Improved Approximation of Storage-Rate Tradeoff for Caching via New Outer Bounds

Avik Sengupta[†], Ravi Tandon[*], T. Charles Clancy[†]
[†] Hume Center & Dept. of Electrical and Computer Engineering
[*] Discovery Analytics Center & Dept. of Computer Science
Virginia Tech, Blacksburg, VA USA
Email: {aviksg, tandonr, tcc}@vt.edu

***Abstract*—Caching is a viable solution for alleviating the severe capacity crunch in modern content centric wireless networks. Parts of popular files are pre-stored in users' cache memories such that at times of heavy demand, users can be served locally from their cache content thereby reducing the peak network load. In this work, we consider a central server assisted caching network where files are jointly delivered to users through multicast transmissions. For such a network, we develop a new information theoretic lower bound on the fundamental cache storage vs. transmission rate tradeoff, which strictly improves upon the best known existing bounds. The new bounds are used to establish the approximate storage vs. rate tradeoff of centralized caching to within a constant multiplicative factor of 8.**

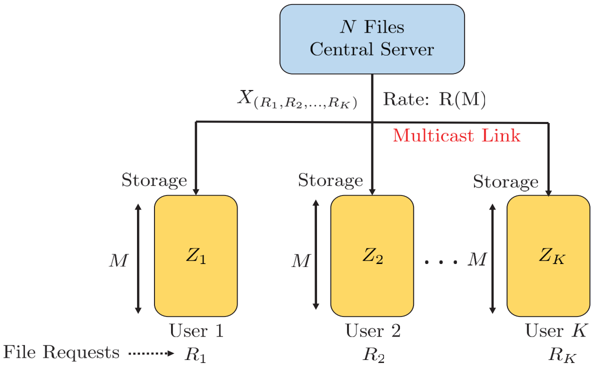

Fig. 1. System Model for Caching in Wireless Networks.

I. Introduction

The dynamics of traffic over wireless networks has undergone a paradigm shift to become increasingly content centric with multimedia content distribution holding precedence. Hence, it is imperative to improve the efficiency of capacity utilization in such networks. Caching is an important tool for facilitating efficient spectrum utilization and reducing network loads at times of peak demand. Parts of popular files are prestored at end users' device memories such that at times of high network load, 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 [1]–[13]. Caching has two phases - (1) the *storage phase* where parts of popular content is placed in users' cache memories and (2) the *delivery phase*, where requested content is delivered by exploiting the local cache storage of users. Consider a caching system with K users and a central server which has a library of N files (denoted by $(F_1, F_2, \ldots, F_N)$, each of size B bits). Each user $k \in 1, \ldots, K$, has a cache storage Z_k of size MB bits. The caches of the users are populated with some function of files based on available cache storage. Once the user requests are revealed, the server delivers content via a shared link to the users. The received transmission in conjunction with the user cache content is capable of decoding the requested files. Figure 1 shows the system model. The fundamental tradeoff for this system is that of cache storage vs. transmission rate (referred to as (M, R) tradeoff).

Recently, Maddah-Ali and Niesen [1]–[4] showed that by jointly designing the storage and delivery phase, and using multicast transmissions to simultaneously deliver content to users, order-wise improvement in the (M, R) tradeoff can be achieved. A novel caching and multicast delivery scheme based on shared content in user caches was presented, whereby a global caching gain was extracted from the system in addition to the traditional local gains. The authors used cut-set based arguments to derive lower bounds on the optimal (M, R) tradeoff and characterized it to within a constant multiplicative factor of 12. Recently, [13] presented an improved achievable scheme for the centralized caching problem in the case of $K > N$ and $M \leq 1/K$ i.e., for very small cache sizes.

In this work, we develop a new converse (lower bound) for the caching problem which better accounts for the content sharing across user caches and file decodability of the multicast transmissions. For this system, we observe that the cut-set based lower bound [1, Theorem 2] is loose expect for very small values of cache storage (when shared content is minimal) and very large values of cache storage (where almost all files can be completely stored in each user's cache). We present a new information theoretic lower bound which is generally tighter than the existing bound in [1] for all values of problem parameters. Using the new lower bound along with the achievable scheme (upper bound) from [1, Theorem 1], we characterize the optimal (M, R) tradeoff for caching to within a constant multiplicative factor of 8. This improves on the current result of 12 in [1, Theorem 3] by a factor of 1.5.

Notation: Let Y_i be a random variable. $Y_{[a:b]}$, where $a < b$ denotes the set of random variables $\{Y_i : i = a, a+1, \ldots, b-1, b\}$. Also, $Y_{[a,b]}$ denotes the set $\{Y_i : i = a, b\}$. $\mathrm{Y}_{[n]}$ denotes a set of any n arbitrary random variables Y_i such that $|Y_{[n]}| = n$. $\mathbb{N}^+$ denotes the set of natural numbers; the function $(x)^+ = \max\{0, x\}$; $\lceil x \rceil$, $\lfloor x \rfloor$ denotes the ceil, floor functions.

II. System Model

The caching network has K users and a library of N files, $F_1, \ldots, F_N$, where each file is of size B bits, for some $B \in \mathbb{N}^+$. Formally, the files F_n are i.i.d and distributed as:

$$F_n \sim \text{Unif}\{1, 2, \ldots, 2^B\}, \quad \forall n = 1, \ldots, N. \tag{1}$$

We next define the key components of the caching problem:

Definition 1 (Storage). *The cache storage phase consists of K caching functions, which map the files $(F_1, \ldots, F_N)$ into the cache content*

$$Z_k \triangleq \phi_k(F_1, \ldots, F_N), \tag{2}$$

for each user $k \in \{1, 2, \ldots, K\}$. The maximum allowable size of the contents of each cache Z_k is MB bits.

Definition 2 (Delivery). *The content delivery phase consists of N^K encoding functions which map the files $(F_1, \ldots, F_N)$ to the multicast transmission*

$$X_{(R_1,\ldots,R_K)} \triangleq \psi_{(R_1,\ldots,R_K)}(F_1, \ldots, F_N), \tag{3}$$

over the shared link in response to the requests $(R_1, \ldots, R_K) \in \{1, 2, \ldots, N^K\}$. Each such transmission has a rate not exceeding RB bits.

Definition 3 (File Decoding). *Once the multicast transmission is received, KN^K decoding functions map the received signal over the shared link $X_{(R_1,\ldots,R_K)}$ and the cache content Z_k to the estimate*

$$\hat{F}_{R_k} \triangleq \mu_{(R_1,\ldots,R_K),k}\left(X_{(R_1,\ldots,R_K)}, Z_k\right), \tag{4}$$

of the requested file F_{R_k} for user $k \in \{1, 2, \ldots, K\}$.

For the (M, R) caching scheme, the probability of error is defined as:

$$P_e \triangleq \max_{(R_1,\ldots,R_K)\in[N]^K} \max_{k\in[K]} \mathbb{P}(\hat{F}_{R_k} \neq F_{R_k}). \tag{5}$$

Definition 4 (Storage vs. Rate Tradeoff). *The storage-rate pair (M, R) is achievable if, for any $\epsilon > 0$, there exists an (M, R) caching scheme for which $P_e \leq \epsilon$. The storage vs. rate tradeoff is defined as:*

$$R^*(M) \triangleq \inf\{R : (M, R) \textit{ is achievable}\}. \tag{6}$$

This work focuses on the lower bound on the optimal (M, R) tradeoff for the caching problem.

III. Main Results and Discussion

We next present our first main result which gives a new lower bound on the optimal (M, R) tradeoff.

Theorem 1. *For any K users and N files, with each user having cache storage size M, where $0 \leq M \leq N$, the optimal content delivery rate $R^*(M)$ is lower bounded by:*

$$R^*(M) \geq R_{LB}(M) \triangleq$$

$$\max_{\substack{s\in\{1,\ldots,K\},\\ \ell\in\{1,\ldots,\lceil \frac{N}{s}\rceil\}}} \frac{1}{\ell}\left\{N - sM - \frac{\mu(N-\ell s)^+}{s+\mu} - (N-K\ell)^+\right\}, \tag{7}$$

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

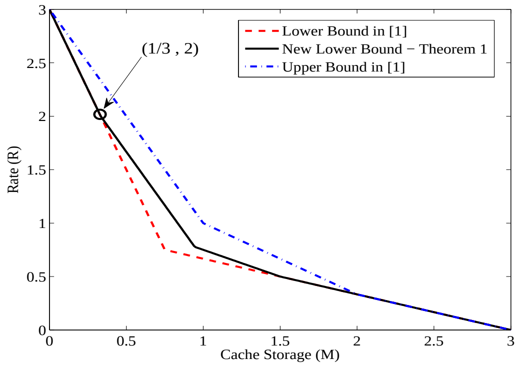

Fig. 2. (M, R) tradeoff for $N = K = 3$.

The proof of Theorem 1 is given in Appendix A. The expression in Theorem 1 has two parameters - s, which is related to the number of user caches, and another parameter ℓ, related to multicast transmissions. Compared to [1, Theorem 2], the additional parameter ℓ adds further flexibility to the lower bound expression and accounts for file decoding through the interaction of caches and transmissions, yielding a generally tighter lower bound for the caching problem. The cut-set based lower bound of [1, Theorem 2] is tight only for very small and large values of cache storage size M. As shown in the sequel, for such values of M, the proposed bound yields the bound in [1, Theorem 2] for specific choices of s and ℓ and is generally tighter for all other values.

Next, we present our second main result which establishes the optimal (M, R) tradeoff of the centralized caching problem to within a constant multiplicative factor.

Theorem 2. *Let $R_{UB}(M)$ be the achievable rate of the centralized caching scheme given in [1, Theorem 1] and let $R_{LB}(M)$ be the lower bound on the optimal rate given in Theorem 1. For any K users, N files, and user cache storage in the range $0 \leq M \leq N$, we have:*

$$Gap = \frac{R_{UB}(M)}{R_{LB}(M)} \leq 8. \tag{8}$$

The proof of Theorem 2 is omitted due to lack of space. The gap achieved by the proposed converse improves on the multiplicative gap of 12 presented in [1, Theorem 3], by a factor of 1.5. In the next section, we present an example to illustrate the new techniques used to obtain Theorem 1.

IV. Intuition Behind Proof of Theorem 1

We consider the case of $K = 3$ users, each with a cache storage M, and $N = 3$ files which we denote by A, B, C. Theorem 1 yields the following lower bounds for different choices of s, ℓ:

$$R^* + 3M \geq 3, \quad s = 3, \ \ell = 1 \tag{9}$$

$$3R^* + M \geq 3, \quad s = 1, \ \ell = 3 \tag{10}$$

$$3R^* + 6M \geq 8, \quad s = 2, \ \ell = 1 \tag{11}$$

$$4R^* + 2M \geq 5, \quad s = 1, \ \ell = 2. \tag{12}$$

The existing lower bounds from [1, Theorem 2] are given by (9) and (10). Theorem 1 provides the additional bounds (11) and (12). Fig. 2 shows that the proposed bound is strictly tighter than the bound in [1, Theorem 2].

Next, we detail the derivation of one of the new bounds, i.e., (11) highlighting the new aspects and techniques. To this end, we consider the requests $(R_1, R_2, R_3) = (A, B, C)$ and $(R_1, R_2, R_3) = (B, C, A)$. It is clear that the first $s = 2$ caches Z_1, Z_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^* and use the other transmission's decoding capability with the caches to derive the bound:

$$\begin{aligned}
3 &\leq H(Z_{[1,2]}, X_{ABC}, X_{BCA}) && (13)\\
&\leq H(Z_{[1,2]}) + H(X_{ABC}, X_{BCA}|Z_{[1,2]}) && (14)\\
&\leq 2M + H(X_{ABC}) + H(X_{BCA}|Z_{[1,2]}, X_{ABC}) && (15)\\
&\leq 2M + R^* + H(X_{BCA}|Z_{[1,2]}, X_{ABC}, A, B) && (16)\\
&\leq 2M + R^* + H(X_{BCA}, Z_3|Z_{[1,2]}, X_{ABC}, A, B) && (17)\\
&\leq 2M + R^* + H(Z_3|Z_{[1,2]}, X_{ABC}, A, B)\\
&\qquad + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B) && (18)\\
&\leq 2M + R^* + H(Z_3|Z_{[1,2]}, A, B)\\
&\qquad + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B, C) && (19)\\
&\leq 2M + R^* + H(Z_3|Z_{[1,2]}, A, B), && (20)
\end{aligned}$$

where (16) follows from the fact that $Z_{[1,2]}$ along with X_{ABC} can decode files A, B and (20) follows from the fact that $H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B, C) = 0$. This is due to the fact that $X_{BCA} = \psi_{(B,C,A)}(A, B, C)$. 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:

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

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

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

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

Summing up (22), (23), (24), and normalizing by 3, we have:

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

We next state Han's Inequality [14, Theorem 17.6.1] on subsets of random variables, which we use for further bounding (25) and deriving the proposed lower bound.

Han's Inequality: Let $\{X_1, X_2, \ldots, X_n\}$ denote a set of random variables. Further, let $\mathsf{X}_{[s]} \subseteq \{X_1, X_2, \ldots, X_n\}$ denote a subset of cardinality s. Then given two subsets $\mathsf{X}_{[r]}, \mathsf{X}_{[m]}$ where $r \geq m$, Han's Inequality states that:

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

where the sums are over all subsets of size r, m respectively. Next, returning to the proof of (25), consider the set of random variables $Z_{[1:3]} = (Z_1, Z_2, Z_3)$ and its subsets $\mathsf{Z}_{[i,j]} \subseteq Z_{[1:3]} : i \neq j, \forall i, j = 1, 2, 3$ of cardinality 2. Applying Han's Inequality for these random variables, with $n = r = 3$ and $m = 2$, we have:

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

$$\Rightarrow \frac{2}{3} H\left(Z_{[1:3]}|A, B\right) \leq \frac{1}{3} \sum_{i,j=1, i\neq j}^{3} H\left(\mathsf{Z}_{[i,j]}|A, B\right). \tag{28}$$

Substituting (28) into (25), we have:

$$\begin{aligned}
3 &\leq 2M + R^* + H(Z_{[1:3]}|A, B) - \frac{2H(Z_{[1:3]}|A, B)}{3} && (29)\\
&\leq 2M + R^* + \frac{1}{3} H(Z_{[1:3]}|A, B) && (30)\\
&\leq 2M + R^* + \frac{1}{3} H(Z_{[1:3]}, C|A, B) && (31)\\
&\leq 2M + R^* + \frac{1}{3}\Big(\underbrace{H(C|A, B)}_{\leq 1} + \underbrace{H(Z_{[1:3]}|A, B, C)}_{=0}\Big) && (32)\\
&\leq 2M + R^* + \frac{1}{3} && (33)\\
&\Rightarrow 3R^* + 6M \geq 8, && (34)
\end{aligned}$$

which is the new bound in (11).

Remark 1. *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 step in (15), a naive upper bound on the term $H(X_{BCA}|Z_{[1,2]}, X_{ABC})$ would be R^*, which would lead to $3 \leq 2M + 2R^*$, which is a loose bound. The main idea is to first observe that given $Z_{[1,2]} = (Z_1, Z_2)$ 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 subsequently leading to the new lower bound. This is the key approach behind Theorem 1 which is a general result and holds for all problem parameters.* ◇

Remark 2. *Recently, [13] 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 [1, Theorem 1]. For $N = K = 3$, the new achievable point $(M, R) = (1/3, 2)$ is highlighted in Figure 2. The lower bound in [1, Th, 2] is shown to be tight only in the regime $0 \leq M \leq 1/K$ for $K \geq N$ in [13]. The lower bound presented in Theorem 1 shows that this is indeed the case and that the new converse is tighter than the cut-set based lower bound for $M > 1/K$. This fact is highlighted in Figure 2, where the proposed lower bound is tighter than the cut-set bound for $M > 1/3$.* ◇

Remark 3. *In [1], the authors characterize the capacity*

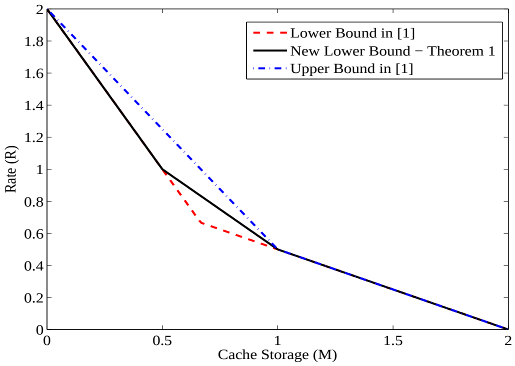

Fig. 3. (M, R) tradeoff for $N = K = 2$.

region for the case of $N = K = 2$ and show that their lower bound, given by $R^* + 2M \ge 2$ and $2R^* + M \ge 2$, is indeed loose. Our proposed lower bound yields the additional bound, $2R^* + 2M \ge 3$, which makes it tighter than the existing bound. From Figure 3 and [1] it can be seen that the proposed converse is equal to the optimal rate for the case of $N = K = 2$. ◇

Remark 4. *We would also like to acknowledge recent independent works of [15]–[17] on the same problem. In particular, [15], [17] also obtain improvements over cut-set bound through different approaches. Moreover, Tian [16] has recently obtained improvements for the specific case of $N = K = 3$ using a computer aided approach.*

V. Conclusion

In this paper, we presented a new information theoretic lower bound for the caching problem in wireless networks. We leveraged Han's Inequality to better model the interaction of user caches and file decoding capability of multicast transmissions to derive lower bounds which are generally tighter than the existing cut-set based bounds. Using the new lower bound, we characterized the cache storage vs. rate tradeoff of the centralized caching problem to within a constant multiplicative factor of 8 for all possible values of problem parameters, thereby improving on the existing result by a factor of 1.5.

Appendix A
Proof of Theorem 1

Let there be a library of $N \in \mathbb{N}^+$ files $\{F_1, F_2, \ldots, F_N\}$, each of size B bits and $K \in \mathbb{N}^+$ users in the content distribution system, with caches $\{Z_1, Z_2, \ldots, Z_K\}$. Let s be an integer such that $s \in \{1, 2, \ldots, K\}$. Consider the first s caches $Z_1, Z_2, \ldots, Z_s$ and a request vector $(R_1, R_2, \ldots, R_s, R_{s+1}, \ldots, R_K) = (1, 2, \ldots, s, \phi, \ldots, \phi)$, where the first s requests are for unique files and last $K - s$ requests can be for arbitrary files. To service this request, the central server makes a multicast transmission $X_1 = \psi(F_1, F_2, \ldots, F_s, \mathsf{F}_{[K-s]})$, where the first s files are unique and the remaining set of files, $\mathsf{F}_{[K-s]}$, to service the remaining $(K - s)$ requests can be arbitrary. This transmission along with the s caches decodes the files $F_1, F_2, \ldots, F_s$. Similarly consider another request, $(R_1, R_2, \ldots, R_s, R_{s+1}, \ldots, R_K) = (s + 1, S + 2, \ldots, 2s, \phi, \ldots, \phi)$, and a resultant multicast transmission $X_2 = \psi(F_{s+1}, F_{s+2}, \ldots, F_{2s}, \mathsf{F}_{[K-s]})$, where the contents of the set $\mathsf{F}_{[K-s]}$ are again arbitrary. This transmission along with the s caches helps in decoding the files $F_{s+1}, F_{s+2}, \ldots, F_{2s}$. Thus considering the transmissions $X_1, X_2, \ldots, X_{\lceil N/s \rceil}$ along with the first s caches $Z_1, Z_2, \ldots, Z_s$, the whole library of files $F_1, F_2, \ldots, F_N$ can be decoded. In the sequel, we consider $B = 1$ without loss of generality. We have:

$$\begin{aligned}
N &\le H\left(Z_{[1:s]}, X_{[1:\lceil N/s\rceil]}\right) && (35)\\
&\le H\left(Z_{[1:s]}\right) + H\left(X_{[1:\lceil N/s\rceil]} | Z_{[1:s]}\right) && (36)\\
&\le sM + H\left(X_{[1:\lceil N/s\rceil]} | Z_{[1:s]}\right) && (37)\\
&\le sM + H\left(X_{[1:\ell]} | Z_{[1:s]}\right) + H\left(X_{[\ell+1:\lceil N/s\rceil]} | Z_{[1:s]}, X_{[1:\ell]}\right) && (38)\\
&\le sM + \ell R^*(M) + H\big(X_{[\ell+1:\lceil N/s\rceil]} | Z_{[1:s]}, X_{[1:\ell]}\big) && (39)\\
&\le sM + \ell R^*(M) + H\left(X_{[\ell+1:\lceil N/s\rceil]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:\ell s]}\right) && (40)\\
&\le sM + \ell R^*(M) \\
&\quad + H\big(X_{[\ell+1:\lceil N/s\rceil]}, Z_{[s+1:s+\mu]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:\ell s]}\big) && (41)\\
&\le sM + \ell R^*(M) + \underbrace{H\left(Z_{[s+1:s+\mu]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:\ell s]}\right)}_{\triangleq \delta} \\
&\quad + \underbrace{H\left(X_{[\ell+1:\lceil N/s\rceil]} | Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:\ell s]}\right)}_{\triangleq \lambda}, && (42)
\end{aligned}$$

where (39) results from bounding the entropy of $\ell \in \{1, 2, \ldots, \lceil N/s \rceil\}$ transmissions given the caches $Z_1, \ldots, Z_s$ by $\ell R^*(M)$, where each transmission is of rate $R^*(M)$. (40) follows from the fact that caches $Z_1, \ldots, Z_s$ with transmissions $X_1, \ldots, X_\ell$ can decode files $F_1, \ldots, F_{\ell s}$. In (41), μ number of caches are introduced into the entropy, where μ is the number of remaining caches which along with caches $Z_1, \ldots, Z_s$ and transmissions $X_1, \ldots, X_\ell$, can decode the remaining $N - \ell s$ files. It is to be noted that all the remaining $K - s$ caches might not be required for decoding all files. Thus we have:

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

<u>**Upper Bound on** δ</u>: We consider the factor δ, from (42) and upper bound it as follows:

$$\begin{aligned}
\delta &= H\left(Z_{[s+1:s+\mu]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:\ell s]}\right) && (44)\\
&\le H\left(Z_{[s+1:s+\mu]} | Z_{[1:s]}, F_{[1:\ell s]}\right) && (45)\\
&= H\left(Z_{[1:s+\mu]} | F_{[1:\ell s]}\right) - H\left(Z_{[1:s]} | F_{[1:\ell s]}\right). && (46)
\end{aligned}$$

Considering all possible subsets of $Z_{[1:s+\mu]}$ having cardinality s, i.e., all possible combination of s caches in (35), and all possible combinations of files in the set $\mathsf{F}_{[K-s]}$ for each transmission $X_{[1:\ell]}$ in (39), we can obtain $\binom{s+\mu}{s}$ different inequalities of the form of (46). Symmetrizing over all the inequalities, we have:

$$\delta \le H\left(Z_{[1:s+\mu]} | F_{[1:\ell s]}\right) - \sum_{i=1}^{\binom{s+\mu}{s}} \frac{H\left(\mathsf{Z}^i_{[s]} | F_{[1:\ell s]}\right)}{\binom{s+\mu}{s}}, \qquad (47)$$

where, $\mathsf{Z}^i_{[s]}$ is the i-th size-s subset of $Z_{[1:s+\mu]}$.

Next, consider $Z_{[1:s+\mu]}$ as the set of random variables

$\{Z_k : k \in 1, \dots, s+\mu\}$ and the subsets $\mathsf{Z}^i_{[s]} \subseteq Z_{[1:s+\mu]}$ $\forall i = 1, \dots, \binom{s+\mu}{s}$. Applying Han's Inequality, from (26), using the conditional entropy of the sets, we have:

$$\frac{1}{\binom{s+\mu}{s+\mu}} \sum_{i=1}^{\binom{s+\mu}{s+\mu}} \frac{H\left(Z_{[1:s+\mu]}|F_{[1:\ell s]}\right)}{s+\mu}$$

$$\leq \frac{1}{\binom{s+\mu}{s}} \sum_{i=1}^{\binom{s+\mu}{s}} \frac{H\left(\mathsf{Z}^i_{[s]}|F_1, \dots, F_{\ell s}\right)}{s} \qquad (48)$$

$$\Rightarrow \frac{s}{s+\mu} H\left(Z_{[1:s+\mu]}|F_{[1:\ell s]}\right) \leq \frac{1}{\binom{s+\mu}{s}} \sum_{i=1}^{\binom{s+\mu}{s}} H\left(\mathsf{Z}^i_{[s]}|F_{[1:\ell s]}\right). \qquad (49)$$

Substituting (49) into (47), we have:

$$\delta \leq H\left(Z_{[1:s+\mu]}|F_{[1:\ell s]}\right) - \frac{s}{s+\mu} H\left(Z_{[1:s+\mu]}|F_{[1:\ell s]}\right) \qquad (50)$$

$$= \frac{\mu}{s+\mu} H\left(Z_{[1:s+\mu]}|F_{[1:\ell s]}\right) \qquad (51)$$

$$\leq \frac{\mu}{s+\mu} H\left(Z_{[1:s+\mu]}, F_{[\ell s+1:N]}|F_{[1:\ell s]}\right) \qquad (52)$$

$$= \frac{\mu}{s+\mu} H\left(F_{[\ell s+1:N]}|F_{[1:\ell s]}\right) + \underbrace{H\left(Z_{[1:s+\mu]}|F_{[1:N]}\right)}_{=0} \qquad (53)$$

$$\leq \frac{\mu}{s+\mu} (N-\ell s)^+, \qquad (54)$$

where (53) follows from the fact that the caches are functions of all the N files in the library.

Upper Bound on λ: To upper bound λ, we consider two cases.

Case 1: $N \leq \ell s + \ell\mu$: All files are decoded by the caches $Z_{[1:s+\mu]}$ and transmissions $X_{[1:\ell]}$ within the conditioning in (42). We have:

$$\lambda = H\left(X_{[\ell+1:\lceil N/s\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:\ell(s+\mu)]}\right) = 0, \qquad (55)$$

In the case when, for $N > K$, fewer than K caches suffices to decode all files with the transmissions within the conditioning in λ i.e. $s+\mu \leq K$, we have:

$$\begin{aligned} &\mu && \leq K - s \\ \Rightarrow\quad &\left\lceil \frac{N-\ell s}{\ell} \right\rceil && \leq K - s \\ \Rightarrow\quad &N && \leq K\ell \\ \Rightarrow\quad &\lambda = (N-K\ell)^+ && = 0. \end{aligned} \qquad (56)$$

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

Case 2: $N > \ell s + \ell\mu$: The case when, even with $s+\mu = K$ caches, all files are not decoded by the caches and transmissions within the conditioning. In this case, $\lambda \neq 0$ and we have:

$$\lambda \leq H\left(X_{[\ell+1:\lceil N/s\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:\ell s+\ell\mu]}\right) \qquad (57)$$

$$\leq H\left(X_{[\ell+1:\lceil N/s\rceil]}, F_{[\ell K+1:N]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:\ell K]}\right) \qquad (58)$$

$$\leq H\left(F_{[\ell K+1:N]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:\ell K]}\right) + \underbrace{H\left(X_{[\ell+1:\lceil N/s\rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:N]}\right)}_{=0} \qquad (59)$$

$$\leq H\left(F_{[\ell K+1:N]}\right) \leq (N-K\ell), \qquad (60)$$

where (58) follows from the fact that $\mu = K - s$ and (59) follows from the fact that the transmissions are functions of the N files. Thus combining (55) and (60), we have:

$$\lambda \leq (N-K\ell)^+. \qquad (61)$$

Substituting (54) and (61) into (42), we have:

$$N \quad \leq sM + \ell R^*(M) + \frac{\mu(N-\ell s)^+}{s+\mu} + (N-K\ell)^+$$

$$\Rightarrow R^*(M) \geq \frac{1}{\ell}\left\{N - sM - \frac{\mu(N-\ell s)^+}{s+\mu} - (N-K\ell)^+\right\}.$$

Optimizing over all parameter values of s, ℓ, μ, we have:

$$R^*(M) \geq R_{LB}(M) \triangleq$$

$$\max_{\substack{s\in\{1,\dots,K\},\\ \ell\in\left\{1,\dots,\left\lceil \frac{N}{s}\right\rceil\right\}}} \frac{1}{\ell}\left\{N - sM - \frac{\mu(N-\ell s)^+}{s+\mu} - (N-K\ell)^+\right\},$$

which completes the proof of Theorem 1. □

References

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[2] M. Maddah-Ali and U. Niesen, "Decentralized Coded Caching Attains Order-Optimal Memory-Rate Tradeoff," *IEEE/ACM Transactions on Networking*, April 2014.

[3] U. Niesen and M. A. Maddah-Ali, "Coded Caching with Nonuniform Demands," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2014, pp. 221–226.

[4] R. Pedarsani, M. Maddah-Ali, and U. Niesen, "Online Coded Caching," in *IEEE ICC*, June 2014, pp. 1878–1883.

[5] A. Sengupta, R. Tandon, and T. C. Clancy, "Fundamental Limits of Caching with Secure Delivery," *Wireless Physical Layer Security Workshop - IEEE ICC*, pp. 771–776, June 2014.

[6] ——, "Decentralized Caching with Secure Delivery," *IEEE International Symposium on Information Theory (ISIT)*, pp. 41–45, July 2014.

[7] ——, "Secure Caching with Non-Uniform Demands," *IEEE Global Wireless Summit (GWS)*, pp. 1–5, May 2014.

[8] ——, "Fundamental Limits of Caching with Secure Delivery," *IEEE Transactions on Information Forensics and Security*, vol. 10, pp. 355–370, Feb 2015.

[9] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning Distributed Caching Strategies in Small Cell Networks," *The Eleventh International Symposium on Wireless Communication Systems (ISWCS)*, pp. 917–921, Aug 2014.

[10] M. Ji, G. Caire, and A. F. Molisch, "Optimal Throughput-Outage Trade-off in Wireless One-Hop Caching Networks," in *IEEE International Symposium on Information Theory (ISIT)*, July 2013, pp. 1461–1465.

[11] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order Optimal Coded Caching-Aided Multicast under Zipf Demand Distributions," in *The Eleventh International Symposium on Wireless Communication Systems (ISWCS)*, 2014.

[12] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *arXiv:1405.5336*, 2014.

[13] Z. Chen, "Fundamental Limits of Caching: Improved Bounds For Small Buffer Users," *arxiv:1407.1935*, August 2014.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*. Hoboken, NJ, USA: Wiley-Interscience, John Wiley and Sons. Inc., 2006.

[15] N. Ajaykrishnan, N. S. Prem, V. M. Prabhakaran, and R. Vaze, "Critical Database Size for Effective Caching," *arXiv:1501.02549*, 2015.

[16] C. Tian, "A note on the fundamental limits of coded caching," *arXiv:1503.00010*, 2015.

[17] H. Ghasemi and A. Ramamoorthy, "Improved Lower Bounds for Coded Caching," *arXiv:1501.06003*, 2015.