Coded Caching with Linear Coded Placement: Exact Tradeoff for the Three User Case

Yinbin Ma and Daniela Tuninetti

University of Illinois Chicago
Work supported in part by NSF Award 1910309

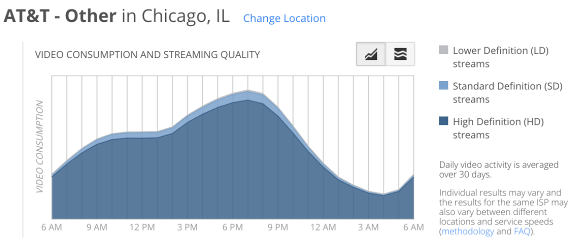
Sept 28, 2023

Networks Information Communications and Engineering Systems Laboratory

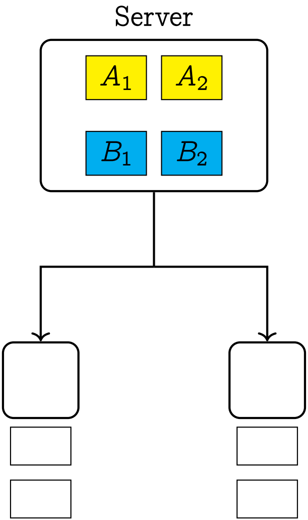
Table of Contents

What is caching?

- Caching: store information locally so as to lighten network traffic load at peak times.
- Cache content:
  - files created ahead of demands (such as videos),
  - distribution of demands is predictable,
  - cache content updated while the network traffic is light.
- Benefit: smooth network traffic during peak times.

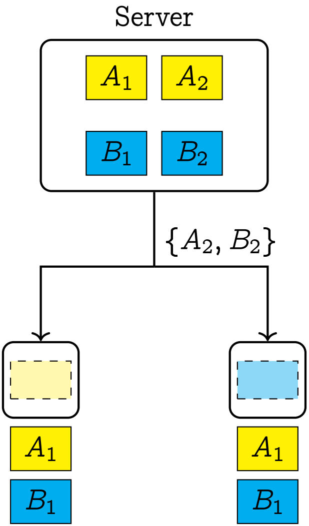


Example



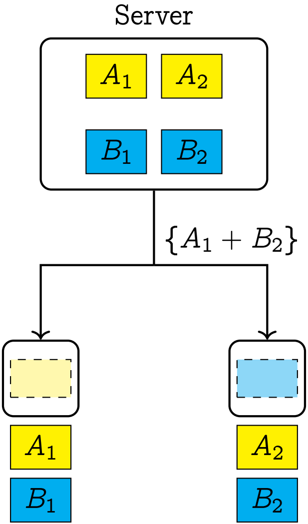
- A central server stores 2 files, A and B.
- An error-free shared link connects the server to 2 users.
- Each user can cache 1 file, and demands a single file from server.

Example



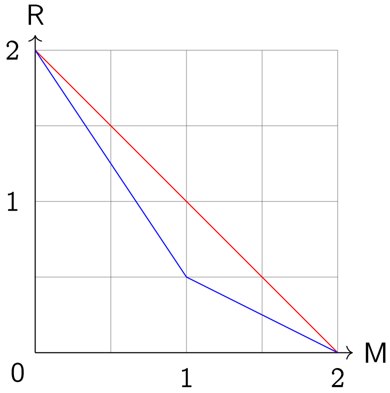
- Users cache pieces of files.
- Once demands are known, the server sends coded messages.
- Example: for d $= [A, B]$, if $X = (A_2, B_2)$ the load is 1.
- This is an *uncoded scheme*, as the X isn't coded.

Example



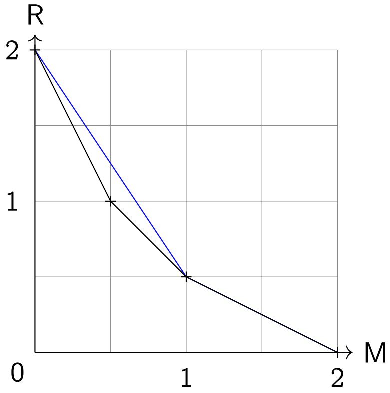
- The first user caches (A_1, B_1) and the second users caches (A_2, B_2).
- Example: for $\mathrm{d} = [A, B]$, if $X = (A_2 + B_1)$ the load is $1/2$.
- User 1 knows B_1, and decodes A_1 from X, thus it can restore A.
- User 2 knows A_2, and decodes B_1 from X, thus it can restore B.
- *Coded deliver* has smaller load than uncoded delivery.

Memory-load Tradeoff Plot for 2 Users and 2 Files



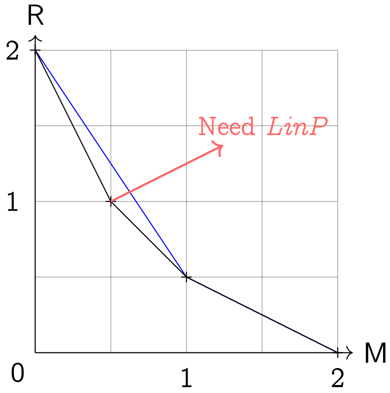
- Red: uncoded scheme.
- Blue: coded scheme with *uncoded placement* [MAN14].

Memory-load Tradeoff Plot for 2 Users and 2 Files



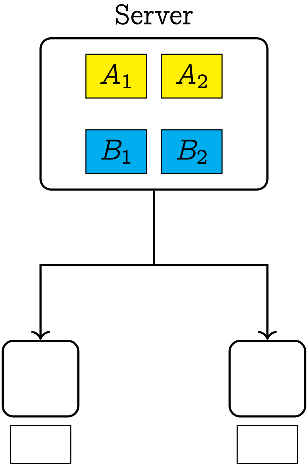
- Blue: coded scheme with *uncoded placement* [MAN14].
- Black: converse [MAN14].

Memory-load Tradeoff Plot for 2 Users and 2 Files



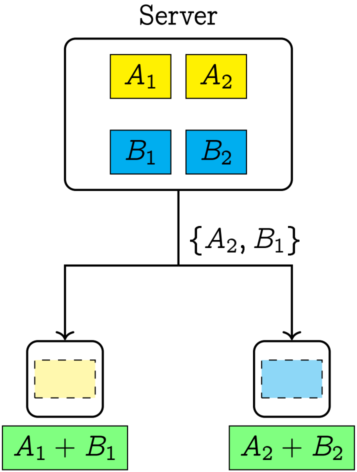
- Blue: coded scheme with *uncoded placement* [MAN14].
- Black: converse [MAN14].
  - [MAN14] shows that $(\mathsf{M}, \mathsf{R}) = (0.5, 1)$ is achievable by *linear coding placement* (LinP).

Example of Linear Placement



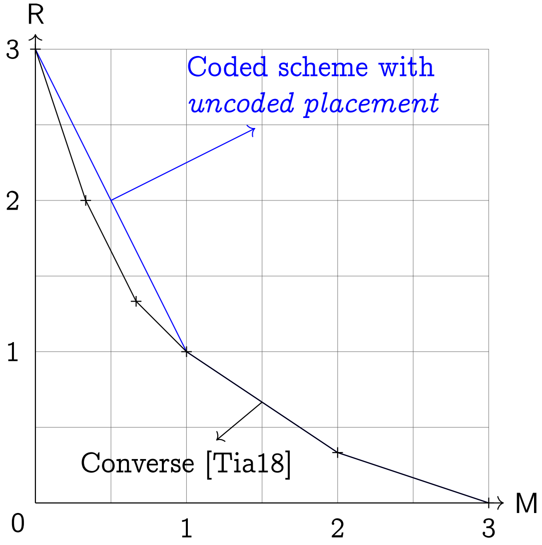
- A central server stores 2 files, A and B.
- An error-free shared link connects the server to 2 users.
- Each user can cache $1/2$ file, and demands a single file from server.

Example of Linear Placement

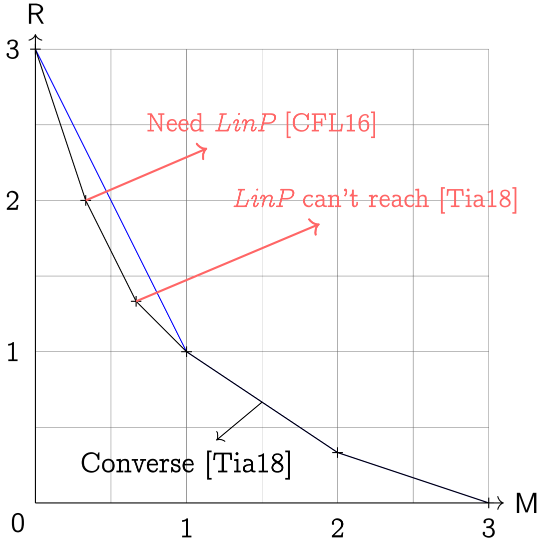


- The first user caches $A_1 + B_1$ and the second users caches $A_2 + B_2$.
- Example: for d $= [A, B]$, if $X = (A_2, B_1)$ the load is 1.
- User 1 can decode A_1 from X, thus it can restore A.
- User 2 can decode B_2 from X, thus it can restore B.
- The placement *coded*.

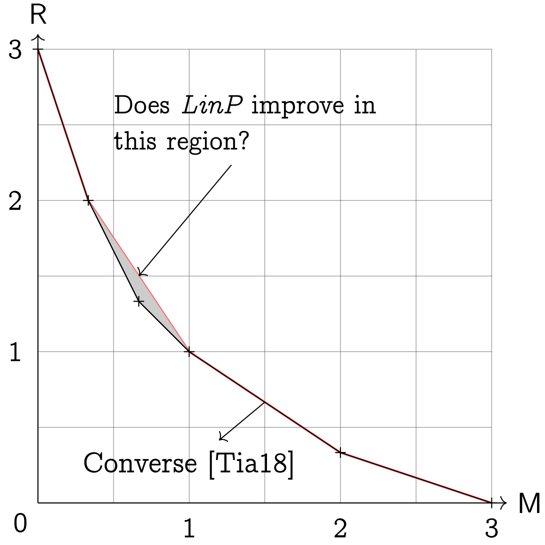
Memory-load Tradeoff Plot for 3 Users and 3 Files



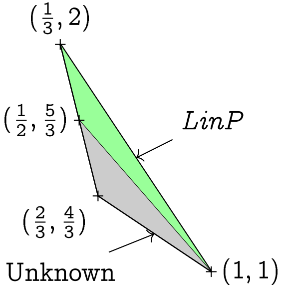
Memory-load Tradeoff Plot for 3 Users and 3 Files



Memory-load Tradeoff Plot for 3 Users and 3 Files

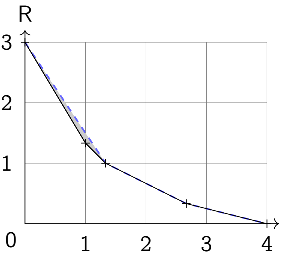


Contributions



- For $\mathsf{N} = \mathsf{K} = 3$, a new optimal point $(\mathsf{M}, \mathsf{R}) = (\frac{1}{2}, \frac{5}{3})$ is found.
- From the capacity of the linear computation broadcast channel with 3 users [YJ22], we derive a converse bound for our coded caching under LinP.
- The gray region for $\mathsf{M} \in [\frac{1}{2}, 1]$ is still open. Only a non-linear coded placement could possibly beat the optimal tradeoff we characterized under LinP.

Contributions



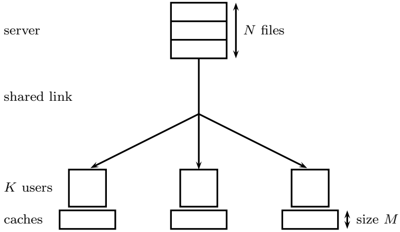
- For $N > K = 3$, uncoded placement is optimal under LinP.
- The optimal placement remains open for $N \in \{4, 5\}$ when $M < N/K$. Converse is from [YMAA18].
- When $N \geq 6$, uncoded placement is optimal [YMAA18].

Questions?

Table of Contents

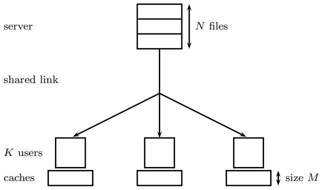
(N, K) Coded Caching Model

- N files with i.i.d uniform B symbols over a finite field.
- An error-free shared link to K cache-aided users.
- Placement: The cache of user $k \in [\mathsf{K}]$, Z_k, can store up to MB symbols and is a function of library.



(N, K) Coded Caching Model

- N files; K users with cache of size MB symbols.
- Delivery:
  - User $k \in [\mathsf{K}]$ requests file $d_k \in [\mathsf{N}]$.
  - The server broadcasts $X(F_{[\mathsf{N}]}, d_{[\mathsf{K}]})$ of size RB symbols.
  - User $k \in [\mathsf{K}]$ must decodes F_{d_k} from X and Z_k.
- Goal: worst-case load



$$\mathsf{R}^\star(\mathsf{M}) = \limsup_{\mathsf{B}\to\infty} \min_{Z_{[\mathsf{K}]},X} \max_{d_{[\mathsf{K}]}} \{\mathsf{R} : \text{all above conditions are satisfied with memory size } \mathsf{M}\}, \forall \mathsf{M} \in [0, \mathsf{N}].$$

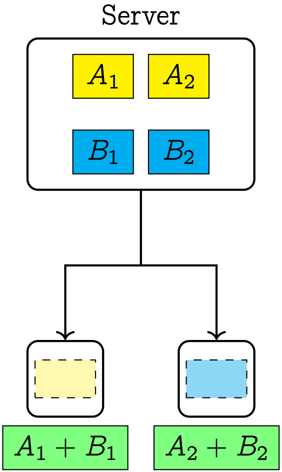
Linear Coding Placement (LinP)

- Let $F = [F_1; \ldots; F_{\mathsf{N}}] \in \mathbb{F}_{\mathsf{q}}^{\mathsf{NB}}$. The cache encoding matrix for user k is $\tilde{\mathrm{E}}_k \in \mathbb{F}_{\mathsf{q}}^{\mathsf{MB}\times\mathsf{NB}}$, i.e.,

$$Z_k = \tilde{\mathrm{E}}_k F \in \mathbb{F}_{\mathsf{q}}^{\mathsf{MB}}.$$

  Note: X need not be linear.
- For example,

$$\begin{aligned} A &= ([1,0] \otimes \mathrm{I}_{\mathsf{B}})F, \\ B &= ([0,1] \otimes \mathrm{I}_{\mathsf{B}})F, \\ Z_1 &= A_1 + B_1 = ([1,0,1,0] \otimes \mathrm{I}_{\mathsf{B}/2})F, \\ Z_2 &= A_2 + B_2 = ([0,1,0,1] \otimes \mathrm{I}_{\mathsf{B}/2})F. \end{aligned}$$



Linear Coding Placement (LinP)

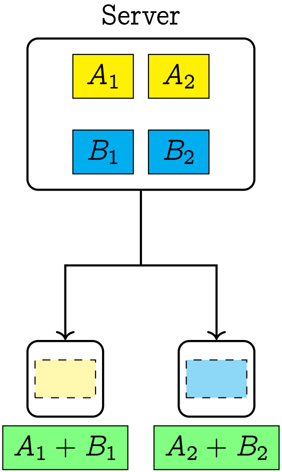
- Let $F = [F_1; \ldots; F_\mathsf{N}] \in \mathbb{F}_\mathsf{q}^{\mathsf{NB}}$. The cache encoding matrix for user k is $\tilde{\mathrm{E}}_k \in \mathbb{F}_\mathsf{q}^{\mathsf{MB}\times\mathsf{NB}}$, i.e.,

$$Z_k = \tilde{\mathrm{E}}_k F \in \mathbb{F}_\mathsf{q}^{\mathsf{MB}}.$$

  Note: X need not be linear.
- Trivially,

$$0.5\mathsf{R}^\star_{\text{Uncoded}} \overset{(a)}{\leq} \mathsf{R}^\star \leq \mathsf{R}^\star_{\text{LinP}} \leq \mathsf{R}^\star_{\text{Uncoded}},$$

  where (a) is proved in [YMAA18].



Uncoded Scheme

Theorem (YMA Scheme [YMAA17])

The lower convex envelop of the following points is achievable

$$(\mathsf{M}_t, \mathsf{R}_t)_{\mathrm{YMA}} = \left(\mathsf{N}\frac{\binom{\mathsf{K}-1}{t-1}}{\binom{\mathsf{K}}{t}}, \frac{\binom{\mathsf{K}}{t+1} - \binom{\mathsf{K}-\min(\mathsf{K},\mathsf{N})}{t+1}}{\binom{\mathsf{K}}{t}} \right), \quad t \in [0:\mathsf{K}].$$

When $\mathsf{N} \geq \mathsf{K}$, it reduces to MAN scheme [MAN14],

$$(\mathsf{M}_t, \mathsf{R}_t)_{\mathrm{MAN}} = \left(\mathsf{N}\frac{t}{\mathsf{K}}, \frac{\mathsf{K}-t}{1+t} \right), \quad t \in [0:\mathsf{K}].$$

Furthermore, $\mathsf{R}_{\mathrm{YMA}} = \mathsf{R}^{\star}_{\mathrm{Uncoded}}$.

Table of Contents

Linear Computation Broadcast Channel ($LCBC$)

A $(\mathsf{q}, \mathsf{r}, \mathsf{K}, \mathrm{E}_{[\mathsf{K}]}, \mathrm{D}_{[\mathsf{K}]})$ LCBC model is as follows.

- A server has $X \in \mathbb{F}_{\mathsf{q}}^{\mathsf{r}}$ uniformly and independently distributed data blocks from $\mathbb{F}_{\mathsf{q}}$ and serves K users.
- For every user j, denote the "cache projection matrix" as $\mathrm{E}_j \in \mathbb{F}_{\mathsf{q}}^{m_j \times \mathsf{r}}$, and the "demand projection matrix" as $\mathrm{D}_j \in \mathbb{F}_{\mathsf{q}}^{n_j \times \mathsf{r}}$, where $m_j, n_j \geq 0$.
- Server sends $\Psi_0(X) \in \mathbb{F}_{\mathsf{q}}^{\Delta}$ to the users.
- User $j \in [\mathsf{K}]$ decodes $y_j := \Psi_j(\Psi_0(X), \mathrm{E}_j X)$ such that $H(\mathrm{D}_j X | y_j) = 0$.
- $\Delta^{\star}(\mathrm{E}_{[\mathsf{K}]}, \mathrm{D}_{[\mathsf{K}]})$ is the smallest Δ to meet all requirements.

LCBC provides the following lower bound for coded caching

$$\mathsf{BR}^{\star}_{\mathrm{LinP}} \geq \min_{\mathrm{E}_{[\mathsf{K}]}} \max_{\mathrm{D}_{[\mathsf{K}]}: \mathrm{D}_j = d_j \otimes I_{\mathsf{B}}} \Delta^{\star}(\mathrm{E}_{[\mathsf{K}]}, \mathrm{D}_{[\mathsf{K}]}).$$

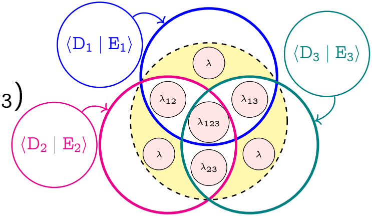
Exact Tradeoff for 3 Users on LCBC

Theorem (LCBC [YJ22])

For $\mathsf{K} = 3$, given $\mathrm{E}_{[3]}$ and $\mathrm{D}_{[3]}$

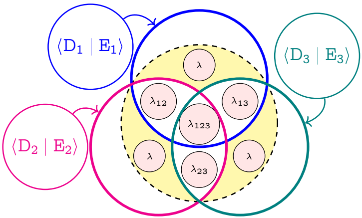
$$\begin{aligned}\Delta^\star = &\, \mathsf{rk}(\mathrm{D}_1 \mid \mathrm{E}_1) + \mathsf{rk}(\mathrm{D}_2 \mid \mathrm{E}_2) + \mathsf{rk}(\mathrm{D}_3 \mid \mathrm{E}_3) \\ &- \max_{\lambda_{(\cdot)}} \{2\lambda_{123} + \lambda_{12} + \lambda_{13} + \lambda_{23} + \lambda\},\end{aligned}$$

where $\lambda_{(\cdot)}$ satisfy some constraints [YJ22].



Exact Tradeoff for 3 Users on LCBC

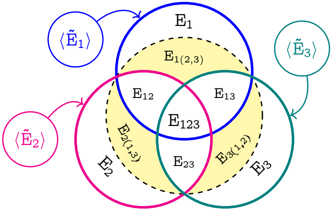
- "(Uncoded load) - (LinP gain)".
- λ_{123} benefits all users, reduces the load by $2\lambda_{123}$.
- λ_{ij} benefits i and j, reduces the load by λ_{ij}.
- Yellow regions are mutually disjoint but two of them contain the remaining one, reduce the load by λ.



$$\Delta^\star = \mathsf{rk}(\mathrm{D}_1 \mid \mathrm{E}_1) + \mathsf{rk}(\mathrm{D}_2 \mid \mathrm{E}_2) + \mathsf{rk}(\mathrm{D}_3 \mid \mathrm{E}_3) - \max_{\lambda_{(\cdot)}}\{2\lambda_{123} + \lambda_{12} + \lambda_{13} + \lambda_{23} + \lambda\}.$$

Design of Cache Encoding Matrix

We partition every $\tilde{\mathrm{E}}_j$ as shows in the right figure.
Let $i \in [3], \{j,\ell\} = [3] \setminus \{i\}$, and $\mathcal{S} \subseteq [3]$,



$$\mathrm{E}_{\mathcal{S}} = \begin{bmatrix} \mathrm{P}^{\mathcal{S}}_{\{1\},1} & 0 & 0 \\ 0 & \mathrm{P}^{\mathcal{S}}_{\{2\},2} & 0 \\ 0 & 0 & \mathrm{P}^{\mathcal{S}}_{\{3\},3} \\ \mathrm{P}^{\mathcal{S}}_{\{1,2\},1} & \mathrm{P}^{\mathcal{S}}_{\{1,2\},2} & 0 \\ \mathrm{P}^{\mathcal{S}}_{\{1,3\},1} & 0 & \mathrm{P}^{\mathcal{S}}_{\{1,3\},3} \\ 0 & \mathrm{P}^{\mathcal{S}}_{\{2,3\},2} & \mathrm{P}^{\mathcal{S}}_{\{2,3\},3} \\ \mathrm{P}^{\mathcal{S}}_{\{1,2,3\},1} & \mathrm{P}^{\mathcal{S}}_{\{1,2,3\},2} & \mathrm{P}^{\mathcal{S}}_{\{1,2,3\},3} \end{bmatrix}, \quad \mathrm{E}_{i(j,\ell)} = \begin{bmatrix} \mathrm{Q}^{i}_{\{1\},1} & 0 & 0 \\ 0 & \mathrm{Q}^{i}_{\{2\},2} & 0 \\ 0 & 0 & \mathrm{Q}^{i}_{\{3\},3} \\ \mathrm{Q}^{i}_{\{1,2\},1} & \mathrm{Q}^{i}_{\{1,2\},2} & 0 \\ \mathrm{Q}^{i}_{\{1,3\},1} & 0 & \mathrm{Q}^{i}_{\{1,3\},3} \\ 0 & \mathrm{Q}^{\mathcal{S}}_{\{2,3\},2} & \mathrm{Q}^{i}_{\{2,3\},3} \\ \mathrm{Q}^{i}_{\{1,2,3\},1} & \mathrm{Q}^{i}_{\{1,2,3\},2} & \mathrm{Q}^{i}_{\{1,2,3\},3} \end{bmatrix}.$$

$(\cdot)^{(\cdot)}_{\mathcal{T},n}$: linear encoding matrix involving $\mathcal{T}$ files for the n^{th} file.

Design of Cache Encoding Matrix

$$\mathrm{E}_{\mathcal{S}} = \begin{bmatrix} \mathrm{P}^{\mathcal{S}}_{\{1\},1} & 0 & 0 \\ 0 & \mathrm{P}^{\mathcal{S}}_{\{2\},2} & 0 \\ 0 & 0 & \mathrm{P}^{\mathcal{S}}_{\{3\},3} \\ \mathrm{P}^{\mathcal{S}}_{\{1,2\},1} & \mathrm{P}^{\mathcal{S}}_{\{1,2\},2} & 0 \\ \mathrm{P}^{\mathcal{S}}_{\{1,3\},1} & 0 & \mathrm{P}^{\mathcal{S}}_{\{1,3\},3} \\ 0 & \mathrm{P}^{\mathcal{S}}_{\{2,3\},2} & \mathrm{P}^{\mathcal{S}}_{\{2,3\},3} \\ \mathrm{P}^{\mathcal{S}}_{\{1,2,3\},1} & \mathrm{P}^{\mathcal{S}}_{\{1,2,3\},2} & \mathrm{P}^{\mathcal{S}}_{\{1,2,3\},3} \end{bmatrix}, \quad \mathrm{E}_{i(j,\ell)} = \begin{bmatrix} \mathrm{Q}^{i}_{\{1\},1} & 0 & 0 \\ 0 & \mathrm{Q}^{i}_{\{2\},2} & 0 \\ 0 & 0 & \mathrm{Q}^{i}_{\{3\},3} \\ \mathrm{Q}^{i}_{\{1,2\},1} & \mathrm{Q}^{i}_{\{1,2\},2} & 0 \\ \mathrm{Q}^{i}_{\{1,3\},1} & 0 & \mathrm{Q}^{i}_{\{1,3\},3} \\ 0 & \mathrm{Q}^{\mathcal{S}}_{\{2,3\},2} & \mathrm{Q}^{i}_{\{2,3\},3} \\ \mathrm{Q}^{i}_{\{1,2,3\},1} & \mathrm{Q}^{i}_{\{1,2,3\},2} & \mathrm{Q}^{i}_{\{1,2,3\},3} \end{bmatrix}.$$

$(\cdot)^{(\cdot)}_{\mathcal{T},n}$: linear encoding matrix involving $\mathcal{T}$ files for the n^{th} file.
The rank of $\mathrm{P}^{\mathcal{S}}_{\mathcal{T},n}$ and $\mathrm{Q}^{i}_{\mathcal{T},n}$ are, WLOG by symmetry

$$\mathsf{rk}(\mathrm{P}^{\mathcal{S}}_{\mathcal{T},n}) = r_{a,b}\mathsf{B}, \quad a = |\mathcal{T}|,\ b = |\mathcal{S}|.$$
$$\mathsf{rk}(\mathrm{Q}^{i}_{\mathcal{T},n}) = q_c\mathsf{B}, \quad c = |\mathcal{T}|.$$

LP for $N \geq K = 3$

When $d = [1, 2, \ldots, K]$, the converse is the LP

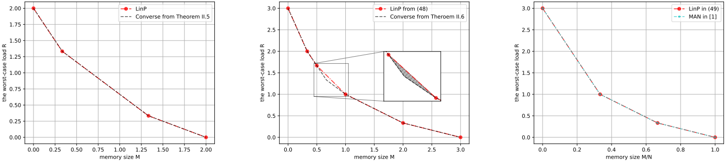
$$\min_{r,q \geq 0} \; 3 - 6r_{1,1} - 8r_{1,2} - 3r_{1,3} - 8r_{2,1} - 9r_{2,2} - 3r_{2,3} - 3r_{3,1} - 3r_{3,2} - r_{3,3} - 4.5q_1 - 6q_2.$$

subject to

$$\sum_{j=1}^{N} \binom{N}{j} (r_{1,j} + 2r_{2,j} + r_{3,j} + q_j) \leq M,$$

$$\sum_{j=1}^{N} \binom{N-1}{j-1} (3r_{1,j} + 3r_{2,j} + r_{3,j} + 2q_j) \leq 1.$$

LP Results



(a) $\mathsf{N} = 2$. (b) $\mathsf{N} = 3$. (c) $\mathsf{N} \geq 4$.

Figure: The memory-load tradeoff under LinP for $\mathsf{K} = 3$ and various N. Figure 1a shows LinP is optimal when $\mathsf{N} = 2$. Figure 1b shows a non-linear coding placement may be needed to close the gap in the grey region when $\mathsf{N} = 3$. Figure 1c shows MAN is optimal under LinP when $\mathsf{N} \geq 4$.

New Optimal Point $(\frac{1}{2}, \frac{5}{3})$ for $\mathrm{N} = \mathrm{K} = 3$

The LP solution shows $r_{1,2} = 1/6, \lambda_{ij} = \lambda = 1/3$.

- $\tilde{\mathrm{E}}_j$ is disjoint and involves exactly two files.
- $3\lambda_{12} + \lambda$ load saving compared to uncoded transmission.

Partition each file into 6 parts, and place

$$Z_1 = \begin{bmatrix} A_1 + B_1 \\ A_2 + C_1 \\ B_2 + C_2 \end{bmatrix}, \ Z_2 = \begin{bmatrix} A_3 + B_3 \\ A_4 + C_3 \\ B_4 + C_4 \end{bmatrix}, \ Z_3 = \begin{bmatrix} A_5 + B_5 \\ A_6 + C_5 \\ B_6 + C_6 \end{bmatrix}.$$

Assume $\mathrm{d} = [1, 2, 3]$, the server transmits

$$X = \begin{pmatrix} A_3, A_6, B_1, B_6, C_1, C_4, \\ C_2 - C_3, B_2 - B_5, A_4 - A_5, \\ B_2 + C_2 + A_4 + C_3 + A_5 + B_5 \end{pmatrix}.$$

New Optimal Point $(\frac{1}{2}, \frac{5}{3})$ for $\mathsf{N} = \mathsf{K} = 3$

Decoding: take User 1 as an example, and same for the others:

$$Z_1 = \begin{bmatrix} A_1 + B_1 \\ A_2 + C_1 \\ B_2 + C_2 \end{bmatrix}, X = \begin{pmatrix} \underline{A_3, A_6}, \underline{B_1}, B_6, \underline{C_1}, C_4, \\ C_2 - C_3, B_2 - B_5, A_4 - A_5, \\ B_2 + C_2 + A_4 + C_3 + A_5 + B_5 \end{pmatrix}.$$

Result:

- obtains A_3, A_6 directly, and extracts A_1, A_2 from its cache.

New Optimal Point $(\frac{1}{2}, \frac{5}{3})$ for $\mathsf{N} = \mathsf{K} = 3$

Decoding: take User 1 as an example, and same for the others:

$$Z_1 = \left[B_2 + C_2\right], X = \left(\frac{C_2 - C_3, B_2 - B_5}{B_2 + C_2 + A_4 + C_3}, \frac{A_4 - A_5,}{+ A_5 + B_5}\right).$$

Result:

- obtains A_3, A_6 directly, and extracts A_1, A_2 from its cache.

New Optimal Point $(\frac{1}{2}, \frac{5}{3})$ for N = K = 3

Decoding: take User 1 as an example, and same for the others:

$$Z_1 = \begin{bmatrix} B_2 + C_2 \\ B_5 + C_3 \end{bmatrix}, X = \begin{pmatrix} A_4 - A_5, \\ \underline{B_2 + C_2} + A_4 + \underline{C_3} + A_5 + \underline{B_5} \end{pmatrix}.$$

Result:

- obtains A_3, A_6 directly, and extracts A_1, A_2 from its cache.
- obtains A_4, A_5 by solving $\begin{bmatrix} A_4 - A_5 \\ A_4 + A_5 \end{bmatrix}$.

Table of Contents

Conclusions

- Our contributions:
  - We derived the exact memory-load tradeoff for $K = 3$ users under linear coding placement.
  - For $N = K = 3$, we discovered a novel optimal point.
  - For $N > K = 3$, we showed that MAN/uncoded placement is optimal under linear coding placement.
- Open problems:
  - Optimal placement for the small memory regime $KM/N < 1$ for $3 = K \leq N \leq 5$,
  - Derive the optimal tradeoff under linear coding placement for arbitrary (N, K).
- Extensions: a new optimal point for $N = K \geq 3$ for $M = 1/(N-1)$; submitted to ICC 2024.

The End

References I

Zhi Chen, Pingyi Fan, and Khaled Ben Letaief.
Fundamental limits of caching: Improved bounds for users with small buffers.
IET Communications, 10(17):2315–2318, 2016.

Mohammad Ali Maddah-Ali and Urs Niesen.
Fundamental limits of caching.
IEEE Transactions on Information Theory, 60(5):2856–2867, 2014.

Chao Tian.
Symmetry, outer bounds, and code constructions: A computer-aided investigation on the fundamental limits of caching.
Entropy, 20(8):603, 2018.

References II

Yuhang Yao and Syed A Jafar.
The capacity of 3 user linear computation broadcast.
arXiv preprint arXiv:2206.10049, 2022.

Qian Yu, Mohammad Ali Maddah-Ali, and A Salman Avestimehr.
The exact rate-memory tradeoff for caching with uncoded prefetching.
IEEE Transactions on Information Theory, 64(2):1281–1296, 2017.

References III

Qian Yu, Mohammad Ali Maddah-Ali, and A Salman Avestimehr.
Characterizing the rate-memory tradeoff in cache networks within a factor of 2.
IEEE Transactions on Information Theory, 65(1):647–663, 2018.