WIRELESS CODED CACHING: A PARADIGM SHIFT IN WIRELESS COMMUNICATIONS

PETROS ELIA

(EURECOM – FRANCE)

Intro

- This tutorial is about a novel use of caching in wireless communication networks

- Using on-board memory at the nodes:
  - NOT to reduce the volume/size of the problem
    - "Prefetch something today so that you don't have to send it tomorrow"
  - BUT to surgically alter the informational structure of networks
    - Use on-board memory to change the network to something faster, simpler, more efficient.

Outline

- Challenges of modern wireless communications
  - The need of a new technology

- Basic elements of coded caching
  - Basic properties
  - Main gains
  - Important variants
  - Main bottlenecks

Outline

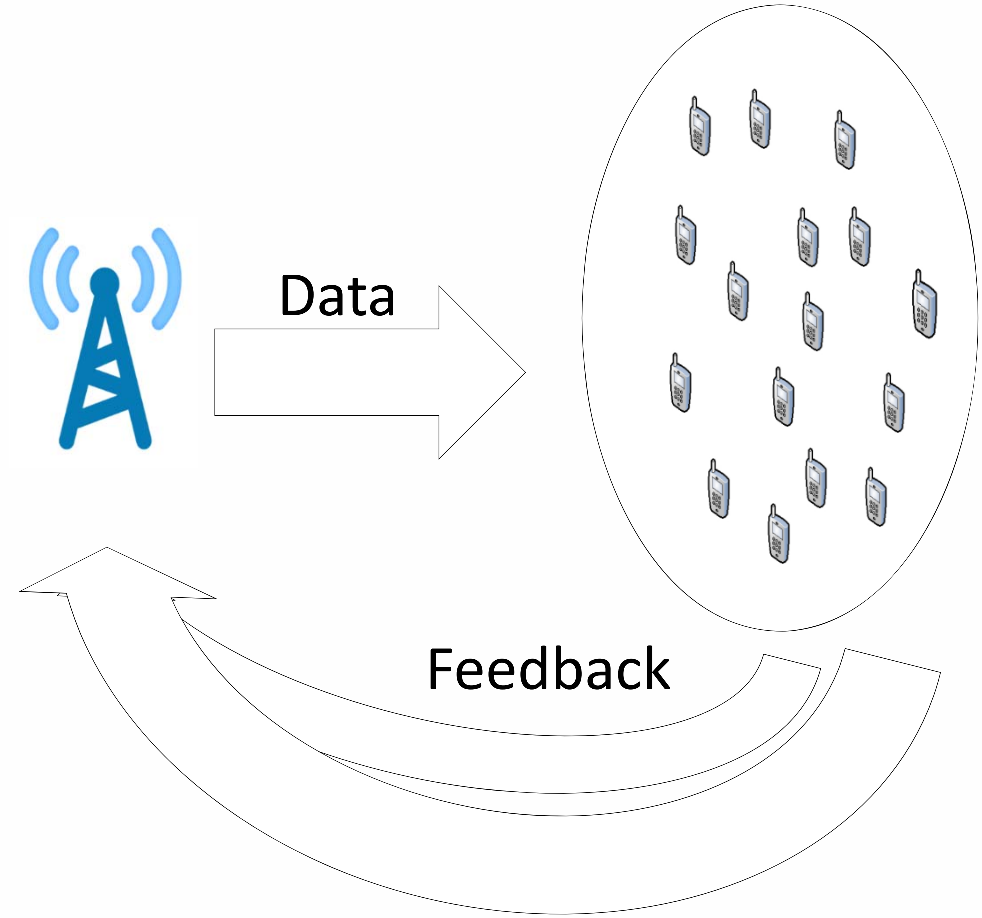
- Need to fuse coded caching with advanced PHY techniques
- Some differences between wired and wireless coded caching
- Coded caching in multi-user MIMO settings
- Coded-caching and feedback

- Coded caching in a variety of wireless networks
  - Femtocaching
  - Caching on the edge
  - Wireless multihop D2D caching networks

- Theoretical and practical open problems

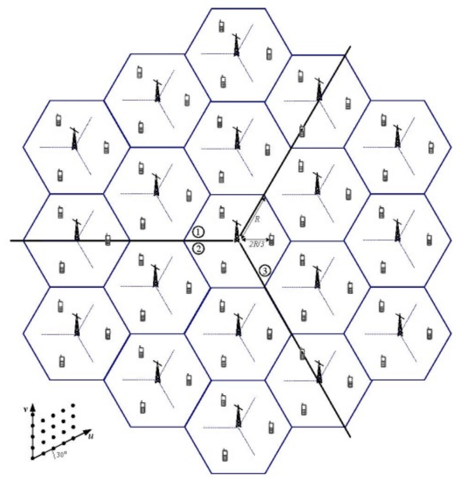
Limitations of Current Communications Paradigms

Main Challenges



- `Feedback': channel-state information (CSI)
  - ➢ The instantaneous strengths of each propagation-path between different nodes
- As K increases, the overhead consumes more and more resources
  - ➢ No room for actual data
  - ➢ Brings current systems and envisioned methods to a halt

Multi-cell Cooperation

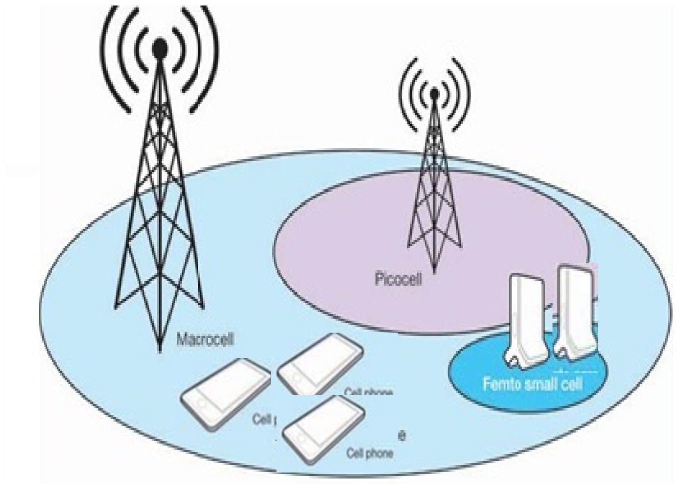


- Even full BS cooperation cannot handle interference
- Spectral efficiency upper bound that is independent of the transmit power
- Cooperation possible only within clusters of limited size (due to CSI)
  - subject to out-of-cluster interference with power similar to in-cluster signals

$$DOF \overset{\text{def}}{=} \frac{Per\ User\ Capacity}{\log SNR} \rightarrow 0 \ \text{ (as } K \text{ increases)}$$

source: Lozano et al. (2012)

Wireless Network Densification



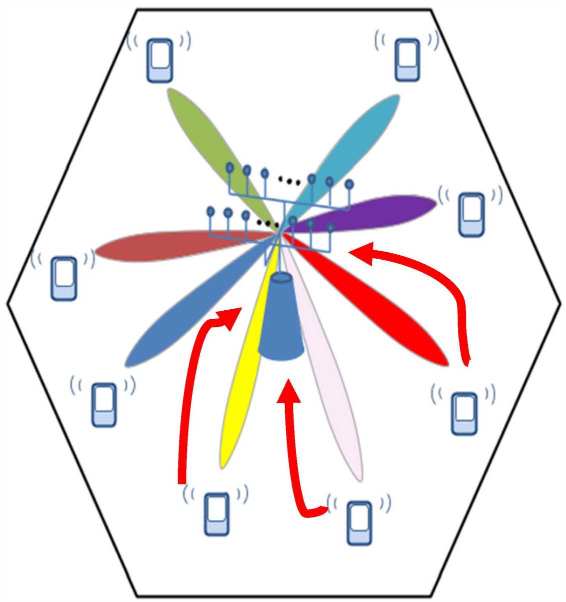
Deployment of more base-stations/access-points per unit area

- Short-range wireless channels different from classical cellular counterparts
  - Exhibit <u>path loss subduction</u> (reduced path loss exponent)
  - <u>Extreme fading</u> (more severe deep-fades)
- SINR decrease after certain densification threshold
- Similar trends are observed for the throughput

⇒ Disruption of densification gains

source: Andrews et al. (2015)

Massive MIMO and mm-Wave



Massive MIMO:

- Gains in spectral efficiency
- Gains reduced by expensive channel estimation
  - Pilot contamination*

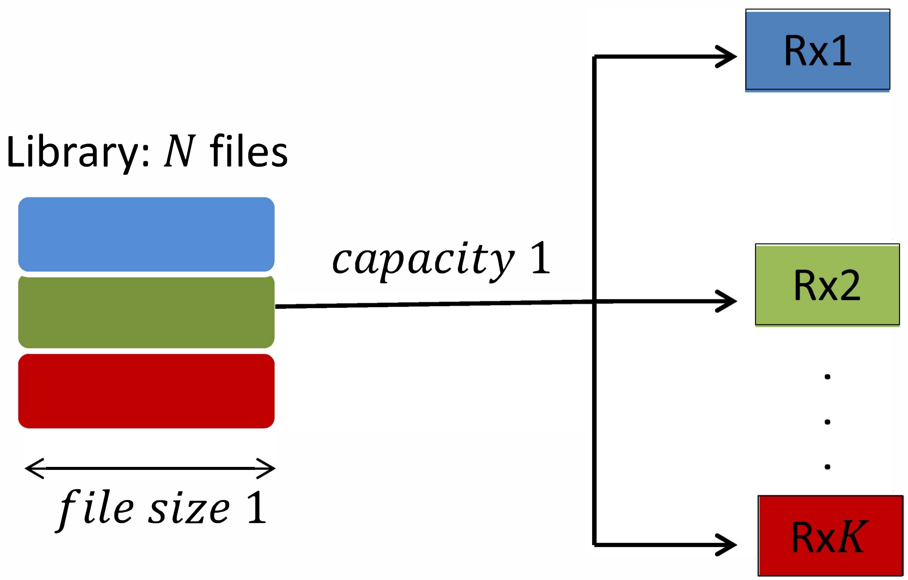
Mm-Wave Communications:

- High frequency results in sparse and easier to estimate channels
- Channels though can fluctuate between sparse and denser
  - think of AoD in urban settings
- Introduces FB delays/overhead
- Also directionality can create signal "holes" for users**

*source: Marzetta (2010)
**source: Rappaport et al. (2014)

Simple Caching

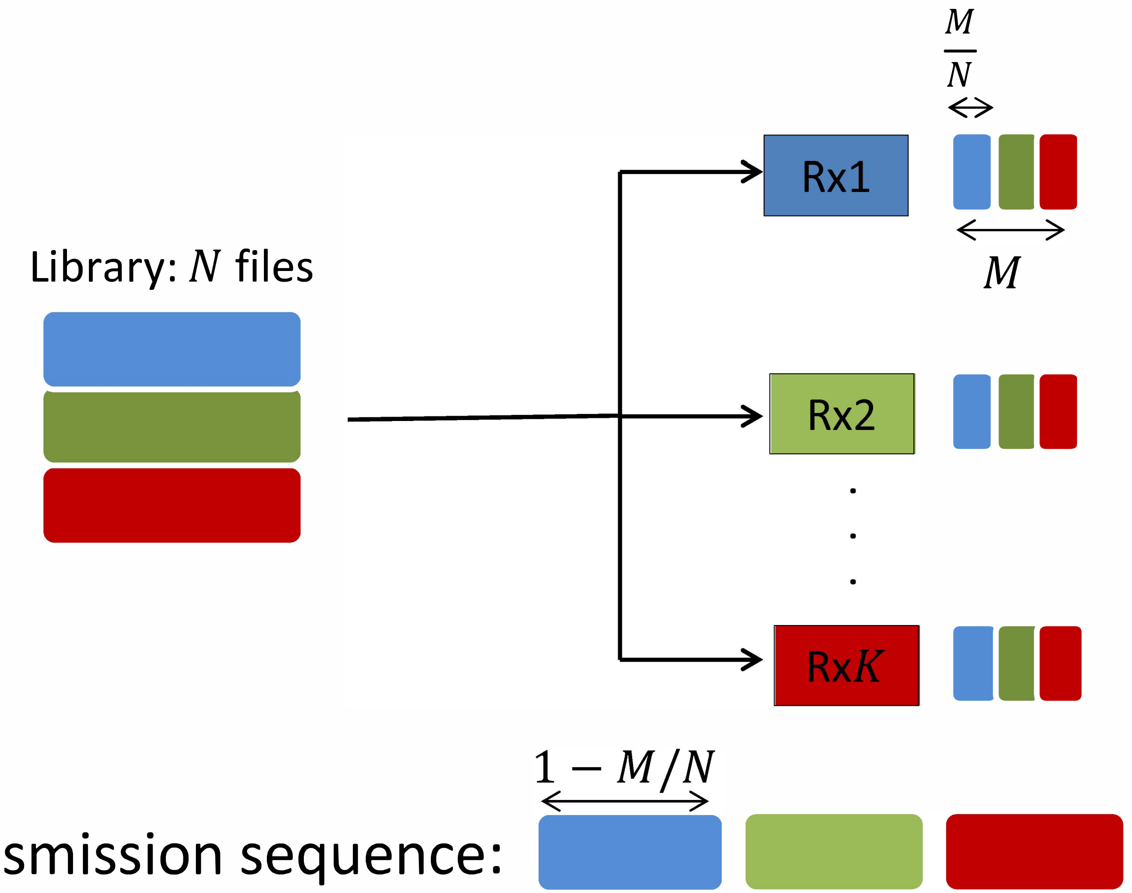
Single stream channel: No caching ($M = 0$)



- Transmission sequence:

$$T = K$$

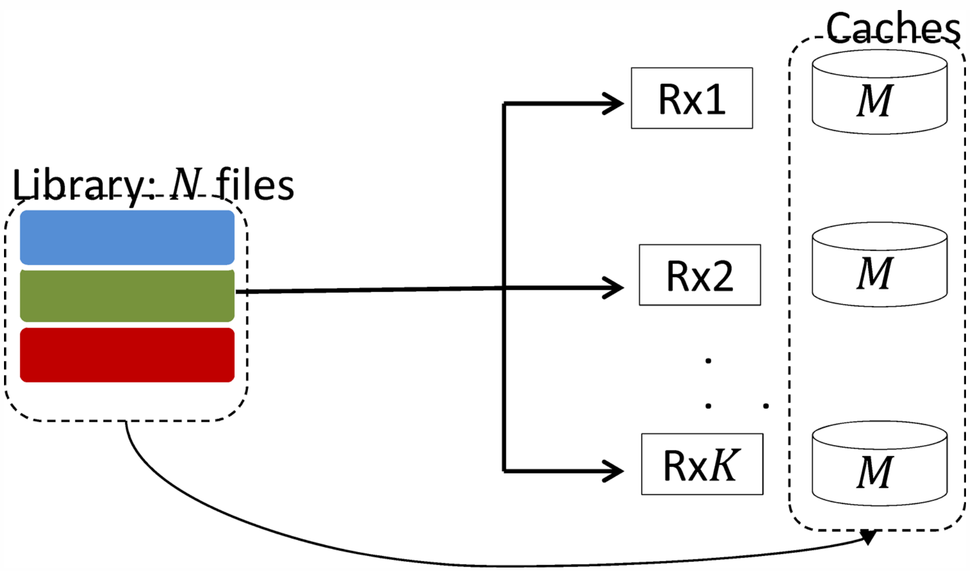
Simple caching (uniform popularity – for now)



- Transmission sequence:
- Local cache gain: $(1 - M/N)$ for each user
- The rate:

$$T = K(1 - M/N) = K(1 - \gamma), \qquad \gamma \stackrel{\text{def}}{=} \frac{M}{N}$$

Basic Parameters

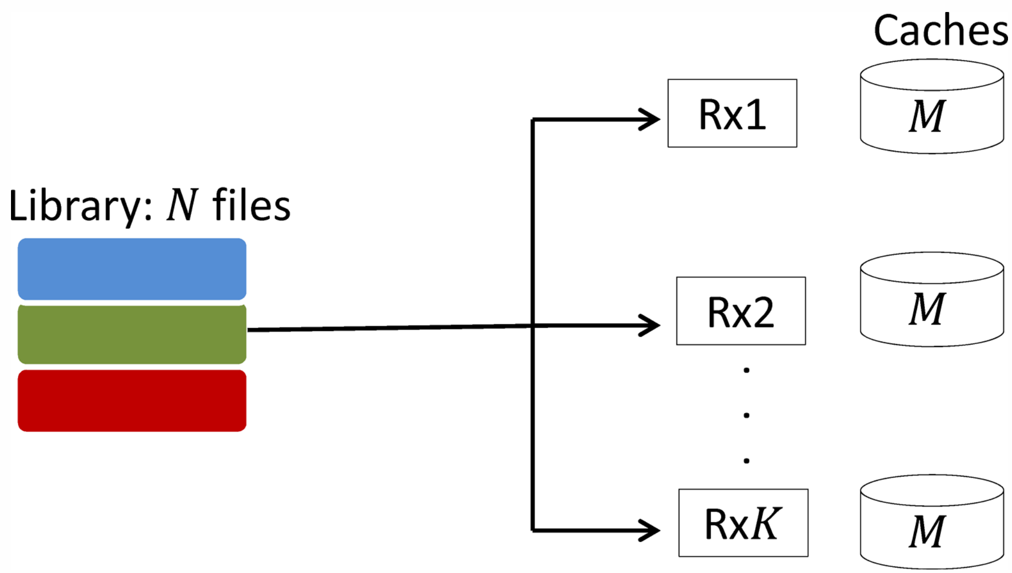


$$\gamma \stackrel{\text{def}}{=} \frac{M}{N} \stackrel{\text{def}}{=} \frac{indivial\ cache\ size}{library\ size}$$

$T(\gamma)$: *duration of delivery phase*

OBJECTIVE: reduce $T(\gamma)$

Coded caching

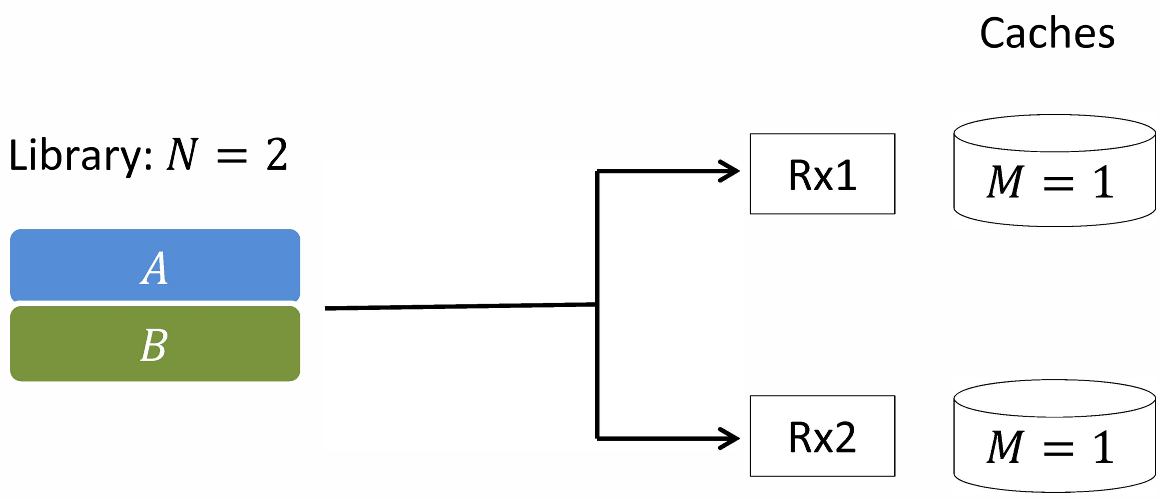


Key breakthrough:

- Cache so that one transmission is useful to many
  - Even if requested files are different
  - Increases multicast opportunities
- Substantial increase in throughput ("worst case")

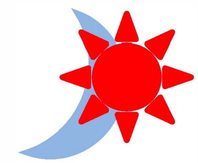
Result: Maddah-Ali, Niesen (2012)

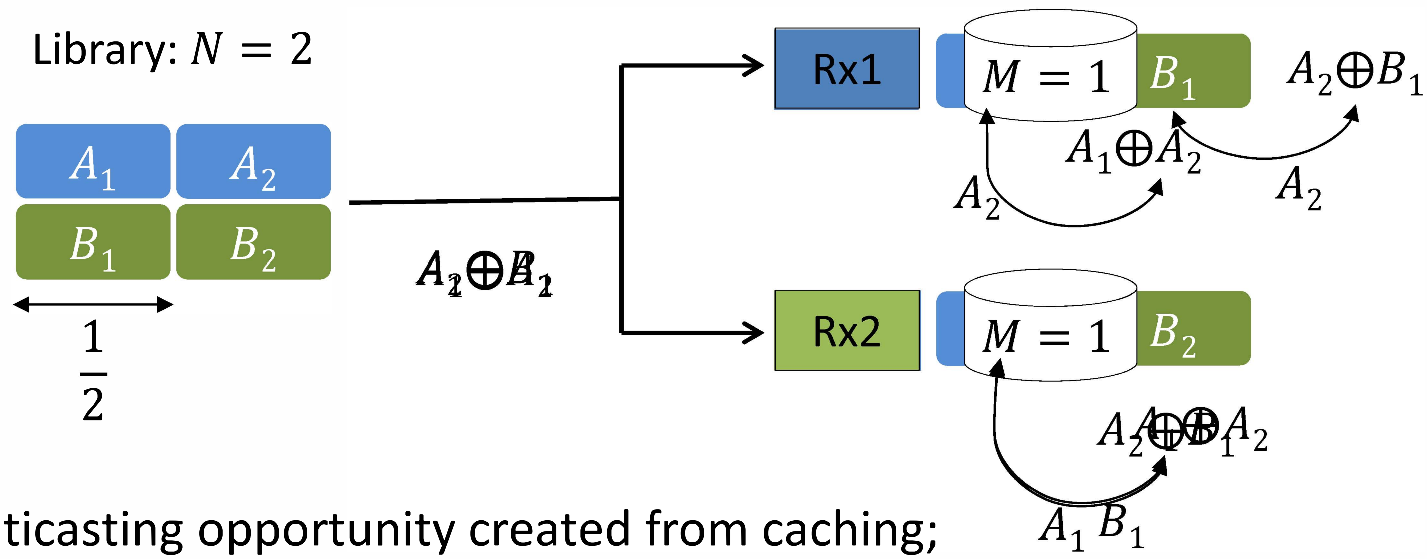
Example: $N = K = 2, M = 1 \quad (\gamma = \frac{1}{2})$



Source: Maddah-Ali, Niesen (2012)

Example: $N = K = 2, M = 1$ $\quad (\gamma = \frac{1}{2})$

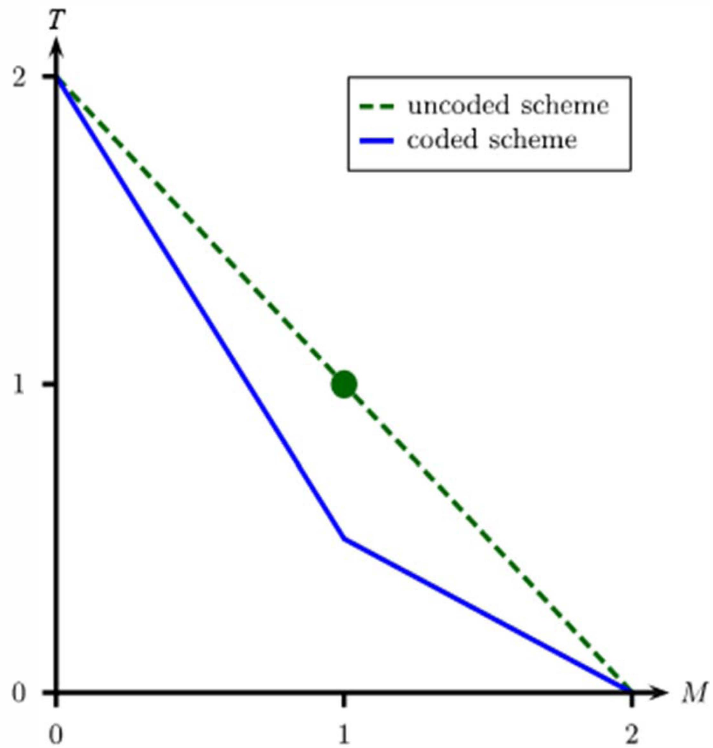




- Multicasting opportunity created from caching;
  - Hard case: distinct requests
  - Easy case: same requests

Source: Maddah-Ali, Niesen (2012)

Comparison: $N = K = 2, M = 1$ $(\gamma = \frac{M}{N} = \frac{1}{2})$



- Uncoded Caching rate:

$$T: K = 2 \rightarrow K(1-\gamma) = 2 \times \frac{1}{2} = 1$$

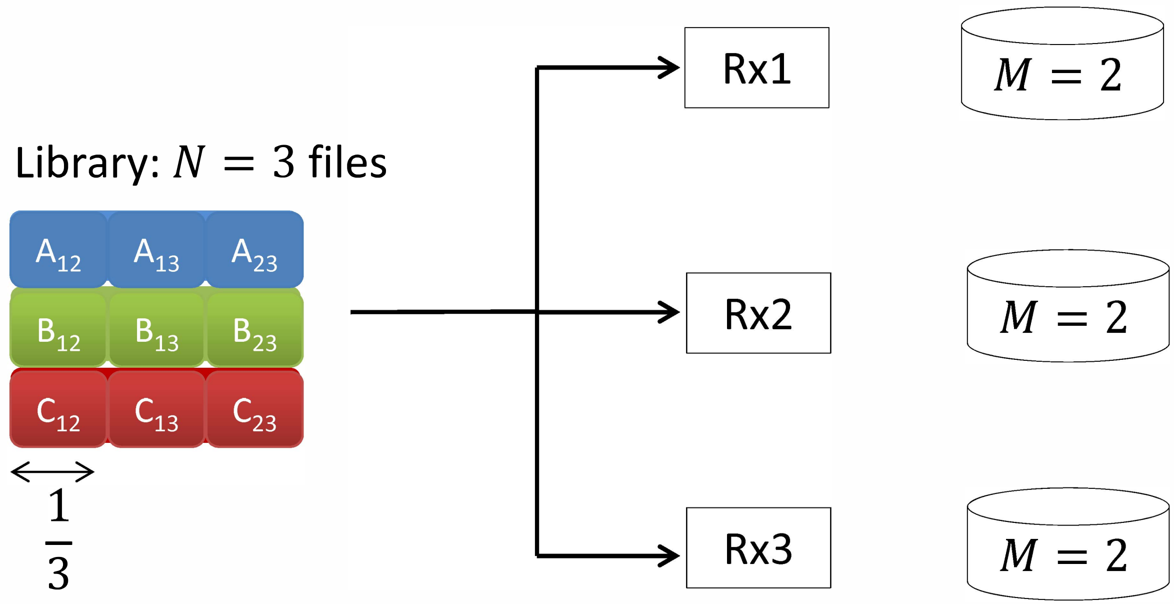
- Coded Caching:

$$T = \frac{1}{2}$$

- For $N = K = 2$ case, optimal rate can be achieved for $M \in [0,1]$

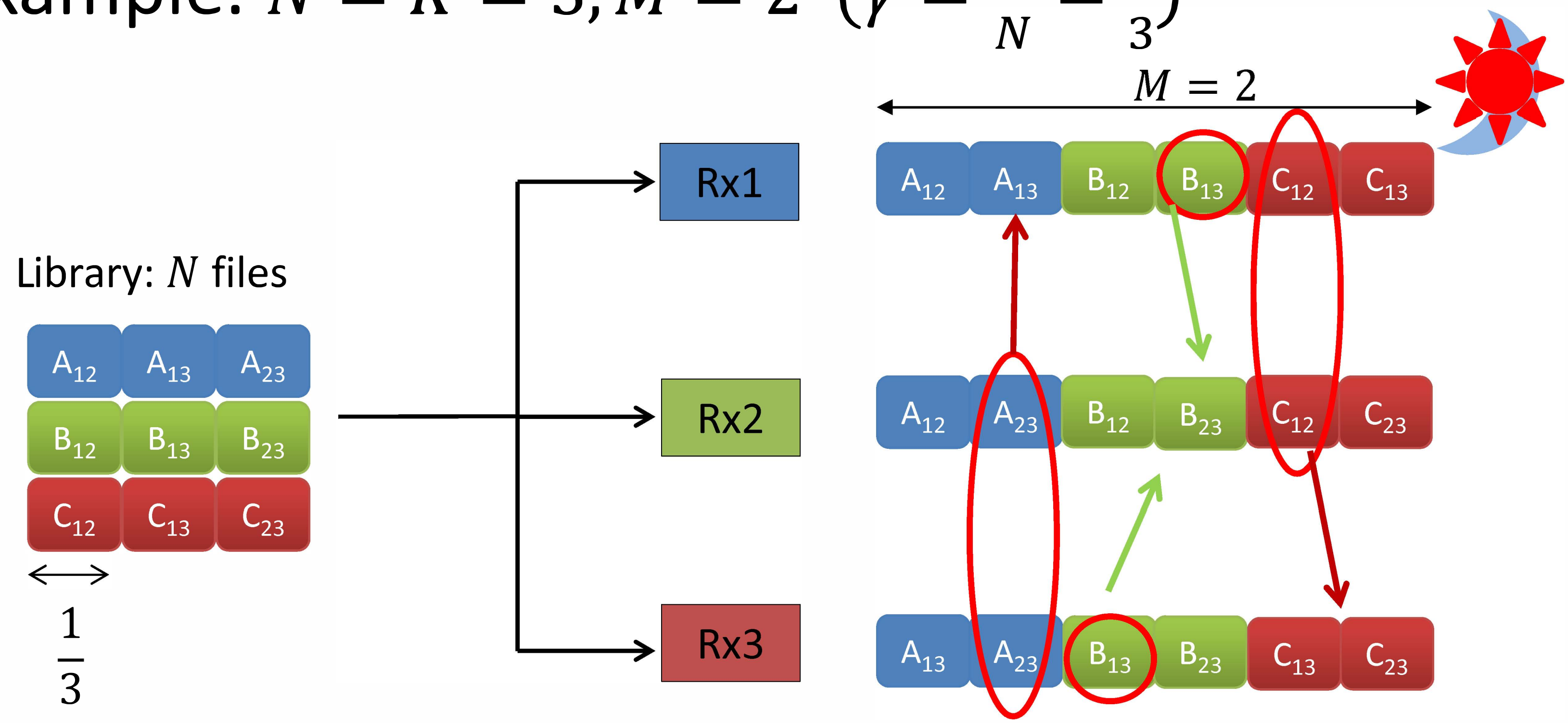
Image source: Maddah-Ali, Niesen (2012)

Another Example: $N = K = 3, M = 2 \quad (\gamma = \frac{2}{3})$



Source: Maddah-Ali, Niesen (2012)

Example: $N = K = 3, M = 2$ $\left(\gamma = \frac{M}{N} = \frac{2}{3}\right)$



- Transmit : $A_{23} \oplus B_{13} \oplus C_{12}$ (a common message for all)

$$T = 1 \times \frac{1}{3} = \frac{1}{3}$$

algorithm: Maddah-Ali, Niesen (2012)

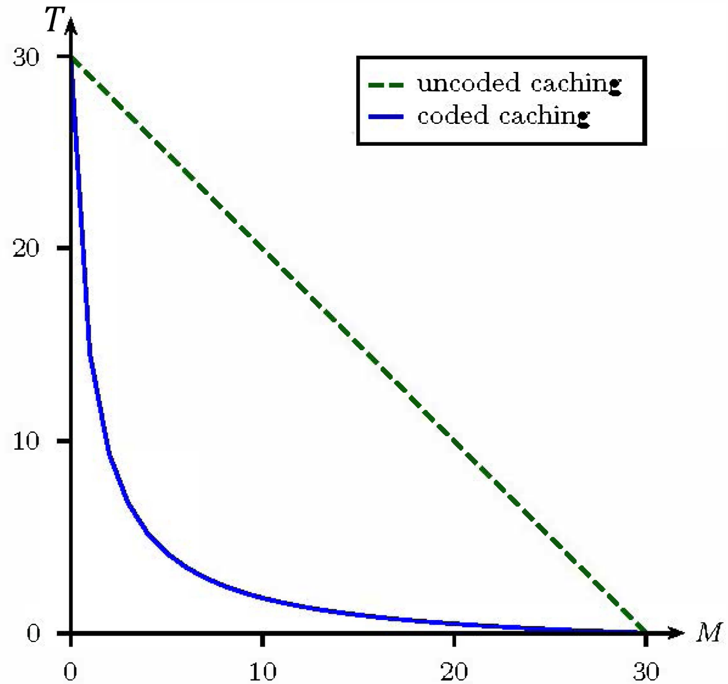
Coded Caching Pseudocode (recall $\gamma \stackrel{\text{def}}{=} \frac{M}{N}$)

- N files in library
- Split each file into $\binom{K}{KM/N} = \binom{K}{K\gamma}$ subfiles
- *Cache: In every* $\frac{MK}{N} = K\gamma$ *set of users, there is one part of each file in common*

- *Request: Each user asks for one file (out of* N*)*
- *Deliver to* $K\gamma + 1$ *users at a time*
  - *Via XORs with* $K\gamma + 1$ *subfiles/summands.* ***Each user (out of the*** $\boldsymbol{K\gamma + 1}$ ***now served) knows all summands except one (its own requested subfile)***
- *Repeat for all possible sets of* $K\gamma + 1$ *users*

Algorithm: Maddah-Ali, Niesen (2012)

Maddah-Ali and Niesen's results



$N = K = 30$

- Uncoded rate (local caching gain) :

$$T = K(1-\gamma)$$

- Coded-caching required :

$$T = \frac{K(1-\gamma)}{1+K\gamma}$$

➢ Coding gain:

$$\text{Gain} \stackrel{\text{def}}{=} \frac{K(1-\gamma)}{T} = 1 + K\gamma$$

Optimal to within a factor 12.

Result and image source: Maddah-Ali, Niesen (2012)

Example:

$K = 10, \gamma = 0.01$ $(K\gamma = 0.1)$:

$T(M) = 9.9$	(only local gain - prefetching)
$T_D(M) = 9.466$	(decentralised caching)
$T_C(M) = 9.0$	(centralised caching)
$T^*(M) \geq 9.0$	(MN optimal bound)

$\Rightarrow$ **Generally small gains when $\boldsymbol{K\gamma < 1}$**

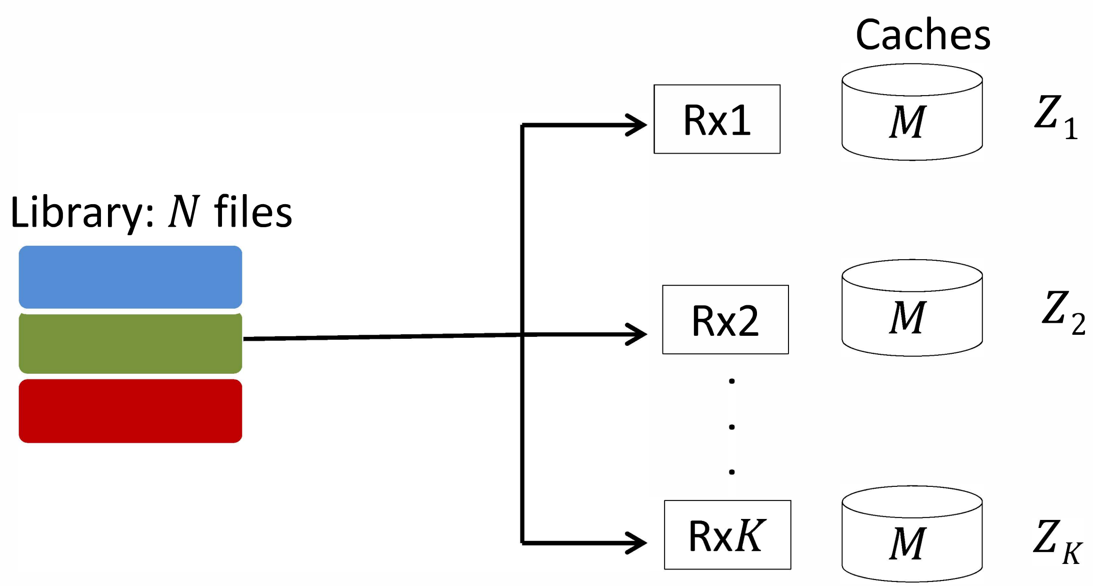
$K = 1000, \gamma = 0.01$ $(K\gamma = 10)$:

$T(M) = 990$	$T_C(M) = 90$
$T_D(M) = 99$	$T^*(M) \geq 25$

Generally large gains when $\boldsymbol{K\gamma > 1}$

Result: Maddah-Ali, Niesen (2013)

On the Optimality of Uncoded Cache-Placement



- Maddah-Ali and Niesen's coded caching is optimal under
  - the constraint of uncoded cache placement
  - the constraint of $N \geq K$

result: Wan et al. (2015)

Bounds to optimal

- Centralized to optimal:

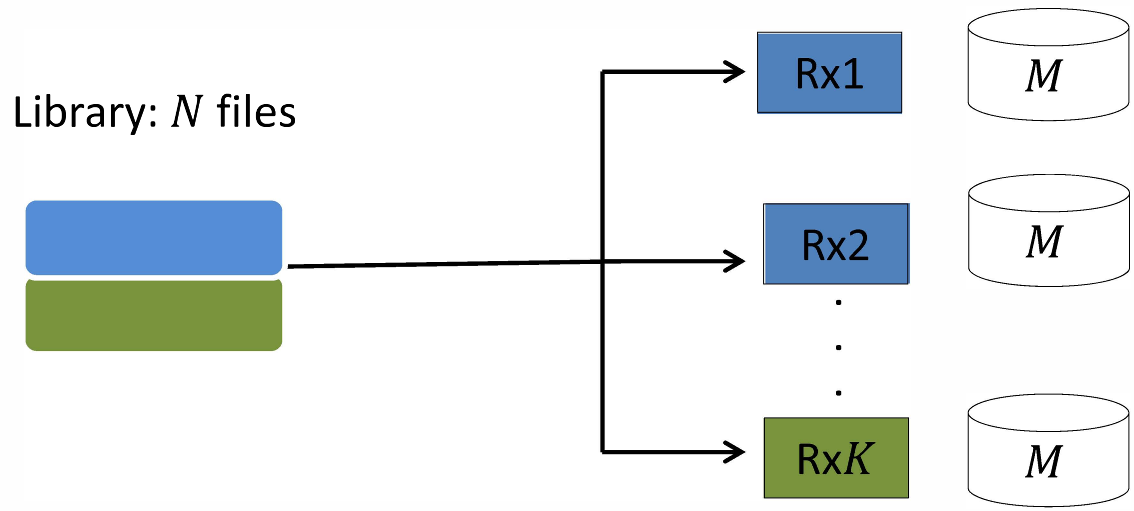
$$1 \leq \frac{T_C(M)}{T^*(M)} \leq 4^{\dagger} \leq 12^{\dagger\dagger}$$

- Decentralised to centralized

$$\frac{T_D(M)}{T_C(M)} \leq 1.5^{*} \leq 4.7^{**} \leq 12^{***}$$

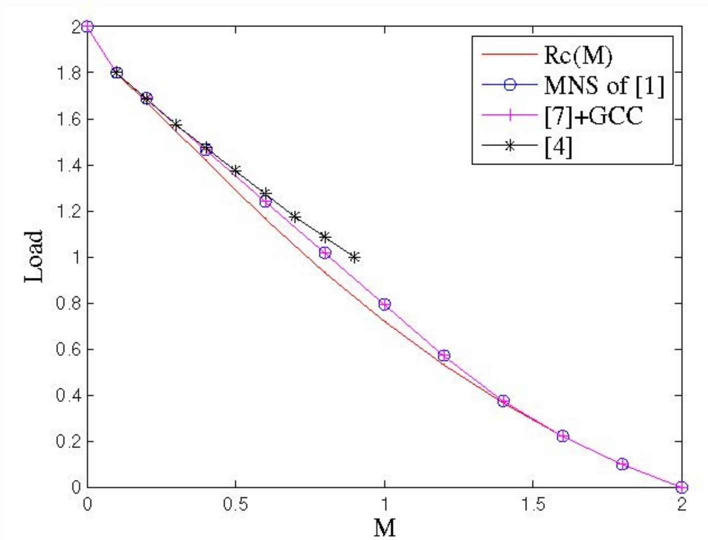
Source ††: Maddah-Ali, Niesen (2013)
Source†: Ghasemi, Ramamoorthy (2015)
Source***: Maddah-Ali, Niesen (2013)
Source**: Lim, Wang and Gastpar (2016)
Source*: Q. Yan, X. Tang, Q. Chen (2016)

Coded vs Traditional Multicasting (More Users than Files)

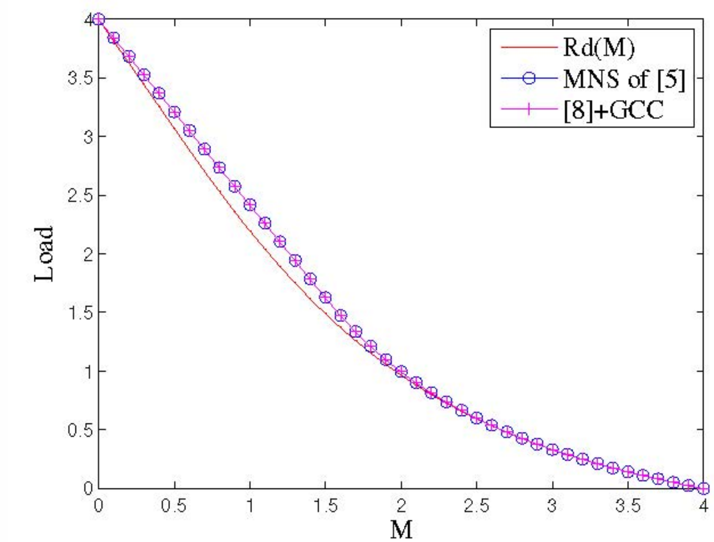


- $N < K$ means that a file may be demanded by multiple users
  - Implies possible additional multicasting opportunities
  - Possible scenario: server to many users, selects few ($N < K$), (equally) popular files
- Original MNS misses this additional (traditional) multicast opportunity
  - because it treats each sub-file demanded by each user as a district sub-file

On Caching with More Users than Files



T(M) vs M (load vs. memory) - centralized system: N = 2 and K = 10



T(M) vs M (load vs. memory) - decentralized system: N = 4 and K = 8

- A novel method* allows for additional (traditional) multicasting opportunities (see also [Chen et al. 2014], and [Sahraei and Gastpar, 2015])

- Optimal under the constraint of uncoded placement and $K > N = 2$
- **Gains are quite limited**
  - ➢ Coded caching automatically `covers' almost all multicasting opportunities!

Result and image source: Wan, Tuninetti and Piantanida (2016)

First Conclusions

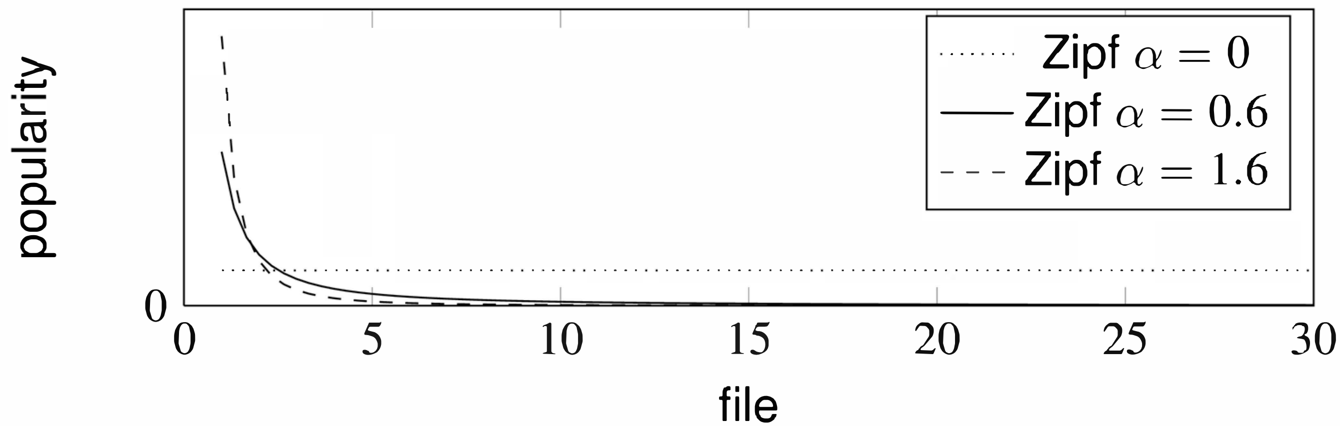
- Significant gain of coded caching
  - Multicasting gain $(K\gamma + 1)$ among users with different demands
- Significant improvement over conventional caching schemes
  - Gains seen when $K\gamma > 1$
  - For large K, then T need not scale as K

$$T \approx \frac{1-\gamma}{\gamma}$$

  - Traditional caching works when $\underline{M \text{ is comparable with } N}$, while coded caching works when $\underline{KM \text{ is comparable with } N}$
- Potential bottlenecks for small γ: T increasing sharply as γ decreases

Coded Caching with Non-uniform Demands

Exploiting File Popularities



- Content popularity is not uniformly distributed
- Could be modelled by a power law / Zipf distribution

Optimal approach for $K = 1$:

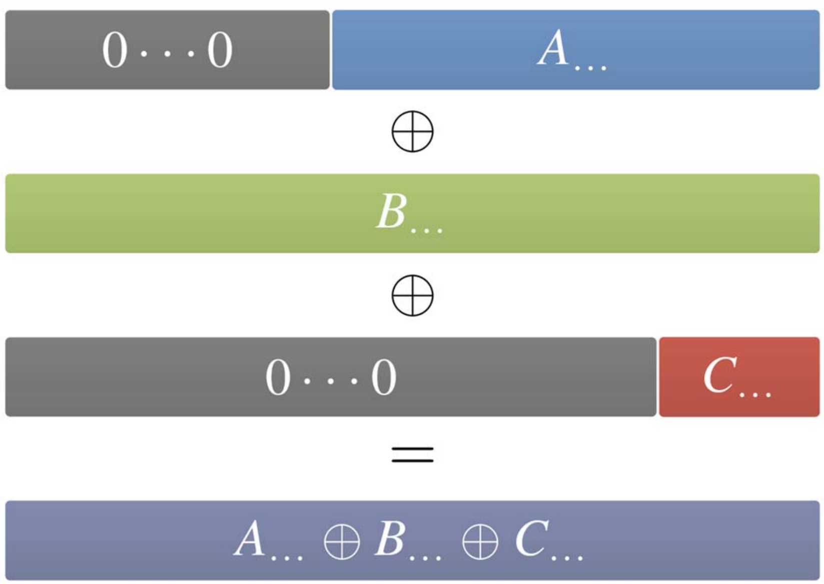
- Least Frequently Used (LFU) eviction policy
- Cache M most popular files
- Can end up with identical caches
  - → No coded-multicasting opportunities

Coded Caching - Non-uniform Demands

Simply assigning a larger cache to more popular files, can result in different subpacket sizes

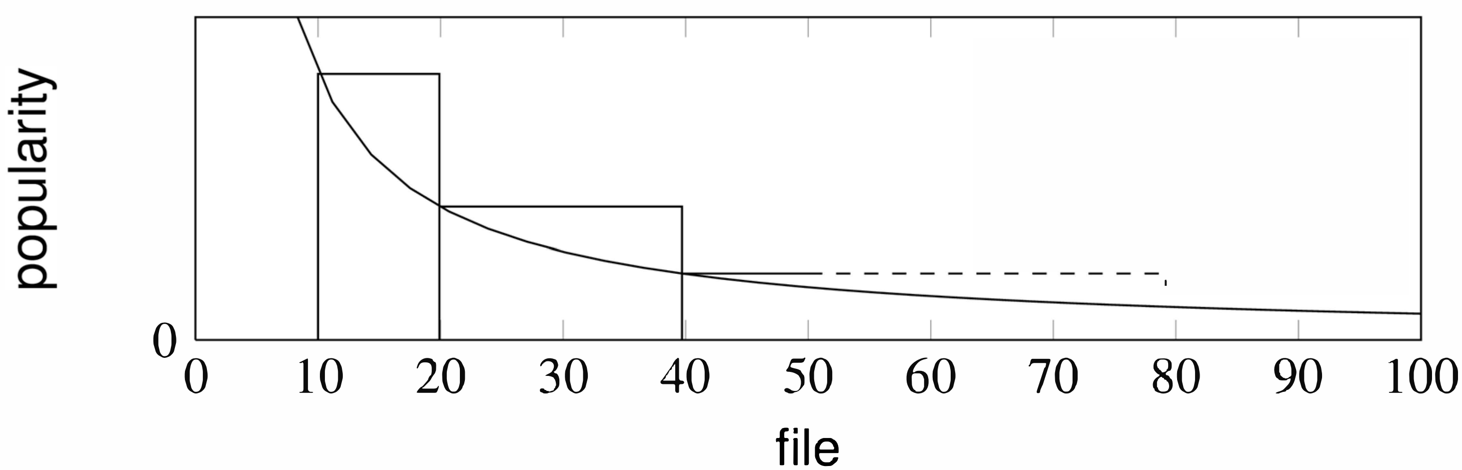
Part sizes must be equal to avoid padding losses

- Biggest subpacket limits rate



Batch Coding for Non-Uniform Demands

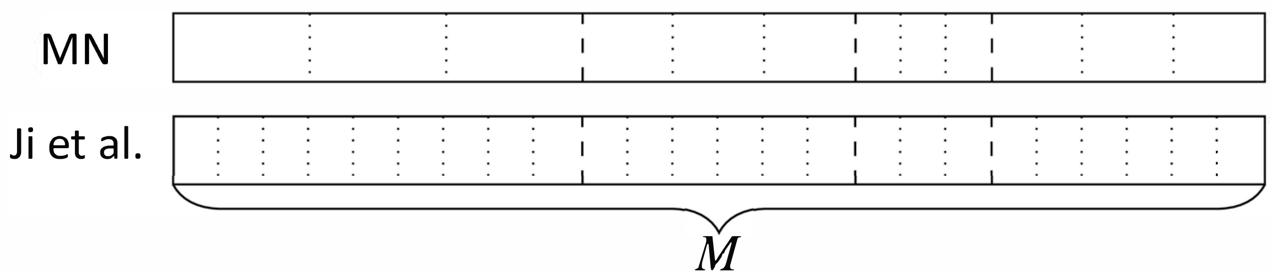
- Separate files into batches of similar popularity
- Cache size allocation is proportional to average batch popularity
- Coded caching for each batch separately
  - Only code among files with same subfile sizes



Source: Maddah-Ali, Niesen (Infocom 2014)

Index-Coding based Scheme for Non-Uniform Demands

- Subfile size same for all files
- Popular files get more subfiles
- Improvement by creating coding opportunities between batches



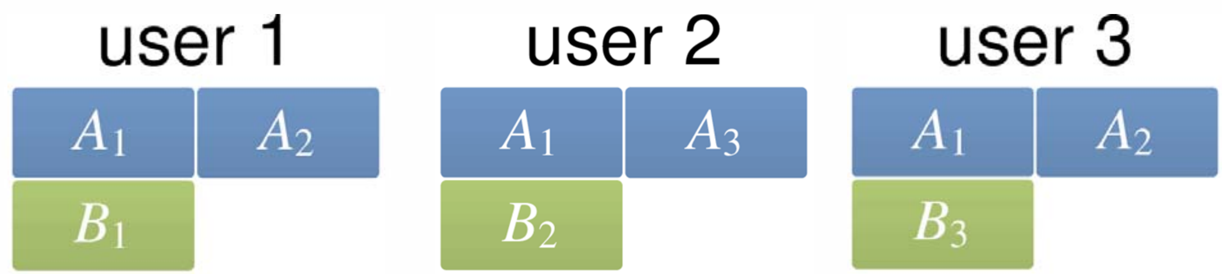
- Delivery uses index coding to combine (XOR) different subfiles
  - graph coloring
  - clique cover

Source: Ji et al. (2015)

Example

- 3 files $\{A, B, C\}$ split into 3 parts each. E.g. $A = \{A_1, A_2, A_3\}$
- Cache distribution $\boldsymbol{p} = \{A = \frac{2}{3}, B = \frac{1}{3}, C = 0\}$

Cache realization $\mathcal{C}$



Request: user1→ A, user2→ B, user3→ C

Queried parts: $\mathcal{Q} = \{A_3, B_1, B_3, C_1, C_2, C_3\}$

Conflict Graph $H_{C,Q}$

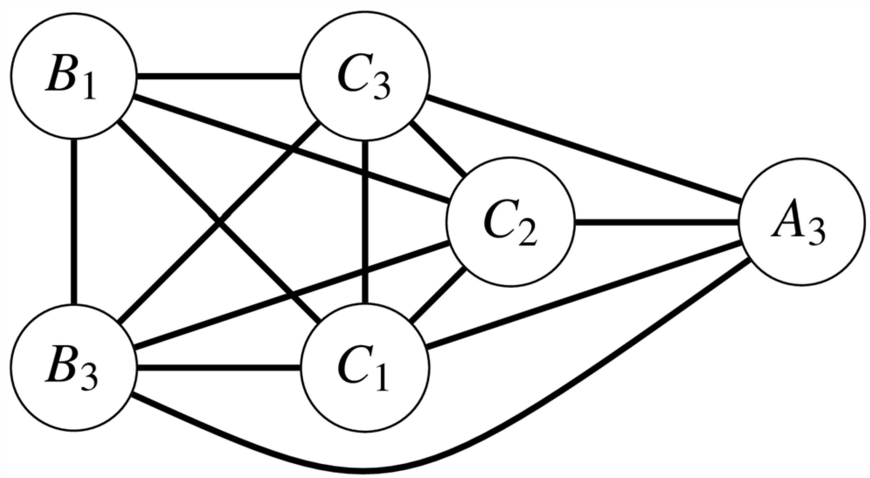
Vertex for each requested subpart ($\in \mathcal{Q}$):

- Replicate if multiple requests of a subfile

Edge if

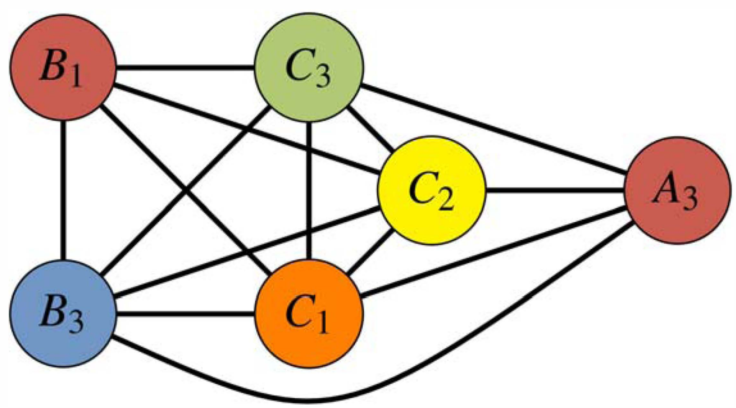
- Not same identity (cannot connect subfile to itself)
- Request(er) not among users caching the other vertex
  - see (A_3, B_1)

Requests: user1→ A, user2→ B, user3→ C
Queried parts: $\mathcal{Q} = \{A_3, B_1, B_3, C_1, C_2, C_3\}$



Graph Coloring $H_{C,Q}$

Connected vertices must have different colors



Transmission

$$T(\gamma) = 5/3$$

Gain

$$\frac{|Q|}{\chi(H_{C,Q})}$$ (χ is chromatic number)

Calculation:

$$\frac{|Q|}{\chi(H_{C,Q})} = \frac{K(1-\gamma)}{T}$$

$$= \frac{3\left(1-\frac{1}{3}\right)}{5/3} = \frac{6}{5}$$

Graph Coloring for non-uniform requests

In general

- NP hard
- exponentially complex

For this particular (coded-caching) coloring problem:

- Greedy constrained coloring used*
  - polynomial complexity in number of users and subfiles

*result: Ji et al. (2015)

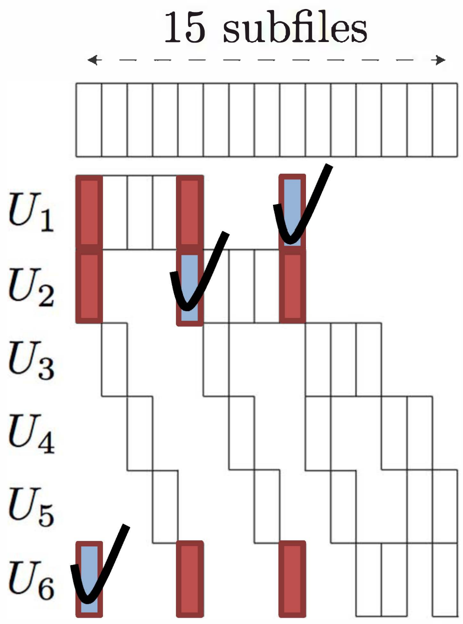
Subpacketization Problem (Motivates Fusing Coded-Caching and PHY)

- Recall need to split each file into $\binom{K}{K\gamma}$ subfiles
- So that:
  - *In every $K\gamma$ caches, there is one part of each file in common*
  - *For each XOR, each of the $K\gamma + 1$ users served knows all subfiles except one (their requested own)*

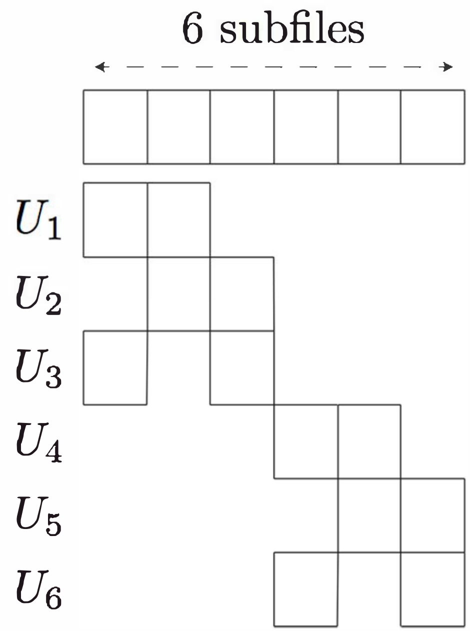
- Introduces intense `sub-packetization' problem
  - Intense file-size problem

Subpacketization constraints

$K = 6, K\gamma = 2$



VS



Users are served $K\gamma$+1 at a time

Users 1,2,6 get content simultaneously

No multicasting between, 1-2-6

No overlaps between 1-6 and 2-6

CC with Bounded File Sizes

- No algorithm with randomized and uncoded placement can escape

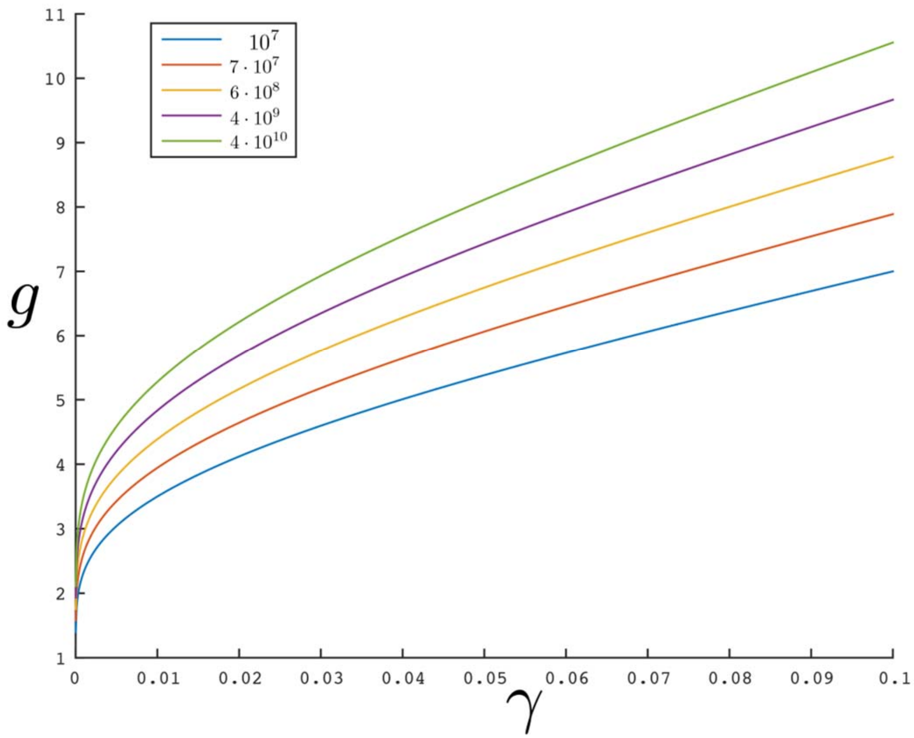
$$g \leq O\left(\frac{\log|F|}{\log\frac{1}{\gamma}}\right)$$

- Conditional bound achieved by Shanmugam et al.
- Also

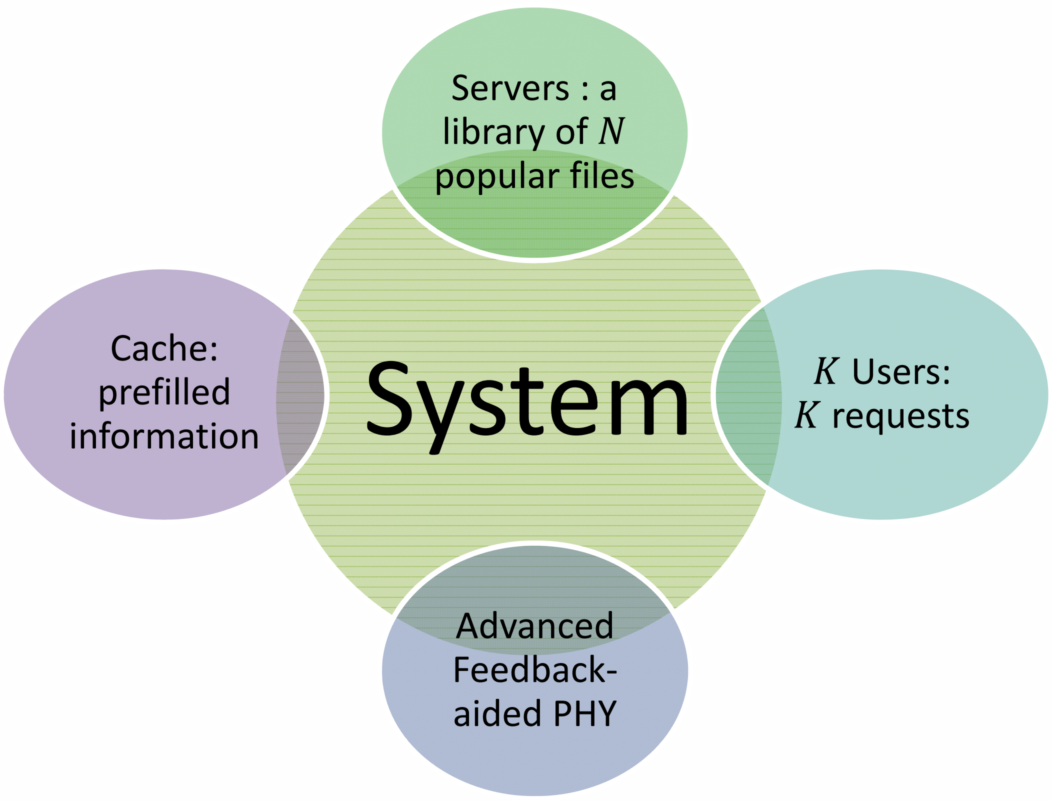
$$(1-\epsilon)R \leq \mathbb{E}\{R\} \leq (1+\epsilon)R$$
$$\Rightarrow |F| = O(K^3 \log K)$$

Results: Shanmugam et al. (2014)

Improved Decentralized Scheme

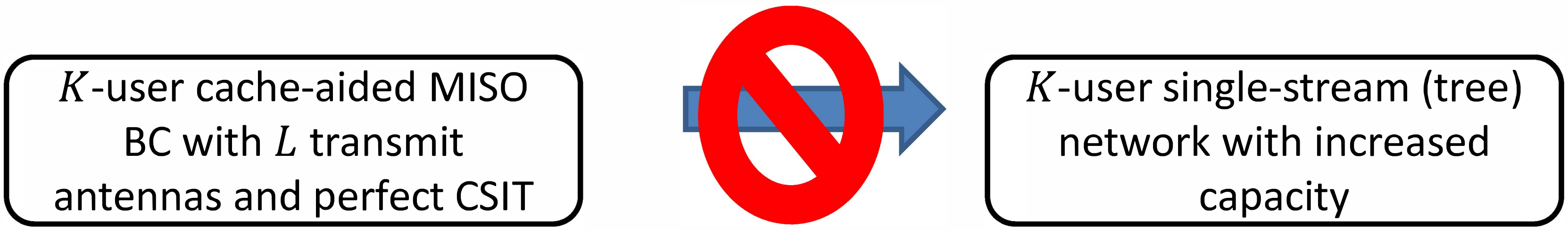


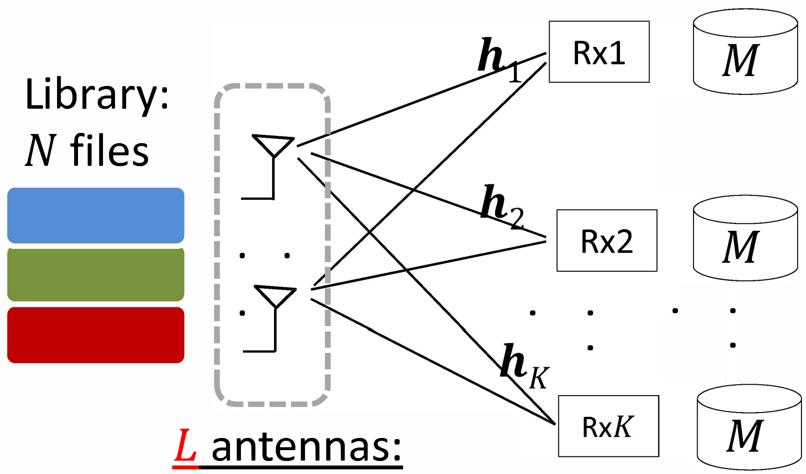
Bottlenecks Introduce Need to Combine Memory and PHY Resources in Wireless Networks

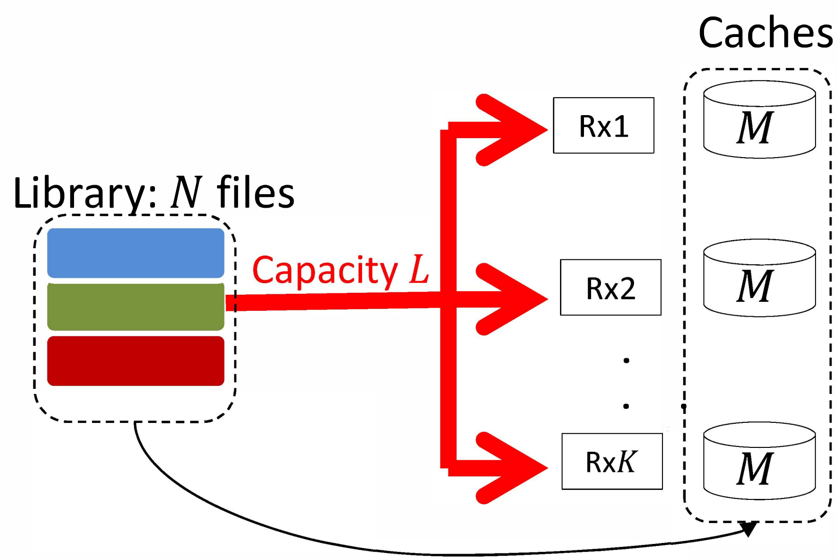


Coded-caching: Non-trivial Application of Single-Stream Coded Caching to Wireless Networks

Example:





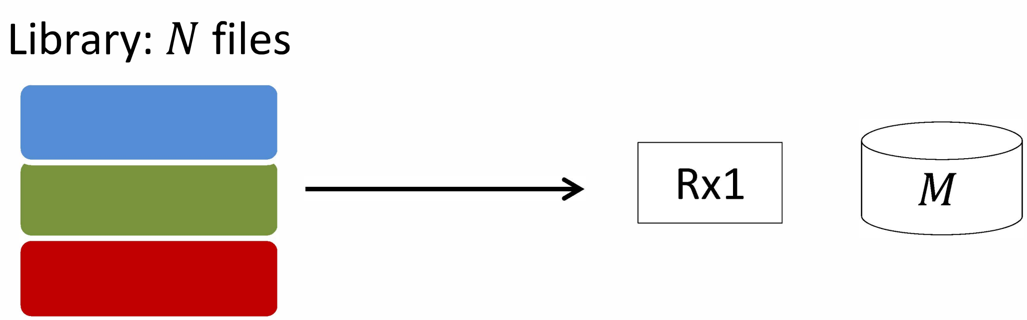


(Cache-aided Degrees of Freedom)

- A equivalent measurement: per-user DoF

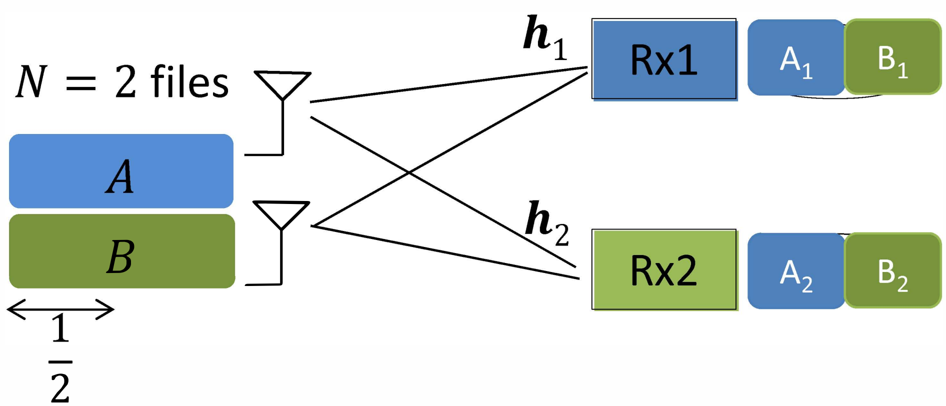
$$d(\gamma) = \frac{1-\gamma}{T} \in [0,1]$$

  - $\gamma = \frac{M}{N}$ is normalized local caching gain: prefilled content
  - $1-\gamma$ excludes local caching gain
  - Captures the joint effect of coded caching and PHY resources



- For one user, the interference-free optimal to serve one file: $T = 1-\gamma$
  - $d(\gamma)$ between 0 and 1 ($d = 1$: Interference-free optimal DoF)

MIMO and Feedback with Coded Caching: Trivial Example $(N = K = 2, M = 1)$

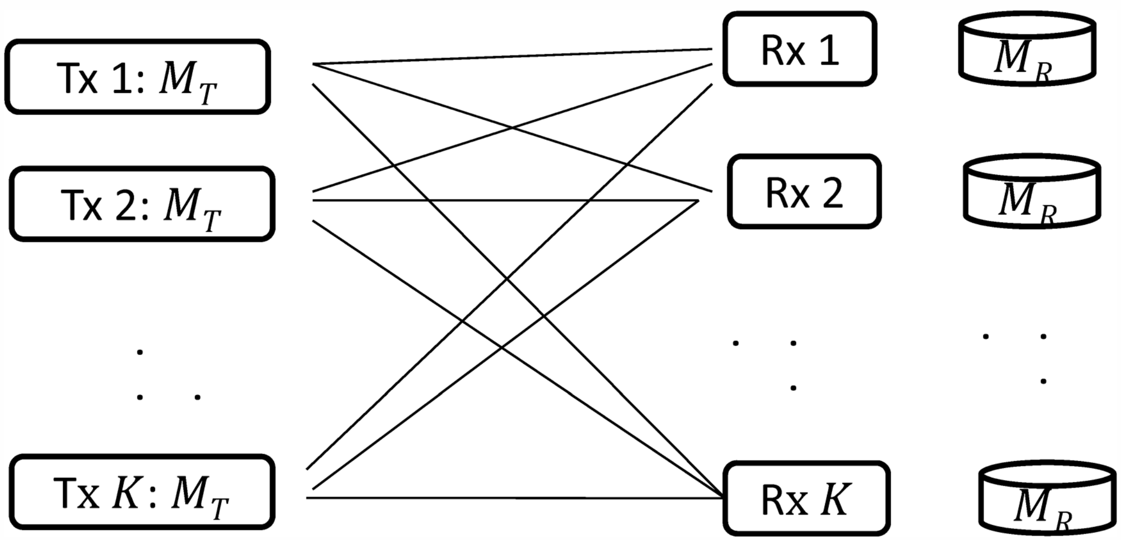


- $A_2 \oplus B_1$ will be delivered
- multicasting phase $\boldsymbol{x}_1 = \begin{bmatrix} A_2 \oplus B_1 \\ 0 \end{bmatrix}$
- $T = \frac{1}{2}$
  - Turns out it is optimal $(T = 1 - \gamma = 1 - \frac{M}{N} = 1 - \frac{1}{2} = \frac{1}{2})$ (same as interference-free)
  - Optimal achieved without CSIT and with just a single antenna

INSIGHT:
Coded caching can reduce need for feedback and multiple antennas, and vice-versa

One Shot Cache-aided Interference channel

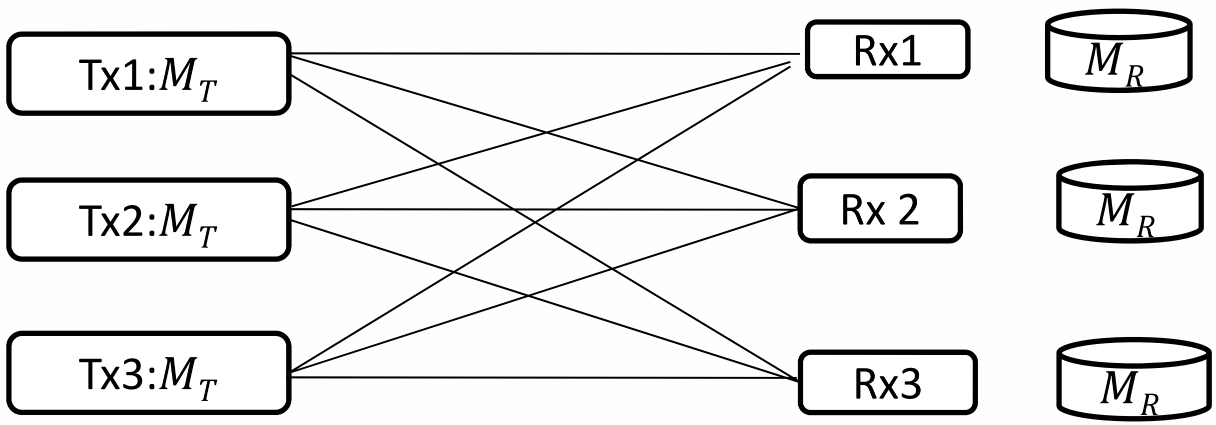
- Cache-aided interference channel
  - K interfering transmitter/ receiver pairs (fully connected)
  - Each transmitter has cache with size $M_T < N$ $(\gamma_T \stackrel{\text{def}}{=} \frac{M_T}{N})$
  - Each receiver has cache with size $M_R < N$ $(\gamma_R \stackrel{\text{def}}{=} \frac{M_R}{N})$



Note:
$KM_T \geq N$

Result: Naderializadeh et al. (2016)

Example: $N = K = 3, M_T = 2, M_R = 1$



- N files: $\mathrm{W}_1 = A, \mathrm{W}_2 = B, \mathrm{W}_3 = C$; $\quad (\gamma_T = \frac{M_T}{N} = \frac{2}{3}, \gamma_R = \frac{M_T}{N} = \frac{1}{3})$
- Split each file into $\binom{K}{K\gamma_T}\binom{K}{K\gamma_R} = \binom{3}{2}\binom{3}{1} = 9$ parts

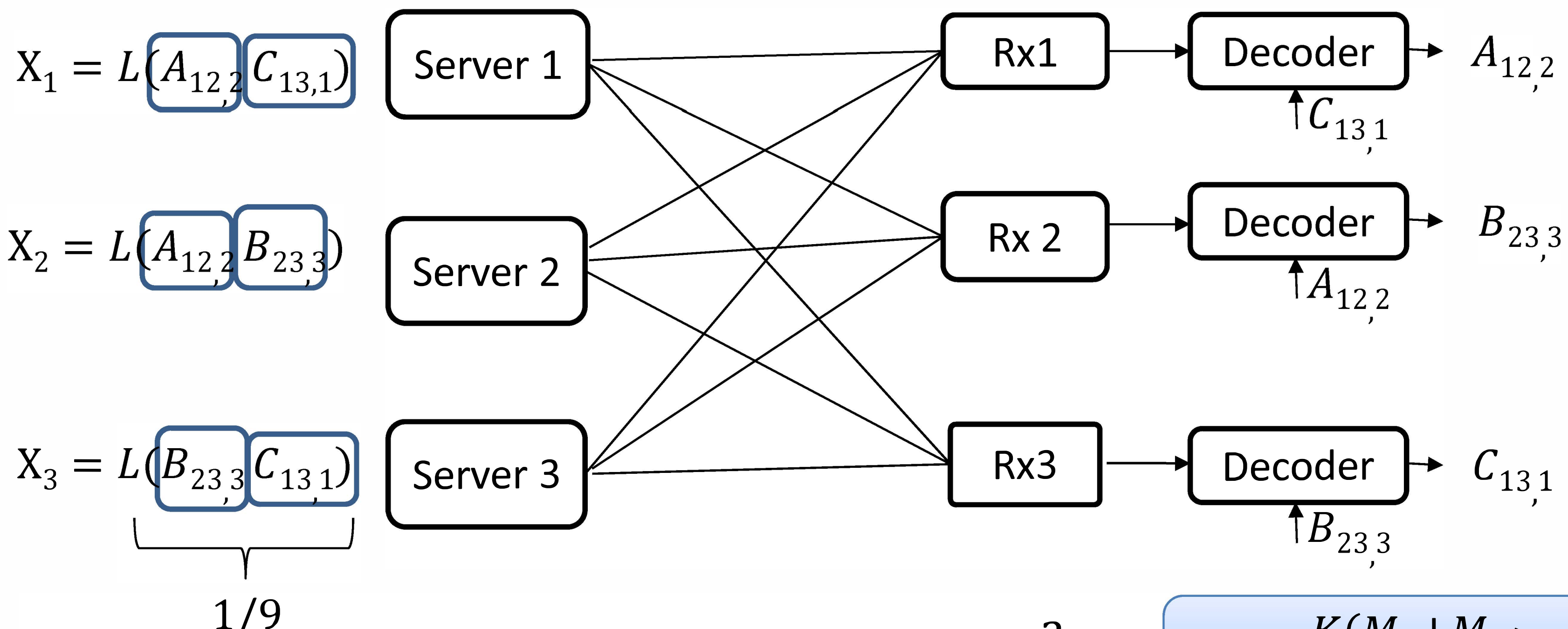
$$A = (A_{12,1}, A_{12,2}, A_{12,3}, A_{13,1}, A_{13,2}, A_{13,3}, A_{23,1}, A_{23,2}, A_{23,3})$$

- Cache Tx 1: $A_{\mathbf{1}2,1}, A_{\mathbf{1}2,2}, A_{\mathbf{1}2,3}, A_{\mathbf{1}3,1}, A_{\mathbf{1}3,2}$
- Cache Rx 1: $A_{12,\mathbf{1}}, A_{13,\mathbf{1}}, A_{23,\mathbf{1}}$

Source: Naderializadeh et al. (2016)

Example: $N = K = 3, M_T = 2, M_R = 1$

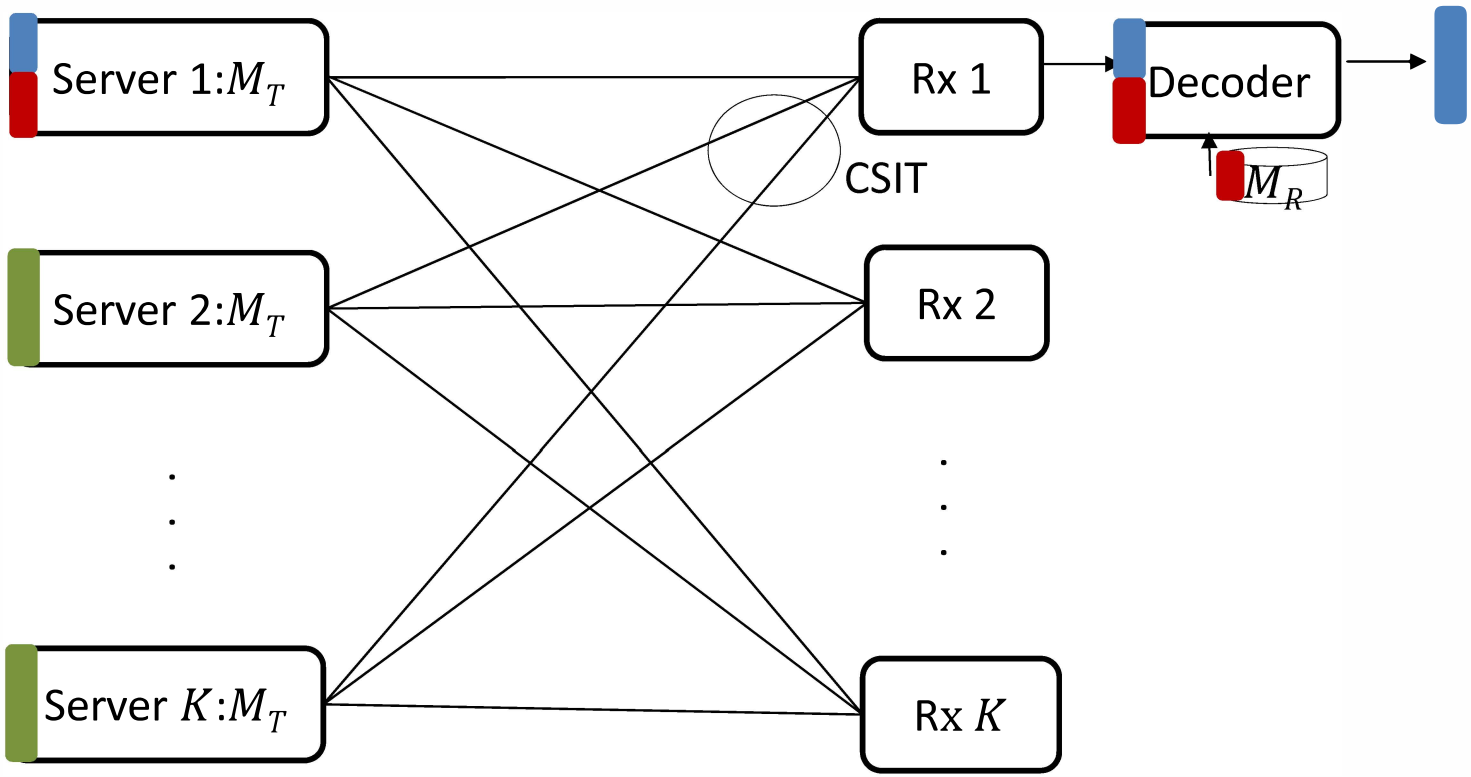
- Rx1 needs: $A_{12,2}, A_{12,3}, A_{13,2}, A_{13,3}, A_{23,2}, A_{23,3}$
- Rx2 needs: $B_{23,3}, B_{13,1}, B_{12,3}, B_{23,1}, B_{13,3}, B_{12,1}$
- Rx3 needs: $C_{13,1}, C_{23,2}, C_{23,1}, C_{12,2}, C_{12,1}, C_{13,2}$



- Other triple symbols are the same: $T = \frac{2}{3} \Rightarrow d_\Sigma = \frac{K(M_T+M_R)}{N} = 3$

Source: Naderializadeh et al. (2016)

Idea for the General Case



- With transmitter cooperation and perfect quality CSIT
  - ➢ interference can be cancelled
- Combining with the caching content
  - ➢ recover the missing information in cache

algorithm: Naderializadeh et al. (2016)

Conclusion – Cache Aided IC (one shot)

- The one-shot linear sum-DoF:

$$d_\Sigma = \min\{\frac{KM_T + KM_R}{N}, K\}$$

$$d(\gamma_T, \gamma_R) = \gamma_T + \gamma_R \leq 1$$

  - Within a factor of 2 of the one-shot linear-DoF optimal
  - Equal contribution of transmitter and receiver caches
  - Linear scaling of DoF with network size
  - Covers single-stream and multi-server cases.

result: Naderializadeh et al. (2016)

Caching and Feedback

Feature: Synergy and interplay between memory and feedback

Background

- In most cases, DoF impact of coded caching:

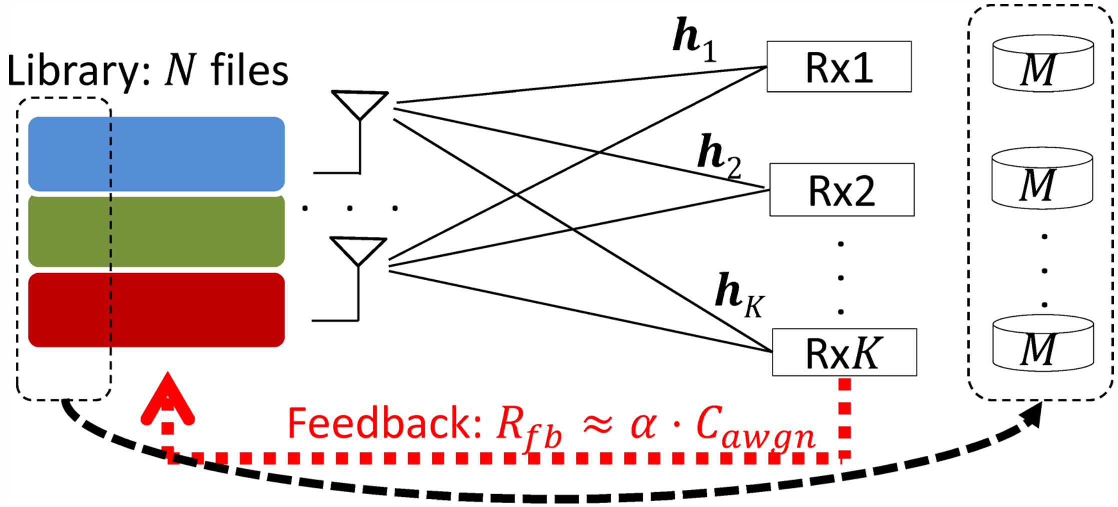
$$d(\gamma) - d(\gamma = 0) = \gamma$$

  - Even in settings with perfect feedback and many antennas

Gains due to caching are $\approx \gamma \approx 10^{-3} \rightarrow 10^{-2}$ (Roberts et al.)

- Are there settings for which the impact of caching is substantially larger?

Cache-aided K-user BC with mixed CSIT



- Delayed CSIT + imperfect-quality current CSIT
- High-SNR <u>current-CSIT</u> quality exponent

$$\alpha = -\lim_{P \to \infty} \frac{\log \mathbb{E}\left[||\boldsymbol{h}_k - \widehat{\boldsymbol{h}}_k||^2\right]}{\log P}, \qquad k \in \{1, \dots, K\}$$

➢ $\alpha = 0$ means ≈no current feedback, and $\alpha = 1$ means perfect CSIT

CSIT/Caching Interplay: MISO BC

Corollary (Zhang-Elia):

$$T(\gamma, \alpha) = \frac{(1-\gamma)\cdot\log(\frac{1}{\gamma})}{\alpha\cdot\log(1/\gamma) + (1-\alpha)(1-\gamma)}$$

Per-user DoF

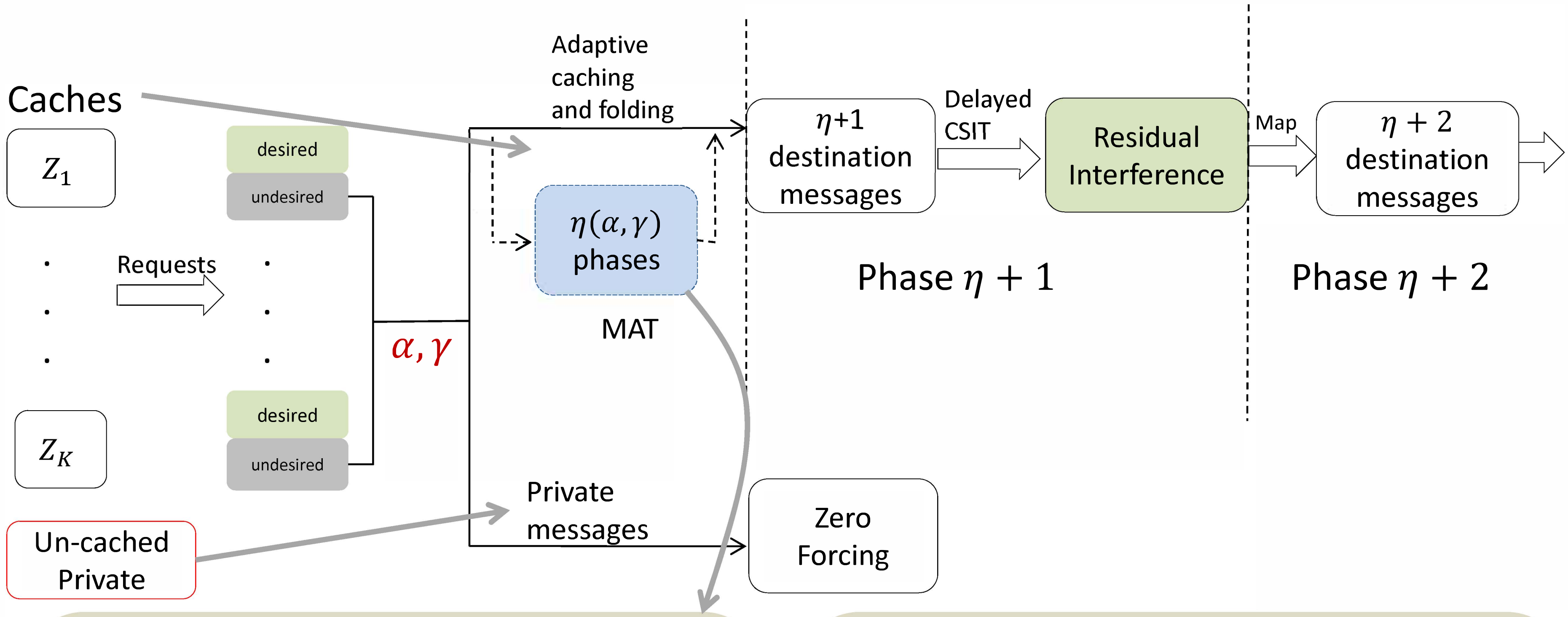
$$d(\gamma, \alpha) = \alpha \ + \ (1-\alpha) \ \ \frac{1-\gamma}{\log\left(\frac{1}{\gamma}\right)}$$

Features:

- additive combination of resources
- Initial offset due to FB (larger K), and then substantial additional boost due to memory

 Under the logarithmic approximation $H_n \approx \log(n)$ (Exact for large K)

Cache-aided Prospective-hindsight Scheme



Feature:

- With delayed CSIT, multicasting is much faster than broadcasting
- Memory boosts broadcasting

Feature: current CSIT increases caching redundancy

$\alpha \uparrow \Rightarrow$ can have more private data

$\Rightarrow$ Less to be cached

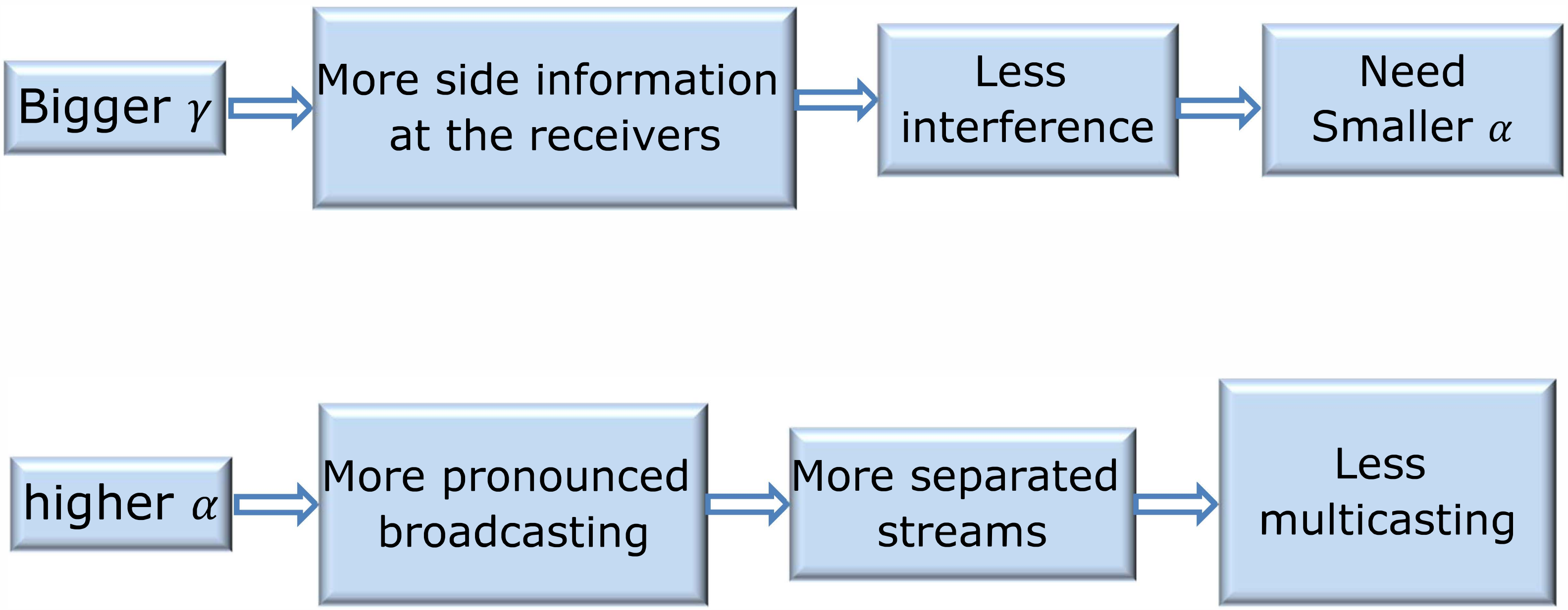
$\Rightarrow$ Caching can be more redundant

$\Rightarrow$ XORS have higher order

$\Rightarrow$ multicast to more users at a time

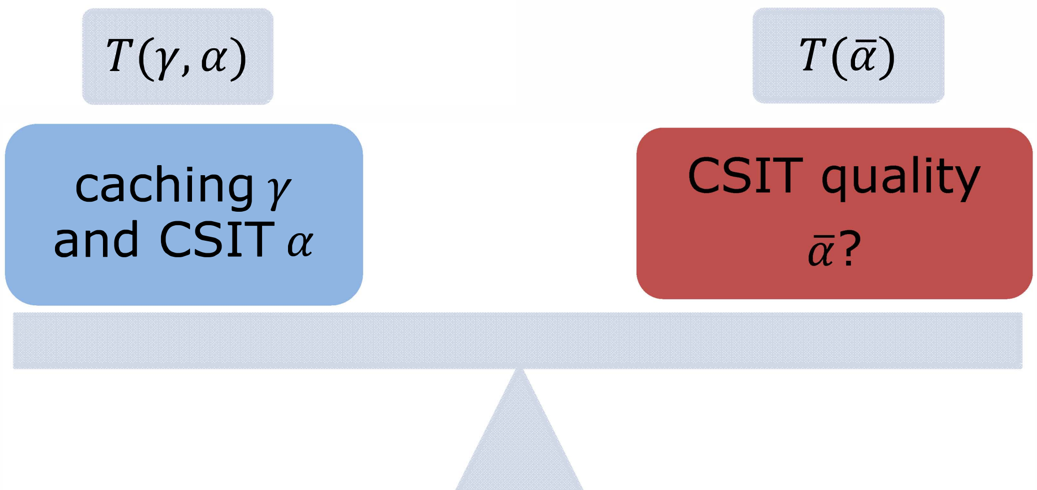
$\Rightarrow$ Much much faster

Intuition: Some Competition between Feedback-Quality and Memory



Cache-aided Feedback Reductions

Example: $\gamma = 10^{-4}, K$ large



- To get the same rate, the required CSIT quality:

$$\bar{\alpha}(\gamma, \alpha) = \alpha + (1-\alpha)\frac{1-\gamma}{\log\left(\frac{1}{\gamma}\right)} = \alpha + 0.11(1-\alpha)$$

cache-aided CSIT reduction

- For example: with $\gamma = 10^{-4}$, then $\bar{\alpha} = 0.2 \rightarrow \alpha = 0.1$
  - ➢ Small cache size, halved CSIT requirement

Using Coded Caching to `Buffer' CSI

Feature: Caching allows for CSIT reductions (and `buffering')

$\boldsymbol{\gamma'_{\alpha} = e^{-\frac{1}{\alpha}}}$ can achieve – without current CSIT – the optimal DoF $d^*(\gamma = 0, \alpha)$ associated to a system with delayed CSIT and α-quality current CSIT.

Example (large $\boldsymbol{K}$)

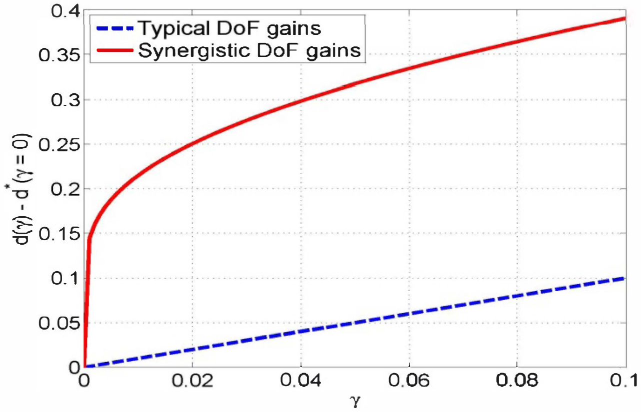
- Assume D-CSIT and α= 1/5. Then

$$\gamma'_{\alpha=\frac{1}{5}} = e^{-5} = 0.0067 \approx \frac{1}{150}$$

$$d^*(\gamma = 0.0067, \alpha = 0) \geq d^*(\gamma = 0, \alpha = 1/5)$$

- The $d^*(\gamma = 0, \alpha = 1/5)$, can be achieved by substituting all current CSIT with DCSIT and coded caching employing $\gamma \approx 1/150$.

Synergistic DoF Gains ($\alpha = 0$)



$$\frac{\delta}{\delta\gamma}\left(d(\gamma) - d(\gamma = 0) \right)\Big|_{\gamma=\frac{1}{K}} \approx \frac{K}{\log^2 K} \quad \text{(for all } K\text{)}$$

vs.

$$\frac{\delta}{\delta\gamma}\left(d(\gamma) - d(\gamma = 0) \right)\Big|_{\gamma=\frac{1}{K}} = \frac{\delta}{\delta\gamma}(\gamma) = 1$$

- **Feature:** CSIT allows for boost from small (reasonable) amounts of caching
- Synergy because PHY and CC exceed sum of two individual compontents

$$\mathrm{d}(\gamma) > \mathrm{d}_{\mathrm{ss}}(\gamma) + \mathrm{d}_{\mathrm{PHY}}(\gamma = 0)$$

- `Exponential' effect of coded caching (for sufficiently large K)
  - A very small $\gamma = e^{-G}$ can offer a very satisfactory

$$d(\gamma = e^{-G}) - d(\gamma = 0) \rightarrow \frac{1}{G}$$

High Impact of Coded Caching

Example

- In a MISO BC system with only delayed CSIT, K antennas and K users:

$$d^*(\gamma = 0) = \frac{1}{H_K} \rightarrow 0 \text{ (as } K \text{ increases)}$$

- A DoF of $d(\gamma \approx \frac{1}{50}) = \frac{1}{4}$ for all K
- A DoF of $d(\gamma \approx \frac{1}{1000}) = \frac{1}{7}$ for all K
- A DoF of $d(\gamma \approx 10^{-5}) > \frac{1}{12}$ for all K

CSIT-Aided Amelioration of the Sub-Packetization Problem

- For CC per-user DoF gain d_G, we needed

$$\binom{K}{K\gamma} = \binom{K}{Kd_G} \quad \text{Sub-packets}$$

- Synergistically, this same DoF gain d_G needs only

$$\binom{K}{Ke^{-1/d_G}} \quad \text{Sub-packets}$$

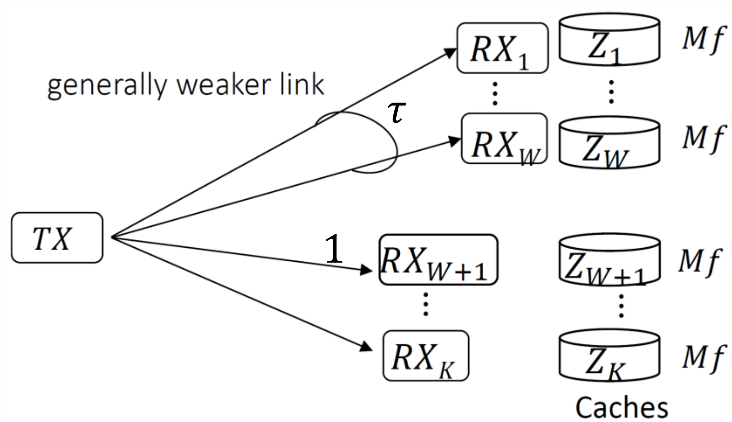
Example (large K): $d_G = \frac{1}{6}$: Then $\binom{K}{K/6} \rightarrow \binom{K}{Ke^{-6}} \approx \binom{K}{K/400}$

Topology (no FB)

Wireless Coded Caching:
A Topological Perspective

Features/Opportunities:
- Topological `holes' to attenuate interference
- XORING on the air
- XORs need not be common
- Interesting relationship between coding gain and local caching gain

Topological SISO BC

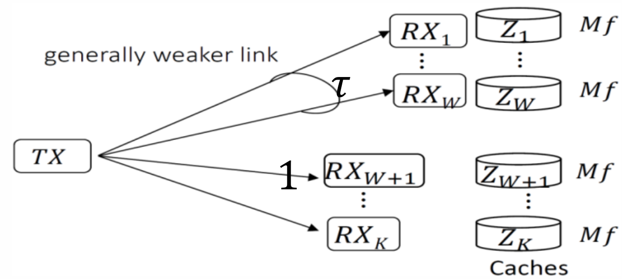


Topologically-uneven wireless SISO K-user BC:

- W weak users with normalized capacity $\tau < 1$
- $K - W$ strong users with normalized capacity $= 1$

- Same cache size per user (M)

System Model



- Recall when $\tau = 1$ (M&N)

$$T(K) = \frac{K(1-\gamma)}{1+K\gamma}, \qquad \gamma = \frac{M}{N}$$

which gives a caching gain

$$g \triangleq \frac{K(1-\gamma)}{T} = K\gamma + 1$$

- Problem: multicasting can suffer from "worst-user" effect

$$d(\gamma) \rightarrow \tau \cdot d(\gamma)$$

- **Opportunity:** Topological `holes' to attenuate interference

- Question: how is the performance affected as τ decreases?

Main Result

Theorem (Zhang-Elia 16):

In the K-user topological cache-aided SISO BC with W weak users,

$$T(\tau) = \begin{cases} \dfrac{T(W)}{\tau}, & 0 \leq \tau \leq \bar{\tau}_{thr} \\ \min\left\{T(K-W) + T(W), \dfrac{\tau_{thr} T(K)}{\tau}\right\}, & 0 \leq \tau \leq \tau_{thr} \\ T(K), & \tau_{thr} \leq \tau \leq 1 \end{cases}$$

is achievable and has a gap-to-optimal

$$\frac{T}{T^*} < 8$$

that is always less than 8.

$$T(N) = \frac{N(1-\gamma)}{1+N\gamma}, \bar{\tau}_{thr} = \frac{T(W)}{T(W)+T(K-W)}, \tau_{thr} = \begin{cases} 1 - \dfrac{\binom{K-W}{K\gamma+1}}{\binom{K}{K\gamma+1}}, & for\ W < K(1-\gamma) \\ 1, & otherwise \end{cases}$$

Topology Threshold

Corollary (Zhang-Elia 16):

There is a threshold

$$\tau_{thr} \approx 1 - \left(1 - \frac{W}{K}\right)^{g_{max}}$$

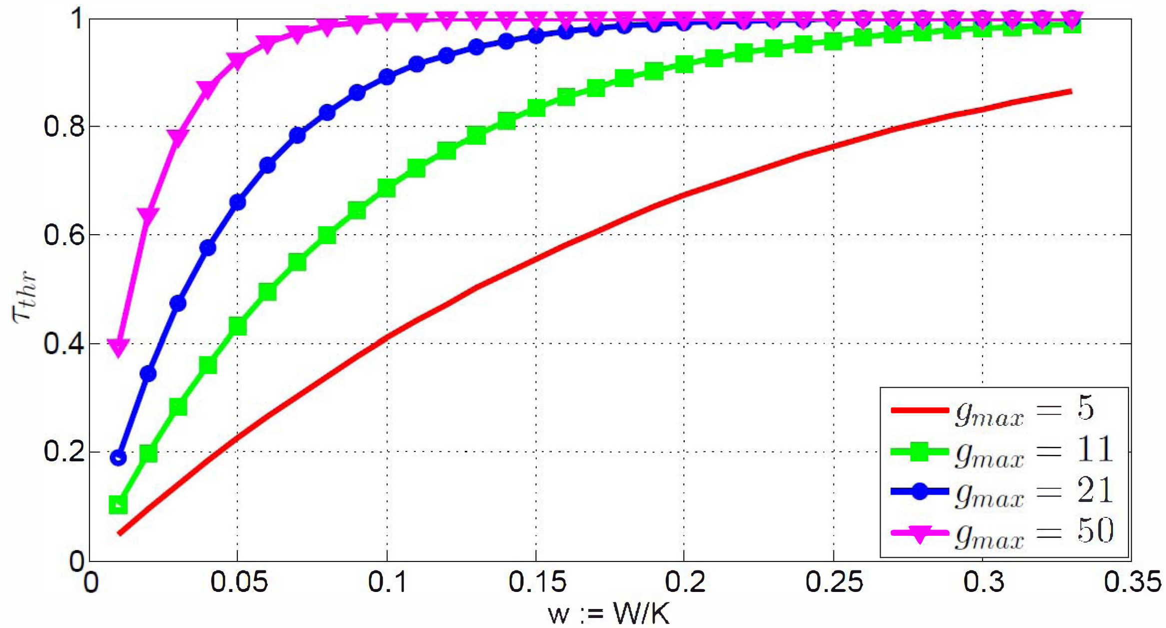
which guarantees full-capacity performance

$$T(\tau \geq \tau_{thr}) = T(K)$$

Recall $g_{max} \stackrel{\text{def}}{=} \mathrm{K}\gamma + 1, w \stackrel{\text{def}}{=} \frac{W}{K}$

$$\tau_{thr} \in \left[1 - (1-w)^{g_{max}}, 1 - \left(1 - w - \frac{w\gamma}{1-\gamma}\right)^{g_{max}}\right]$$

Topology Threshold

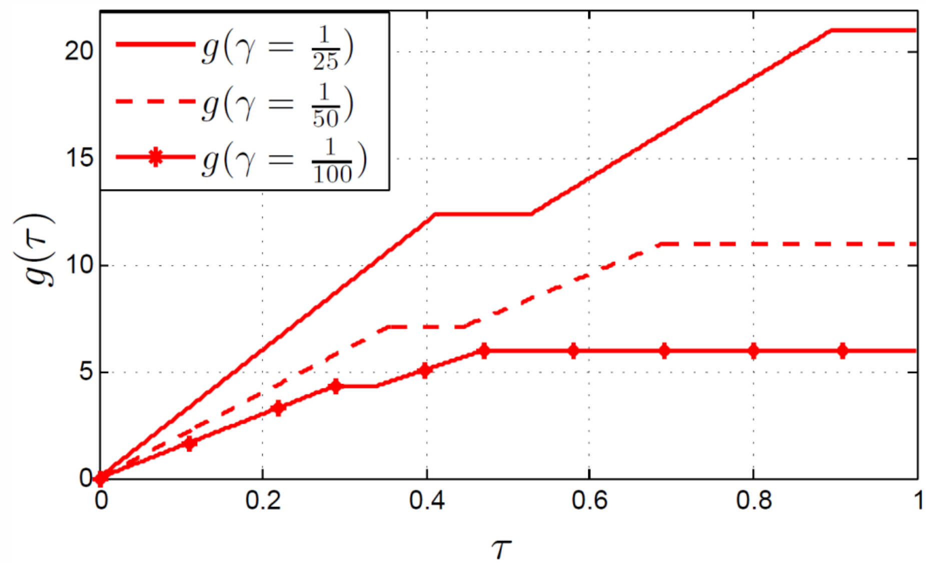


- τ_{thr} corresponding to distinct values for gains g_{max}
- E.g., for $g_{max} = 5$ and $w = 0.1$, then $\tau_{thr} \approx 0.4$

Coded-caching Gain

- Coded-caching gain under topology setting

$$g(\tau) \triangleq \frac{K(1-\gamma)}{T} \in [0, g_{max}]$$

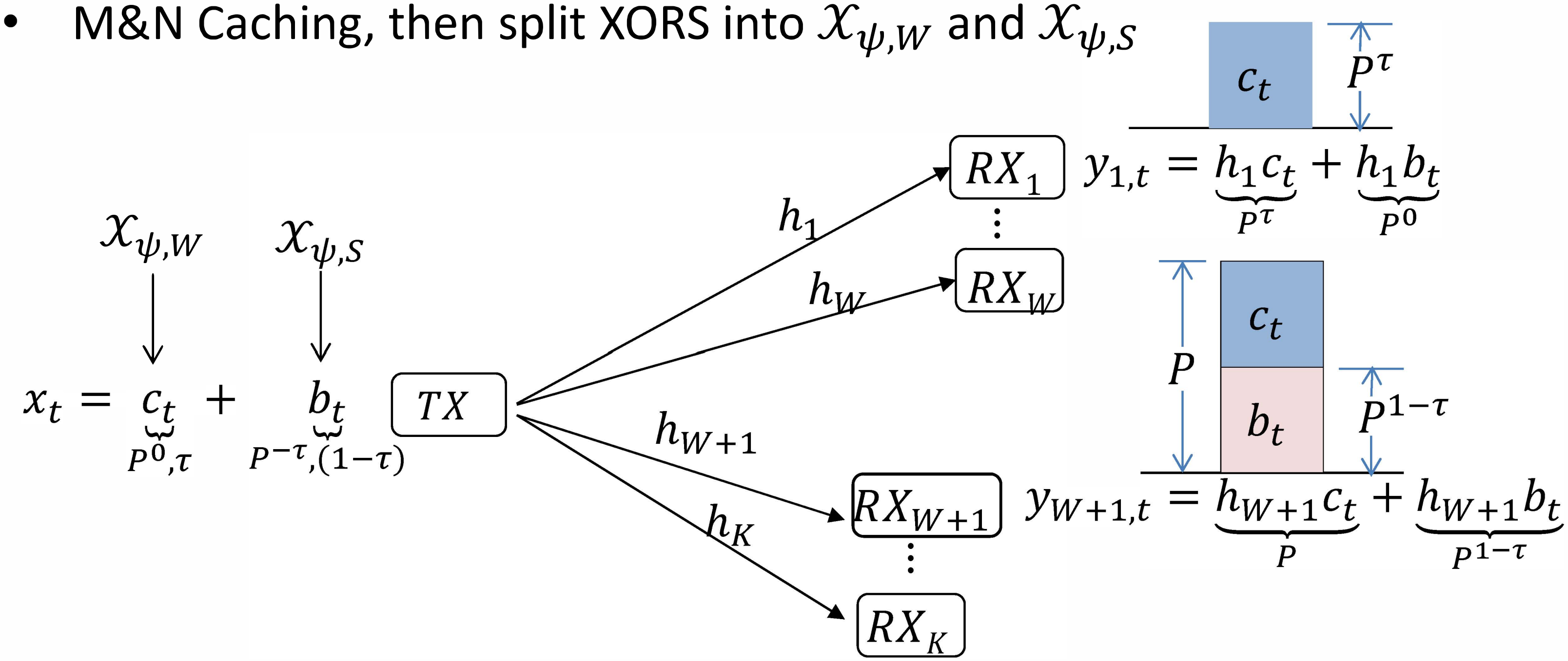


The caching gain for $K = 500, W = 50$

- The horizontal lines denote the maximum gain g_{max} corresponding to $\tau = 1$
- Demonstrate how these can be achieved even with lesser link capacities.

Intuition of the schemes

- M&N Caching, then split XORS into $\mathcal{X}_{\psi,W}$ and $\mathcal{X}_{\psi,S}$



- Interference $\mathcal{X}_{\psi,S}$ hidden from weak users due to topology
  - Treat strong users ($\mathcal{X}_{\psi,S}$) while slowly serving weak ($\mathcal{X}_{\psi,W}$)
  - Transmission rate can be kept (in some cases) at 1 (as if all strong)
  - This ameliorates the negative effects of uneven topology

(Insight)

- For large K (actually for large envisioned gains),… we are in trouble

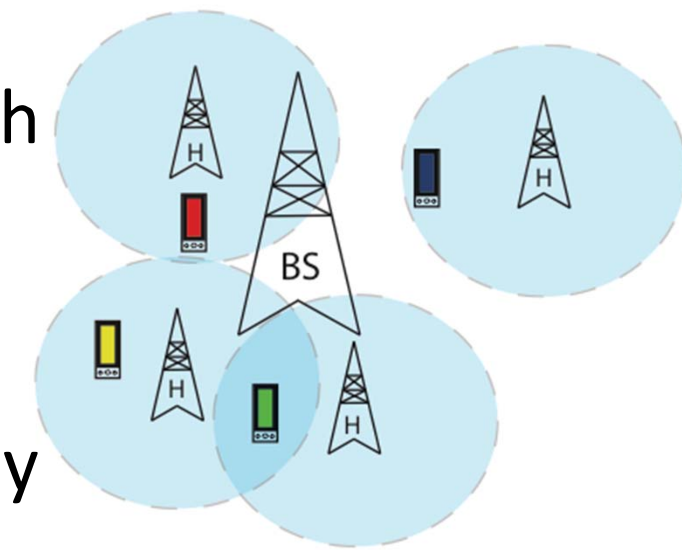
- Else, `worst-user’ effect can be ameliorated
  - **Feature: Sometimes strong users can lift the performance of the weak users** without any penalties on the overall (worst-case) T

Caching in More Involved Networks

Caching at the Edge

FemtoCaching

- 1 BS delivering content to users
- Introduce helper nodes with caches
- Caches are filled with different <u>whole</u> files
- Content follows a popularity distribution
- Users connect to helpers if their requested file is present

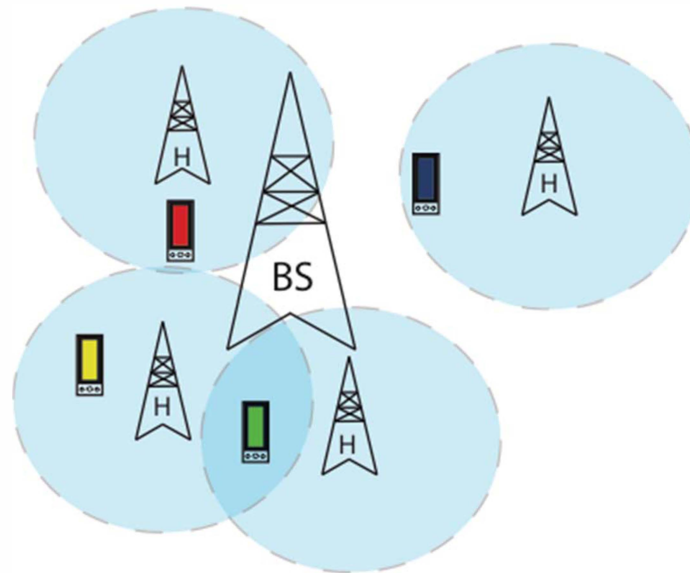


Setting: Golrezaei et al. (2012)

Femtocaching

Main Goal:
Reduce the usage of the BS-UE link

Main Question:
Which files to cache and where?



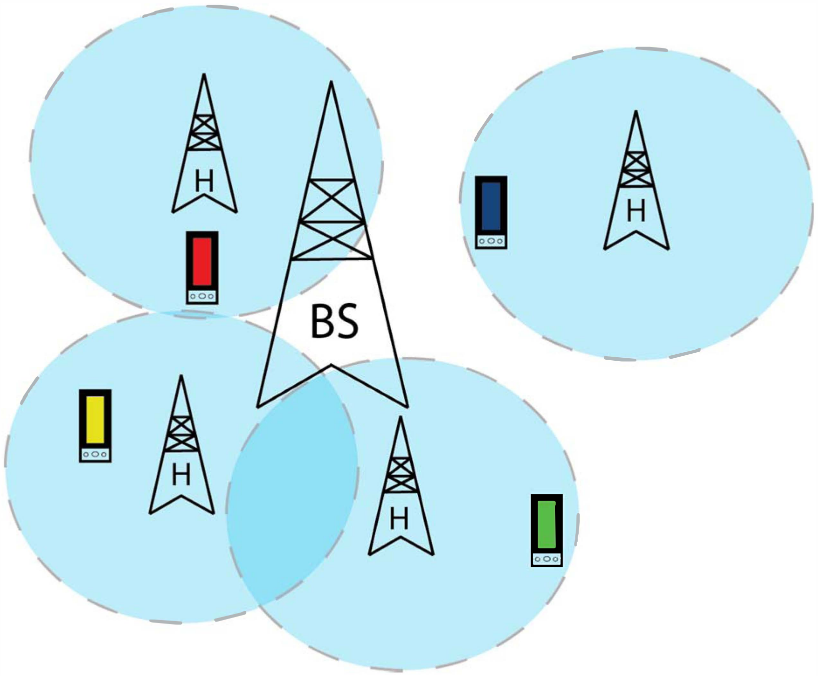
Setting: Golrezaei et al. (2012)

FemtoCaching

Results

- If each user is attached to a single helper, then:

Optimal Solution:
Cache the most popular content



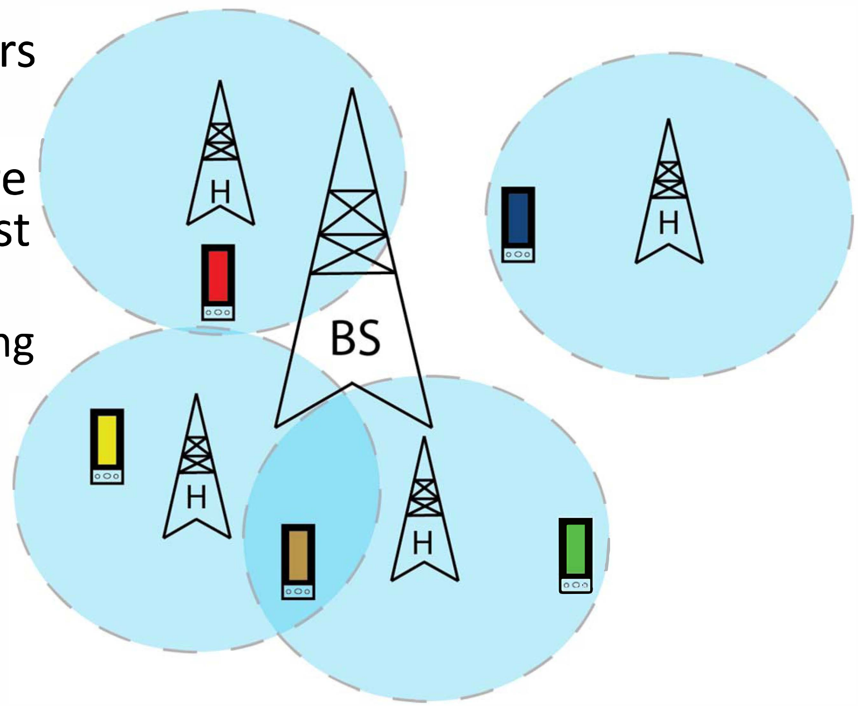
Setting: Golrezaei et al. (2012)

FemtoCaching

Results

- If users could be served by multiple helpers
- Main idea: If 2 or more helper-nodes share 1 or more users, then cache more than just the most popular files
  - Increase the union of caches of neighboring helpers

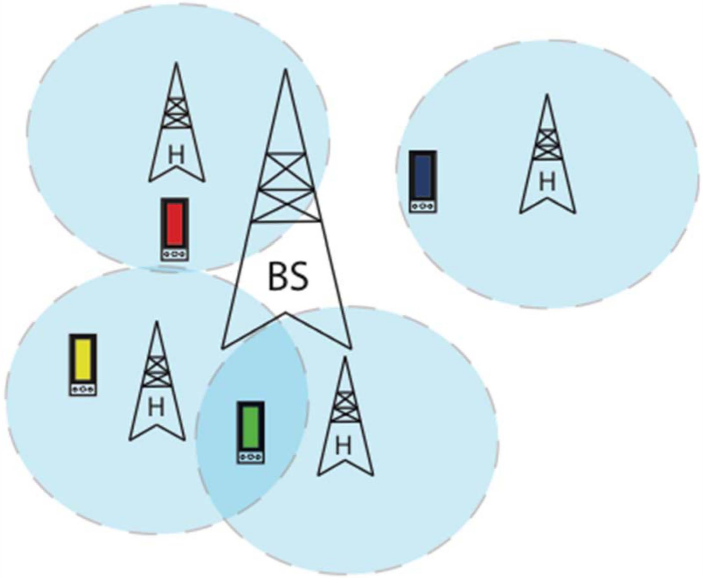
Optimal Caching Solution:
NP-complete



Source: Golrezaei et al. (2012)

FemtoCaching

- Greedy algorithm is 2 from optimal in terms of
  - Hit probability
  - Using the knowledge of user positions



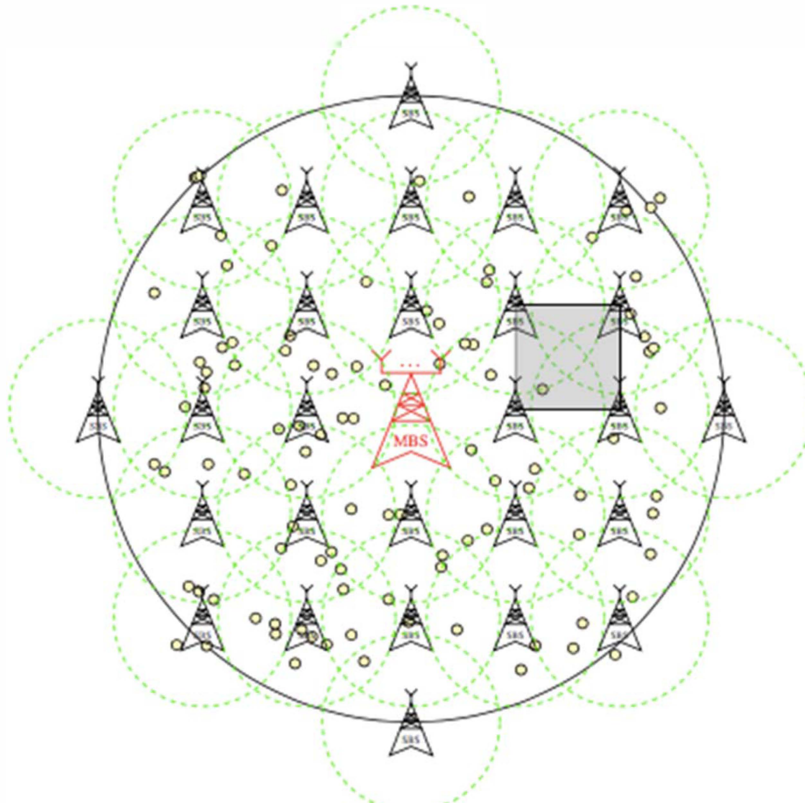
Main Result (simulation):
4-5 times more users served simultaneously

Result contributed substantially to the revival of caching

Result: Golrezaei et al. (2012)

RS-coded Caching at the Edge

- Main BS with all content
- Helper BSs with fraction of content cached
- Users requesting files
- Users can connect to multiple helper BSs, and to the main BS if necessary

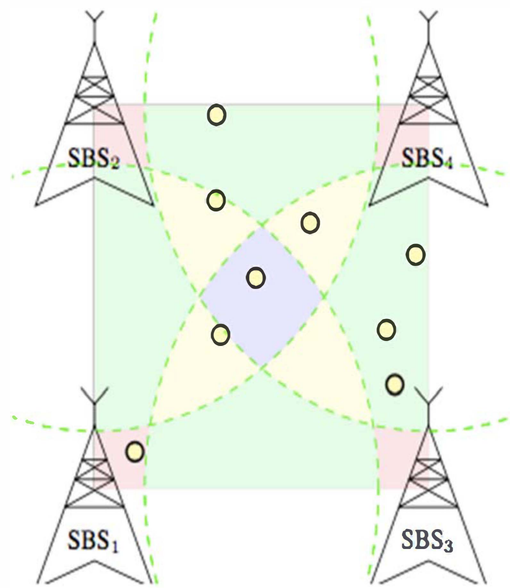


Altman, Avrachenkov, Goseling (2014)
Also Bioglio, Gabry, Land (2015)(image source)

RS-coded Caching at the Edge

- N files, each split in D subfiles
- RS code each file: D→D' subfiles
- Each helper BS gets at least one element of each codeword
  - No overlaps/no content repetition
- User needs only D (out of D') elements of a codeword (RS)
- Look for subfiles in neighboring BS
- The rest from main BS
- Effort – reduce (remaining) amount of information leaving main BS

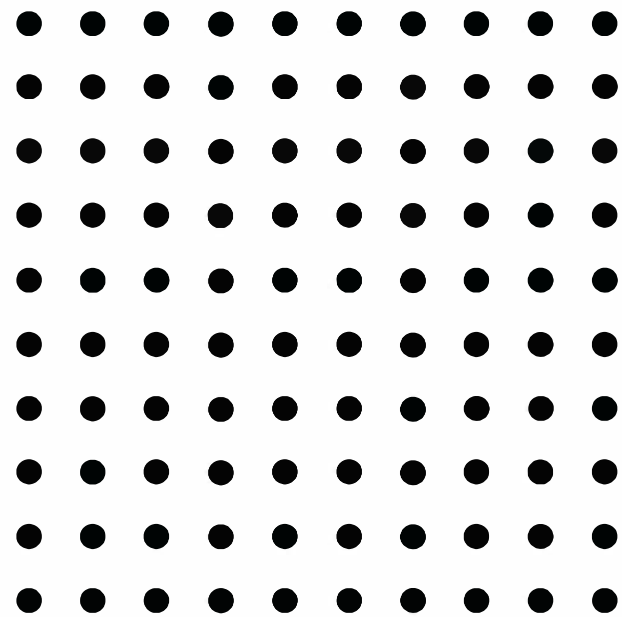
- Simulation results as a function of:
  - radius of vicinity (more HBSs per user)
  - Cache size (increases D')
    - Increases chance to get file from HBs



Altman, E., Avrachenkov, K. and Goseling (2014)
Also Bioglio, Gabry, Land (2015)(image source)

Fundamental Limits of Caching in Wireless D2D Networks

- Users are positioned in a grid
- N files
- γ: fraction of each file pre-cached at each node
- Next day, users can ask for anything
- Variable Tx Radii: with and w/o spatial reuse



- Both decentralized and deterministic cases

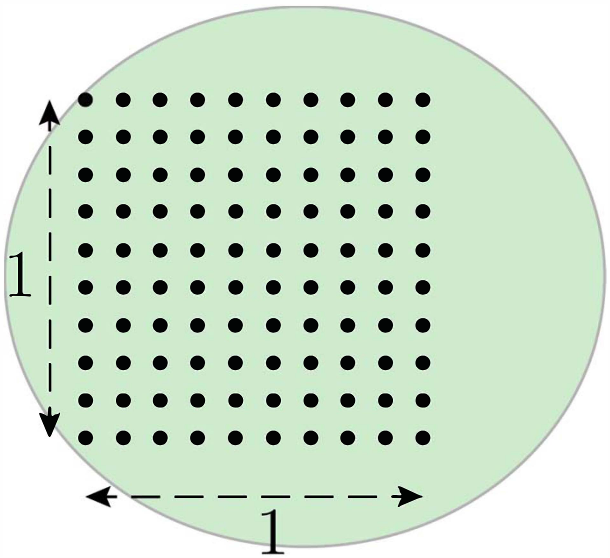
Goal: Delivery of the requested content

79

Setting: Ji, Caire, Molisch (2013)

D2D - No Spatial Reuse Model

- Radius covers the whole network
- One user communicates at a time
- Performance:



$$T(\gamma) = \frac{K(1-\gamma)}{K\gamma} = \frac{1-\gamma}{\gamma}$$

(order optimal)

Result: Ji, Caire, Molisch (2013)

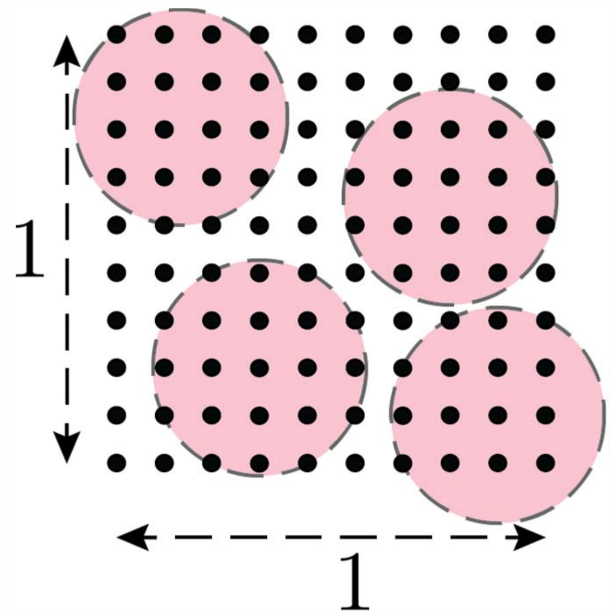
80

D2D - Spatial Reuse Model

- Small radius ensures simultaneous transmissions
- Radius is fixed and same for all users
- Users are clustered and exchange content inside the same cluster
- Radius/memory is big enough to ensure all the library is available inside a cluster
- Performance

$$T(\gamma) = \frac{1-\gamma}{\gamma}$$

(order optimal)



INSIGHT: Multicasting and Spatial Reuse are competing resources

Result: Ji, Caire, Molisch (2013)

D2D - Placement Schemes

Deterministic Placement Scheme

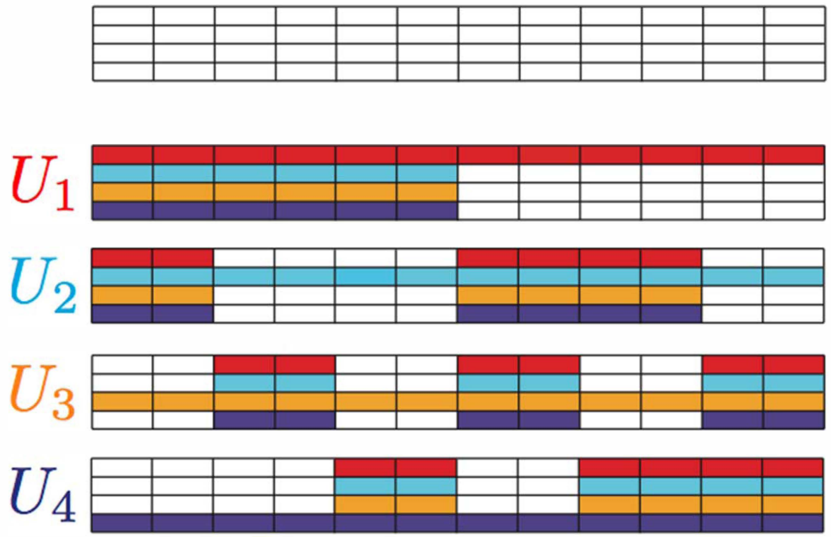
- A variation of original MN, with $K\gamma \cdot \binom{K}{K\gamma}$ subpacketization
- In the case of Spatial Reuse the subpacketization level is reduced compared to MN

Decentralized Placement Scheme

- Files are encoded through an MDS code
- Ensures, with high probability that all the content exists in the network
- Achieves order optimality
- Is considered more practical

82

D2D – Deterministic Delivery Example



Initial Placement with $K\gamma = 2$

$\psi = 123$

User 1 Serves 2 & 3

User 3 Serves 1 & 2

User 2 Serves 1 & 3

$\psi = 124$

User 4 Serves 1 & 2

User 2 Serves 1 & 4

User 1 Serves 2 & 4

$\psi = 134$

User 4 Serves 1 & 3

User 3 Serves 1 & 4

User 1 Serves 3 & 4

$\psi = 234$

User 4 Serves 2 & 3

User 3 Serves 2 & 4

User 2 Serves 3 & 4

83

Setting: Ji, Caire, Molisch (2013)

General Conclusions

Caching in wireless: a set of different challenges

- Several salient features when caching is for wireless
- Certain non-separability between caching and PHY
- Feedback and topology are unexplored frontiers in caching for wireless.
  - Among many interesting differentiating ingredients
- Interesting tradeoffs, synergies, and opportunities

Addressed Misconceptions

- Where to install memory
  - ➢ No need of deploying too many caches due to its costly nature
  - ➢ Now, much higher gains though. Change of mind?

- The differences between wireless and wired caching
  - ➢ Caching is an upper layer problem
  - ➢ Fusion is fascinating, and very powerful

Open Problems and Future Directions

- Different measures of performance (beyond rate, capacity, delay, DoF, etc)
  - Infuse this approach with network-theoretic considerations!!
- Subpacketization bottleneck
  - Perhaps look into coded placement
- Fusing PHY and CC to improve performance and subpacketization
  - Need to boost DoF gains for small γ
  - Under subpacketization constraints
  - Need to explore new cache-endowed powerful PHY resources
- CC in different network topologies.
  - Topologies affect FB, interference, and multicasting capabilities (all connected)
  - Currently worst-channel user `brings down’ the rest. Can this be ameliorated?

Open Problems and Future Directions

- Caching with secure communications (e.g. https)
  – Public key encryption changes files differently at different receivers

- Cost of cache placement
  – Mainly have assumed zero-cost placement
  – Updating is also an issue (see `Online coded caching')

- What is the best way to utilize file popularity and user behavior
  – Open problem and could be key in unlocking CC for commercial use

- Computational and implementation complexity (subpacketization, clique-finding, cache-allocation)

THANKS FOR YOUR ATTENTION!

**Looking for Postdocs and PhD students

References

Limitations of current communication paradigms:

- Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019. White Paper.
- J. Andrews, X. Zhang, G. Durgin, A. Gupta, “Are we approaching the fundamental limits of wireless network densification?,” ArXiv e-prints, Dec. 2015.
- L. Zheng and D. N. C. Tse, “Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," IEEE Trans. Information Theory, Feb. 2002.
- A. Lozano, R. W. Heath Jr., and J. G. Andrews, “Fundamental limits of cooperation,” IEEE Trans. Information Theory, Sep. 2013.
- M. A. Maddah-Ali and D. N. C. Tse, “Completely stale transmitter channel state information is still very useful,” IEEE Trans. Information Theory, Jul. 2012.
- A. Singh, P. Elia, J. Jaldén, “Achieving a vanishing SNR-gap to exact lattice decoding at a subexponential complexity,” IEEE Trans. Information Theory, Jun. 2012.
- A. J. Fehske, et al., “The global footprint of mobile communications: The ecological and economic perspective,” IEEE Communications Magazine, Aug. 2011.
- T. M. Cover and J. A. Thomas, Elements of information theory. John Wiley & Sons, 2012.

References

Feedback communication theoretic solutions:

- V.R. Cadambe, S.A. Jafar, "Interference alignment and degrees of freedom of the K-User interference channel," IEEE Trans. Information Theory, Aug. 2008.
- J. Chen and P. Elia, "Toward the performance versus feedback tradeoff for the two-user MISO broadcast channel," IEEE Trans. Information Theory, Dec. 2013.
- J. Chen, P. Elia, and S. Jafar, "On the two-user MISO broadcast channel with alternating CSIT: A topological perspective," IEEE Trans. Information Theory, Aug. 2015.
- R. Tandon, S. A. Jafar, S. Shamai, and H. V. Poor, "On the synergistic benefits of alternating CSIT for the MISO broadcast channel,'' IEEE Trans. Information Theory, Jul. 2013.
- C. S. Vaze, S. Karmakar, and M. K. Varanasi, "On the generalized degrees of freedom region of the MIMO interference channel with no CSIT," IEEE International Symposium on Information Theory (ISIT), Aug. 2011.
- J. Chen, S. Yang, A. Ozgur, A. Goldsmith, "Achieving Full DoF in Heterogeneous Parallel Broadcast Channels with Outdated CSIT", IEEE Trans. Information Theory, Sep. 2014
- P. de Kerret, D. Gesbert, J. Zhang and P. Elia, "Optimal sum-DoF of the K-user MISO BC with current and delayed feedback," ArXiv e-prints, Apr. 2016

References

Single-stream Coded caching:

- M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," IEEE Trans. Information Theory, May 2014.
- U. Niesen and M. Maddah-Ali, "Coded caching with non-uniform demands," in Computer Communications Workshops (INFOCOM WKSHPS), May 2014.
- M. Maddah-Ali and U. Niesen, "Decentralized caching attains order-optimal memory-rate tradeoff," Allerton Conference, Monticello, IL, Oct. 2013 – see also ArXiv e-prints, 2013.
- K.Wan, D. Tuninetti, and P. Piantanida, "On Caching with More Users than Files", ArXiv e-prints, Jan. 2016.
- K. Shanmugam, M. Ji, A. M.Tulino, J. Llorca, and A. G. Dimakis, "Finite length analysis of caching-aided coded multicasting," ArXiv e-prints, Aug. 2015.
- Z. Chen, P. Fan, and K.B. Letaief, "Fundamental Limits of Caching: Improved Bounds For Small Buffer Users", ArXiv e-prints, Nov. 2015.
- M. Ji, A. M. Tulino, J. Llorca, G. Caire, "Order-Optimal Rate of Caching and Coded Multicasting with Random Demands", ArXiv e-prints, Feb. 2015.
- S. Sahraei, M. Gastpar, "Multi-Library Coded Caching", ArXiv e-prints, Jan. 2016

References

Single-stream Coded caching:

- M. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” IEEE Trans. Information Theory, May 2014.
- M. Mohammadi Amiri, D. Gunduz, “Fundamental Limits of Caching: Improved Delivery Rate-Cache Capacity Trade-off”, ArXiv e-prints, Apr. 2016.
- R. Pedarsani, M. Ali Maddah-Ali, U. Niesen, “Online Coded Caching”, ArXiv e-prints, Nov. 2013.
- J. Hachem, N. Karamchandani, S. Diggavi, “Effect of Number of Users in Multi-Level Coded Caching”, ArXiv e-prints, Apr. 2015
- J. Zhang, X. Lin, X. Wang, “Coded Caching under Arbitrary Popularity Distributions”, Information Theory and Applications Workshop (ITA), Feb. 2015
- C. Wang, S. Lim, M. Gastpar, “Information-Theoretic Caching: Sequential Coding for Computing”, ArXiv e-prints, Feb. 2016
- U. Niesen, M. Maddah-Ali, “Coded Caching for Delay-Sensitive Content”, ArXiv e-prints, Jul. 2014.
- S. Bidokhti, M. Wigger, R. Timo, “Noisy Broadcast Networks with Receiver Caching”, ArXiv e-prints, May. 2016.

References

Extensions of Coded Caching in different settings:

- R. Timo and M. A. Wigger, “Joint cache-channel coding over erasure broadcast channels,” ArXiv e-prints, May 2015.
- A. Ghorbel, M. Kobayashi, S. Yang, “Cache-Enabled Broadcast Packet Erasure Channels with State Feedback”, Allerton Conference, Monticello, IL, Oct. 2015.
- N. Karamchandani, U. Niesen, M. Maddah-Ali, S. Diggavi, “Hierarchical Coded Caching”, ArXiv e-prints, Jun 2014.
- K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, G. Caire"FemtoCaching: Wireless video content delivery through distributed caching helpers", ArXiv e-prints, Sep 2013
- M. Ji, G. Caire, A. F. Molisch, “Fundamental limits of distributed caching in D2D wireless networks”, ArXiv e-prints, Apr 2013
- Y. Ugur, Z. Awan, A. Sezgin, "Cloud Radio Access Networks With Coded Caching", ArXiv e-prints Feb 2016
- S. Lim, C. Wang, M. Gastpar, “Information Theoretic Caching: The Multi-User Case”, ArXiv e-prints Apr 2016

References

- V. Bioglio, F. Gabry, I. Land, "Optimizing MDS Codes for Caching at the Edge", ArXiv e-prints Sep 2015.
- Altman, E., Avrachenkov, K. and Goseling, J. “Distributed storage in the plane”. In Networking Conference, 2014 IFIP (pp. 1-9). IEEE.
- Avrachenkov, K., Bai, X. and Goseling, J., 2016. “Optimization of Caching Devices with Geometric Constraints,” arXiv preprint arXiv:1602.03635.
- B. Perabathini, E. Başţuğ, M. Kountouris, M. Debbah, A. Conte, “Caching at the Edge: a Green Perspective for 5G Networks”, ArXiv e-prints Mar 2015.
- M. Deghel, E. Bastug, M. Assaad, and M. Debbah, “On the benefits of edge caching for MIMO interference alignment,” in Signal Processing Advances in Wireless Communications (SPAWC), Jun. 2015.
- A. Liu, V.t K. N. Lau, “Cache-Enabled Opportunistic Cooperative MIMO for Video Streaming in Wireless Systems”, IEEE Trans. Signal Processing, Jan. 2014.

References

Coded Caching with Feedback:

- S. P. Shariatpanahi, A. S. Motahari, and B. H. Khalaj, "Multi-server coded caching," ArXiv e-prints, Aug. 2015.
- A. Liu, V. K. N. Lau, "Mixed-Timescale Precoding and Cache Control in Cached MIMO Interference Network", IEEE Trans. Signal Processing, Dec. 2013.
- Maddah-Ali and Niesen, "Cache-Aided Interference Channels", 2015.
- J. Zhang, F. Engelmann and P. Elia, "Coded caching for reducing CSIT-feedback in wireless communications," Allerton Conference, Monticello, IL, Oct. 2015.
- J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback," ArXiv e-prints, Nov. 2015.
- J. Zhang and P. Elia, "The synergistic gains of coded caching and delayed feedback," ArXiv e-prints, Apr. 2016
- N. Naderializadeh, M. Maddah-Ali and A. Avestimehr, "Fundamental Limits of Cache-Aided Interference Management", ArXiv e-prints, Apr. 2015.
- G. Paschos, E. Baştuğ, I. Land, G. Caire, M. Debbah, "Wireless Caching: Technical Misconceptions and Business Barriers", ArXiv e-prints, Jan. 2016.
- M. Wigger, R. Timo, S. Shamai (Shitz), "Complete Interference Mitigation Through Receiver-Caching in Wyner's Networks", ArXiv e-prints, May. 2016.

References

Outer bounds of Coded Caching

- H. Ghasemi and A. Ramamoorthy, “Improved lower bounds for coded caching,” ArXiv e-prints, Jan. 2015.
- R. Timo and M. A. Wigger, “Joint cache-channel coding over erasure broadcast S. Lim, C. Wang, M. Gastpar, “Information Theoretic Caching: The Multi-User Case”, ArXiv e-prints Apr. 2016
- K. Wan, D. Tuninetti, P. Piantanida, “On the Optimality of Uncoded Cache Placement”, ArXiv e-prints, Nov. 2015.
- A. N., N. S. Prem, V. M. Prabhakaran, R. Vaze, “Critical Database Size for Effective Caching”, ArXiv e-prints, Jan. 2015.
- Q. Yan, X. Tang, Q. Chen, "On the Gap Between Decentralized and Centralized Coded Caching Schemes," Arxiv May 2016.