left(\frac{\min\{m_i,\tilde{m}_r\}}{C_r}+\frac{\min\{m_i,\tilde{m}_{r+1}\}}{C_{r+1}}\right)-\sum_{i=1}^{r-1}\left(\frac{\min\{m_i,m_r\}}{C_r}+\frac{\min\{m_i,m_{r+1}\}}{C_{r+1}}\right), \quad (90)$$

$$=\left(\frac{1}{C_{r+1}}-\frac{1}{C_r}\right)\sum_{i=1}^{r-1}(\min\{m_i,\alpha+\delta\}-\min\{m_i,\alpha\}), \quad (91)$$

$$=\delta(r-1)\left(\frac{1}{C_{r+1}}-\frac{1}{C_r}\right). \quad (92)$$

Thus, $\Theta_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})-\Theta_{\mathfrak{A},\mathfrak{D}}(\tilde{\boldsymbol{m}})=r\delta\left(\frac{1}{C_{r+1}}-\frac{1}{C_r}\right)\leq 0$, as $C_{r+1}\geq C_r$. ∎

Using Lemma 2, (87) can be simplified to (93).

Lemma 3: For $C_1 \leq \cdots \leq C_K$ and $m_{tot} \leq 1$, optimization problem (39) reduces to

$$\min_{\boldsymbol{m}} \sum_{k=1}^{K} \frac{1 - k\, m_k}{C_k} \tag{93a}$$

$$\textit{s.t.} \sum_{k=1}^{K} m_k \leq m_{tot}, \tag{93b}$$

$$0 \leq m_{k+1} \leq m_k, \ \forall\, k \in [K-1]. \tag{93c}$$

Equivalently, the optimal memory allocation for (93) is obtained by solving

$$\max_{\boldsymbol{m}} \sum_{k=1}^{K} \frac{k\, m_k}{C_k} \tag{94a}$$

$$\text{s.t.} \sum_{k=1}^{K} m_k \leq m_{\text{tot}}, \tag{94b}$$

$$0 \leq m_{k+1} \leq m_k, \ \forall\, k \in [K-1]. \tag{94c}$$

Finally, the optimal memory allocation in Theorem 5 is obtained by solving the dual of the linear program in (94).

REFERENCES

[1] *Cisco VNI Forecast and Methodology, 2015–2020*, Cisco, San Jose, CA, USA, Jun. 2016.

[2] K. C. Almeroth and M. H. Ammar, "The use of multicast delivery to provide a scalable and interactive video-on-demand service," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 6, pp. 1110–1122, Aug. 1996.

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

[4] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," in *Proc. IEEE ITW*, Sep. 2016, pp. 161–165.

[5] K. Wan, D. Tuninetti, and P. Piantanida. (2017). "A novel index coding scheme and its application to coded caching." [Online]. Available: https://arxiv.org/abs/1702.07265

[6] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr. (2016). "The exact rate-memory tradeoff for caching with uncoded prefetching." [Online]. Available: https://arxiv.org/abs/1609.07817

[7] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for small buffer users," *IET Commun.*, vol. 10, no. 17, pp. 2315–2318, Nov. 2016.

[8] M. M. Amiri and D. Gündüz, "Fundamental limits of coded caching: Improved delivery rate-cache capacity tradeoff," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 806–815, Feb. 2017.

[9] J. Gómez-Vilardebó. (2016). "Fundamental limits of caching: Improved bounds with coded prefetching." [Online]. Available: https://arxiv.org/abs/1612.09071

[10] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4388–4413, Jul. 2017.

[11] S. H. Lim, C.-Y. Wang, and M. Gastpar, "Information-theoretic caching: The multi-user case," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7018–7037, Nov. 2017.

[12] C. Tian and K. Zhang. (2017). "Fundamental limits of coded caching: From uncoded prefetching to coded prefetching." [Online]. Available: https://arxiv.org/abs/1704.07901

[13] C.-Y. Wang, S. S. Bidokhti, and M. Wigger, "Improved converses and gap-results for coded caching," in *Proc. IEEE ISIT*, Jun. 2017, pp. 2428–2432.

[14] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr. (2017). "Characterizing the rate-memory tradeoff in cache networks within a factor of 2." [Online]. Available: https://arxiv.org/abs/1702.04563

[15] M. Ji *et al.*, "On the fundamental limits of caching in combination networks," in *Proc. IEEE SPAWC*, Jun./Jul. 2015, pp. 695–699.

[16] A. A. Zewail and A. Yener, "Combination networks with or without secrecy constraints: The impact of caching relays," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1140–1152, Jun. 2018.

[17] K. Wan, M. Ji, P. Piantanida, and D. Tuninetti, "Caching in combination networks: Novel multicast message generation and delivery by leveraging the network topology," in *Proc. IEEE ICC*, May 2018, pp. 1–6.

[18] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[19] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Device-to-device coded caching with heterogeneous cache sizes," in *Proc. IEEE ICC*, May 2018, pp. 1–6.

[20] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[21] R. Timo and M. Wigger, "Joint cache-channel coding over erasure broadcast channels," in *Proc. IEEE ISWCS*, Aug. 2015, pp. 201–205.

[22] A. Ghorbel, M. Kobayashi, and S. Yang, "Content delivery in erasure broadcast channels with cache and feedback," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6407–6422, Nov. 2016.

[23] S. S. Bidokhti, M. Wigger, and A. Yener. (2017). "Benefits of cache assignment on degraded broadcast channels." [Online]. Available: https://arxiv.org/abs/1702.08044

[24] M. M. Amiri and D. Gündüz, "Cache-aided content delivery over erasure broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 370–381, Jan. 2018.

[25] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

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

[27] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.

[28] Q. Yang and D. Gündüz, "Coded caching and content delivery with heterogeneous distortion requirements," *IEEE Trans. Inf. Theory*, vol. 64, no. 6, pp. 4347–4364, Jun. 2018.

[29] P. Hassanzadeh, E. Erkip, J. Llorca, and A. Tulino, "Distortion-memory tradeoffs in cache-aided wireless video delivery," in *Proc. IEEE Allerton*, Sep. 2015, pp. 1150–1157.

[30] A. M. Ibrahim, A. A. Zewail, and A. Yener, "On coded caching with heterogeneous distortion requirements," in *Proc. IEEE ITA*, Feb. 2018, pp. 1–9.

[31] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1146–1158, Feb. 2017.

[32] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3923–3949, Jun. 2017.

[33] A. Ramakrishnan, C. Westphal, and A. Markopoulou, "An efficient delivery scheme for coded caching," in *Proc. IEEE ITC*, Sep. 2015, pp. 46–54.

[34] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," in *Proc. IEEE ITA*, Feb. 2015, pp. 98–107.

[35] S. Jin, Y. Cui, H. Liu, and G. Caire. (2017). "Structural properties of uncoded placement optimization for coded delivery." [Online]. Available: https://arxiv.org/abs/1707.07146

[36] U. Niesen and M. A. Maddah-Ali, "Coded caching for delay-sensitive content," in *Proc. IEEE ICC*, Jun. 2015, pp. 5559–5564.

[37] V. Ravindrakumar, P. Panda, N. Karamchandani, and V. M. Prabhakaran, "Private coded caching," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 685–694, Mar. 2018.

[38] A. Sengupta, R. Tandon, and T. C. Clancy, "Fundamental limits of caching with secure delivery," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 2, pp. 355–370, Feb. 2015.

[39] A. A. Zewail and A. Yener, "Fundamental limits of secure device-to-device coded caching," in *Proc. IEEE Asilomar*, Nov. 2016, pp. 1414–1418.

[40] S. Wang, W. Li, X. Tian, and H. Liu. (2015). "Coded caching with heterogenous cache sizes." [Online]. Available: https://arxiv.org/abs/1504.01123

[41] M. M. Amiri, Q. Yang, and D. Gündüz, "Decentralized caching and coded delivery with distinct cache capacities," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4657–4669, Nov. 2017.

[42] A. Sengupta, R. Tandon, and T. C. Clanc, "Layered caching for heterogeneous storage," in *Proc. IEEE Asilomar*, Nov. 2016, pp. 719–723.

[43] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Centralized coded caching with heterogeneous cache sizes," in *Proc. IEEE WCNC*, Mar. 2017, pp. 1–6.
[44] A. M. Daniel and W. Yu. (2017). "Optimization of heterogeneous coded caching." [Online]. Available: https://arxiv.org/abs/1708.04322
[45] A. Tang, S. Roy, and X. Wang, "Coded caching for wireless backhaul networks with unequal link rates," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 1–13, Jan. 2018.
[46] A. F. Dana, R. Gowaikar, R. Palanki, B. Hassibi, and M. Effros, "Capacity of wireless erasure networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 789–804, Mar. 2006.
[47] F. Arbabjolfaei, B. Bandemer, Y.-H. Kim, E. Şaşoğlu, and L. Wang, "On the capacity region for index coding," in *Proc. IEEE ISIT*, Jul. 2013, pp. 962–966.

Abdelrahman M. Ibrahim (S'07) received the B.Sc. degree in electrical engineering from Alexandria University, Alexandria, Egypt, in 2011, and the M.Sc. degree in wireless communications from Nile University, Giza, Egypt, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA. His Ph.D. dissertation is focused on data storage and energy management in emerging networks. He was an Exchange Research Assistant with Sabanci University, Istanbul, Turkey, from 2013 to 2014, funded by the Marie Curie International Research Exchange Scheme. He is currently a Graduate Research Assistant with the Department of Electrical Engineering, The Pennsylvania State University. His research interests include data storage systems, cache-aided networks, green communications, and resource allocation in wireless networks.

Ahmed A. Zewail (S'07–M'19) received the B.Sc. degree in electrical engineering from Alexandria University, Alexandria, Egypt, in 2011, and the M.Sc. degree in wireless communications from Nile University, Giza, Egypt, in 2013. He is currently pursuing the Ph.D. degree with the Wireless Communications and Networking (WCAN) Laboratory, The Pennsylvania State University, University Park, PA, USA. He was ranked the fourth of his class in his B.Sc. degree. His graduation project was about warehouse management systems using RFID, which was funded by Siemens, NTRA, Vodafone, BA, and Vision Solutions. His master's thesis focused on the capacity and degrees of freedom of relay networks. He is currently a Research Assistant with the Wireless Communications and Networking (WCAN) Laboratory, The Pennsylvania State University. His Ph.D. dissertation is focused on secrecy guarantees in emerging networks, e.g., untrusted relay networks and cache-aided networks. His current research interests include network information theory, cache-aided networks, wireless communications, and physical layer security. He received first place in the INDAC-Siemens 2011 Competition.

Aylin Yener (S'91–M'01–SM'14–F'15) received the B.Sc. degree in electrical and electronics engineering and the B.Sc. degree in physics from Bogazici University, Istanbul, Turkey, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Wireless Information Network Laboratory (WINLAB), Rutgers University, New Brunswick, NJ, USA. She is a Distinguished Professor of electrical engineering at The Pennsylvania State University, University Park, PA, USA, where she joined the faculty as an Assistant Professor in 2002. Since 2017, she has been a Dean's Fellow of the College of Engineering at The Pennsylvania State University. She was a Visiting Professor of electrical engineering at Stanford University (2016–2018) and a Visiting Associate Professor in the same department (2008–2009). Her current research interests are in information security, green communications, caching systems, and more generally in the fields of information theory, communication theory, and networked systems. She received the NSF CAREER Award in 2003, the Best Paper Award in communication theory from the IEEE International Conference on Communications in 2010, the Penn State Engineering Alumni Society (PSEAS) Outstanding Research Award in 2010, the IEEE Marconi Prize Paper Award in 2014, the PSEAS Premier Research Award in 2014, the Leonard A. Doggett Award for Outstanding Writing in Electrical Engineering at Penn State in 2014, and the IEEE Women in Communications Engineering Outstanding Achievement Award in 2018. She is a Distinguished Lecturer for the IEEE Information Theory Society (2019–2020), the IEEE Communications Society (2018–2020), and the IEEE Vehicular Technology Society (2017–2019).

Dr. Yener is serving as the Vice President of the IEEE Information Theory Society in 2019. Previously, she was the second Vice President (2018), member of the Board of Governors (2015–2018), and the Treasurer (2012–2014) of the IEEE Information Theory Society. She served as the Student Committee Chair for the IEEE Information Theory Society (2007–2011), and was the Co-Founder of the Annual School of Information Theory in North America in 2008. She was a Technical (Co)-Chair for various symposia/tracks at the IEEE ICC, PIMRC, VTC, WCNC, and Asilomar in 2005, from 2008 to 2014, and 2018. Previously, she served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS (2009–2012), an Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING (2017–2018), and an Editor and an Editorial Advisory Board Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2001–2012). She also served a Guest Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY in 2011, and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in 2015. Currently, she serves as a Senior Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.