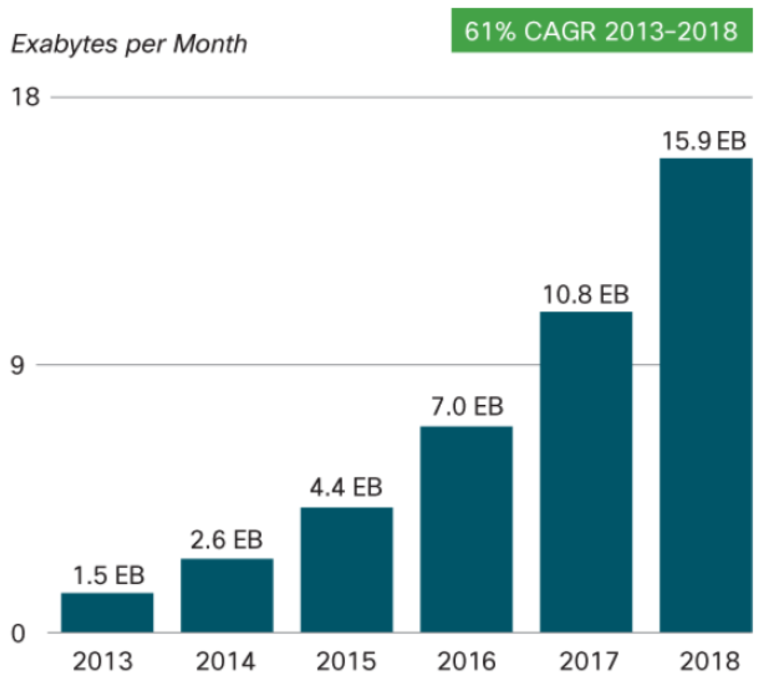
Performance Limits of Coded Caching under Heterogeneous Settings

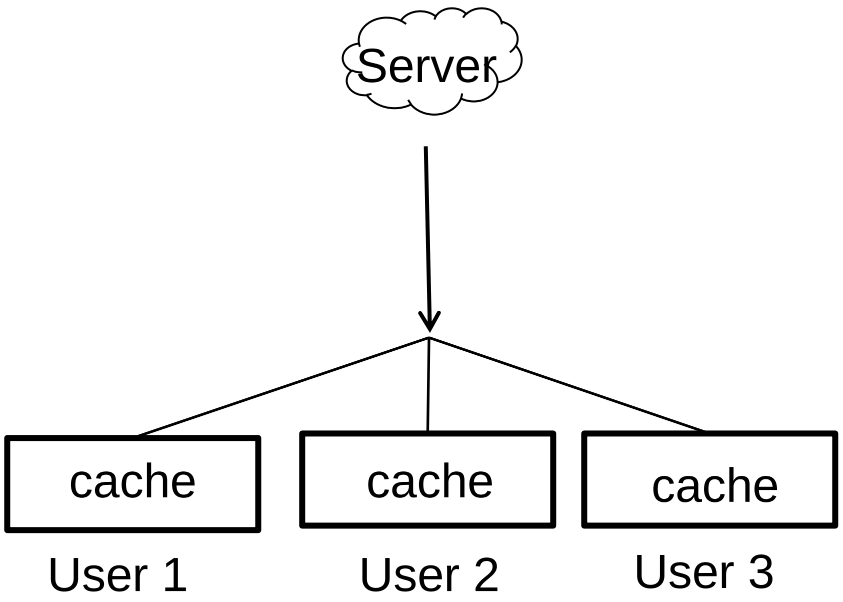
Xiaojun Lin

Associate Professor, Purdue University

Joint work with **Jinbei Zhang (SJTU), Chih-Chun Wang (Purdue) and Xinbing Wang (SJTU)**

The Importance of Caching





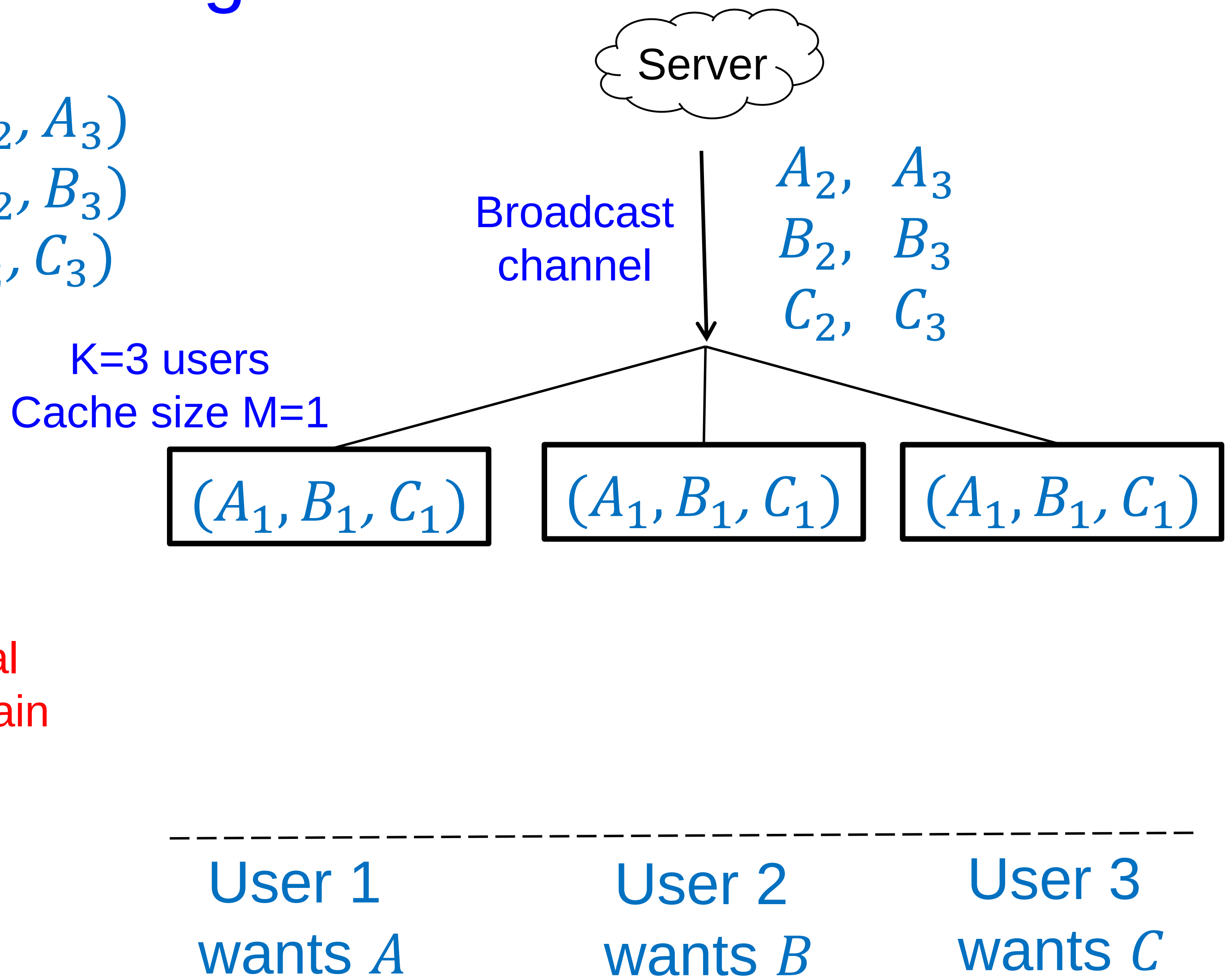
Source: Cisco VNI Mobile, 2014

- Data traffic continues to grow at significant rates
- A major fraction (60-80%) of traffic will be generated by multimedia content, such as video
- Caching is important for reducing backhaul requirement in serving large volumes of content that multiple users are interested in

Traditional (Uncoded) Caching: Individual cache size needs to be large

N=3 files (unit-size):

$$A = (A_1, A_2, A_3)$$
$$B = (B_1, B_2, B_3)$$
$$C = (C_1, C_2, C_3)$$



Back-haul Requirement:

- Uncoded Caching

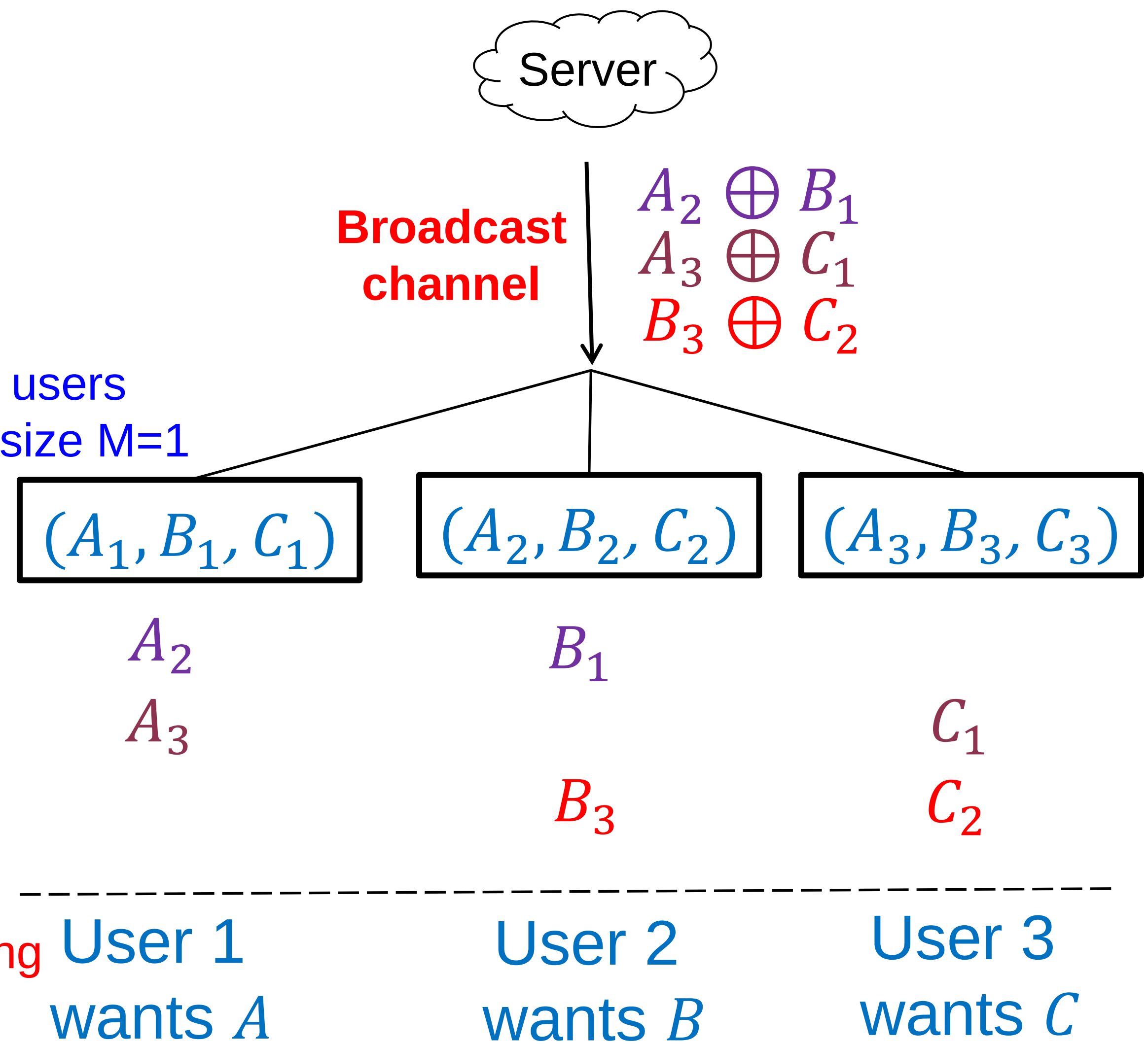
$$K \cdot \left(1 - \frac{M}{N}\right) = 2$$

Individual caching gain

Coded Caching: Global Caching Gains

N=3 files: $A = (A_1, A_2, A_3)$
$B = (B_1, B_2, B_3)$
$C = (C_1, C_2, C_3)$

K=3 users
Cache size M=1



Back-haul Requirement:

- Uncoded Caching

$$K \cdot \left(1 - \frac{M}{N}\right) = 2$$

- Coded Caching [1]

$$K \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{1}{1 + \frac{KM}{N}} = 1$$

Global caching gain

[1] Fundamental Limits of Caching, M. Maddah-Ali and U. Niesen, IEEE Trans. Inf. Theory, 2014.

Homogeneous vs Heterogeneous Settings

- [Maddah-Ali and Niesen '14] shows that the worst-case transmission rate $\boldsymbol{K \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{1}{1+\frac{KM}{N}}}$ is at most a ***constant factor*** (12x) away from the (information-theoretic) minimum possible

- Generalized to
  - Decentralized/probabilistic caching schemes [Maddah-Ali and Niesen `14]
  - Hierarchical caching [Karamchandani et al `14]
  - Online caching [Pedarsani et al `13]

- These studies assume a ***homogeneous*** setting where all files are equally important and are with the same parameters

Homogeneous vs Heterogeneous Settings

- In practice, heterogeneity arises naturally

- In ***homogeneous*** settings, all files are cached uniformly

- In ***heterogeneous*** settings:
  - Should ***more popular*** files be cached more aggressively [Niesen and Maddah-Ali `14, Ji et al `14, Hachem et al `14]?
  - Should ***larger*** files be cached more aggressively?

Our Contribution

- Coded caching needs to be adapted in different ways to different aspects of heterogeneity

- Heterogeneous ***popularity***:
  - Only files above a popularity ***threshold*** are cached
  - However, all popular files are cached ***uniformly*** (similar to [Ji et al '14])
  - We show ***constant-factor*** bounds that are independent of the popularity distribution
- Heterogeneous ***file-sizes***
  - (Roughly) ***quadratically*** more content is cached for larger files
  - We show ***logarithmic-factor*** bounds

- While the new achievable schemes are quite intuitive, the corresponding ***lower bounds*** are more involved and reveal useful insights

Outline

- *Coded Caching under Arbitrary Popularity Distributions*
  - System Model
  - Achievable Bounds and Intuitions
  - Lower Bounds
- Coded Caching under Distinct File Sizes
  - System Model
  - Achievable Bounds and Intuitions
  - Lower Bounds
- Conclusion and Discussions

Network Model: Heterogeneous Popularity

- Server with a broadcast channel
- K users: cache size M
- N (unit-size) files:

  $$\mathcal{F} = \{F_1, \dots, F_N\}$$

  Popularity (decreasing):

  $$\mathcal{P} = \{p_1, \dots, p_N\}$$

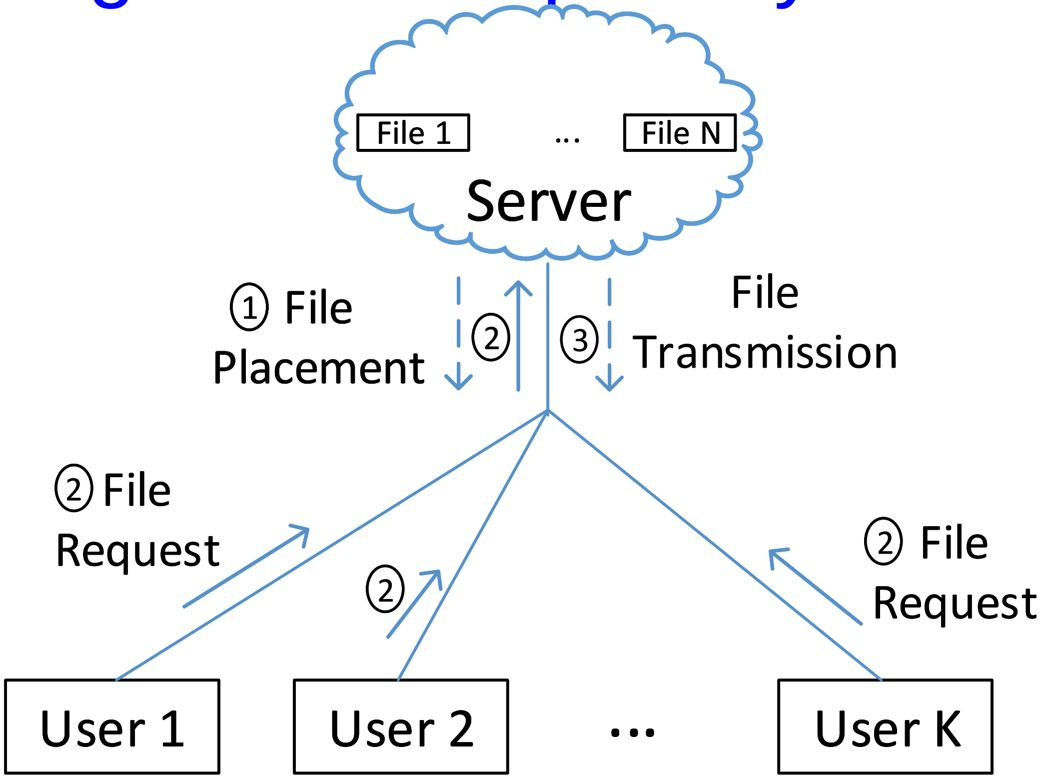
- Random request W_i

  $$W_i = \{f_{i1}, \dots, f_{iK}\}, f_{ik} \in \mathcal{F}$$

  Rate for serving W_i is $r(W_i)$

- Expected rate:

$$R(K, \mathcal{F}, \mathcal{P}) = \sum_{i=1}^{N^K} r(W_i) P(W_i)$$



Average-Case vs. Worst-Case

- Expected rate: $$\sum_{i=1}^{N^K} r(W_i)P(W_i)$$
- Worst-case rate [Maddah-Ali and Niesen `14]:

$$\max_{i=1,\ldots,N^K} r(W_i)$$

- Obviously:

$$\sum_{i=1}^{N^K} r(W_i)P(W_i) \leq \max_{i=1,\ldots,N^K} r(W_i)$$

- However, ***constant-factor results do NOT carry over*** from the worst case to the average case

$$\frac{\max_{i=1,\ldots,N^K} r(W_i)}{\max_{i=1,\ldots,N^K} r^*(W_i)} \leq c \quad \not\Longrightarrow \quad \frac{\sum_{i=1}^{N^K} r(W_i)P(W_i)}{\sum_{i=1}^{N^K} r^*(W_i)P(W_i)} \leq c$$

Related Work on the Average Case

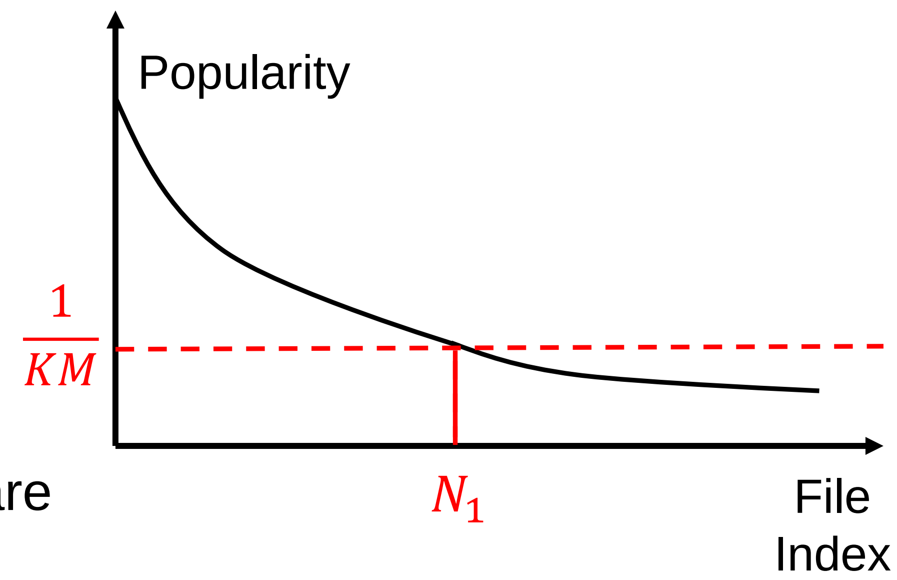
- U. Niesen, and M.A. Maddah-Ali, "Coded Caching with Nonuniform Demands", arXiv:1308.0178v2 [cs.IT], Mar. 2014.
  - Divide the files into ***groups***
  - The gap between the lower bound and the achievable (upper) bound increases with # of groups (unbounded)
- J. Hachem, N. Karamchandani and S. Diggavi, "Multi-level Coded Caching", arXiv:1404.6563 [cs.IT], Apr. 2014.
  - Popularity has ***multiple levels***
  - The gap increases with # of levels (unbounded)
- M. Ji, A. Tulino, J. Llorca and G. Caire, "On the Average Performance of Caching and Coded Multicasting with Random Demands", arXiv:1402.4576v2 [cs.IT], Jul. 2014.
  - ***Zipf*** popularity distribution $p_i \propto \frac{1}{i^\alpha}$
  - The gap increases with $\frac{1}{\alpha-1}$ when $\alpha > 1$ (unbounded)

Our Main Results

- ***Constant-factor gap*** between the lower bound (R_{lb}) and the achievable (upper) bound (R_{ub}) of the expected backhual transmission rate:

$$R_{ub} \leq 87R_{lb} + 2$$

- The achievable bound (R_{ub}) is attained by a simple coded caching scheme similar to [Ji et al '14]
  - Perform coded caching only among the ***most popular N_1 files***
  - However, all N_1 popular files are treated ***uniformly***

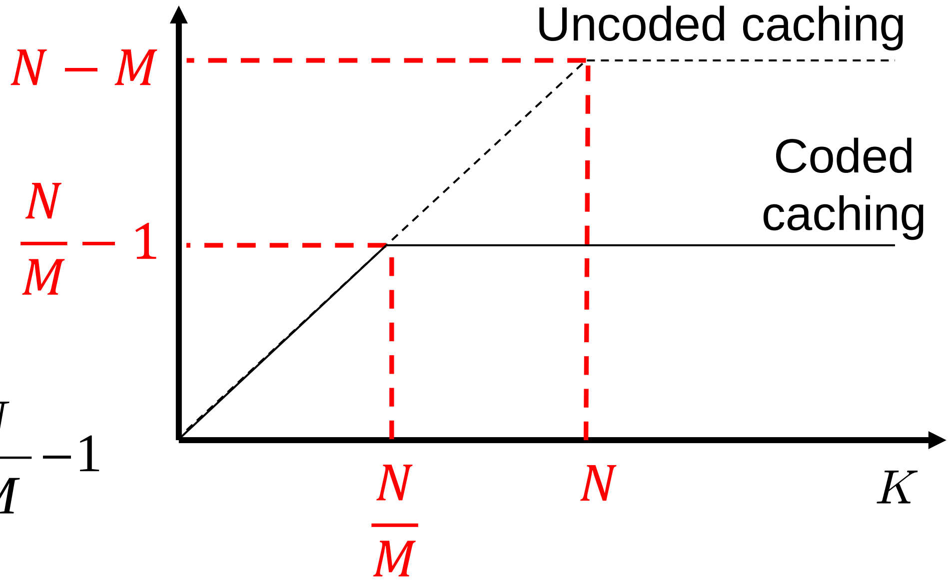


- The key step is to show a matching lower bound

Arbitrary Popularity Distribution!

Main Intuition: An “Insensitivity” Property

- The “best” worst-case rate for serving N files can be achieved by uniform caching [Maddah-Ali and Niesen ’14]



$$K \bullet (1-\frac{M}{N}) \bullet \frac{1}{1+\frac{KM}{N}} \approx \frac{N}{M}(1-\frac{M}{N}) = \frac{N}{M}-1$$

  whenever $K >> N/M$

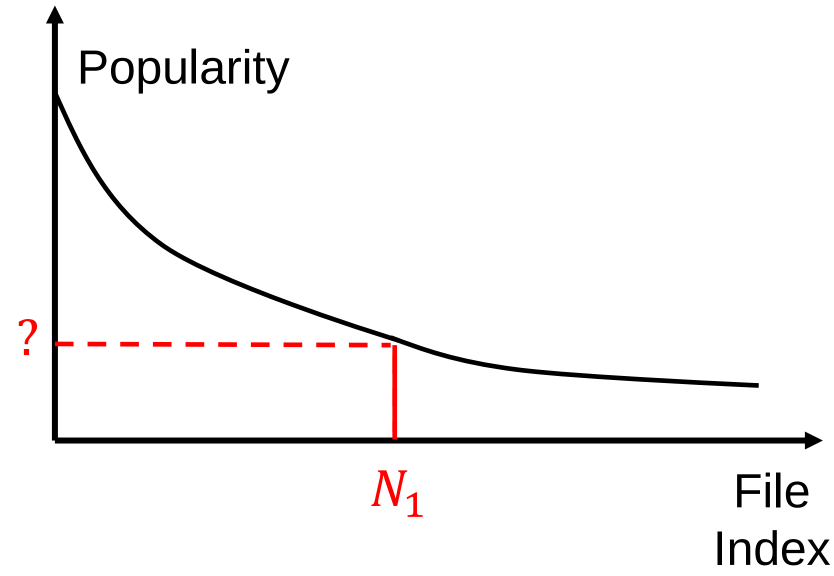
- ***Key Insight:*** Beyond K=N/M, the above rate is *independent of the number of users K*
- Due to its global caching gain, coded caching significantly reduce the threshold for this insensitivity to arise

Main Intuition: Average Case

- Consider the following scheme:
  - Only perform coded caching among most “popular” files 1 to N_1
- The average transmission rate for the “popular” files will be upper-bounded by the worst-case rate:

$$K' \bullet (1-\frac{M}{N_1}) \bullet \frac{1}{1+\frac{K'M}{N_1}} \approx \frac{N_1}{M}(1-\frac{M}{N_1}) = \frac{N_1}{M} - 1$$

whenever $K' \gg N_1/M$



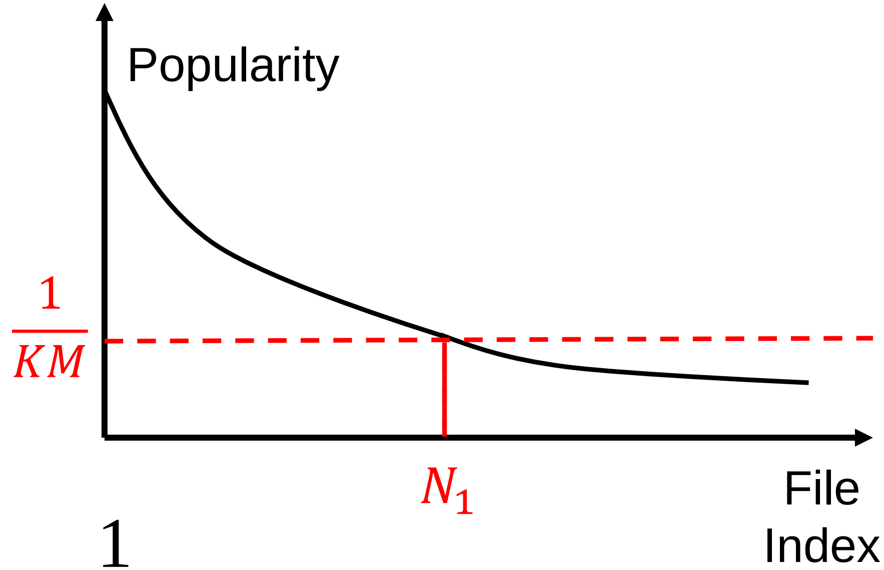
- If these files are indeed very popular, K' will be large. Thus, the expected rate will likely be close to this upper bound $\frac{N_1}{M} - 1$

Once a file is “popular”, its popularity does not matter!

Achievable Bound

- An achievable rate:

$$\underbrace{[\frac{N_1}{M}-1]}_{\text{“Popular”}}+\underbrace{\sum_{i>N_1} Kp_i}_{\text{“Unpopular”}}$$



- The minimum occurs at

$$\frac{1}{M} \approx Kp_{N_1} \iff p_{N_1} \approx \frac{1}{KM}$$

- **Proposition 1:** Assume $M \geq 2$. There exists an achievable scheme whose average transmission rate satisfies:

$$R(K,\mathcal{F},\mathcal{P}) \leq R_{ub} = \left[\frac{N_1}{M} - 1\right]_+ + \sum_{i>N_1} Kp_i$$

where N_1 satisfies $p_{N_1} \geq \frac{1}{KM}$ and $p_{N_1+1} < \frac{1}{KM}$.

Outline

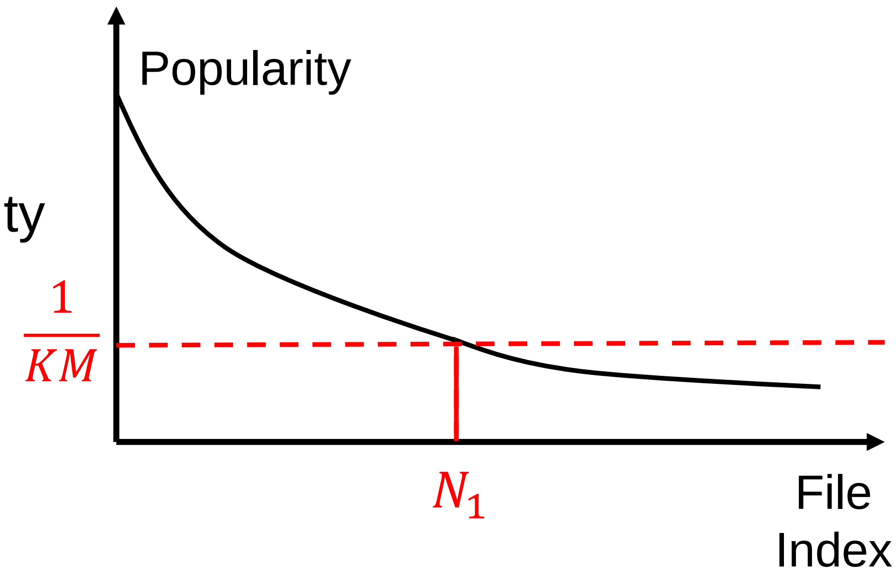
- Coded Caching under Arbitrary Popularity Distributions
  - System Model
  - Achievable Bounds and Intuitions
  - *Lower Bounds*
- Coded Caching under Distinct File Sizes
  - System Model
  - Achievable Bounds and Intuitions
  - Lower Bounds
- Conclusion and Discussions

Lower Bound: Statement

Proposition 2:

- Assume that $M \geq 2$. Let N_1 be the least popular file with popularity no smaller than $\frac{1}{KM}$, i.e.,

  $p_{N_1} \geq \frac{1}{KM}$ and $p_{N_1+1} < \frac{1}{KM}$



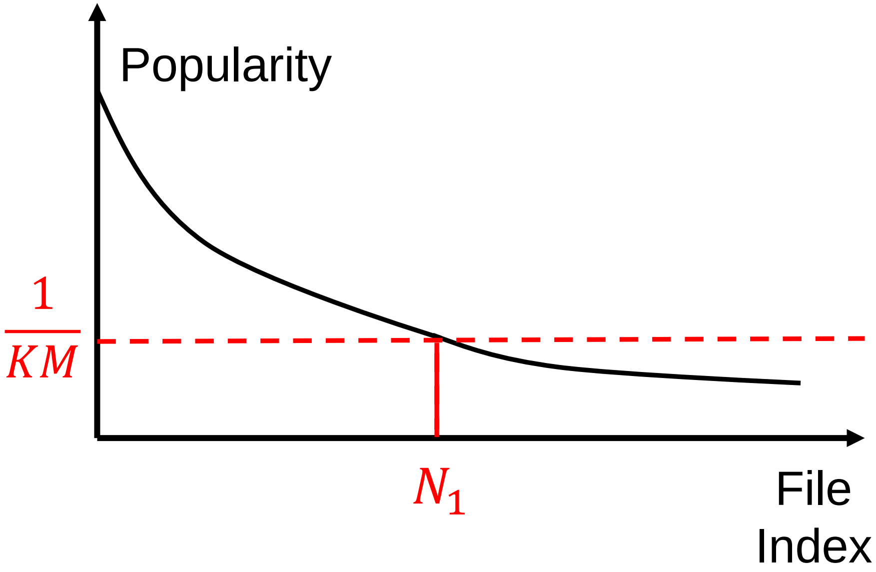
- For all possible coded caching schemes, the average transmission rate is lower bounded by

$$R(K,\mathcal{F},\mathcal{P}) \geq R_{lb} = \max\{\underbrace{\frac{1}{29}\left[\frac{N_1}{M}-1\right]_+}_{\substack{\text{Lower bound for}\\ \text{serving “popular files”}\\ 1 \text{ to } N_1}}, \underbrace{\frac{1}{58}\left[\sum_{i>N_1} Kp_i - 2\right]_+}_{\substack{\text{Lower bound for}\\ \text{serving “unpopular files”}\\ N_1+1 \text{ to } N}}\}$$

Lower Bound: Challenges

Need to show:

- For popular files, popularity does not matter
  - Reduce to uniform popularity $1/KM$ and use stochastic dominance



- With uniform popularity for all popular files, their worst-case rate and average-case rate are on the same order [Niesen and Maddah-Ali `14]
- For unpopular files, it is a good idea not to use any caching

Popular Files: Popularity Does Not Matter

$\mathcal{F}, \mathcal{P}$ | F_1, p_1 $\cdots$ F_{N_1}, p_{N_1} | F_{N_1+1}, p_{N_1+1} $\cdots$ F_N, p_N | $R(K, \mathcal{F}, \mathcal{P})$

Remove all unpopular files

IV

$\mathcal{F}_1, \mathcal{P}_1$ | F_1, p_1 $\cdots$ F_{N_1}, p_{N_1} | ∅ (empty file), $1 - \sum_{i=1}^{N_1} p_i$ | $R(K, \mathcal{F}_1, \mathcal{P}_1)$

Reduce all popularity to
the lowest value p_{N_1}

IV

$\mathcal{F}_1, \mathcal{P}_2$ | F_1, p_{N_1} $\cdots$ F_{N_1}, p_{N_1} | ∅ (empty file), $1 - N_1 p_{N_1}$ | $R(K, \mathcal{F}_1, \mathcal{P}_2)$

System 2

System 2 $(K, \mathcal{F}_1, \mathcal{P}_2)$: Worst-case vs. Average-Case

- Each of the *K* user request one of the N_1 popular files with equal probability $p_{N_1} \geq \frac{1}{KM}$
- The average number of users requesting popular files is $KN_1 p_{N_1} \geq \frac{N_1}{M}$
- ***Property 1***: with reasonable probability, the *number of users* K_r requesting popular files is no smaller than $\left\lfloor \frac{N_1}{M} \right\rfloor$. More precisely,

$$P\left(K_r \geq \left\lfloor \frac{N_1}{M} \right\rfloor\right) \geq 0.5$$

- ***Property 2:*** with reasonable probability, *the number of distinct files* K_d requested is no smaller than $0.5K_r$. More precisely

$$P\left(K_d \geq \left\lfloor \frac{1}{2} K_r \right\rfloor \middle| K_r\right) \geq 0.56$$

$$\Rightarrow \quad P\left(K_d \geq \left\lfloor \frac{1}{2} \left\lfloor \frac{N_1}{M} \right\rfloor \right\rfloor\right) \geq 0.28$$

denoted as K_3

Popular Files: Further Reduction from System 2

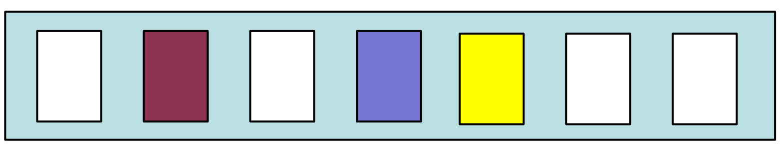
System 2
$R(K, \mathcal{F}_1, \mathcal{P}_2)$

F_1, p_{N_1}	$\cdots$	F_{N_1}, p_{N_1}	$\emptyset$(empty file), $1 - N_1 p_{N_1}$

If $K_d \geq K_3$,
i.e., the number of distinct files requested is no smaller than K_3
(This happens with probability ≥ 0.28)

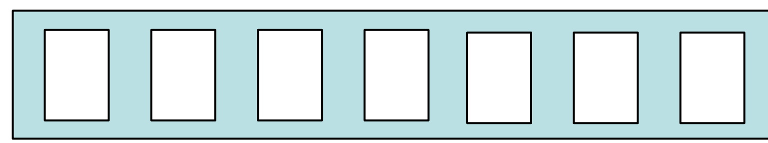
If $K_d < K_3$,

Users 1 2 … … … K



Exactly K_3 users requests K_3 distinct files.
Each pattern is equally likely.

1 2 … … … K



No files are requested

For the left hand side: Minimizing the average case will be equivalent to minimizing the worst case

$$R \geq \frac{1}{8}\left[\frac{N_1}{M} - 1\right]_+ \Rightarrow R(K, \mathcal{F}, \mathcal{P}) \geq \frac{0.28}{8}\left[\frac{N_1}{M} - 1\right]_+ \geq \frac{1}{29}\left[\frac{N_1}{M} - 1\right]_+$$

Lower Bound: Unpopular Files

$$R(K,\mathcal{F},\mathcal{P}) \geq \max\{\underbrace{\frac{1}{29}\left[\frac{N_1}{M}-1\right]_+}_{\text{Bound for serving "popular files"}}, \underbrace{\frac{1}{58}\Big[\sum_{i>N_1} Kp_i - 2\Big]_+}_{\text{Bound for serving "unpopular files"}}\}$$

$F_1, p_1 \quad \cdots \quad F_{N_1}, p_{N_1} \quad | \quad F_{N_1+1}, p_{N_1+1} \quad \cdots \quad F_i, p_i \quad F_j, p_j \quad \cdots \quad F_N, p_N$ $\qquad R(K,\mathcal{F},\mathcal{P})$

Remove all popular files $\qquad \geq$

$\emptyset, 1-\sum_{i=N_1+1}^{N} p_i \quad | \quad F_{N_1+1}, p_{N_1+1} \quad \cdots \quad F_i, p_i \quad F_j, p_j \quad \cdots \quad F_N, p_N$ $\qquad R(K,\mathcal{F}_3,\mathcal{P}_3)$

Merge any two unpopular files into one

$\emptyset, 1-\sum_{i=N_1+1}^{N} p_i \quad | \quad F_{N_1+1}, p_{N_1+1} \quad \cdots \quad V_k, p_i + p_j \quad \cdots \quad F_N, p_N$

Lower Rate

Unpopular Files: Reduction to System 2

$\emptyset, 1-\sum_{i=N_1+1}^{N} p_i$ | F_{N_1+1}, p_{N_1+1} $\cdots$ $V_k, p_i + p_j$ $\cdots$ F_N, p_N

Merge files until the sum popularity is just above 1/KM

$\emptyset, 1-\sum_{i=1}^{N_2} p_i$ | V_1, v_1 V_2, v_2 $\cdots$ V_{N_2}, v_{N_2} $R(K,\mathcal{F}_4,\mathcal{P}_4)$

$$\frac{2}{KM} > v_i \geq \frac{1}{KM},$$
$$N_2 \geq \left(\textstyle\sum_{i>N_1} p_i\right) KM/2$$

Reduce all popularity to 1/KM

$\geq$

$\emptyset, 1-\frac{N_2}{KM}$ | $V_1, \frac{1}{KM}$ $V_1, \frac{1}{KM}$ $\cdots$ $V_{N_2}, \frac{1}{KM}$ $R(K,\mathcal{F}_4,\mathcal{P}_5)$

This is exactly like System 2!

$$R(K,\mathcal{F},\mathcal{P}) \geq \frac{1}{29}\left[\frac{N_2}{M}-1\right]_+ \geq \frac{1}{29}\left[\frac{\sum_{i>N_1} K p_i}{2}-1\right]_+$$

Constant Factor

- We have shown the lower bound:

$$R(K,\mathcal{F},\mathcal{P}) \geq R_{lb} = \max\{\frac{1}{29}\left[\frac{N_1}{M} - 1\right]_+, \frac{1}{58}[\sum_{i>N_1} Kp_i - 2]_+\}$$

- Recall the achievable bound

$$R(K,\mathcal{F},\mathcal{P}) \leq R_{ub} = \left[\frac{N_1}{M} - 1\right]_+ + \sum_{i>N_1} Kp_i$$

- Constant-factor:

$$R_{ub} \leq 87 R_{lb} + 2$$

Arbitrary Popularity Distribution!

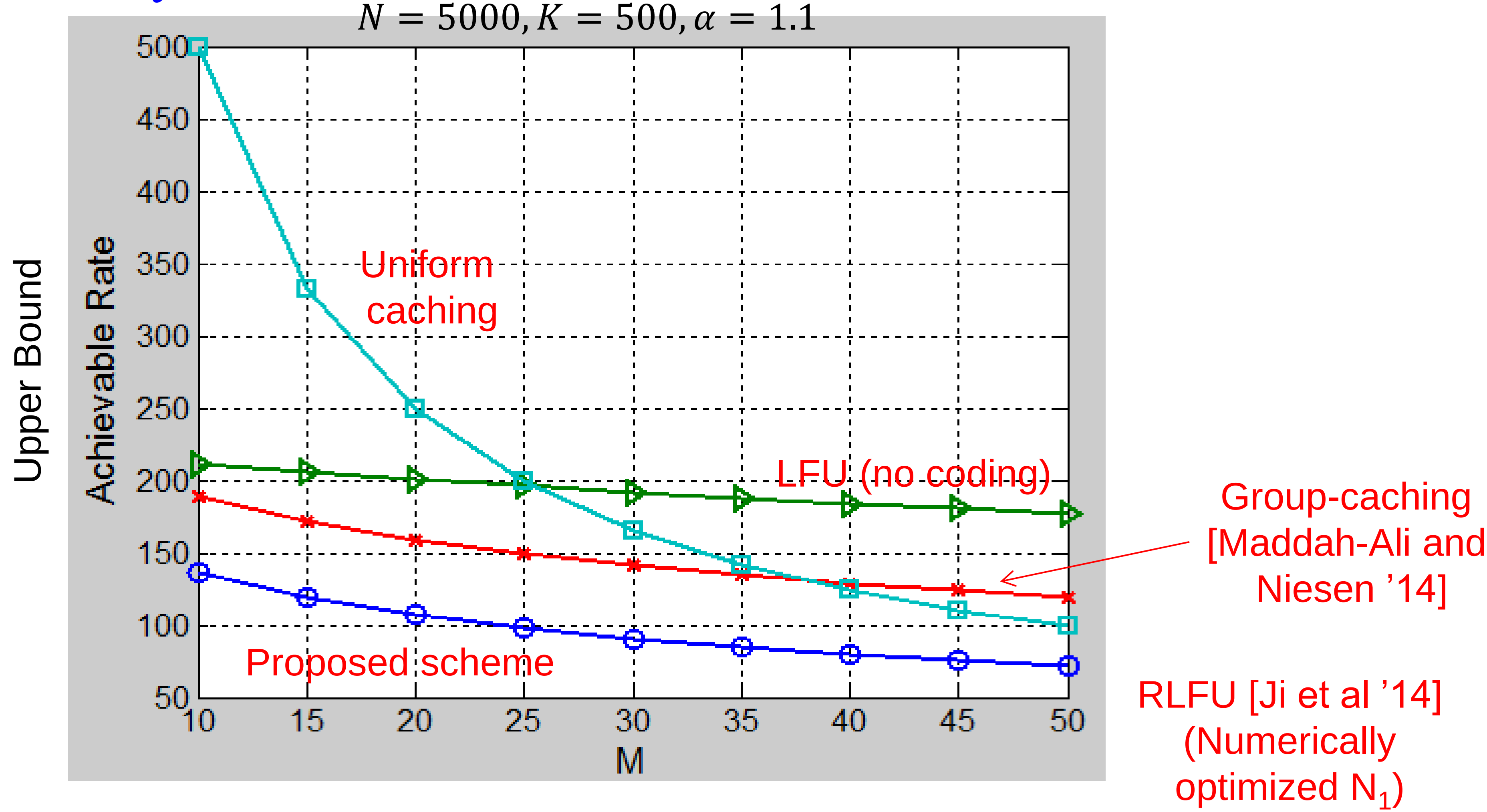
Numerical Comparison

- **LFU** [Lee et al ’01]:
  - Cache the M most popular contents (No coding)
- **Uniform-caching** [Maddah-Ali and Niesen ’14]:
  - Randomly cache $\frac{M}{N}$ portion of every content, regardless of popularity
- **Group-caching** [Niesen and Maddah-Ali, ’14]:
  - Divide the files into groups with similar popularity; perform coded caching within each group
  - Include an additional cache-allocation optimization
- **RLFU (Random LFU)** [Ji et al ’14]:
  - Assume Zipf popularity distribution; perform coded caching among the most popular files 1 to N_1
  - Numerically optimize N_1 based on some upper bound

- For uniform-caching, group-caching and RLFU, we plot the upper bound on the average transmission rate

Numerical Results: Zipf Distribution

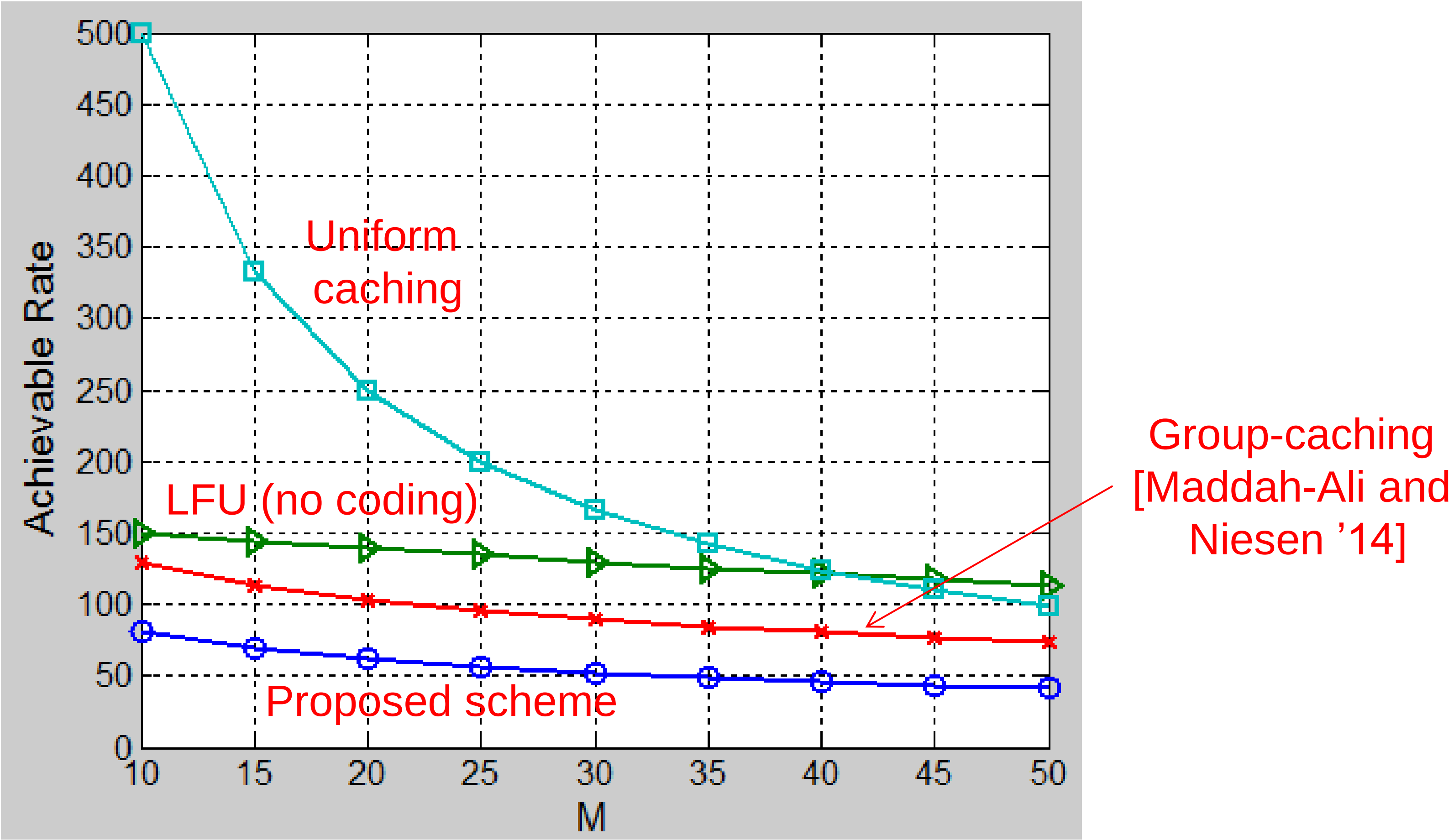
$p_i \propto \frac{1}{i^\alpha}$ and $\alpha = 1.1$



Note: For RLFU, Group-Caching, Uniform-Caching, we plot the upper bound

Non-Zipf Distribution

- Zipf-Mandelbrot law distribution
  - ➢ $p_i \propto \frac{1}{(i+r)^{\alpha}}, \alpha = 1.4, r = 2, N = 5000, K = 500$



Summary: Arbitrary Popularity Distributions

- We study the expected transmission rate of coded caching ***under arbitrary popularity distributions***
- We obtain achievable bounds that differ from the information-theoretic lower bound by at most a ***constant factor*** (except for a small additive term)

- ***Threshold*** Structure:
  - Perform coded caching only among popular files $\geq p_{N_1} \approx \frac{1}{KM}$
  - However, all popular files are cached uniformly
  - Similar to [Ji et al '14], but use a different N_1

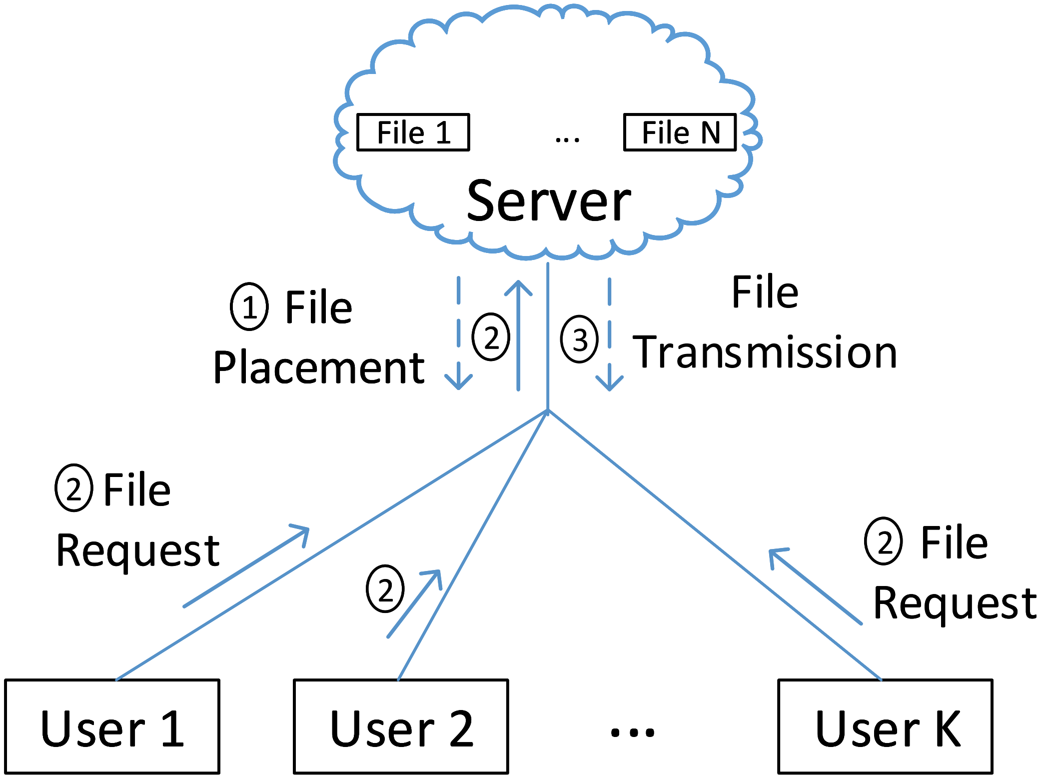
Outline

- Coded Caching under Arbitrary Popularity Distributions
  - System Model
  - Achievable Bounds and Intuitions
  - Lower Bounds
- *Coded Caching under Distinct File-Sizes*
  - System Model
  - Achievable Bounds and Intuitions
  - Lower Bounds
- Conclusion and Discussions

Network Model: Distinct File Sizes

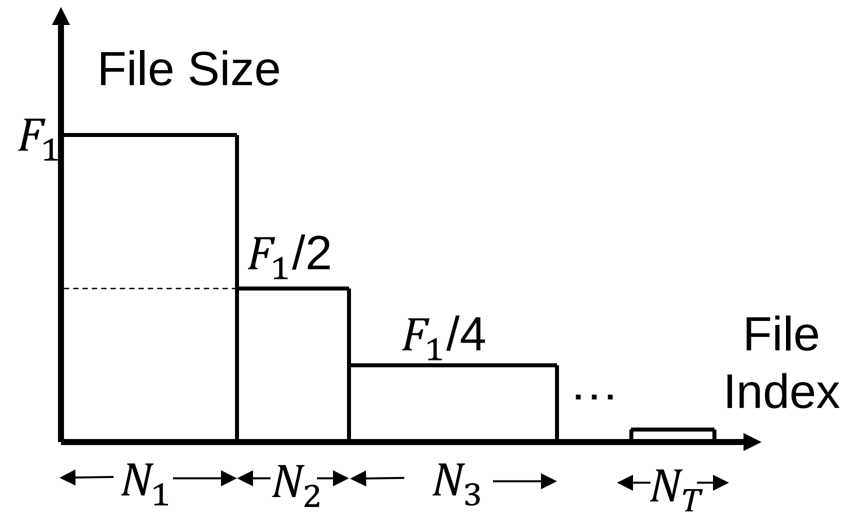
- Server with a broadcast channel
- K users: cache size M
- N files: $\mathbb{F} = \{\mathcal{F}_1, \dots, \mathcal{F}_N\}$
  File-size (non-increasing): $|\mathcal{F}_i| = F_i$
  $F_i \geq F_j$, if $i \leq j$
- Request pattern:
  $W_i = \{f_{i1}, \dots, f_{iK}\}, f_{ik} \in \mathbb{F}$
  Rate for W_i is $r(W_i)$
- **Worst-case rate**:

$$R = \max_i r(W_i)$$



Power-of-2 Simplification

- File-sizes differ by power-of-2 factors



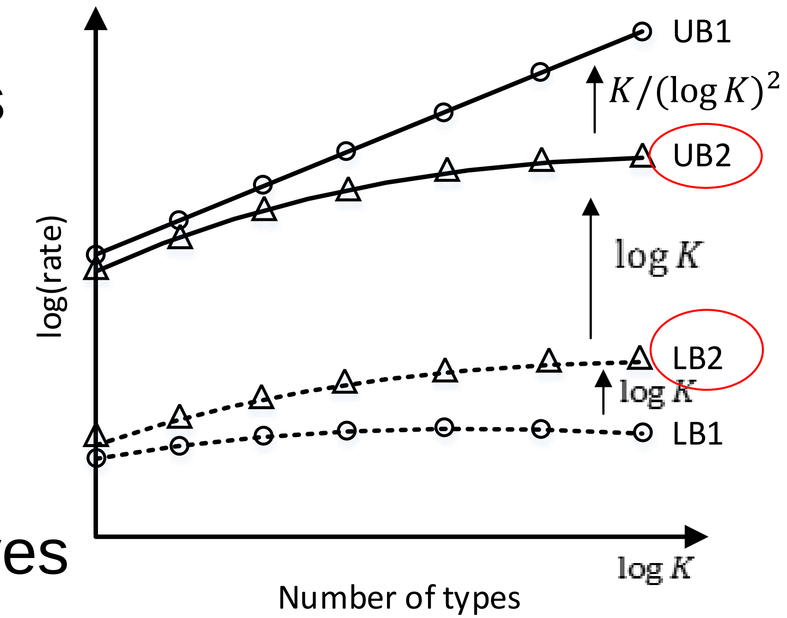
- l-th type: $F_l = F_1/2^{l-1}$
  - N_l files of type l
- T distinct types
  - The total number of files $\sum_{l=1}^{T} N_l = N$
- $\bar{T} = \min\{T, \log_2 K\}$
  - Files of type $l > \bar{T}$ can be virtually neglected
  - Their sizes ≤ F_1/K

Our Main Results

- ***Logarithmic-factor gap*** between the lower bound (R_{LB2}) and the achievable (upper) bound (R_{UB2}) for the worst-case transmission rate:

$$R_{UB2} \leq 32 \log_2 K \cdot R_{LB2} + 22$$

- The achievable bound (R_{UB2}) is attained by caching larger files more aggressively
  - ***Quadratically*** more content is cached for larger files
- The key step is to show a tighter lower bound, which involves careful use of entropy inequalities



Recall “Insensitivity” under Unit File-size

- The worst-case rate with uniform file size of 1 [Maddah-Ali & Niesen `14] is given by
  - $K \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{1}{1+\frac{KM}{N}} \approx \frac{N}{M} - 1 \approx \frac{N}{M}$, when K >> $\frac{N}{M}$ and M<<N
- Each user caches every “bit” of each file with probability $q = \frac{M}{N}$
  - Worst-case rate $\approx \frac{N}{M} = \frac{1}{q}$
- When the uniform file-size is F , these numbers become:
  - Caching probability $q = \frac{M}{NF}$,
  - Worst-case rate:

$$KF \cdot \left(1 - \frac{M}{NF}\right) \cdot \frac{1}{1+\frac{KM}{NF}} \approx \frac{NF^2}{M} = \frac{F}{q},$$ when q << 1 and K >> $1/q$

Two Achievable Schemes (UB1 vs UB2)

Achievable Scheme 1 (UB1)

- All files are cached with an equal probability q
  - ***Linearly*** more content is cached for larger files
- Cache constraint:

$$q \sum_{l=1}^{\bar{T}} N_l F_l = M \text{ with}$$
$$\bar{T} = \min(T, \log_2 K)$$

$$\begin{aligned} R_{UB1} &\geq O(\max_l \frac{F_l}{q}) \\ &= O(\frac{F_1 \sum_{l=1}^{\bar{T}} N_l F_l}{M}) \end{aligned}$$

Achievable Scheme 2 (UB2)

- Let the caching probability $q_l = F_l/c$
  - ***Quadratically*** more content is cached for larger files
- Cache constraint:

$$\sum_{l=1}^{\bar{T}} q_l N_l F_l = M \Longrightarrow \sum_{l=1}^{\bar{T}} N_l F_l^2 = cM$$

$$\begin{aligned} R_{UB2} &\cdot \sum_{l=1}^{\bar{T}} \frac{F_l}{q_l} + K F_{\bar{T}+1} \\ &= \bar{T}c + K F_{\bar{T}+1} \\ &\cdot (\bar{T}+1) \frac{\sum_{l=1}^{\bar{T}} N_l F_l^2}{M} \end{aligned}$$

Two Achievable Schemes (UB1 vs UB2)

Achievable Scheme 1 (UB1)

$$R_{UB1} \geq O\left(\frac{F_1 \sum_{l=1}^{\bar{T}} N_l F_l}{M}\right)$$

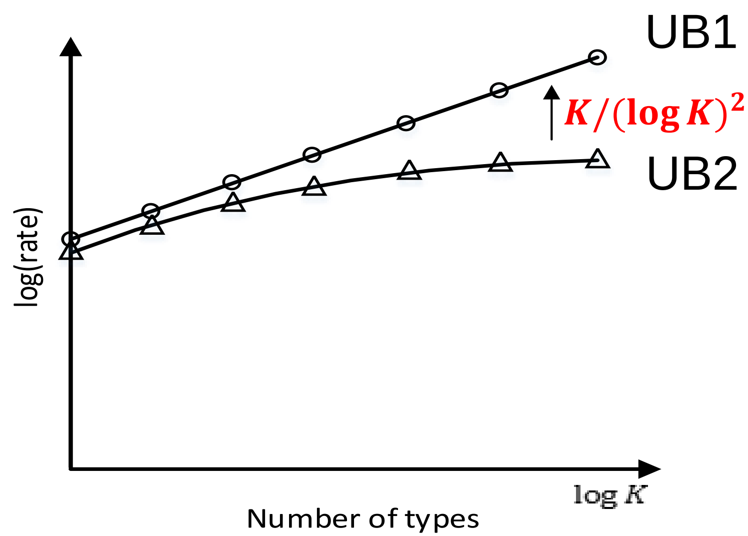
Achievable Scheme 2 (UB2)

$$R_{UB2} \cdot (\bar{T}+1)\frac{\sum_{l=1}^{\bar{T}} N_l F_l^2}{M}$$

- Consider $T = \log_2 K$, $N_{l+1} = 4N_l$

$$R_{UB1} \geq K\frac{N_1 F_1^2}{8M}$$

$$R_{UB2} \cdot (\bar{T}+1)\bar{T}\frac{N_1 F_l^2}{M}$$

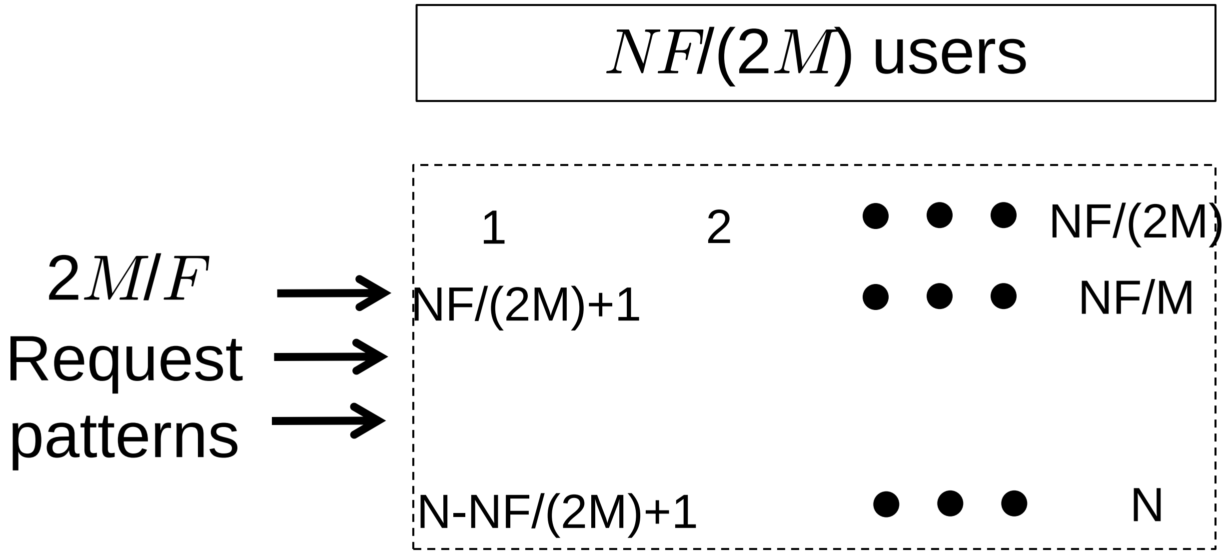


Critical to cache *quadratically* more content for larger files!

Outline

- Coded Caching under Arbitrary Popularity Distributions
  - System Model
  - Achievable Bounds and Intuitions
  - Lower Bounds
- Coded Caching under Distinct File-Sizes
  - System Model
  - Achievable Bounds and Intuitions
  - *Lower Bounds*
- Conclusion and Discussions

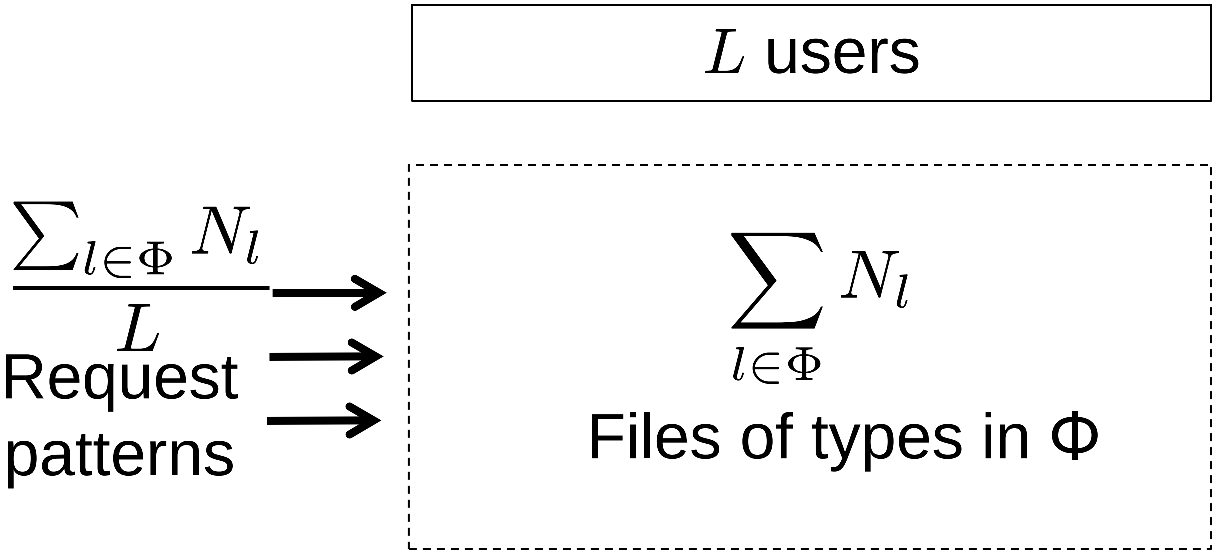
Lower Bound: Uniform File-size



- In order to serve all request patterns, we must have [Maddah-Ali & Niesen `14]

$$\frac{2M}{F}R^*(\mathbf{F}) + \frac{NF}{2M}M \geq NF \quad \Longrightarrow \quad R^*(\mathbf{F}) \geq \frac{NF^2}{4M} = \frac{F}{4q}$$

Lower Bound: First Try



- In order to serve all request patterns, we must have

$$\left\lceil \frac{\sum_{l\in\Phi} N_l}{L} \right\rceil \cdot R^*(\mathbb{F}) + L \cdot M \geq \sum_{l\in\Phi} N_l F_l$$

- Maximizing over L and Φ

$$R^*(\mathbb{F}) \geq \max_{\Phi} \frac{\left(\sum_{l\in\Phi} N_l F_l\right)^2}{4M \sum_{l\in\Phi} N_l} \qquad \textbf{(LB1)}$$

LB1 vs UB2

Lower Bound 1 (LB1)

$$R \geq R_{LB1} = \max_{\Phi} \frac{\left(\sum_{l\in\Phi} N_l\, F_l\right)^2}{4M \sum_{l\in\Phi} N_l}$$

- Consider $N_{l+1} = 4N_l$

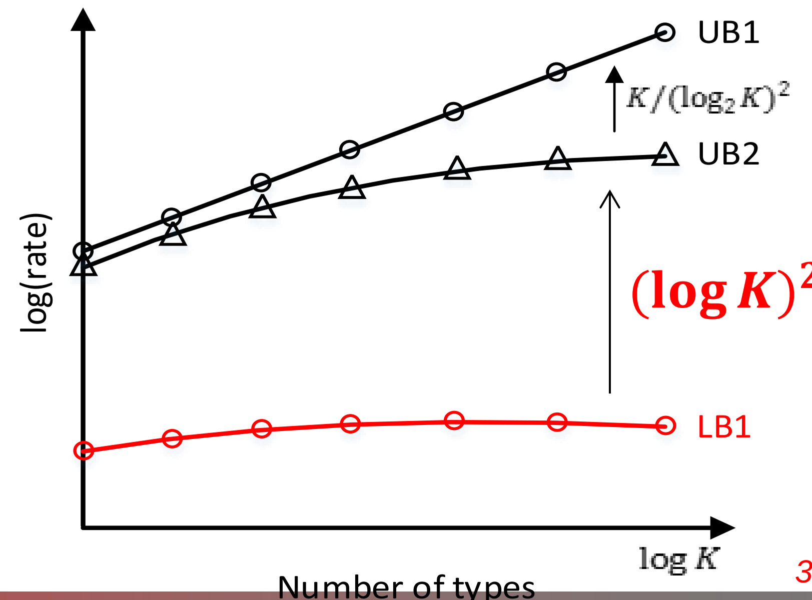
$$R_{LB1} \cdot \frac{N_1 F_1^2}{4M}$$

LB1 fails to account for heterogeneous caching probabilities!

Achievable Bound 2 (UB2)

$$R \cdot \quad R_{UB2} \quad = \quad (\bar{T}+1)\frac{\sum_{l=1}^{\bar{T}} N_l F_l^2}{M}$$

$$R_{UB2} \quad \geq \quad (\bar{T}+1)\bar{T}\frac{N_1 F_1^2}{M}$$



An Improved Lower Bound (LB2)

Proposition 3: Under

- Assumption 1: $\frac{2M}{F_1}$ and $\frac{N_l F_l}{2M}$ are integers for all $1 \le l \le \bar{T}$
- Assumption 2: $\sum_{l=1}^{\bar{T}} \frac{N_l F_l}{2M} \le K$

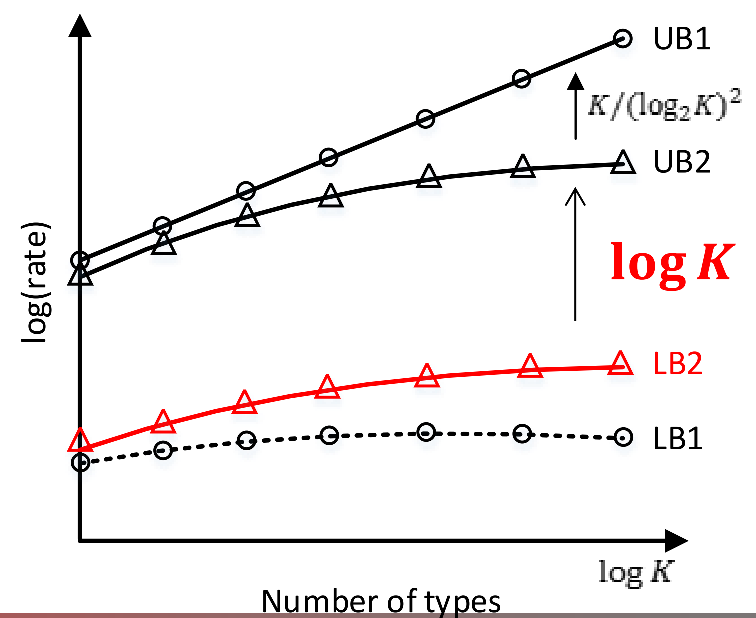
We must have

$$R^*(\mathbb{F}) \geq \sum_{l=1}^{\bar{T}} \frac{N_l F_l^2}{4M}$$

(LB2)

Compared with UB2:

$$R_{UB2} \quad \cdot \quad (\bar{T}+1)\frac{\sum_{l=1}^{\bar{T}} N_l F_l^2}{M}$$



Intuition for LB2 (Two types: $F_2 = F_1/2$)

$$R^*(\mathbb{F}) \ge \sum_{l=1}^{\bar{T}} \frac{N_l F_l^2}{4M} = \frac{N_1F_1^2}{4M} + \frac{N_2F_2^2}{4M} \qquad \textbf{(LB2)}$$

- Assumption 2 ensures that $s_1+s_2 = \frac{N_1F_1}{2M} + \frac{N_2F_2}{2M} \le K$

$s_1 = \frac{N_1F_1}{2M}$ users (U_1)

$s_2 = \frac{N_2F_2}{2M}$ users (U_2)

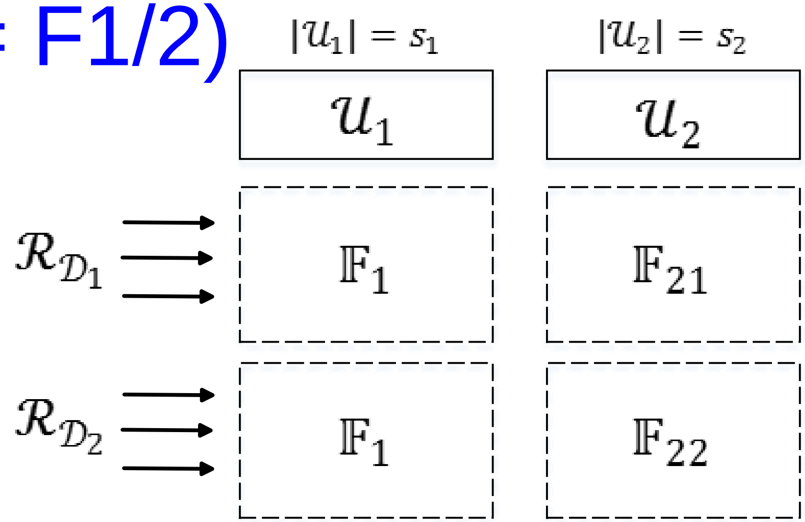
$\frac{N_1}{s_1} = \frac{2M}{F_1}$ patterns $\rightarrow$ N_1 files of type 1 (F_1)

$\frac{N_1}{s_1} = \frac{2M}{F_1}$ patterns $\rightarrow$ N_1 files of type 1 (F_1)

$$R^*(\mathbb{F}) \ge \frac{N_1F_1^2}{4M}$$

N_2 files of type 2 (F_2)

Proof: (Two types: F2 = F1/2)



- $s_1 = \frac{N_1 F_1}{2M}, \quad s_2 = \frac{N_2 F_2}{2M}$
- $H(\mathcal{M}_1) = 0.5H(\mathbb{F}_1) = \frac{N_1 F_1}{2}$
- $H(\mathcal{M}_2) = 0.5H(\mathbb{F}_2) = \frac{N_2 F_2}{2}$

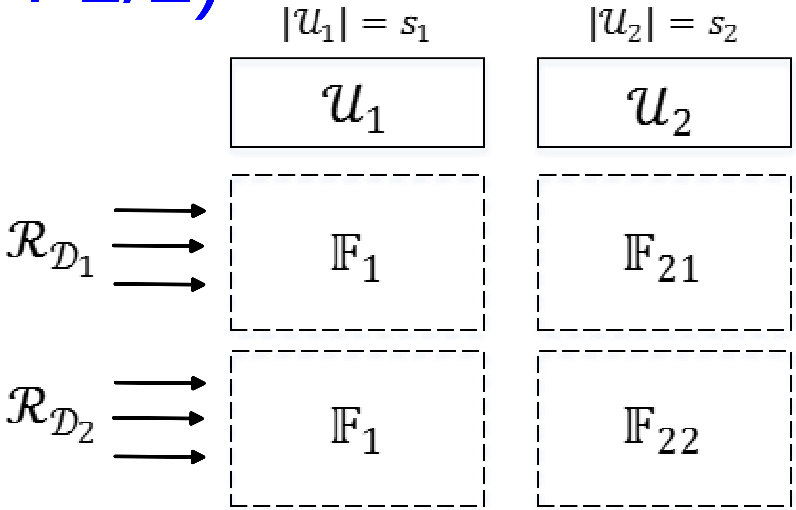
$$\begin{aligned}
\frac{2N_1}{s_1} R^*(\mathbb{F}) &\geq H(\mathcal{R}_{\mathcal{D}_1}) + H(\mathcal{R}_{\mathcal{D}_2}) \\
&= H(\mathcal{R}_{\mathcal{D}_1}|\mathbb{F}_1) + I(\mathcal{R}_{\mathcal{D}_1};\mathbb{F}_1) + H(\mathcal{R}_{\mathcal{D}_2}|\mathbb{F}_1) + I(\mathcal{R}_{\mathcal{D}_2};\mathbb{F}_1) \\
&\geq H(\mathcal{R}_{\mathcal{D}_1} \cup \mathcal{R}_{\mathcal{D}_2}|\mathbb{F}_1) + I(\mathcal{R}_{\mathcal{D}_1};\mathbb{F}_1) + I(\mathcal{R}_{\mathcal{D}_2};\mathbb{F}_1) \\
&= I(\mathcal{R}_{\mathcal{D}_1} \cup \mathcal{R}_{\mathcal{D}_2};\mathbb{F}_2|\mathbb{F}_1) + I(\mathcal{R}_{\mathcal{D}_1};\mathbb{F}_1) + I(\mathcal{R}_{\mathcal{D}_2};\mathbb{F}_1).
\end{aligned}$$

$$I(\mathcal{R}_{\mathcal{D}_1};\mathbb{F}_1) \geq \frac{N_1 F_1}{2}, \; I(\mathcal{R}_{\mathcal{D}_2};\mathbb{F}_1) \geq \frac{N_1 F_1}{2}$$

$$I(\mathcal{R}_{\mathcal{D}_1} \cup \mathcal{R}_{\mathcal{D}_2};\mathbb{F}_2|\mathbb{F}_1) \geq \frac{N_2 F_2}{2}$$

Proof: (Two types: F2 = F1/2)

$$s_1 = \frac{N_1 F_1}{2M}, \quad s_2 = \frac{N_2 F_2}{2M}$$



- From $$\frac{2N_1}{s_1} R^*(\mathbb{F}) \geq N_1 F_1 + \frac{N_2 F_2}{2}$$

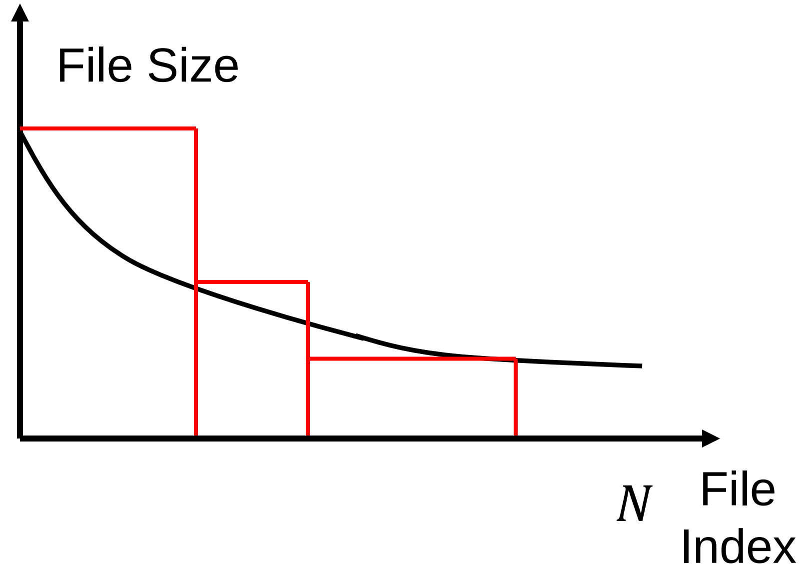
$$\Rightarrow \quad \frac{4M}{F_1} R^*(\mathbb{F}) \geq N_1 F_1 + \frac{N_2 F_2}{2}$$

$$\Rightarrow \quad R^*(\mathbb{F}) \geq \frac{N_1 F_1^2}{4M} + \frac{N_2 F_2 F_1}{8M}$$

$$\Rightarrow \quad R^*(\mathbb{F}) \geq \frac{N_1 F_1^2}{4M} + \frac{N_2 F_2^2}{4M}$$

(LB2)

Beyond Power-of-2



- Original file-set:
  $\mathbb{F} = \{\mathcal{F}_1, \ldots, \mathcal{F}_N\}, |\mathcal{F}_i| = F_i$

- Upper-quantized version

$$\mathbb{F}^{UB} = \left\{\mathcal{F}_i^{UB} | F_i^{UB} = F_1 \cdot 2^{-\left\lfloor \log_2 \frac{F_1}{F_i} \right\rfloor}\right\}$$

- Lower-quantized version

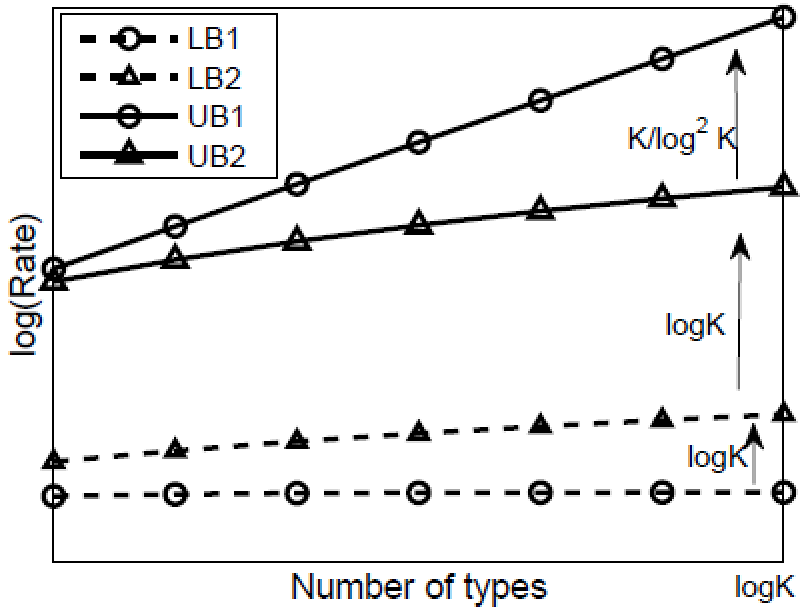
$$\mathbb{F}^{LB} = \left\{\mathcal{F}_i^{LB} | F_i^{LB} = F_1 \cdot 2^{-\left\lfloor \log_2 \frac{F_1}{F_i} \right\rfloor - 1}\right\}$$

- $R^*(\mathbb{F}^{LB}) \leq R^*(\mathbb{F}) \leq R^*(\mathbb{F}^{UB})$

- Under certain conditions (Assumption 2), $R^*(\mathbb{F}^{LB})$ is in a constant gap with $R^*(\mathbb{F}^{UB})$

Comparisons

•With Assumptions 1 & 2



Rate	LB2/LB1	Gain	UB1/UB2	Gain
2 types	4/3.6	11%	15.3757/13.2523	16%
3 types	5/3.8462	30%	21.2593/16.2196	31%

General Result (without Assumptions 1 and 2):

$$R_{UB2} \leq (32\log_2 K + 22) R_{LB2}$$

Conclusions

- Heterogeneous ***popularity***:
  - A simple ***threshold-based*** policy (similar to [Ji et al ’14]): files above a popularity threshold are cached uniformly
  - We show ***constant-factor*** gap that is ***independent*** of the popularity distribution
- Heterogeneous ***file-sizes***
  - ***Quadratically*** more content is cached for larger files
  - We show ***logarithmic-factor*** gap
- While the new achievable schemes are quite intuitive, the corresponding ***lower bounds*** more involved and reveal useful insights

Potential Future Directions

- Refinements:
  - Combining heterogeneous popularity and file-sizes?
  - Reduce the logarithmic factor?
- Other aspects of heterogeneity:
  - Heterogeneous cache size?
  - The role of cache location?
- Wireless environments (HetNet):
  - Is coded caching still helpful?
- Practical considerations:
  - Low-complexity coding/transmission schemes
  - Real-time transmissions [Niesen and Maddah-Ali `15]

Thank you!

- J. Zhang, X. Lin and X. Wang, “Coded Caching under Arbitrary Popularity Distributions," in *ITA Workshop,* UCSD, February 2015.
- J. Zhang, X. Lin, C.-C. Wang and X. Wang , “Coded Caching for Files with Distinct File Sizes,” in ISIT, June 2015 (to appear).