Improved Approximation of Storage-Rate Tradeoff for Caching With Multiple Demands

Avik Sengupta, *Member, IEEE*, and Ravi Tandon, *Senior Member, IEEE*

***Abstract*— Caching at the network edge has emerged as a viable solution for alleviating the severe capacity crunch in modern content centric wireless networks by leveraging network load-balancing in the form of localized content storage and delivery. In this paper, we 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: 1) centralized content delivery, where user demands at each transmission interval are serviced by the central server via multicast transmissions; and 2) device-to-device assisted distributed delivery, where users multicast to each other in order to service file demands. For such cache-aided networks, we present new results on the fundamental cache storage versus transmission rate tradeoff. Specifically, we develop a new technique for characterizing information theoretic lower bounds on the storage-rate tradeoff and show that the new lower bounds are strictly tighter than cut-set bounds from literature. Furthermore, using the new lower bounds, we improve the constant factor approximation of the optimal storage-rate tradeoff for cache-aided systems under both delivery models.**

***Index Terms*— Caching, edge networks, centralized content delivery, device-to-device, Han's inequality.**

I. Introduction

THE dynamics of traffic over wireless networks has undergone a paradigm shift to become increasingly content centric with high volume multimedia content (e.g., video) distribution holding precedence. Therefore, efficient utilization of network resources is imperative in such networks for improving capacity. With the proliferation of cheap storage at the network edge (e.g. at user devices and small cell base stations), caching has emerged as an important tool for facilitating efficient load balancing and maximal resource utilization for future 5G wireless networks [1]. Parts of popular files are pre-stored at the edge caches such that at times of high network load [1]–[4], the local content can be leveraged to reduce the over-the-air transmission rates. Caching and complimentary file delivery in wireless networks has been the subject of a wealth of recent research as evidenced by the results in [2] and [5]–[18]. Caching generally works in two phases - (*a*) the *cache storage phase* where parts of popular content is placed in users' cache memories by a central server e.g., an LTE eNodeB in modern cellular networks and (*b*) the *file delivery phase*, where requested content is delivered by exploiting local cache storage. A caching and content delivery policy is said to be *feasible* if all users can *decode* their requested content from the received transmissions and their local cache storage with an arbitrarily low probability of error. Cache placement happens over a much larger time-scale than the file request and delivery phase or a *transmission interval*, and therefore needs to be agnostic to user demands. The fundamental tradeoff in such cache-aided systems is between the cache storage and the delivery rate.

Recently, Maddah-Ali and Niesen [2], [5]–[7] showed that by jointly designing the storage and delivery phases, order-wise improvement in the *worst-case* delivery rate can be achieved for any given size of cache storage for the case when users demand only one file at every transmission interval and assuming *uniform file popularity* i.e., every file is equally likely to be requested. The proposed schemes extract a *global caching gain*, in addition to the traditional *local caching gain*, by distributing common content across users' caches and subsequently designing *centralized coded multicast transmissions* which leverage this shared content to reduce delivery rates. The authors used cut-set based arguments to derive an information theoretic lower bound on the optimal storage-rate tradeoff and characterized it to within a constant multiplicative factor of 12 for worst-case user demands under uniform file popularity. An new lower bound as well as an improved characterization of the optimal storage-rate tradeoff to within a factor of 8 was presented in our previous work in [19]. The case when users demand *multiple* files at each transmission interval was initially studied in [20] for the case of worst-case user demands, where the authors proposed the first known cut-set lower based bound for this setting. Alternately, the study of cache-aided systems with centralized content delivery under non-uniform file popularity was undertaken in [3], [6], [21], and [22] where upper and lower bounds on the *expected* delivery rate was considered instead of the worst-case rate. Another line of recent work in [23]–[28] extends the notion of caching and centralized content delivery to case of wireless interference channels.

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

Manuscript received June 14, 2016; revised September 27, 2016 and December 12, 2016; accepted January 27, 2017. Date of publication February 6, 2017; date of current version May 13, 2017. This work was presented in part at the IEEE International Symposium of Information Theory (ISIT), Hong Kong, June 2015 and at the Information Theory and Applications Workshop (ITA), UCSD, Feb 2015. The work of A. Sengupta was done when the author was with Wireless@VT, Virginia Tech, Blacksburg, VA 24060 USA. The work of R. Tandon was supported by the US National Science Foundation Grant CAREER-1651492. The associate editor coordinating the review of this paper and approving it for publication was W. Chen.

A. Sengupta is with the Next Generation and Standards Group at Intel Corp., Santa Clara, CA 95054 USA (e-mail: avik.sengupta@intel.com).

R. Tandon is with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721 USA (e-mail: tandonr@email.arizona.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.2017.2664815

0090-6778 © 2017 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.

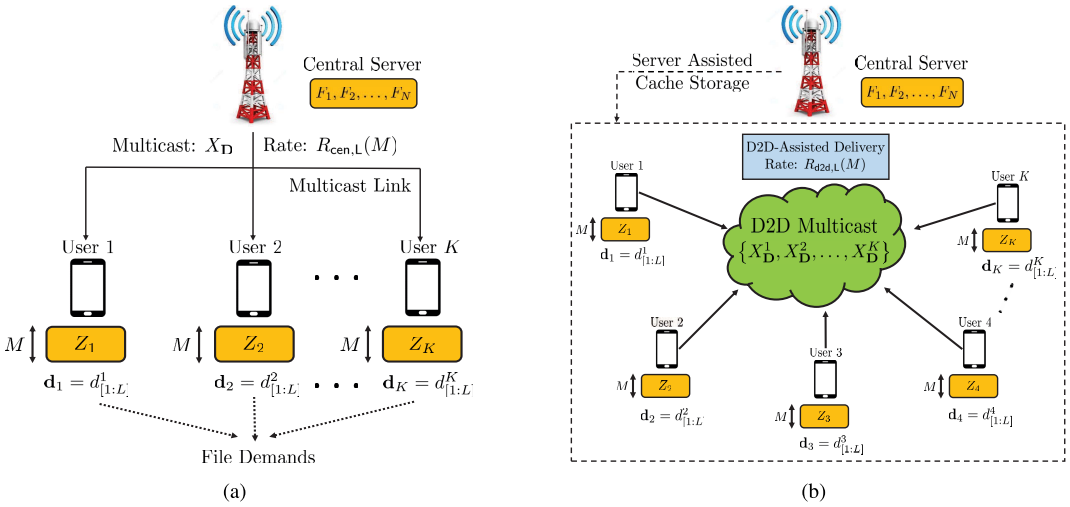


Fig. 1. System Model for cache-aided network with (a) centralized content delivery where the requested content is delivered via multicast transmission by the central server; and (b) D2D-assisted content delivery where each device multicasts to all the other devices using the contents placed in the device cache by the central server.

in [12] whereby the delivery phase is relegated to the users instead of a central server. While the cache placement phase is centralized for both delivery models, the main difference is in the *distributed nature of multicast transmissions* for the case of D2D-assisted content delivery. In the centralized delivery model of [2], 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. Ji *et al.* [12], 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 [12] 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 tradeoff. In our prior work in [29], we improved on the cut-set bound and showed that the achievable scheme in [12] is within a constant multiplicative factor of 8 from the optimal by leveraging the new bounds. 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 to the best of the authors' knowledge.

Main Contributions: In this work, we consider the worst-case delivery rate for cache-aided systems under uniform file popularity as in [2], [5], [12], and [20] and present fundamental results on the storage vs. rate trade-off for centralized as well as D2D-assisted file delivery. The main contributions of the paper 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 [2, Th. 2] and [20, Th. 5] for centralized content delivery for all values of problem parameters. The proposed technique also yields improved lower bounds for D2D-assisted content delivery when each user demands multiple files at each transmission interval[1]
- Using the new lower bounds, we characterize the optimal storage-rate tradeoff to within a constant multiplicative factor of 11 for centralized delivery and 10 for D2D-assisted delivery improving on previous results .

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.

II. System Model and Preliminary Results

In this section, we introduce the system model for file storage and delivery in cache-aided systems. We then present achievable schemes for the case of multiple file demands in each transmission interval which are based on schemes which treat each of the L sets of K user file demands independently as a single per-user file demand case for multicast delivery.

A. System Model

We consider a cache-aided network (see Fig. 1) with K users and a library of N files, $F_{[1:N]}$, where each file is of

[1]We acknowledge that the cut-set bound in [20, Th. 5] for centralized content delivery with multiple demands is also a valid lower bound for D2D-assisted content delivery. The proposed technique in this paper however strictly improves on this lower bound by explicitly accounting for the constraints of D2D-assisted delivery.

size B bits, for $B \in \mathbb{N}^+$. Formally, the files F_n are i.i.d. and distributed as:

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

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 is centralized and consists of K caching functions which map the files $F_{[1:N]}$ into the cache content

$$Z_k \triangleq \phi_k\Big(F_{[1:N]}\Big), \tag{2}$$

for each user $k \in [1:K]$. For cache-aided systems with centralized content delivery the cache storage constraint is such that $H(Z_k) \in [0, MB]$.[2] For the case of D2D-assisted delivery, an additional storage constraint is that all caches should be collectively capable of storing the entire library $F_{[1:N]}$ i.e., $KM \geq N$ and $H(Z_k) \in [NB/K, MB]$.[3] 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

$$X_{\mathbf{D}} \triangleq \psi_{\mathbf{D}}\big(F_1, \ldots, F_N\big), \tag{3}$$

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

$$X_{\mathbf{D}} = \Big\{ \Big(X^1_{\mathbf{D}}, \ldots, X^K_{\mathbf{D}}\Big) : X^k_{\mathbf{D}} = \psi^k_{\mathbf{D}}\left(Z_k\right), \forall k \in [1:K] \Big\}. \tag{4}$$

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.,

$$H(X_{\mathbf{D}}) \leq \sum_{k=1}^{K} H\left(X^k_{\mathbf{D}}\right) \leq \sum_{k=1}^{K} R_k B \leq RB. \tag{5}$$

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

$$\widehat{F}_{\mathbf{d}_k} \triangleq \mu_{\mathbf{D},k}\Big(X_{\mathbf{D}}, Z_k\Big). \tag{6}$$

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

$$P_e \triangleq \max_{\mathbf{D},\ k\in[1:K],\ d\in\mathbf{d}_k} \mathbb{P}\left(\hat{F}_d \neq F_d\right), \tag{7}$$

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 Tradeoff): 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 tradeoffs are defined as

$$R^*_{\mathsf{cen,L}}(M) \triangleq \inf\left\{R_{\mathsf{cen,L}} : (M, R_{\mathsf{cen,L}}) \text{ is achievable}\right\}; \tag{8}$$

$$R^*_{\mathsf{d2d,L}}(M) \triangleq \inf\left\{R_{\mathsf{d2d,L}} : (M, R_{\mathsf{d2d,L}}) \text{ is achievable}\right\}. \tag{9}$$

B. Preliminary Results

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

1) Centralized Delivery With Multiple Demands: An achievable scheme for caching with centralized delivery was first proposed in [2] for the case of single ($L = 1$) user requests. 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 $M \in \frac{N_t}{K}$ files for any $t \in [0:K]$, an achievable rate for centralized content delivery is given by

$$R_{\mathsf{cen,L}}(M) = K\left(1 - \frac{M}{N}\right)\min\left(\frac{L}{1 + KM/N}, \frac{N}{K}\right), \tag{10}$$

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

Proof: The delivery rate in (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 [2, Th. 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. □

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 [12] 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 $M \in \frac{N_t}{K}$ files for any $t \in [0:K]$ with $KM \geq N$,

[2]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 $X_{\mathbf{D}}$.

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

an achievable rate for D2D-assisted content delivery is given by

$$R_{\mathsf{d2d,L}}(M) = \min\left\{\frac{LN}{M}\left(1-\frac{M}{N}\right), N\right\}, \tag{11}$$

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

Proof: The delivery rate in (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 [12, Th. 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 [20], 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 III, 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 sets of multiple demands independently.

III. 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.

A. Centralized Content Delivery

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

Theorem 1: For any N files and K users, each having a cache size of $M \in [0, N]$, the optimal centralized content delivery rate $R^*_{\mathsf{cen,L}}(M)$ is lower bounded as

$$\max_{\substack{s\in[1:\min\{\lceil N/L\rceil, K\}],\\ \ell\in[1:\lceil N/(Ls)\rceil]}} \frac{N - sM - \frac{\mu(N-L\ell s)^+}{s+\mu} - (N-KL\ell)^+}{\ell}, \tag{12}$$

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 1 is given in Appendix A. The expression in Theorem 1 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 [20, Th. 5], the additional parameter ℓ adds further flexibility to the lower bound expression and accounts for file decoding through the interaction of caches and transmissions, 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 IV-A for more details). The cut-set based lower bound of [20, Th. 5] 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. 2(a) where, in addition to the achievable rate from Section II-B, 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 tradeoff can be obtained by use of the proposed lower bound.

Theorem 2: 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:

$$\frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq 11. \tag{13}$$

The proof of Theorem 2 is provided in Appendix B. This result improves on the gap of 18 between the achievable scheme and the cut-set bound in [20, Th. 5]. 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 [20] 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:

$$\max_{\substack{s\in[1:K],\\ \ell\in[1:\lceil N/s\rceil]}} \frac{N - sM - \frac{\mu(N-\ell s)^+}{s+\mu} - (N-K\ell)^+}{\ell}, \tag{14}$$

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

Corollary 1 follows by setting $L = 1$ in Theorem 1 and was originally presented in [19].

Remark 1: The new bounds strictly improve on the cut-set lower bounds presented in [2, Th. 2] as shown in Fig. 2(b). Using the achievable rate from [2, Th. 1] (also shown in Fig. 2(b)) and the lower bound in (14), the approximation of the optimal storage-rate tradeoff can be improved to within a factor of 8 as compared to 12 [2, Th. 3]. The proofs are omitted for brevity and are presented in detail in [30].

B. 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

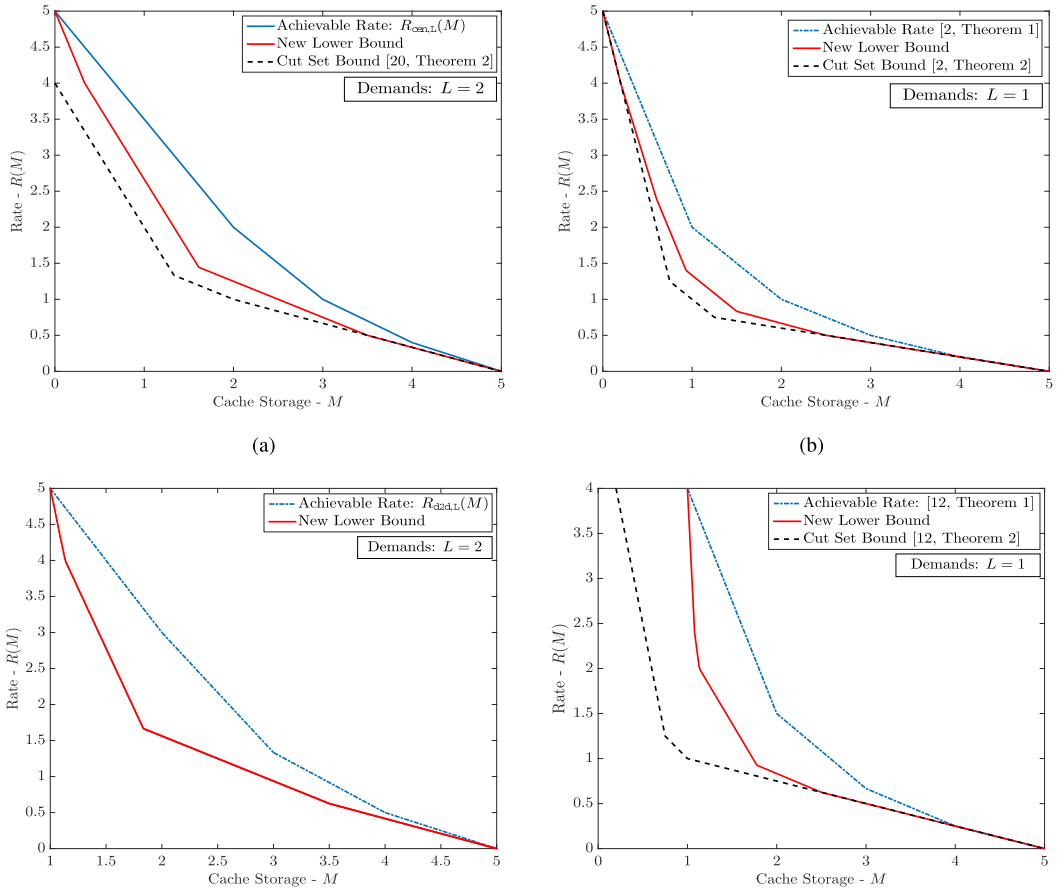


Fig. 2. Storage-rate trade-off for centralized content delivery with $N = K = 5$ and (a) $L = 2$ demands per user; and (b) $L = 1$ demand per user; Storage-rate tradeoff for D2D-assisted content delivery with $N = K = 5$ and (c) $L = 2$ demands per user; and (d) $L = 1$ demand per user.

result which gives a new lower bound on the optimal storage-rate tradeoff.

Theorem 3: 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

$$\max_{\substack{s\in[1:\min\{\lceil N/L\rceil, K\}],\\ \ell\in[1:\lceil N/(Ls)\rceil]}} \frac{N - sM - \frac{\mu}{s+\mu}(N - L\ell s)^+}{\ell\left(\frac{K-s}{K}\right)}, \qquad (15)$$

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

The proof of Theorem 3 is presented in Appendix C. Similar to Theorem 1, 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. 2(c) shows the lower bound in (15) and the upper bound $R_{\mathsf{d2d,L}}(M)$ given in (11). Leveraging the proposed lower bound, we present our second main result in the following theorem.

Theorem 4: 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

$$\frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq 10. \qquad (16)$$

The proof of Theorem 4 is presented in Appendix D. The result shows that treating each of the L sets of user demands as a single demand case as outlined in Lemma 2 is order-optimal and yields a constant factor approximation of the optimal storage-rate trade-off for D2D-assisted content delivery with multiple demands per user. We note here that based on our proposed approach in Appendix D, an order-optimality result can be proved for the cut-set lower bound for centralized delivery in [20, Th. 5] (which is also a valid lower bound for D2D-assisted delivery) but with a higher constant gap.

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:

$$\max_{\substack{s\in[1:K],\\ \ell\in[1:\lceil N/s\rceil]}} \frac{N - sM - \left(\frac{\mu}{s+\mu}\right)(N - \ell s)^+}{\ell\left(\frac{K-s}{K}\right)}, \qquad (17)$$

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

Corollary 2 follows by setting $L = 1$ in Theorem 3 and was originally presented in [29].

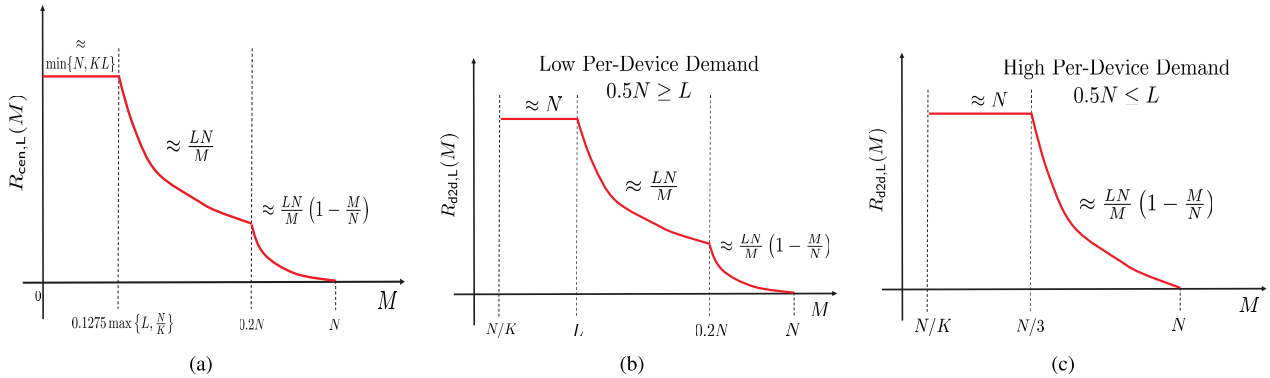


Fig. 3. A representation of the order-optimal approximations to the delivery rate for schemes which treat each of the L sets of K user demands independently as a single per-user demand case for (a) centralized content delivery, which is used in the proof of Theorem 2; and $(b)-(c)$ for D2D-assisted content delivery with low and high per-device demands, which are used in the proof of Theorem 4.

Remark 2: Compared to the cut-set bound in [12, Th. 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 (17). Furthermore, the bound in [12] 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 (17) is tight and yields the achievable rate in [12, Th. 1] as shown in Fig. 2(d).

Remark 3: Using the lower bound in Corollary 2, the optimal storage-rate tradeoff for the case of single demands per user can be approximated to within a factor of 8 and was first presented in [29, Th. 3]. The proof is omitted for brevity and presented in detail [30].

Remark 4: 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 2 and 4, we use approximations to the achievable rates presented in Lemmas 1 and 2. These approximations are highlighted in Fig. 3. 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(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 B and D.

IV. Case Studies

In this section, we present two case studies to illustrate the new techniques used to obtain the lower bounds in Theorems 1 and 3. For ease of exposition, we consider the special case of $L = 1$ 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.

A. Centralized Content Delivery: Intuition Behind Proof of Theorem 1

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:

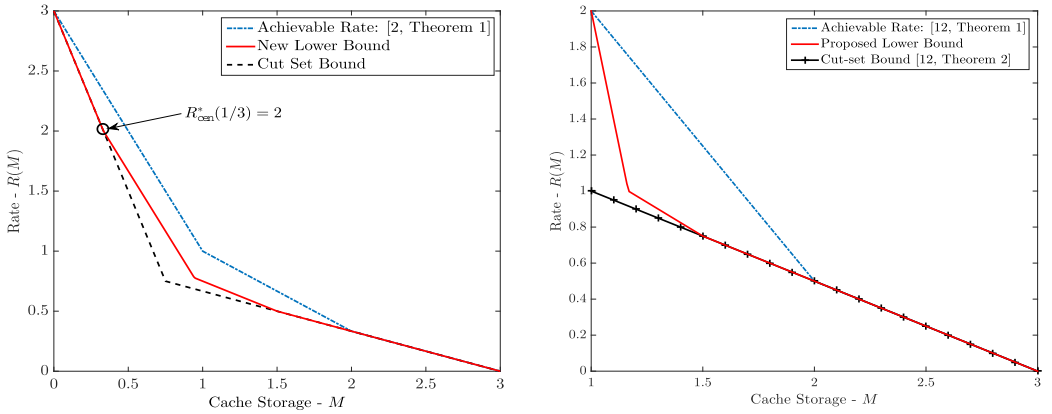
$$\begin{aligned} 3R^*_{\text{cen}} + 6M &\geq 8, \quad s = 2, \ \ell = 1; \\ 4R^*_{\text{cen}} + 2M &\geq 5, \quad s = 1, \ \ell = 2 \end{aligned} \tag{18}$$

$$\begin{aligned} R^*_{\text{cen}} + 3M &\geq 3, \quad s = 3, \ \ell = 1; \\ 3R^*_{\text{cen}} + \ M &\geq 3, \quad s = 1, \ \ell = 3. \end{aligned} \tag{19}$$

The existing lower bounds from [2, Th. 2] are given by (19). The proposed approach provides the additional bounds in (18), thereby yielding tighter lower bounds than [2, Th. 2] as shown in Fig. 4(a). Next, we detail the derivation of the first bound in (18) 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

$$\begin{aligned} 3B &\leq H(Z_{[1,2]}, X_{ABC}, X_{BCA}) \\ &\leq H(Z_{[1,2]}) + H(X_{ABC}, X_{BCA}|Z_{[1,2]}) \\ &\leq 2MB + H(X_{ABC}) + H(X_{BCA}|Z_{[1,2]}, X_{ABC}) \\ &\overset{(a)}{\leq} 2MB + R^*_{\text{cen}}B + H(X_{BCA}|Z_{[1,2]}, X_{ABC}, A, B) \\ &\leq 2MB + R^*_{\text{cen}}B + H(X_{BCA}, Z_3|Z_{[1,2]}, X_{ABC}, A, B) \\ &\leq 2MB + R^*_{\text{cen}}B + H(Z_3|Z_{[1,2]}, X_{ABC}, A, B) \\ &\quad + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B) \\ &\leq 2MB + R^*_{\text{cen}}B + H(Z_3|Z_{[1,2]}, A, B) \\ &\quad + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B, C) \\ &\overset{(b)}{\leq} 2MB + R^*_{\text{cen}}B + H(Z_3|Z_{[1,2]}, A, B), \end{aligned} \tag{20}$$



(a) (b)

Fig. 4. Storage-rate tradeoff for $N = K = 3$ and $L = 1$ with (a) centralized content delivery and (b) D2D-assisted content delivery.

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 (20), we have:

$$H(Z_3|Z_{[1,2]}, A, B) = H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{21}$$

Using (21) in (20), we have:

$$3B \leq 2MB + R^*_{\text{cen}}B + H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{22}$$

Now considering all possible subsets of $Z_{[1:3]}$ with cardinality 2, in the RHS of (22), we have:

$$3B \leq 2MB + R^*_{\text{cen}}B + H(Z_{[1:3]}|A, B) - H(Z_{[2,3]}|A, B) \tag{23}$$

$$3B \leq 2MB + R^*_{\text{cen}}B + H(Z_{[1:3]}|A, B) - H(Z_{[1,3]}|A, B). \tag{24}$$

Summing (22)-(24), and normalizing by 3, we have:

$$3B \leq 2MB + R^*_{\text{cen}}B + H(Z_{[1:3]}|A, B) - \sum_{\substack{i,j=1,\\ i\neq j}}^{3} \frac{H(Z_{[i,j]}|A, B)}{3}. \tag{25}$$

We next state Han's Inequality [31, Th. 17.6.1] on subsets of random variables, which we use for further upper bounding (25) 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

$$\frac{1}{\binom{n}{r}} \sum_{\mathsf{Y}_{[r]}:|\mathsf{Y}_{[r]}|=r} \frac{H\left(\mathsf{Y}_{[r]}\right)}{r} \leq \frac{1}{\binom{n}{m}} \sum_{\mathsf{Y}_{[m]}:|\mathsf{Y}_{[m]}|=m} \frac{H\left(\mathsf{Y}_{[m]}\right)}{m}, \tag{26}$$

where the sums are over all subsets of cardinality r, m respectively. Next, from (25), 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 (26), we have:

$$\frac{2H\left(Z_{[1:3]}|A, B\right)}{3} \leq \sum_{i,j=1,i\neq j}^{3} \frac{H\left(Z_{[i,j]}|A, B\right)}{3}. \tag{27}$$

Substituting (27) into (25), we have:

$$\begin{aligned} 3B &\leq 2MB + R^*_{\text{cen}}B + H(Z_{[1:3]}|A, B) - \frac{2}{3}H(Z_{[1:3]}|A, B) \\ &\leq 2MB + R^*_{\text{cen}}B + \frac{1}{3}H(Z_{[1:3]}|A, B) \\ &\leq 2MB + R^*_{\text{cen}}B + \frac{1}{3}H(Z_{[1:3]}, C|A, B) \\ &\leq 2MB + R^*_{\text{cen}}B + \frac{1}{3}\left(\underbrace{H(C|A, B)}_{\leq B} + \underbrace{H(Z_{[1:3]}|A, B, C)}_{=0}\right) \\ &\leq 2MB + R^*_{\text{cen}}B + \frac{1}{3}B. \end{aligned} \tag{28}$$

Rearranging (28), we get the new lower bound given by the first inequality in (18). The second bound in (18) 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 (20)-(28).

Remark 5: 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 (20), 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 (23), and (24), similar to (22). 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 1 which is a general result and holds for all problem parameters.

Remark 6: Recently [16]–[18] proposed caching and delivery schemes which improve upon the original multicasting scheme presented in [2, Th. 1]. Specifically, [16] 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 [2, Th. 1]. For $N = K = 3$, the new achievable point $(M, R) = (1/3, 2)$ is highlighted in Fig. 4(a). The lower bound in [2, Th. 2] is shown to be tight only in the regime $0 \leq M \leq 1/K$ for $K \geq N$ in [16]. 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. 4(a).

Remark 7: Maddah-Ali and Niesen [2] characterize the optimal storage-rate tradeoff for the case of $N = K = 2$ and show that the cut-set lower bound, given by $R^*_{\mathsf{cen}} + 2M \geq 2$ and $2R^*_{\mathsf{cen}} + M \geq 2$, is indeed loose by deriving an additional tighter lower bound $2R^*_{\mathsf{cen}} + 2M \geq 3$ 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 [2, Appendix] is discussed only for the case of $N = K = 2$, whereas our approach is a more general one for any N, K.

B. D2D-Assisted Content Delivery: Intuition Behind Proof of Theorem 3

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:

$$R^*_{\mathsf{d2d}} + 6M \geq 8, \quad s = 2, \ \ell = 1 \tag{29}$$

$$8R^*_{\mathsf{d2d}} + 6M \geq 15, \quad s = 1, \ \ell = 2 \tag{30}$$

$$2R^*_{\mathsf{d2d}} + \ M \geq 3, \quad s = 1, \ \ell = 3, \tag{31}$$

where (31) also recovers the cut set bound in [12, Th. 2]. Fig. 4(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 $X_{ABC} = \left\{X^1_{ABC}, X^2_{ABC}, X^3_{ABC}\right\}, X_{BCA} = \left\{X^1_{BCA}, X^2_{BCA}, X^3_{BCA}\right\}$. From the sum-rate constraint of the multicast transmissions in (5), we have

$$H(X_{ABC}) \leq \sum_{k=1}^{3} H(X^k_{ABC}) \leq R^*_{\mathsf{d2d}} B,$$
$$H\left(X^k_{ABC}\right) \leq R^*_{\mathsf{d2d}} B/3, \quad \forall k \in \{1, 2, 3\}, \tag{32}$$

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.

$$\begin{aligned}
3B &\leq H(Z_{[1,2]}, X_{ABC}, X_{BCA}) \\
&\leq H(Z_{[1,2]}) + H(X_{ABC}, X_{BCA}|Z_{[1,2]}) \\
&\leq H(Z_{[1,2]}) + H(X_{ABC}|Z_{[1,2]}) + H(X^3_{BCA}|Z_{[1,2]}, X_{ABC}) \\
&\overset{(a)}{\leq} 2MB + H(X^3_{ABC}) + H(X_{BCA}|Z_{[1,2]}, X_{ABC}) \\
&\leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(X_{BCA}|Z_{[1,2]}, X_{ABC}, A, B) \\
&\leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(X_{BCA}, Z_3|Z_{[1,2]}, X_{ABC}, A, B) \\
&\leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(Z_3|Z_{[1,2]}, X_{ABC}, A, B) \\
&\quad + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B) \\
&\overset{(b)}{\leq} 2MB + R^*_{\mathsf{d2d}} B/3 + H(Z_3|Z_{[1,2]}, A, B),
\end{aligned} \tag{33}$$

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:

$$H(Z_3|Z_{[1,2]}, A, B) = H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{34}$$

Using (34) in (33), we have:

$$3B \leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{35}$$

Again, considering all possible subsets of $Z_{[1:3]}$ having cardinality 2, in the RHS of (35), we have

$$3B \leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(Z_{[1:3]}|A, B) - H(Z_{[2,3]}|A, B). \tag{36}$$

$$3B \leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(Z_{[1:3]}|A, B) - H(Z_{[1,3]}|A, B). \tag{37}$$

Symmetrizing over the inequalities in (35)-(37), we have:

$$3B \leq 2MB + \frac{R^*_{\mathsf{d2d}} B}{3} + H(Z_{[1:3]}|A, B) - \sum_{\substack{i,j=1,\\ i \neq j}}^{3} \frac{H(Z_{[i,j]}|A, B)}{3}. \tag{38}$$

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 (26)), we have from (35)

$$\begin{aligned}
3B &\leq 2MB + R^*_{\mathsf{d2d}} B/3 + H(Z_{[1:3]}|A, B) - \frac{2H(Z_{[1:3]}|A, B)}{3} \\
&\leq 2MB + R^*_{\mathsf{d2d}} B/3 + \frac{H(Z_{[1:3]}, C|A, B)}{3} \\
&\leq 2MB + R^*_{\mathsf{d2d}} B/3 + B/3.
\end{aligned} \tag{39}$$

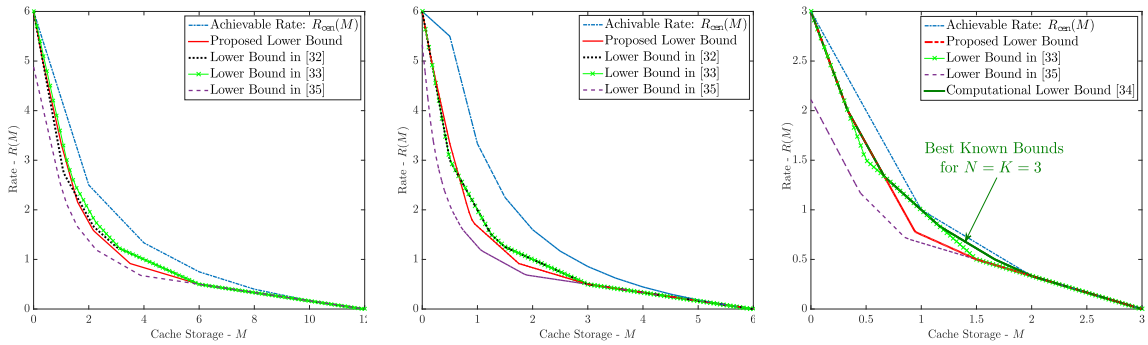


Fig. 5. Comparisons with parallel results for the case of centralized content delivery with $L = 1$ for a cache-aided system with (a) $N = 12, K = 6$; (b) $N = 6, K = 12$ and (c) $N = K = 3$.

Rearranging (39), we get the new lower bound in (29). 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 $2R^*_{\text{d2d}}/3$ 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 (30). 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 $2R^*_{\text{d2d}}/3$ as before, thereby recovering the cut set bound in (31). 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 (29) and comparing with the upper bound from [12, Th. 1] yields $R^*_{\text{d2d}}(1) = 2$ i.e., the achievable scheme proposed in [12] is optimal at $M = 1$.

V. Comparisons With Independent Results

We acknowledge the recent independent contributions from [32]–[36] on developing converse results for cache-aided systems. Wan et al. [32] derive a new converse bound based on index coding for the case of centralized content delivery with $L = 1$, which shows that the achievable scheme in [2] is optimal if *uncoded* cache placement is assumed. Improvements over the cut-set bound are also obtained for the case when $L = 1$ [33] and $L \geq 1$ [34] for centralized delivery and for the case of $L = 1$ for D2D-assisted delivery in [34], through different approaches than ours. The lower bounding approach adopted in these papers are inspired by the method adopted in [2, 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 [34], 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 [34] yields a better bound. To this end, in Fig. 5(a) and 5(b), we plot the results in [33] and [34] for $L = 1$. It can be seen that our bounds are better than [33] for the case of low cache memory for both cases and supercedes [34] in the second case, again for low cache storage. Note however, that unlike the simple form of our bound, the algorithm used in [34] 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.

Ajaykrishnan *et al.* [33] do not derive a constant gap result, however, Ghasemi and Ramamoorthy [34] show a constant gap of 4 to the achievable rate in [2, Th. 1]. We emphasize here that the analyses to obtain multiplicative gaps (as in Theorems 2 and 4) 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 [36]. The proposed lower bounds are generally always looser than the cut-set bounds for the case of centralized content delivery with $L = 1$ and by extension than the bounds presented in this paper as shown in Fig. 5. However, the authors leverage the structure of the bounds to approximate the storage-rate tradeoff to within a constant multiplicative factor of 4.7. We note here that the analysis presented in this paper 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 [34] and [36] in terms of approximately characterizing the optimal storage-rate tradeoff.

Finally, Tian [35] 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. 5(c). Our proposed method recovers the bound $6M + 3R^*_{\text{cen}} \geq 8$, while the approach in [33] and [34] recovers the bound $M + R^*_{\text{cen}} \geq 2$. However, it is unclear whether the bounds $12M + 18R^*_{\text{cen}} \geq 29$ and $3M + 6R^*_{\text{cen}} \geq 8$ 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.

VI. Conclusion

In this paper, we presented a new technique for deriving information theoretic lower bounds for 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 tradeoff 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.

Appendix A
Proof of Theorem 1

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

$$\begin{aligned}\mathbf{D}_1 &= (\mathbf{d}_{[1:s]}, \mathbf{d}_{[s+1:K]}) \\ &= \Big(\underbrace{[1:L], [L+1:2L], \ldots, [L(s-1)+1:Ls]}_{=\,\mathbf{d}_{[1:s]}}, \phi\Big),\end{aligned} \tag{40}$$

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,

$$\begin{aligned}\mathbf{D}_2 = \ \Big([Ls+1:L(s+1)], [L(s+1)+1:L(s+2)], \\ \ldots, [L(2s-1):2Ls], \phi\Big),\end{aligned} \tag{41}$$

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:

$$\begin{aligned}N &\le I\left(F_{1:N}; Z_{[1:s]}, X_{[1:\lceil N/(Ls)\rceil]}\right) \le H\left(Z_{[1:s]}, X_{[1:\lceil N/(Ls)\rceil]}\right) \\ &\le H\left(Z_{[1:s]}\right) + H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) \\ &\le sM + H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) \\ &\le sM + H\left(X_{[1:\ell]}|Z_{[1:s]}\right) + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}, X_{[1:\ell]}\right) \\ &\overset{(a)}{\le} sM + \ell R^*_{\mathsf{cen,L}}(M) \\ &\quad + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right) \\ &\overset{(b)}{\le} sM + \ell R^*_{\mathsf{cen,L}}(M) \\ &\quad + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}, Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right) \\ &\le sM + \ell R^*_{\mathsf{cen,L}}(M) + \underbrace{H\left(Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right)}_{\triangleq \delta} \\ &\quad + \underbrace{H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right)}_{\triangleq \lambda},\end{aligned} \tag{42}$$

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:

$$\mu = \min\left\{\left\lceil \frac{N - L\ell s}{L\ell} \right\rceil, K - s\right\} = \min\left\{\lceil N/(L\ell)\rceil, K\right\} - s, \tag{43}$$

where the last equality follows since s is an integer. Next, we obtain upper bounds on the two terms δ and λ in (42).

Upper Bound on $\boldsymbol{\delta}$: We consider the factor δ, from (42) and upper bound it as follows:

$$\begin{aligned}\delta &= H\left(Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right) \\ &\le H\left(Z_{[s+1:s+\mu]}|Z_{[1:s]}, F_{[1:L\ell s]}\right) \\ &= H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) - H\left(Z_{[1:s]}|F_{[1:L\ell s]}\right).\end{aligned} \tag{44}$$

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 (42), we can obtain $\binom{s+\mu}{s}$ different inequalities of the form of (44). Symmetrizing over all the inequalities, we have:

$$\delta \le H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) - \sum_{i=1}^{\binom{s+\mu}{s}} \frac{H\left(\mathsf{Z}^i_{[s]}|F_{[1:L\ell s]}\right)}{\binom{s+\mu}{s}}, \tag{45}$$

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 (26), we have:

$$\frac{s}{s+\mu} H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) \le \frac{1}{\binom{s+\mu}{s}} \sum_{i=1}^{\binom{s+\mu}{s}} H\left(\mathsf{Z}^i_{[s]}|F_{[1:L\ell s]}\right). \tag{46}$$

Substituting (46) into (45), we have:

$$\begin{aligned}\delta &\le H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right) - \frac{s}{s+\mu}H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right)\\ &= \frac{\mu}{s+\mu}H\left(Z_{[1:s+\mu]}|F_{[1:L\ell s]}\right)\\ &\le \frac{\mu}{s+\mu}H\left(Z_{[1:s+\mu]}, F_{[L\ell s+1:N]}|F_{[1:L\ell s]}\right)\\ &= \frac{\mu}{s+\mu}\left(H\left(F_{[L\ell s+1:N]}|F_{[1:L\ell s]}\right) + \underbrace{H\left(Z_{[1:s+\mu]}|F_{[1:N]}\right)}_{=0}\right)\\ &\overset{(a)}{\le} \frac{\mu}{s+\mu}(N-L\ell s)^+, \end{aligned} \tag{47}$$

where step (a) follows from the fact that the caches are functions of all N files in the library.

Upper Bound on $\boldsymbol{\lambda}$: To upper bound λ, we observe from the last step in (42) 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.,

$$\lambda = H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell(s+\mu)]}\right). \tag{48}$$

In order to characterize the upper bound on λ, we consider two cases as follows.

- ***Case 1*** $\boldsymbol{(N \le L\ell(s+\mu))}$ **:** All files are decoded by the caches $Z_{[1:s+\mu]}$ and transmissions $X_{[1:\ell]}$ within the conditioning for the term λ in (42). We have

$$\lambda = H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:N]}\right) = 0, \tag{49}$$

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. $s+\mu \le K$, we have:

$$KL\ell \ge L\ell(s+\mu) \ge N, \quad \text{i.e., } \lambda = (N-KL\ell)^+ = 0. \tag{50}$$

It can also be easily seen that for the case of $K \ge N$, $\lambda = (N-KL\ell)^+ = 0$ since $\ell, L \ge 1$.

- ***Case 2*** $\boldsymbol{(N > L\ell(s+\mu))}$ **:** 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 (42). In this case, $\lambda \ne 0$ and we have:

$$\begin{aligned}\lambda &= H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}\right)\\ &\le H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}, F_{[KL\ell+1:N]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}\right)\\ &\le H\left(F_{[KL\ell+1:N]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}\right)\\ &\quad + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:N]}\right)\\ &\overset{(a)}{\le} H\left(F_{[KL\ell+1:N]}\right) \le (N-KL\ell), \end{aligned} \tag{51}$$

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 (49) and (51), we can compactly bound λ as:

$$\lambda \le (N-KL\ell)^+. \tag{52}$$

Substituting (47) and (52) into (42), we have:

$$N \le sM + \ell R^*_{\mathsf{cen,L}}(M) + \frac{\mu}{s+\mu}(N-L\ell s)^+ + (N-KL\ell)^+ \tag{53}$$

Rearranging (53), we obtain the following lower bound on the optimal rate $R^*_{\mathsf{cen,L}}(M)$ as

$$\frac{1}{\ell}\left\{N - sM - \frac{\mu}{s+\mu}(N-L\ell s)^+ - (N-KL\ell)^+\right\}. \tag{54}$$

Optimizing over all parameter values of s, ℓ, completes the proof of Theorem 1.

Appendix B
Proof of Theorem 2

From Theorem 1, 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

$$\begin{aligned}\frac{\mu}{\mu+s} &= \frac{\min\left\{\left\lceil \frac{N}{L\ell}\right\rceil, K\right\} - s}{\min\left\{\left\lceil \frac{N}{L\ell}\right\rceil, K\right\}} \le 1 - \frac{s}{\left\lceil \frac{N}{L\ell}\right\rceil} = 1 - \frac{s}{\left\lceil \frac{N}{L\left\lceil \frac{\beta N}{Ls}\right\rceil}\right\rceil}\\ &\le 1 - \frac{s}{\left\lceil \frac{s}{\beta}\right\rceil} \le 1 - \frac{s}{\frac{s}{\beta}+1}\\ &= 1 - \frac{\beta}{1+\frac{\beta}{s}} \le 1 - \frac{\beta}{1+\beta} = \frac{1}{1+\beta}, \end{aligned} \tag{55}$$

where the last inequality follows from the fact that $s \ge 1$. Substituting (55) into (12), we have:

$$\begin{aligned}&R^*_{\mathsf{cen,L}}(M)\\ &\ge \frac{N - sM - \frac{1}{1+\beta}\left(N - L\left\lceil \frac{\beta N}{Ls}\right\rceil s\right)^+ - \left(N - KL\left\lceil \frac{\beta N}{Ls}\right\rceil\right)^+}{\lceil (\beta N)/(Ls)\rceil}\\ &\ge \frac{\left(\frac{2\beta}{1+\beta}\right)N - sM - N\left(1 - K\frac{\beta}{s}\right)^+}{\lceil (\beta N)/(Ls)\rceil}. \end{aligned} \tag{56}$$

Next, we consider two cases, namely (i) $\min\left\{\frac{N}{L}, K\right\} \le 10$; and (ii) $\min\left\{\frac{N}{L}, K\right\} \ge 11$.

- ***Case 1*** $\boldsymbol{\left(\min\left\{\frac{N}{L}, K\right\} \le 10\right)}$*:* For this case, setting $s = 1$ and $\beta = 1$ in (56), we have the following form on the lower bound,

$$R^*_{\mathsf{cen,L}}(M) \ge \frac{N\left(1 - \frac{M}{N}\right)}{\lceil N/L\rceil} \tag{57}$$

Consider first, the case when $\frac{N}{L} \le K$. From (10), we have the following upper bound on the achievable rate

$$R_{\mathsf{cen,L}}(M) \le \min\{N, KL\}\left(1 - \frac{M}{N}\right) \le N\left(1 - \frac{M}{N}\right). \tag{58}$$

Therefore, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \le \left\lceil \frac{N}{L}\right\rceil \le 10. \tag{59}$$

Next, consider the case when $K \le \frac{N}{L}$. Again, from (10), we have the following upper bound on the achievable rate

$$R_{\mathsf{cen,L}}(M) \le \min\{N, KL\}\left(1 - \frac{M}{N}\right) \le KL\left(1 - \frac{M}{N}\right). \tag{60}$$

Again, setting $s = 1$ and $\beta = 1$ in (56), we have

$$R^*_{\mathsf{cen,L}}(M) \geq \frac{N\left(1-\frac{M}{N}\right)}{\frac{N}{L}+1} = \frac{L\left(1-\frac{M}{N}\right)}{1+\frac{L}{N}} \geq \frac{KL\left(1-\frac{M}{N}\right)}{1+K} \tag{61}$$

Therefore, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq K+1 \leq 10+1 = 11. \tag{62}$$

- ***Case 2*** $\left(\min\left\{\frac{N}{L}, K\right\} \geq 11\right)$*:* 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*** $(0 \leq M \leq 1.275\max\{L, N/K\})$*:* 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 (56), we have

$$\begin{aligned}
&R^*_{\mathsf{cen,L}}(M) \\
&\geq \frac{\left(\frac{2\times 0.9649}{1+0.9649}\right) - s\frac{M}{N} - \left(1-K\frac{0.9649}{s}\right)^+}{\frac{0.9649}{Ls}+\frac{1}{N}} \\
&= \frac{1}{\frac{0.9649}{L\lfloor 0.3049\min\{N/L,K\}\rfloor}+\frac{1}{N}}\Bigg[\left(\frac{2\times 0.9649}{1+0.9649}\right) \\
&\quad - \lfloor 0.3049\min\{N/L, K\}\rfloor \frac{1.275\max\{L, N/K\}}{N} \\
&\quad - \left(1-K\frac{0.9649}{\lfloor 0.3049\min\{N/L, K\}\rfloor}\right)^+\Bigg] \\
&\overset{(a)}{\geq} \frac{1}{\frac{0.9649}{L(0.3049\min\{N/L,K\}-1)}+\frac{1}{N}}\Bigg[\left(\frac{2\times 0.9649}{1+0.9649}\right) \\
&\quad -(0.3049\times 1.275)\frac{\min\{N/L, K\}\max\{L, N/K\}}{N} \\
&\quad - \left(1-K\frac{0.9649}{0.3049\min\{N/L, K\}}\right)^+\Bigg] \\
&\geq \frac{1}{0.9649+\frac{L\left(0.3049\min\left\{\frac{N}{L},K\right\}-1\right)}{N}}\Bigg[L\min\left\{\frac{N}{L}, K\right\} \\
&\quad \times \left(0.3049-\frac{1}{\min\left\{\frac{N}{L}, K\right\}}\right)\left\{\left(\frac{2\times 0.9649}{1+0.9649}\right)\right. \\
&\qquad \left. - (0.3049\times 1.275) - \left(1-\frac{0.9649}{0.3049}\right)^+\right\}\Bigg] \\
&\overset{(b)}{\geq} \frac{1}{0.9649+0.3049}\Bigg[\min\{N, KL\}\left(0.3049-\frac{1}{10+1}\right) \\
&\qquad \times \left\{\left(\frac{2\times 0.9649}{1+0.9649}\right) - (0.3049\times 1.275)\right. \\
&\qquad \left. - \left(1-\frac{0.9649}{0.3049}\right)^+\right\}\Bigg] \\
&\geq \frac{\min\{N, KL\}}{10},
\end{aligned} \tag{63}$$

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 (10), we have

$$R_{\mathsf{cen,L}}(M) \leq \min\{N, KL\}\left(1-\frac{M}{N}\right) \leq \min\{N, KL\}. \tag{64}$$

Therefore for *Regime 1*, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq 10. \tag{65}$$

- ***Regime 2*** $(1.275\max\{L, N/K\} < M \leq 0.2N)$*:* For this regime, setting $s = \left\lfloor 0.442\frac{N}{M}\right\rfloor \in [1 : \min\{N/L, K\}]$[4] and $\ell = \left\lceil \frac{0.984N}{Ls}\right\rceil$, from (56), we have

$$\begin{aligned}
&R^*_{\mathsf{cen,L}}(M) \\
&\geq \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - s\frac{M}{N} - \left(1-K\frac{0.984}{s}\right)^+}{\frac{0.984}{Ls}+\frac{1}{N}} \\
&= \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - \left\lfloor 0.442\frac{N}{M}\right\rfloor\frac{M}{N} - \left(1-K\frac{0.984}{\left\lfloor 0.442\frac{N}{M}\right\rfloor}\right)^+}{\frac{0.984}{L\left\lfloor 0.442\frac{N}{M}\right\rfloor}+\frac{1}{N}} \\
&\overset{(a)}{\geq} \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - 0.442\frac{N}{M}\frac{M}{N} - \left(1-\frac{0.984}{0.442}\frac{KM}{N}\right)^+}{\frac{0.984}{L\left(0.442\frac{N}{M}-1\right)}+\frac{1}{N}} \\
&\overset{(b)}{\geq} \frac{1}{0.984+\frac{0.442}{1.275}\frac{L}{M}}\Bigg[\frac{LN}{M}\left(0.442-\frac{M}{N}\right)\left\{\left(\frac{2\times 0.984}{1+0.984}\right)\right. \\
&\quad \left. - 0.442 - \left(1-\frac{0.984}{0.442}\times 1.275\right)^+\right\}\Bigg] \\
&\overset{(c)}{\geq} \frac{1}{0.984+\frac{0.442}{1.275}}\Bigg[\frac{LN}{M}(0.442-0.2)\left\{\left(\frac{2\times 0.984}{1+0.984}\right)\right. \\
&\quad \left. - 0.442 - \left(1-\frac{0.984\times 1.275}{0.442}\right)^+\right\}\Bigg] \\
&\geq \frac{LN}{10M},
\end{aligned} \tag{66}$$

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 (10), we have

$$R_{\mathsf{cen,L}}(M) \leq \frac{KL\left(1-\frac{M}{N}\right)}{1+\frac{KM}{N}} \leq \frac{LN}{M}\left(1-\frac{M}{N}\right) \leq \frac{LN}{M}. \tag{67}$$

[4]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

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq 10. \tag{68}$$

- ***Regime 3 ($0.2N < M \leq N$)*:** In this regime, setting $s = 1$ and $\ell = \left\lceil \frac{N}{L} \right\rceil$ in (56), we note that in this case, $\mu = 0$ and $(N - K\ell)^+ = 0$. Thus, we have

$$R^*_{\mathsf{cen,L}}(M) \geq \frac{N\left(1 - \frac{M}{N}\right)}{\frac{N}{L} + 1} \geq \frac{\left(1 - \frac{M}{N}\right)}{\frac{1}{L} + \frac{1}{N}}. \tag{69}$$

From (10), we have

$$R_{\mathsf{cen,L}}(M) \leq \frac{KL\left(1 - \frac{M}{N}\right)}{1 + \frac{KM}{N}} \leq \frac{LN}{M}\left(1 - \frac{M}{N}\right). \tag{70}$$

Therefore for *Regime 3*, we have

$$\text{Gap} = \frac{R_{\mathsf{cen,L}}(M)}{R^*_{\mathsf{cen,L}}(M)} \leq \frac{LN}{M}\left(\frac{1}{L} + \frac{1}{N}\right) \leq \frac{2N}{M} \leq \frac{2}{0.2} \leq 10 \tag{71}$$

Combining (59), (62), (65), (68) and (71), completes the proof of Theorem 2.

Appendix C
Proof of Theorem 3

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

$$\begin{aligned}\mathbf{D}_1 &= \left(\mathbf{d}_{[1:s]}, \mathbf{d}_{[s+1:K]}\right) \\ &= \Big(\underbrace{\{1:L\}, \{L+1:2L\}, \ldots, \{L(s-1)+1:Ls\}}_{=\,\mathbf{d}_{[1:s]}}, \phi\Big),\end{aligned} \tag{72}$$

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

$$\begin{aligned}X_1 = \Big\{&X^1_{(\mathbf{d}_{[1:s]},\phi)}, \ldots, X^s_{(\mathbf{d}_{[1:s]},\phi)}, X^{s+1}_{(\mathbf{d}_{[1:s]},\phi)}, \\ &X^{s+2}_{(\mathbf{d}_{[1:s]},\phi)}, \ldots, X^K_{(\mathbf{d}_{[1:s]},\phi)}\Big\},\end{aligned} \tag{73}$$

composed of K device multicast transmissions, which, along with the s device caches decodes the files $F_{[1:Ls]}$. Similarly consider another demand vector,

$$\begin{aligned}\mathbf{D}_2 = \Big(&\{Ls+1 : L(s+1)\}, \{L(s+1)+1 : L(s+2)\}, \\ &\ldots, \{L(2s-1) : 2Ls\}, \phi\Big).\end{aligned} \tag{74}$$

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 (5) i.e.,

$$H\left(X^k_{\mathbf{D}}\right) \;\leq\; R^*_{\mathsf{d2d,L}}(M)/K, \quad \forall k \in [1:K]. \tag{75}$$

Considering $B = 1$ without loss of generality, we have:

$$\begin{aligned}N &\leq I\left(F_{1:N}; Z_{[1:s]}, X_{[1:\lceil N/(Ls)\rceil]}\right) \leq H\left(Z_{[1:s]}, X_{[1:\lceil N/(Ls)\rceil]}\right) \\ &\leq H\left(Z_{[1:s]}\right) + H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) \\ &\leq sM + H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) \\ &\leq sM + H\left(X_{[1:\ell]}|Z_{[1:s]}\right) + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}, X_{[1:\ell]}\right) \\ &\overset{(a)}{\leq} sM + \frac{\ell(K-s)}{K} R^*_{\mathsf{d2d,L}}(M) \\ &\quad + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right) \\ &\overset{(b)}{\leq} sM + \frac{\ell(K-s)}{K} R^*_{\mathsf{d2d,L}}(M) \\ &\quad + H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}, Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right) \\ &\leq sM + \frac{\ell(K-s)}{K} R^*_{\mathsf{d2d,L}}(M) \\ &\quad + \underbrace{H\left(Z_{[s+1:s+\mu]}|Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right)}_{\triangleq\, \delta} \\ &\quad + \underbrace{H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell s]}\right)}_{\triangleq\, \lambda},\end{aligned} \tag{76}$$

where in step (a), the second term follows from (75) 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:

$$H\left(X_{[1:\lceil N/(Ls)\rceil]}|Z_{[1:s]}\right) = 0, \tag{77}$$

since transmissions are functions of all K caches. As a result, the second step in (76) yields the minimum storage constraint for D2D-assisted delivery $KM \geq N$. Next we upper bound the terms δ, λ in (76) 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 (44)-(47) in Appendix A, yielding the upper bound

$$\delta \leq \frac{\mu}{s + \mu}(N - L\ell s)^+. \tag{78}$$

Upper Bound on $\boldsymbol{\lambda}$: We next derive an upper bound on the factor λ in (76) and consider two distinct cases as follows.

- ***Case 1 ($N \leq L\ell(s + \mu)$)*:** 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 (76), i.e., $L\ell(s+\mu) \geq N$. Thus, we have

$$\lambda = H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:s+\mu]}, F_{[1:N]}\right) = 0, \tag{79}$$

which follows from the fact that the transmissions are functions of all N files.

• ***Case 2 ($N > L\ell(s+\mu)$)*** *:* 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:

$$\begin{aligned}\lambda &= H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:K]}, F_{[1:KL\ell]}\right)\\ &\leq H\left(X_{[\ell+1:\lceil N/(Ls)\rceil]}|Z_{[1:K]}\right) = 0,\end{aligned} \tag{80}$$

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 (79) and (80) we have:

$$\lambda = 0. \tag{81}$$

Substituting (78) and (81) into (76) and optimizing over all parameter values of s, ℓ, completes the proof of Theorem 3.

Appendix D
Proof of Theorem 4

From Theorem 3, 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 (55) in Appendix B. Using this in (15) from Theorem 3, we have

$$\begin{aligned}R^*_{\mathsf{d2d,L}}(M) &\geq \frac{N - sM - \frac{1}{1+\beta}\left(N - L\left\lceil \frac{\beta N}{Ls}\right\rceil s\right)^+}{\left\lceil \frac{\beta N}{Ls}\right\rceil \left(\frac{K-s}{K}\right)}\\ &\geq \frac{N\left(\frac{2\beta}{1+\beta} - s\frac{M}{N}\right)}{\left\lceil \frac{\beta N}{Ls}\right\rceil \left(\frac{K-s}{K}\right)}\end{aligned} \tag{82}$$

In order to facilitate the proof of Theorem 4, 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 ($0.5N \geq L$)****:* 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 ($N/K \leq M \leq L$)*** **:** 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 ($N < K$)***: 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 (82), we have

$$R^*_{\mathsf{d2d,L}}(M) \geq (1 - M), \tag{83}$$

where we have used the fact that $L = 1$ when $N = 1$. Again considering the upper bound in (11), we have $R_{\mathsf{d2d,L}} \leq 1$. Using the upper and the lower bounds, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}}{R^*_{\mathsf{d2d,L}}} \leq \frac{1}{1-M} \leq \frac{1}{1-0.5} = 2. \tag{84}$$

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 (82), we have

$$\begin{aligned}&R^*_{\mathsf{d2d,L}}(M)\\ &\geq \frac{N\left(1 - \left\lceil \frac{N}{L}\right\rceil \frac{M}{N}\right)}{\left\lceil \frac{N}{L\left\lceil \frac{N}{L}\right\rceil}\right\rceil \frac{K - \left\lceil \frac{N}{L}\right\rceil}{K}} \geq N\left(1 - \left(\frac{N}{L} + 1\right)\frac{M}{N}\right)\\ &= N\left(1 - \left(\frac{1}{L} + \frac{1}{N}\right)M\right) \overset{\text{(a)}}{\geq} N\left(1 - \frac{3}{2} \times 0.5\right),\end{aligned} \tag{85}$$

where step (a) follows from the fact that $N \geq 2$ and $L \geq 1$. Again, from the upper bound in (11), we have $R_{\mathsf{d2d,L}} \leq N$. Using this, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}}{R^*_{\mathsf{d2d,L}}} \leq \frac{1}{1 - \frac{3}{2} \times 0.5} = 4. \tag{86}$$

We next consider the sub-regime $0.5 \leq M \leq L$. In this regime, setting $s = \left\lfloor 0.5\frac{N}{M} \right\rfloor \in [1:K]$[5] and $\beta = 1$, from the lower bound in (82), we have

$$\begin{aligned}R^*_{\mathsf{d2d,L}} &\geq \frac{N\left(1 - \left\lfloor 0.5\frac{N}{M}\right\rfloor \frac{M}{N}\right)}{\left\lceil \frac{N}{L\left\lfloor 0.5\frac{N}{M}\right\rfloor}\right\rceil \frac{K - \left\lfloor 0.5\frac{N}{M}\right\rfloor}{K}} \geq \frac{N(1-0.5)}{\left\lceil \frac{N/L}{\left\lfloor 0.5\frac{N}{L}\right\rfloor}\right\rceil}\\ &\overset{\text{(a)}}{\geq} \frac{N(1-0.5)}{3},\end{aligned} \tag{87}$$

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 (11), we have $R_{\mathsf{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq \frac{3}{1-0.5} = 6. \tag{88}$$

– ***Sub-case 2 ($N \geq K$)****:* 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,

[5]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 $0.5N/M \geq 0.5N/L \geq 1$.

setting $s = \left\lfloor 0.5\frac{N}{M} \right\rfloor \in [1:K]$[6] and $\beta = 1$, from the lower bound in (82), we have

$$R^*_{\mathsf{d2d,L}} \geq \frac{N\left(1-\left\lfloor 0.5\frac{N}{M}\right\rfloor \frac{M}{N}\right)}{\left\lceil \frac{N}{L\left\lfloor 0.5\frac{N}{M}\right\rfloor}\right\rceil \frac{K-\left\lfloor 0.5\frac{N}{M}\right\rfloor}{K}} \geq \frac{N(1-0.5)}{\left\lceil \frac{N/L}{\left\lfloor 0.5\frac{N}{L}\right\rfloor}\right\rceil} \overset{(a)}{\geq} \frac{N(1-0.5)}{3}, \tag{89}$$

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 (11), we have $R_{\mathsf{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq \frac{3}{1-0.5} = 6. \tag{90}$$

- ***Regime 2* ($L \leq M \leq 0.2N$) :** For this regime, setting $s = \left\lfloor 0.51\frac{N}{M}\right\rfloor \in [1:K]$[7] and $\ell = \left\lceil \frac{0.984N}{Ls}\right\rceil$, from (82), we have

$$\begin{aligned} R^*_{\mathsf{d2d,L}}(M) &\geq \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - s\frac{M}{N}}{\left(\frac{0.984}{Ls}+\frac{1}{N}\right)\left[\frac{K-s}{K}\right]} \overset{(a)}{=} \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - \left\lfloor 0.51\frac{N}{M}\right\rfloor \frac{M}{N}}{\frac{0.984}{L\left\lfloor 0.51\frac{N}{M}\right\rfloor}+\frac{1}{N}} \overset{(b)}{\geq} \frac{\left(\frac{2\times 0.984}{1+0.984}\right) - 0.51\frac{N}{M}\frac{M}{N}}{\frac{0.984}{L\left(0.51\frac{N}{M}-1\right)}+\frac{1}{N}} \\ &\geq \frac{\frac{LN}{M}\left(0.51-\frac{M}{N}\right)\left\{\left(\frac{2\times 0.984}{1+0.984}\right)-0.51\right\}}{0.984+0.51\left(\frac{L}{M}-\frac{L}{N}\right)} \\ &\overset{(c)}{\geq} \frac{\frac{LN}{M}(0.51-0.2)\left\{\left(\frac{2\times 0.984}{1+0.984}\right)-0.51\right\}}{0.984+0.51} \geq \frac{LN}{10M}, \end{aligned} \tag{91}$$

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 (11), we have

$$R_{\mathsf{d2d,L}}(M) \leq \frac{LN}{M}\left(1-\frac{M}{N}\right) \leq \frac{LN}{M}. \tag{92}$$

Therefore for *Regime 2*, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq 10. \tag{93}$$

- ***Regime 3* ($0.2N \leq M \leq N$) :** In this regime, setting $s = 1$ and $\beta = 1$ in (82), we have

$$R^*_{\mathsf{d2d,L}}(M) \geq \frac{N\left(1-\frac{M}{N}\right)}{\frac{N}{L}+1} \geq \frac{\left(1-\frac{M}{N}\right)}{\frac{1}{L}+\frac{1}{N}}. \tag{94}$$

Again, considering the upper bound in (11), we have

$$R_{\mathsf{d2d,L}}(M) \leq \frac{LN}{M}\left(1-\frac{M}{N}\right). \tag{95}$$

Therefore for *Regime 3*, we have

$$\begin{aligned} \text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} &\leq \frac{LN}{M}\left(\frac{1}{L}+\frac{1}{N}\right) \overset{(a)}{\leq} \frac{LN}{M}\times\frac{2}{L} \\ &\leq \frac{2N}{M} \leq \frac{2}{0.2} \leq 10, \end{aligned} \tag{96}$$

where step (a) follows from the fact that $L \leq N$.

- ***Case 2* ($0.5N \leq L$)**: 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* ($N/K \leq M \leq N/3$) :** For this regime, setting $s = 1$ and $\ell = \left\lceil \frac{0.5N}{Ls}\right\rceil$, from (82), we have

$$\begin{aligned} R^*_{\mathsf{d2d,L}}(M) &\geq \frac{N\left(\left(\frac{2\times 0.5}{1+0.5}\right)-\frac{M}{N}\right)}{\left\lceil\frac{0.5N}{L}\right\rceil\left(\frac{K-1}{K}\right)} \geq \frac{N\left(\left(\frac{2\times 0.5}{1+0.5}\right)-\frac{M}{N}\right)}{\frac{0.5N}{L}+1} \\ &\overset{(a)}{\geq} \frac{N\left(\left(\frac{2\times 0.5}{1+0.5}\right)-\frac{1}{3}\right)}{2}, \end{aligned} \tag{97}$$

where step (a) follows by using the lower bound $L \geq 0.5N$. Again, considering the upper bound in (11), we have $R_{\mathsf{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} \leq \frac{2}{\left(\frac{2\times 0.5}{1+0.5}\right)-\frac{1}{3}} \leq 6. \tag{98}$$

- ***Regime 2* ($N/3 \leq M \leq N$) :**
In this regime, setting $s = 1$ and $\beta = 1$ in (82), we have

$$R^*_{\mathsf{d2d,L}}(M) \geq \frac{N\left(1-\frac{M}{N}\right)}{\frac{N}{L}+1} \geq \frac{\left(1-\frac{M}{N}\right)}{\frac{1}{L}+\frac{1}{N}}. \tag{99}$$

From (11), we have

$$R_{\mathsf{d2d,L}}(M) \leq \frac{LN}{M}\left(1-\frac{M}{N}\right). \tag{100}$$

Therefore for *Regime 3*, we have

$$\begin{aligned} \text{Gap} = \frac{R_{\mathsf{d2d,L}}(M)}{R^*_{\mathsf{d2d,L}}(M)} &\leq \frac{LN}{M}\left(\frac{1}{L}+\frac{1}{N}\right) \\ &\leq \frac{LN}{M}\times\frac{2}{L} \leq \frac{2N}{M} \leq \frac{2}{1/3} = 6 \end{aligned} \tag{101}$$

Finally, combining (84), (86), (88), (90), (93), (96), (98) and (101), completes the proof of Theorem 4.

[6]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 $0.5N/M \geq 0.5N/L \geq 1$.

[7]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 $\lfloor 0.5N/M\rfloor \in [1:K]$.

REFERENCES

[1] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

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

[3] M. Ji, A. M. Tulino, J. Llorca, and G. Caire. (2015). "Order-optimal rate of caching and coded multicasting with random demands." [Online]. Available: http://arxiv.org/abs/1502.03124

[4] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.

[5] 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. [Online]. Available: http://dx.doi.org/10.1109/TNET.2014.2317316

[6] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2014, pp. 221–226.

[7] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 836–845, Apr. 2016.

[8] 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.

[9] A. Sengupta, R. Tandon, and T. C. Clancy, "Secure caching with non-uniform demands," in *Proc. IEEE Global Wireless Summit (GWS)*, May 2014, pp. 1–5.

[10] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning distributed caching strategies in small cell networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2014, pp. 917–921.

[11] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 1461–1465.

[12] 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.

[13] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2014, pp. 56–60.

[14] J. Hachem, N. Karamchandani, and S. Diggavi. (2014). "Content caching and delivery over heterogeneous wireless networks." [Online]. Available: http://arxiv.org/abs/1404.6560

[15] S. P. Shariatpanahi, A. S. Motahari, and B. H. Khalaj. (2015). "Multi-server coded caching." [Online]. Available: http://arxiv.org/abs/1503.00265

[16] Z. Chen, P. Fan, and K. B. Letaief. (Aug. 2014). "Fundamental limits of caching: improved bounds for small buffer users." [Online]. Available: https://arxiv.org/abs/1407.1935

[17] M. M. Amiri and D. Gunduz, "Fundamental limits of coded caching: Improved delivery rate-cache capacity trade-off," *IEEE Trans. Commun.*, to be published.

[18] K. Wan, D. Tuninetti, and P. Piantanida. (2016). "On caching with more users than files." [Online]. Available: http://arxiv.org/abs/1601.06383

[19] A. Sengupta, R. Tandon, and T. C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 1691–1695.

[20] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Caching and coded multicasting: Multiple groupcast index coding," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 881–885.

[21] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2015, pp. 98–107.

[22] M. Ji, A. Tulino, J. Llorca, and G. Caire, "Caching-aided coded multicasting with multiple random requests," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr./May 2015, pp. 1–5.

[23] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 809–813.

[24] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr. (Feb. 2016). "Fundamental limits of cache-aided interference management." [Online]. Available: http://arxiv.org/pdf/1602.04207v1

[25] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. 50th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2016, pp. 320–325.

[26] A. Sengupta, R. Tandon, and O. Simeone, "Cloud RAN and edge caching: Fundamental performance trade-offs," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2016, pp. 1–5.

[27] A. Sengupta, R. Tandon, and O. Simeone, "Pipelined fronthaul-edge content delivery in fog radio access networks," in *Proc. IEEE Globecom Workshops, Emerg. Technol. 5G Wireless Cellular Netw. (GC Wkshps)*, Washington, DC, USA, 2016, pp. 1-6, doi: 10.1109/GLOCOMW.2016.7849002. [online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7849002&isnumber=7848793

[28] A. Sengupta, R. Tandon, and O. Simeone. (2016). "Cloud and cache-aided wireless networks: Fundamental latency trade-offs." [Online]. Available: http://arxiv.org/abs/1605.01690

[29] A. Sengupta and R. Tandon, "Beyond cut-set bounds-the approximate capacity of D2D networks," in *Proc. Inf. Theory Appl. (ITA)*, Feb. 2015, pp. 78–83.

[30] A. Sengupta and R. Tandon. (Jun. 2016). "Improved approximation of storage-rate tradeoff for caching with multiple demands." [Online]. Available: http://arxiv.org/abs/1606.04202

[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[32] K. Wan, D. Tuninetti, and P. Piantanida. (2015). "On the optimality of uncoded cache placement." [Online]. Available: https://arxiv.org/abs/1511.02256

[33] N. Ajaykrishnan, N. S. Prem, V. M. Prabhakaran, and R. Vaze, "Critical database size for effective caching," in *Proc. 21st Nat. Conf. Commun. (NCC)*, 2015, pp. 1–6.

[34] H. Ghasemi and A. Ramamoorthy. (2015). "Improved lower bounds for coded caching." [Online]. Available: https://arxiv.org/abs/1501.06003

[35] C. Tian. (2015). "A note on the fundamental limits of coded caching." [Online]. Available: https://arxiv.org/abs/1503.00010

[36] C.-Y. Wang, S. H. Lim, and M. Gastpar. (2016). "A new converse bound for coded caching." [Online]. Available: http://arxiv.org/abs/1601.05690

Avik Sengupta (S'10–M'17) received the B.Tech. degree in electronics and communication engineering from the St.Thomas' College of Engineering and Technology, West Bengal University of Technology, Kolkata, India, in 2010, the M.S. degree in electrical and computer engineering from Kansas State University, Manhattan, KS, USA, in 2012, and the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2016. He is currently a 5G Wireless Systems Engineer with the Next Generation and Standards (NGS) Group, Intel Corp., Santa Clara, CA, USA. He has previously held internship positions with Qualcomm Inc., Huawei R&D (Futurewei), and Blackberry (Advanced Technology Group). His research interests include wireless communication systems with an emphasis on low latency content distribution and caching in wireless networks, information theory, and the applications of machine learning in cellular systems.

Ravi Tandon (S'03–M'09–SM'16) received the B.Tech. degree in electrical engineering from IIT Kanpur, Kanpur, India, in 2004, and the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park (UMCP), College Park, MD, USA, in 2010. From 2010 to 2012, he was a Post-Doctoral Research Associate with the Department of Electrical Engineering, Princeton University. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Arizona. Prior to joining the University of Arizona in fall 2015, he was a Research Assistant Professor with Virginia Tech with positions in the Bradley Department of ECE, the Hume Center for National Security and Technology, and the Discovery Analytics Center in the Department of Computer Science. His current research interests include information theory and its applications to wireless networks, communications, security and privacy, distributed storage systems, machine learning, and data mining. He was a co-recipient of the Best Paper Award at IEEE GLOBECOM 2011. He was nominated for the Graduate School Best Dissertation Award, and also for the ECE Distinguished Dissertation Fellowship Award at UMCP. He received the NSF CAREER Award in 2017.