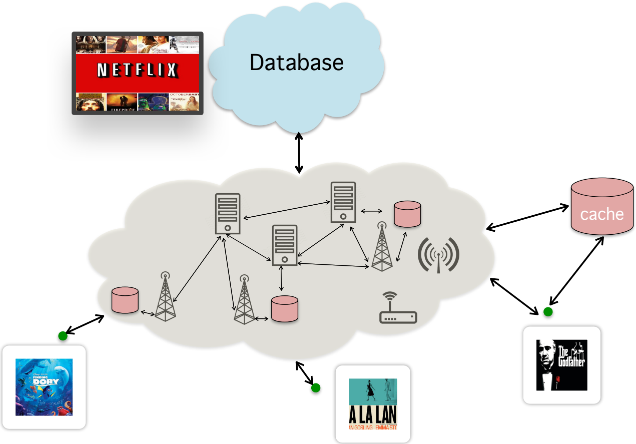
Caching Networks: Low-Subpacketization Schemes and Improved Delivery Methods

Michèle Wigger
Telecom ParisTech

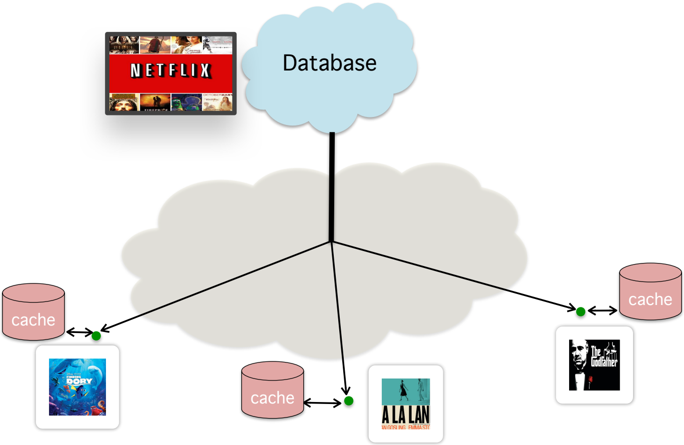
Joint work with S. Bidokhti-Saeedi, R. Timo, Q. Yan, S. Yang, A. Yener

8th Spanish Workshop on Signal Processing, Communication and Information
14 September 2018

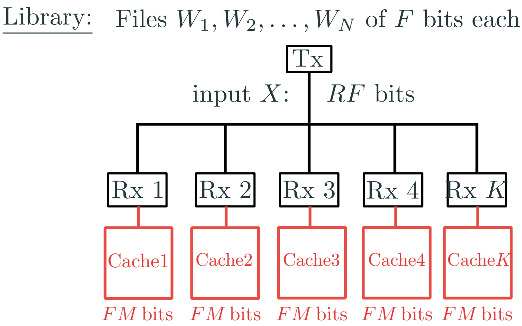
Caching in Networks



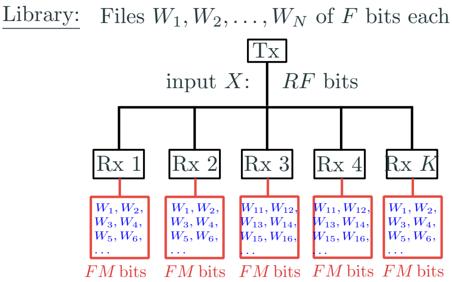
One-To-Many Caching Network



A Simple Network

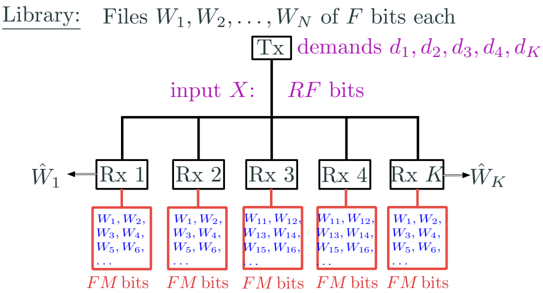


A Simple Network



1) Placement phase: Tx fills caches without knowing which receiver demands which message: $Z_k = g_k(W_1, \ldots, W_N)$

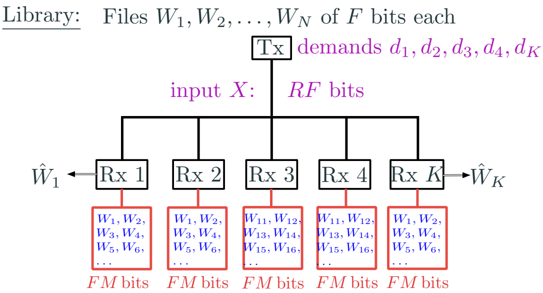
A Simple Network



2) Delivery phase:

- Receiver k wants file W_{d_k} $\rightarrow$ sends demand d_k to transmitter
- Tx sends input $X = f(W_1, \ldots, W_N, d_1, \ldots, d_K)$
- Rx k produces $\hat{W}_k = \varphi_k(X, Z_k, d_1, \ldots, d_K)$.

A Simple Network



Goal: $\hat{W}_k = W_{d_k}$ for all $k = 1, \ldots, K$

Fundamental Rate-Memory Tradeoff $R^\star(M)$

$$R^\star(M) \quad := \quad \min\big\{R\colon\ \text{such that for } (R, M) \text{ each Rx } k \text{ can learn } W_{d_k}\big\}$$

Some properties:

- $R^\star(M)$ is decreasing in M.
- $R^\star(M)$ is bounded above by $\min\{N, K\}$. Moreover:
$$R^\star(M = 0) = \min\{N, K\}.$$
- $R^\star(M)$ is nonnegative. Moreover:
$$R^\star(M) = 0, \qquad \forall M \geq N.$$

Traditional Uncoded Scheme for K Receivers

- Split W_d into $(W_d^{(1)}, W_d^{(2)})$ of sizes $F\frac{M}{N}$ and $F\left(1-\frac{M}{N}\right)$ bits
- For $d = 1, \ldots, N$: cache part $W_d^{(1)}$ *at all* rxs
- Deliver part $W_d^{(2)}$ for each demanded message W_d.
  - If $K \geq N$, in the worst case:

$$X = \left(W_1^{(2)},\ W_2^{(2)}, \ldots,\ W_N^{(2)}\right).$$

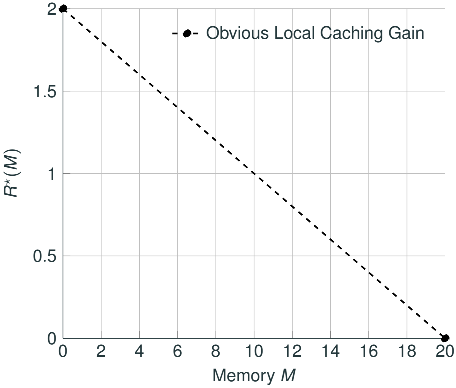
  - If $K < N$, in the worst case:

$$X = \left(W_{d_1}^{(2)},\ W_{d_2}^{(2)}, \ldots,\ W_{d_K}^{(2)}\right).$$

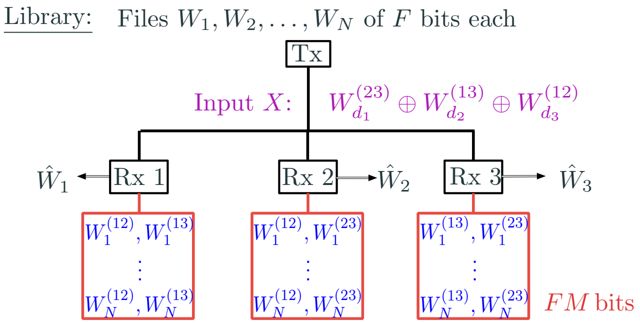
Trivial Upper Bound on $R^\star(M)$

$$R^\star(M) \leq \min\{K, N\}\left(1 - \frac{M}{N}\right).$$

- $N = 20$ files and $K = 2$ Users



Coded caching for $K = 3$ Receivers, Parameter $t = 2$

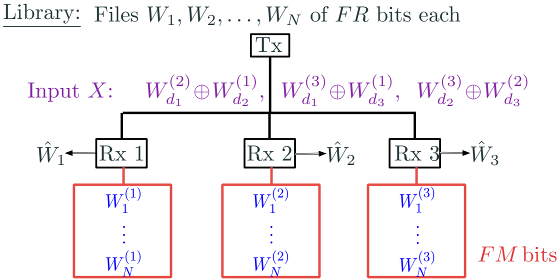


- Split W_d into three parts ($W_d^{(12)}, W_d^{(13)}, W_d^{(23)}$) each of $\frac{F}{3}$ bits

Achieves Rate-Memory Pair $M = \frac{2N}{3}$ and $R = \frac{1}{3}$.

[1] M. A. Maddah-Ali, U. Niesen, “Fundamental Limits of Caching.” *IT-Trans 2014*

Coded caching for $K = 3$ Receivers, Parameter $t = 1$

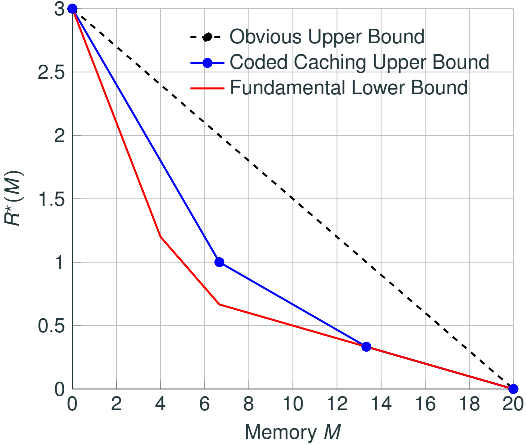


- Split W_d into three parts $(W_d^{(1)}, W_d^{(2)}, W_d^{(3)})$ each $\frac{F}{3}$ bits

Achieves Rate-Memory Pair $M = \frac{N}{3}$ and $R = 1$.

[1] M. A. Maddah-Ali, U. Niesen, "Fundamental Limits of Caching." *IT-Trans 2014*

Bounds for 3 Users ($N = 20$ files)



Lower bound on $R^\star(M)$

Theorem

$$R^\star(M) \geq \max\left\{ \max_{\ell\in\bar{\mathcal{N}}}\left[\ell - \mathrm{M}\frac{\ell^2}{N}\right],\ \max_{\ell\in\bar{\mathcal{N}}}\left[\ell - \mathrm{M}\sum_{k=1}^{\ell}\frac{k}{N-k+1}\right]\right\}$$

[2] C.-Y. Wang et al. “Improved Converses and Gap-Results for Coded Caching”, IT-Trans 2018.

[3] Q. Yu, “Characterizing the rate-memory tradeoff in cache networks within a factor of 2,” ISIT 2017.

Proof of Lower Bound on $R^\star(M)$

- Fix $\ell = 1, \ldots, \min\{K, N\}$ and demands $(d_1, \ldots, d_\ell)$:

$$\begin{aligned} FR &\geq H(X) \geq I(X; W_{d_1}, \ldots, W_{d_\ell}, Z_1, \ldots, Z_\ell) \geq \ldots \\ &= H(W_{d_1}, \ldots, W_{d_\ell}) - \sum_{k=1}^{\ell} I(W_{d_k}; Z_1, \ldots, Z_k | W_{d_1}, \ldots, W_{d_{k-1}}) \end{aligned}$$

- Average over all demand vectors $(d_1, \ldots, d_\ell)$:

$$R \geq \ell - \sum_{k=1}^{\ell} \underbrace{\frac{1}{\binom{N}{\ell}\ell!} \sum_{(d_1,\ldots,d_\ell)} \frac{1}{F} I(W_{d_k}; Z_1, \ldots, Z_k | W_{d_1}, \ldots, W_{d_{k-1}})}_{=:\alpha_k}$$

- By the chain rule, Han's inequality, and counting arguments:

$$\sum_{k=1}^{\ell} \alpha_k \leq \min\left\{ \frac{\ell^2}{N} M, \ \sum_{k=1}^{\ell} \frac{kM}{N-k+1} \right\}.$$

Coded Caching for K Users (Maddah-Ali &Niesen IT-Trans 2014)

- Parameter $t \in \{1, \ldots, K-1\}$

- *Placement:* Split each W_d into $\binom{K}{t}$ parts and save each part at a different subset of receivers
  Let for each size-t subset $\mathcal{G}$ denote $W_d^{\mathcal{G}}$ the part of W_d placed in caches of all receivers in $\mathcal{G}$.

- *Delivery transmission:* For each set $\mathcal{S} = \{s_1, \ldots, s_{t+1}\}$, send

$$W_{\text{XOR},\mathcal{S}} := \bigoplus_{\ell=1}^{t+1} W_{d_{s_\ell}}^{(\mathcal{S}\backslash\{s_\ell\})}$$

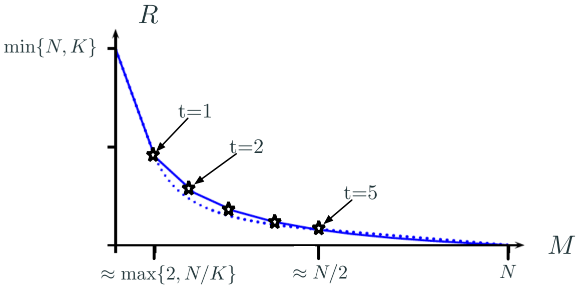
- *Delivery reception:* Receiver s_j has stored in its cache memory

$$W_{d_{s_\ell}}^{(\mathcal{S}\backslash\{s_\ell\})}, \qquad \forall \ell \in \{1, \ldots, j-1, j+1, \ldots, t\}.$$

  So, with $W_{\text{XOR},\mathcal{S}}$ it can recover $W_{d_{s_j}}^{(\mathcal{S}\backslash\{s_j\})}$ and $W_{d_{s_j}}$.

Performance of Coded Caching

- $K = 6$



Coded Caching Upper Bound

For all $M \in \frac{1}{N} \cdot \{0, 1, \ldots, K-1, K\}$:

$$R^{\star}(M) \leq \min\left\{K\left(1 - \frac{M}{N}\right)\left(1 + \frac{MK}{N}\right)^{-1},\ N\left(1 - \frac{M}{N}\right)\right\}.$$

Optimality and Improvements

- Optimal under uncoded placement
- Coded placement can improve performance
- In general within a factor of 2.009 from optimal

Main Problem: Subpacketization Level

Large files required that can be split into $\binom{K}{t}$ packets

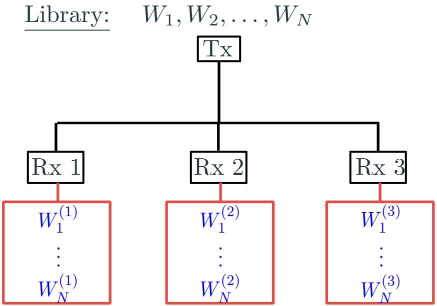
$\rightarrow$ Use placement delivery arrays (PDAs) to find solution

[4] K. Wan et al. "On the optimality of uncoded cache placement", ITW 2016.

[5] Q. Yu et al, "The exact rate-memory tradeoff for caching with uncoded prefetching" IT-Trans 2018.

[6] J. Gomez Vilardebo, "A novel centralized coded caching scheme with coded prefetching" JSAC on Comm.

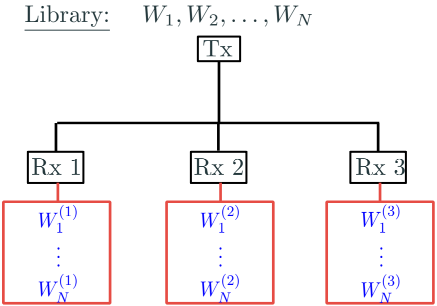
PDA-Example for coded caching with $K = 3$ and $t = 1$



packet\Rx	Rx1	Rx2	Rx3
$W^{(1)}$			
$W^{(2)}$			
$W^{(3)}$			

- PDA represents both placement and delivery

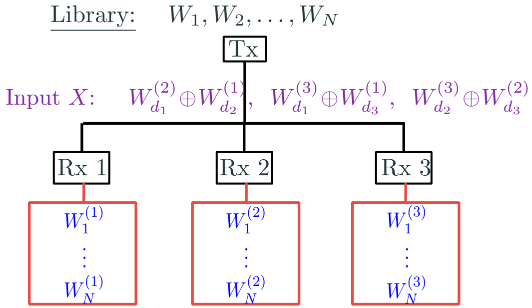
PDA-Example for coded caching with $K = 3$ and $t = 1$



packet\Rx	Rx1	Rx2	Rx3
$W^{(1)}$	$*$		
$W^{(2)}$		$*$	
$W^{(3)}$			$*$

- PDA represents both placement and delivery

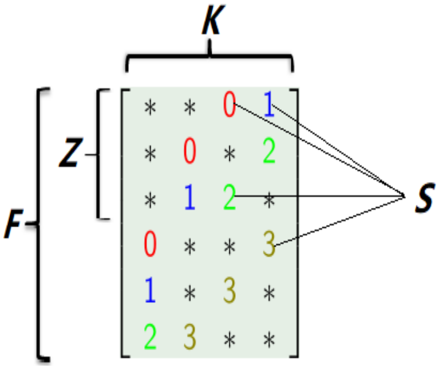
PDA-Example for coded caching with $K = 3$ and $t = 1$



packet\Rx	Rx1	Rx2	Rx3
$W^{(1)}$	$*$	1	2
$W^{(2)}$	1	$*$	3
$W^{(3)}$	2	3	$*$

- PDA represents both placement and delivery

Definition: (K, F, Z, S) PDA



- Any two non-star symbols in each row/column are distinct;
- If $p_{a,b} = p_{c,d} = s \neq *$, then $p_{a,d} = p_{b,c} = *$;
- Regular PDAs: each symbol s occurs g times (coding gain)

---

[7] Q. Yan et al. "On the placement delivery array design for centralized coded caching schemes," *IT-Trans 2017*

Connection between PDA and Caching Networks

Theorem

Given a (K, F, Z, S) PDA.

- *The corresponding caching scheme has $M = \frac{NZ}{F}$ and $R = \frac{S}{F}$.*

[8] M. Cheng et al. , “Coded caching schemes with low rate and subpacketizations”, Arxiv.

Connection between PDA and Caching Networks

Theorem

Given a (K, F, Z, S) *PDA.*

- *The corresponding caching scheme has* $M = \frac{NZ}{F}$ *and* $R = \frac{S}{F}$*.*
- *The rate* $R = \frac{S}{F}$ *cannot be smaller than*
  $\frac{S}{F} \overset{(a)}{\geq} K\left(1 - \frac{M}{N}\right)\left(1 + K\frac{M}{N}\right)^{-1}$

[8] M. Cheng et al. , “Coded caching schemes with low rate and subpacketizations”, Arxiv.

Connection between PDA and Caching Networks

Theorem

Given a (K, F, Z, S) PDA.

- *The corresponding caching scheme has $M = \frac{NZ}{F}$ and $R = \frac{S}{F}$.*
- *The rate $R = \frac{S}{F}$ cannot be smaller than*
  $\frac{S}{F} \overset{(a)}{\geq} K\left(1 - \frac{M}{N}\right)\left(1 + K\frac{M}{N}\right)^{-1}$
- *If equality holds in (a), then $F \overset{(b)}{\geq} \binom{K}{\frac{KM}{N}}$*

[8] M. Cheng et al. , “Coded caching schemes with low rate and subpacketizations”, Arxiv.

Connection between PDA and Caching Networks

Theorem

Given a (K, F, Z, S) PDA.

- *The corresponding caching scheme has* $M = \frac{NZ}{F}$ *and* $R = \frac{S}{F}$*.*
- *The rate* $R = \frac{S}{F}$ *cannot be smaller than* $\frac{S}{F} \overset{(a)}{\geq} K\left(1 - \frac{M}{N}\right)\left(1 + K\frac{M}{N}\right)^{-1}$
- *If equality holds in* (a)*, then* $F \overset{(b)}{\geq} \binom{K}{\frac{KM}{N}}$
- *Maddah-Ali and Niesen's coded caching achieves equality in* (a) *and* (b)

[8] M. Cheng et al. , "Coded caching schemes with low rate and subpacketizations", Arxiv.

Constructions of PDA for low subpacketization schemes

Theorem

For any $q, m \in \mathbb{N}^+$, $q \geq 2$, *there exists a* $\big(q(m+1), q^m, q^{m-1}, q^{m+1} - q^m\big)$ *PDA with rate* $R = q - 1$.

Theorem

Given $q, m \in \mathbb{N}^+$, $q \geq 2$, *there exists a* $(q(m+1), (q-1)q^m, (q-1)^2 q^{m-1}, q^m)$ *PDA with rate* $R = 1/(q-1)$.

---

[7] Q. Yan et al. "On the placement delivery array design for centralized coded caching schemes," *IT-Trans 2017*

Comparison with an Example ($K = 6, \frac{M}{N} = \frac{1}{2}$)

$$
\begin{bmatrix}
* & * & * & 0 & 1 & 2 \\
* & * & 0 & * & 3 & 4 \\
* & * & 1 & 3 & * & 5 \\
* & * & 2 & 4 & 5 & * \\
* & 0 & * & * & 6 & 7 \\
* & 1 & * & 6 & * & 8 \\
* & 2 & * & 7 & 8 & * \\
* & 3 & 6 & * & * & 9 \\
* & 4 & 7 & * & 9 & * \\
* & 5 & 8 & 9 & * & * \\
0 & * & * & * & 10 & 11 \\
1 & * & * & 10 & * & 12 \\
2 & * & * & 11 & 12 & * \\
3 & * & 10 & * & * & 13 \\
4 & * & 11 & * & 13 & * \\
5 & * & 12 & 13 & * & * \\
6 & 10 & * & * & * & 14 \\
7 & 11 & * & * & 14 & * \\
8 & 12 & * & 14 & * & * \\
9 & 13 & 14 & * & * & *
\end{bmatrix}
\quad \textit{versus} \quad
\begin{bmatrix}
* & 0 & * & 2 & * & 1 \\
1 & * & * & 3 & 0 & * \\
2 & * & 0 & * & * & 3 \\
* & 3 & 1 & * & 2 & *
\end{bmatrix}
$$

R=0.7

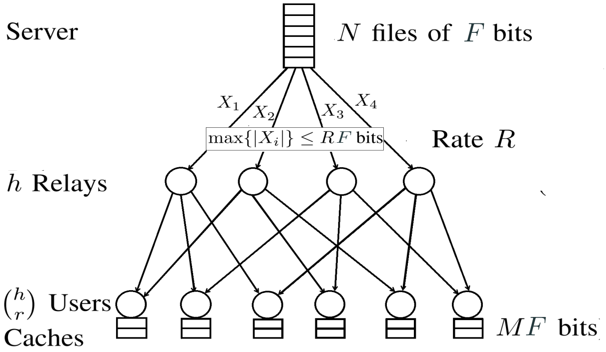
R= 1

General Performance Comparison, $K = q(m+1)$

	Performance	Maddah-Ali-Niesen scheme	New scheme
$\frac{M}{N} = \frac{1}{q}$	g	$\frac{K}{q} + 1$	$\frac{K}{q}$
	R	$\frac{K}{K+q}(q-1)$	$q-1$
	F	$\sim \frac{q}{\sqrt{2\pi K(q-1)}} \cdot q^{\frac{K}{q}} \cdot \left(\frac{q}{q-1}\right)^{K(1-\frac{1}{q})}$	$q^{\frac{K}{q}-1}$
$\frac{M}{N} = \frac{q-1}{q}$	g	$K\frac{q-1}{q} + 1$	$K\frac{q-1}{q}$
	R	$\frac{K}{q+K(q-1)}$	$\frac{1}{q-1}$
	F	$\sim \frac{q}{\sqrt{2\pi K(q-1)}} \cdot q^{\frac{K}{q}} \cdot \left(\frac{q}{q-1}\right)^{K(1-\frac{1}{q})}$	$(q-1)q^{\frac{K}{q}-1}$

Both constructions: $\lim_{K\to\infty} \frac{R_{\text{MN}}}{R_{\text{New}}} = 1, \qquad \lim_{K\to\infty} \frac{F_{\text{MN}}}{F_{\text{New}}} = \infty.$

Combination Networks

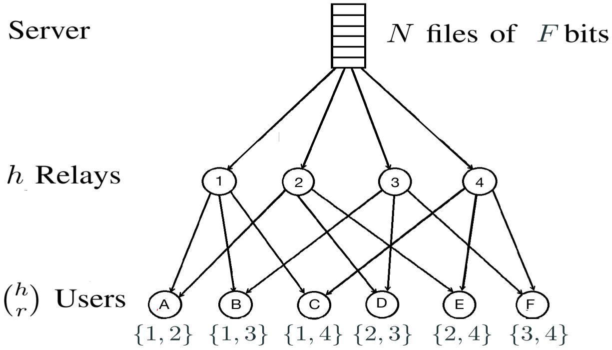


- Individual links from server to relays
- Each user connected to a different subset of *r* relays
- Relays simply forward incoming information (shared link model)
- Rate-memory tradeoff $R^\star(M)$

First Thoughts

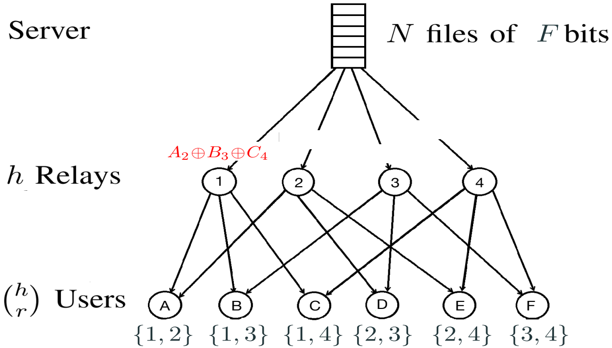
- Can use schemes from single-link model and route packets through network $\rightarrow$ same packet can occupy multiple resources!
- Design packets that can be routed over a single relay
- $R^\star(N) = 0$ and $R^\star(0) = \frac{K}{h}$ if $K < N$
- Traditional uncoded caching $R^\star(M) \leq \frac{K}{h}\left(1 - \frac{M}{N}\right)$ if $K < N$

Designing a Scheme using PDAs



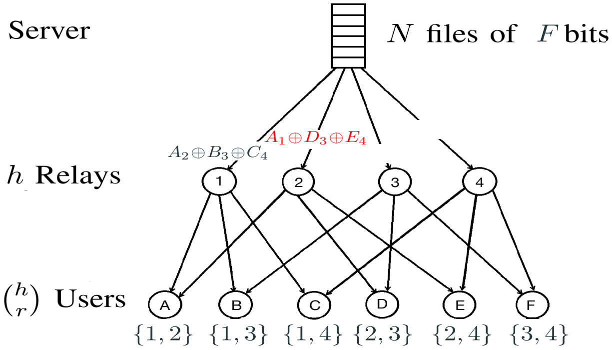
packet\Rx	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
L_1	1	2	3	*	*	*
L_2	0	*	*	2	3	*
L_3	*	0	*	1	*	3
L_4	*	*	0	*	1	2

Designing a Scheme using PDAs



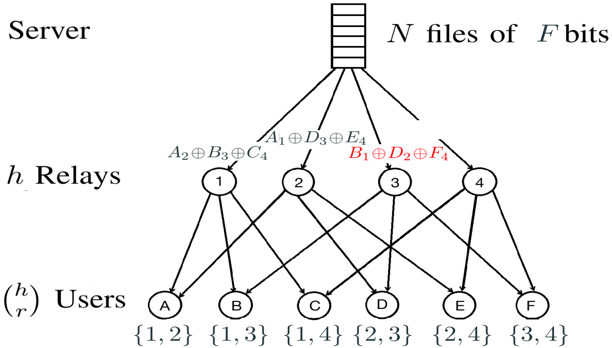
packet\Rx	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
L_1	1	2	3	*	*	*
L_2	0	*	*	2	3	*
L_3	*	0	*	1	*	3
L_4	*	*	0	*	1	2

Designing a Scheme using PDAs



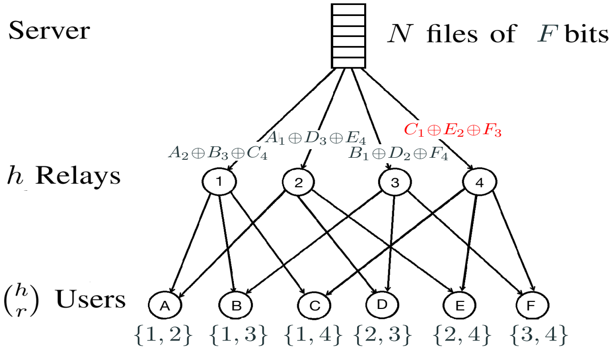
packet\Rx	{1, 2}	{1, 3}	{1, 4}	{2, 3}	{2, 4}	{3, 4}
L_1	1	2	3	*	*	*
L_2	0	*	*	2	3	*
L_3	*	0	*	1	*	3
L_4	*	*	0	*	1	2

Designing a Scheme using PDAs



packet\Rx	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
L_1	1	2	3	*	*	*
L_2	0	*	*	2	3	*
L_3	*	0	*	1	*	3
L_4	*	*	0	*	1	2

Designing a Scheme using PDAs



packet\Rx	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
L_1	1	2	3	*	*	*
L_2	0	*	*	2	3	*
L_3	*	0	*	1	*	3
L_4	*	*	0	*	1	2

Combinational-PDA $\longleftrightarrow$ Combination Network

Definition

A PDA is called *combinational PDA (C-PDA)*, if its columns can be labeled by relay subsets of cardinality r, in a way that for any integer s, the labels of all columns containing s have nonempty intersection.

Theorem

Given a (K, F, Z, S) *C-PDA. For any* (h, r) *combination network with* $K = \binom{h}{r}$*, it holds that*

$$R^\star\left(M = \frac{N \cdot Z}{F}\right) \leq \frac{S}{Fh}.$$

C-PDA Schemes Optimal for Large Cache Memories

Theorem

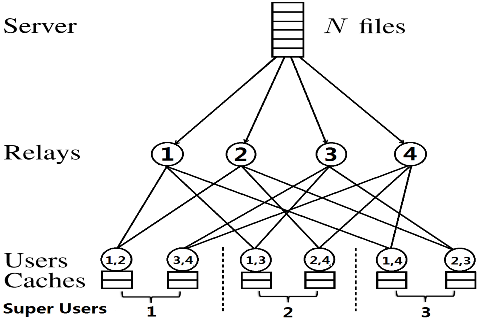
For (h, r)-combination network:

$$R^\star(M) = \frac{1}{r}\left(1 - \frac{M}{N}\right), \quad M \in \left[N\frac{K-h+r-1}{K},\ N\right].$$

Achieved with subpacketization level $F = \binom{h}{r-1}$ when $M = N\frac{K-h+r-1}{K}$.

- Example PDA from before achieves this performance

Resolvable Combination Networks $r|h$



Network is resolvable if $r|h$. Then users can be partitioned s.t.:

- Any subset of users connects to all relays
- Different users of a subset connect to different relays

How to Exploit Resolvability

- Let relay *i* serve the single user in each subset connected to it
- Let each relay act as a server in a single-shared link
- Design a PDA for each relay

Example: $h = 4, r = 2$

$\{1,2\}$	$\{3,4\}$	$\{1,3\}$	$\{2,4\}$	$\{1,4\}$	$\{2,3\}$
*	*	1	4	2	5
1	7	*	*	3	6
2	8	3	6	*	*
*	*	7	10	11	8
4	10	*	*	12	9
5	11	9	12	*	*

How to Exploit Resolvability

- Let relay *i* serve the single user in each subset connected to it
- Let each relay act as a server in a single-shared link
- Design a PDA for each relay

Example: $h = 4, r = 2$

$$\begin{bmatrix} * & 1 & 2 \\ 1 & * & 3 \\ 2 & 3 & * \end{bmatrix} \Rightarrow$$

$\{1,2\}$	$\{3,4\}$	$\{1,3\}$	$\{2,4\}$	$\{1,4\}$	$\{2,3\}$
*	*	1	4	2	5
1	7	*	*	3	6
2	8	3	6	*	*
*	*	7	10	11	8
4	10	*	*	12	9
5	11	9	12	*	*

Transforming PDAs into C-PDAs for Resolvable Networks

- Replicate a $(\binom{h}{r}\frac{r}{h}, \tilde{F}, \tilde{Z}, \tilde{S})$ PDA a number of $\frac{h}{r} \cdot r$ times
- Distribute the columns of the replica PDAs so that the columns of each symbol s have non-empty intersection.

Theorem

Given a $(\binom{h}{r}\frac{r}{h}, \tilde{F}, \tilde{Z}, \tilde{S})$ PDA. There exists a (K, F, Z, S) C-PDA, for a resolvable (h, r)-combination network (i.e., where $r|h$) with

$$K = \binom{h}{r}, \quad F = r\tilde{F}, \quad Z = r\tilde{Z}, \quad \text{and} \quad S = h\tilde{S}.$$

Pair $(M = N\frac{\tilde{Z}}{\tilde{F}}, R = \frac{\tilde{S}}{\tilde{F}r})$ achieved with subpacketization level $F = r\tilde{F}$.

Transforming PDAs into C-PDAs for Resolvable Networks

- Replicate a $(\binom{h}{r}\frac{r}{h}, \tilde{F}, \tilde{Z}, \tilde{S})$ PDA a number of $\frac{h}{r} \cdot r$ times
- Distribute the columns of the replica PDAs so that the columns of each symbol s have non-empty intersection.

Theorem

Given a $(\binom{h}{r}\frac{r}{h}, \tilde{F}, \tilde{Z}, \tilde{S})$ PDA. There exists a (K, F, Z, S) C-PDA, for a resolvable (h, r)-combination network (i.e., where $r|h$) with

$$K = \binom{h}{r}, \quad F = r\tilde{F}, \quad Z = r\tilde{Z}, \quad \text{and} \quad S = h\tilde{S}.$$

Pair $(M = N\frac{\tilde{Z}}{\tilde{F}}, R = \frac{\tilde{S}}{\tilde{F}r})$ achieved with subpacketization level $F = r\tilde{F}$.

Using the Proposed PDAs with Low Subpacketization Level

Theorem

For resolvable networks (r|h), $M \in \{\frac{1}{q} \cdot N : q \in \mathbb{N}^+, q \geq 2\}$*,*

$$R^\star(M) \leq R_{\text{LSub1}} \triangleq \frac{1}{r} \cdot \left(\frac{N}{M} - 1\right).$$

is achievable with $F_{\text{LSub1}} \triangleq r\left(\frac{N}{M}\right)^{\lceil \frac{KMr}{Nh} \rceil - 1}$*.*

Theorem

For resolvable networks (r|h), $M \in \{\frac{q-1}{q} \cdot N : q \in \mathbb{N}^+, q \geq 2\}$*,*

$$R^\star(M) \leq R_{\text{LSub2}} \triangleq \frac{1}{r} \cdot \left(\frac{N}{M} - 1\right).$$

with $F_{\text{LSub2}} \triangleq \frac{rM}{N-M} \cdot \left(\frac{N}{N-M}\right)^{\left\lceil \frac{Kr}{h}\left(1-\frac{M}{N}\right)\right\rceil - 1}$*.*

Comparison with Known Schemes

- L. Tang and A. Ramamoorthy: Coded caching adapted to resolvable networks.

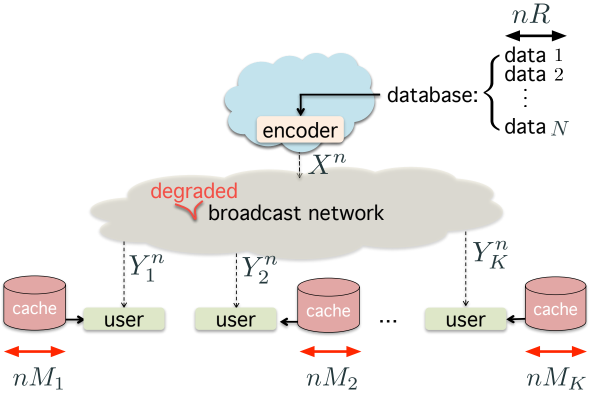
$$\lim_{K\to\infty} \frac{R_{\text{TR}}}{R_{\text{LSub1}}} = 1 \qquad \text{or} \qquad \lim_{K\to\infty} \frac{R_{\text{TR}}}{R_{\text{LSub2}}} = 1.$$

$$\lim_{K\to\infty} \frac{F_{\text{TR}}}{F_{\text{LSub1}}} = \infty \qquad \text{or} \qquad \lim_{K\to\infty} \frac{F_{\text{TR}}}{F_{\text{LSub2}}} = \infty.$$

---

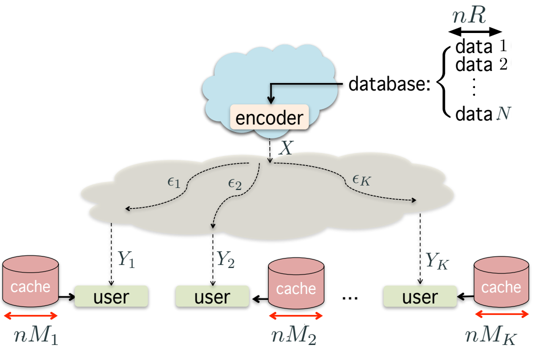
[10] L. Tang and A. Ramamoorthy “Coded caching for networks with resolvability property,” ISIT 2016.

Noisy Broadcast Channel and Heterogeneous Cache Sizes



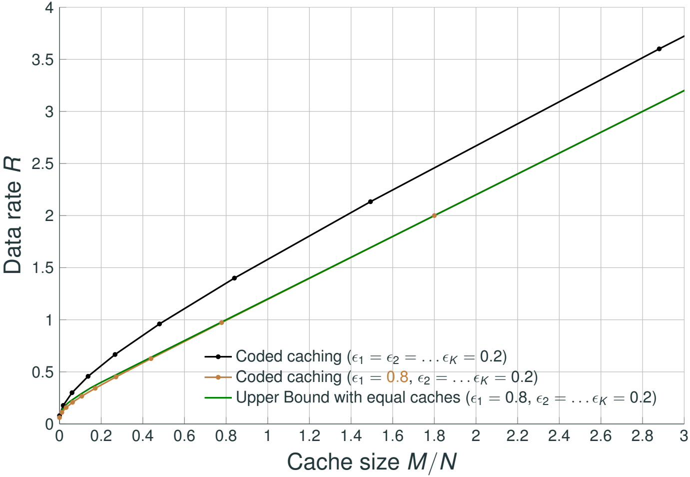
- Under all possible demands, files need to be sent reliably
- Largest data-rate R in function of cache rates $M_1, \ldots, M_K$?

Example: An Erasure Broadcast Network

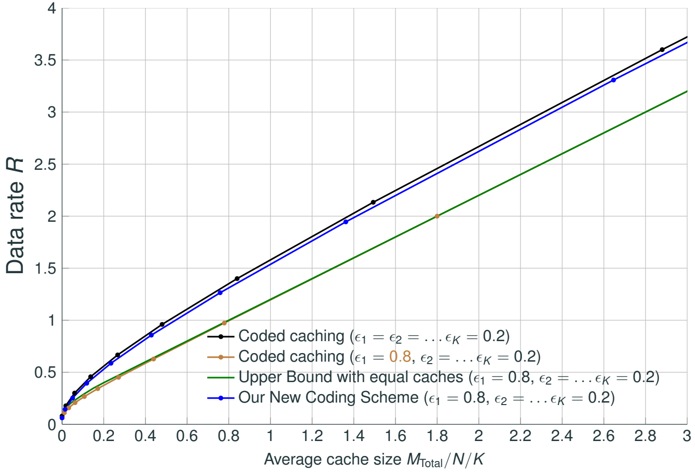


- Binary input X
- Output $Y_k = \begin{cases} X & \text{with probability } 1 - \epsilon_k \\ ? & \text{with probability } \epsilon_k \end{cases}$
- $1 \geq \epsilon_1 \geq \epsilon_2 \geq \epsilon_3 \geq \ldots \geq \epsilon_K \geq 0$

Single Weak Receiver Degrades Performance

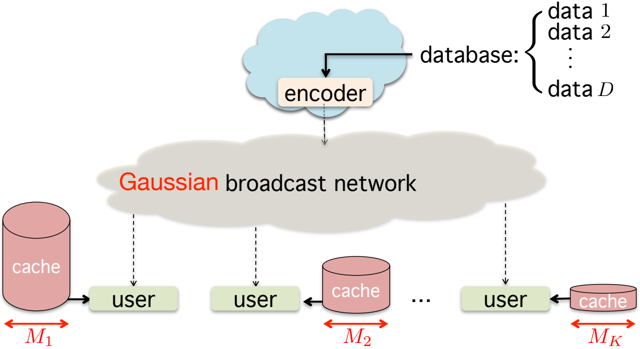


Performance when Cache Memory can be Freely Assigned



- Careful cache assignment + new coding allows to mitigate loss!

Cache Assignment

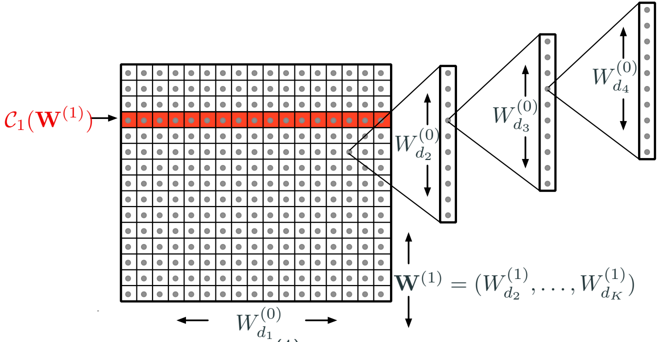


- Power constraint P and noise variances $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_K^2$
- Cache memories: $M_1 + \ldots + M_K \leq M_{\text{Total}}$

$R(M_{\text{Total}})$?

All Cache to Weakest User & Superpos. Piggyback Coding

- Placement: Cache $\{W_d^{(1)}\}_{d=1}^N$ at Rx 1, where $W_d = (W_d^{(0)}, W_d^{(1)})$
- Delivery: Send $W_{d_1}^{(0)}, W_{d_2}, W_{d_3}, \ldots, W_{d_K}$ using following code

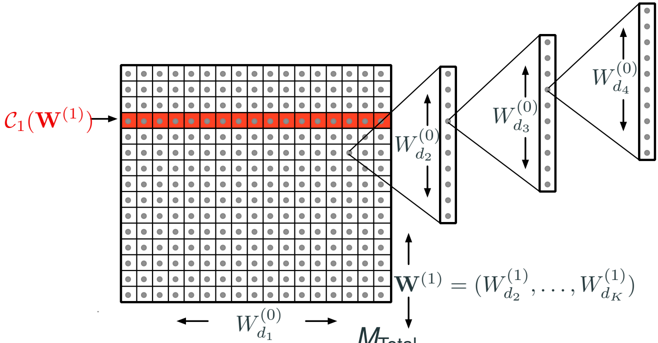


- Receiver 1 knows $\mathbf{W}^{(1)}$ and restricts decoding to single row!

→ Piggybacking $\mathbf{W}^{(1)}$ provides *virtual cache access* for Rxs $2-K$

All Cache to Weakest User & Superpos. Piggyback Coding

- Placement: Cache $\{W_d^{(1)}\}_{d=1}^{N}$ at Rx 1, where $W_d = (W_d^{(0)}, W_d^{(1)})$
- Delivery: Send $W_{d_1}^{(0)}, W_{d_2}, W_{d_3}, \ldots, W_{d_K}$ using following code



- Achieves $R^\star = \underbrace{C_0}_{\text{no-cache}} + \underbrace{\frac{M_{\text{Total}}}{N}}_{\text{perfect caching gain}}, \qquad M_{\text{Total}} \leq M^{(S)}$

Generalized Coded Caching for $K = 3$ and $t = 2$

- Split $W_d = \Big(\; W_d^{(12)}, \qquad W_d^{(13)}, \qquad W_d^{(23)} \Big)$
  of rates $\frac{1}{2}\log\left(1+\frac{P}{\sigma_3^2}\right) \geq \frac{1}{2}\log\left(1+\frac{P}{\sigma_2^2}\right) \geq \frac{1}{2}\log\left(1+\frac{P}{\sigma_1^2}\right)$

- Placement: Store $\left\{ W_d^{(ij)} \right\}$ in cache memories of Rxs i and j

- Delivery:
  - Rx 1 requires $W_{d_1}^{(23)}$; Rx 2 requires $W_{d_2}^{(13)}$; Rx 3 requires $W_{d_3}^{(12)}$
  - Send Gaussian codeword $x^n\left(W_{d_1}^{(23)}, W_{d_2}^{(13)}, W_{d_3}^{(12)}\right)$

  - Decoding at Rx 1 based on restricted codebook

$$\mathcal{C}_1(W_{d_2}^{(13)}, W_{d_3}^{(12)}) := \left\{ x^n\left(w, W_{d_2}^{(13)}, W_{d_3}^{(12)}\right) \right\}_{w=1}^{2^{nR^{(23)}}}$$

Generalized Coded Caching Performance for K receivers

Theorem

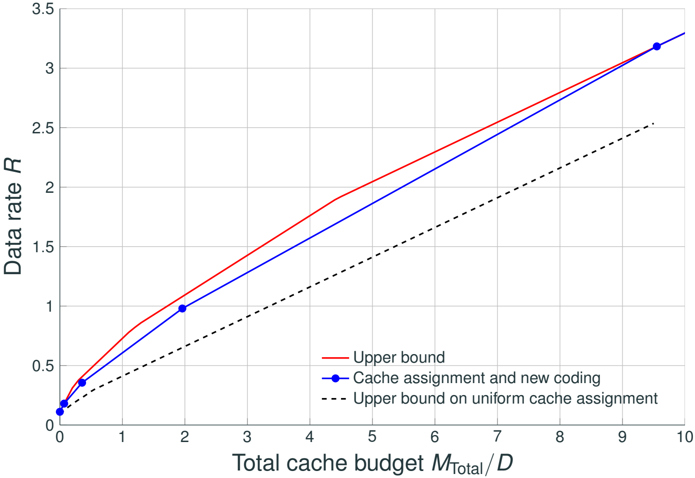
For any $t = 1, \ldots, K-1$, the following memory-rate pair is achievable

$$\frac{M^{(t)}_{Total}}{D} = tR^{(t)}$$

$$R^{(t)} = \frac{\sum_{\ell=1}^{\binom{K}{t}} \prod_{k \in \mathcal{G}_\ell^{(t),c}} \frac{1}{2}\log\left(1 + \frac{P}{\sigma_k^2}\right)}{\sum_{j=1}^{\binom{K}{t+1}} \prod_{k \in \mathcal{G}_j^{(t+1),c}} \frac{1}{2}\log\left(1 + \frac{P}{\sigma_k^2}\right)}$$

where $\mathcal{G}_1^{(t)}, \ldots, \mathcal{G}_{\binom{K}{t}}^{(t)}$ denote all size-t subsets of $\{1, \ldots, K\}$

Bounds on the Rate-Memory Tradeoff



Gaussian broadcast network $\sigma_1^2 = 4, \sigma_2^2 = 2, \sigma_3^2 = 1, \sigma_1^2 = 0.5$

Exact Results: From Global to Local Caching Gain

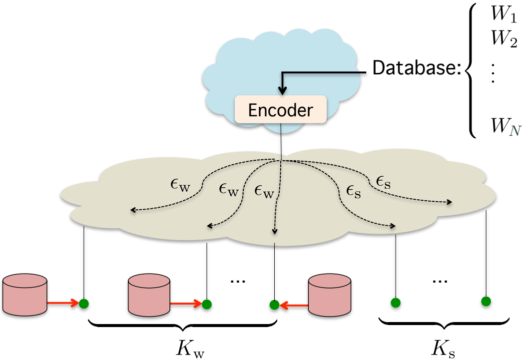
- Small total cache budget:
  - all cache memory to weakest receiver
  - superposition piggyback coding

$$R^\star = \underbrace{C_0}_{\text{no-cache}} + \underbrace{\frac{M_{\text{Total}}}{N}}_{\text{perfect caching gain}}, \qquad M_{\text{Total}} \leq M^{(S)}$$

- Large total cache budget:
  - the more cache memory the weaker the receiver
  - generalized coded caching

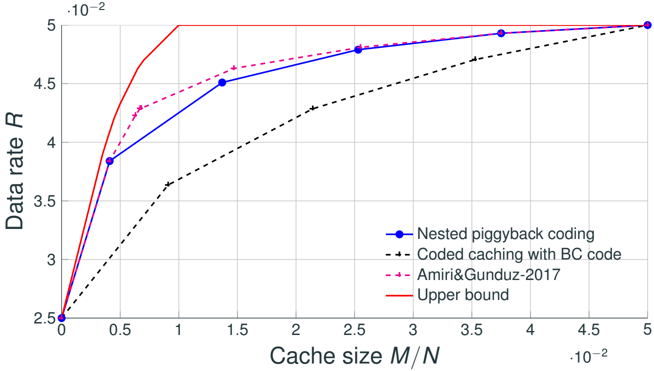
$$R^\star = \underbrace{\sum_{k=1}^{K} \frac{1}{2} \log\left(1 + \frac{P}{\sigma_k^2}\right)}_{K \text{ point-to-point links}} + \underbrace{\frac{1}{K} \frac{M_{\text{Total}} - M^{(\text{L})}}{N}}_{\text{only local caching gain}}, \qquad M_{\text{Total}} \geq M^{(\text{L})}$$

A Setup with Fixed Cache Assignment



- K_w weak receivers with erasure probability ϵ_w
- K_s strong receivers with erasure probability $\epsilon_s < \epsilon_w$
- Cache memories of size nM bits only at weak receivers

Benefits of Joint-Cache Channel Coding



- 4 weak, and 16 strong users, $\epsilon_w = 0.8$ and $\epsilon_s = 0.2$

Some Related Works and Further Discussions

- Additional libraries with higher resolution information (Cacciapuoti, Caleffi, Ji, Llorca, Tulino-2016)
- Fading broadcast channels (Zhang&Elia-2016)
- Broadcast channels with feedback (Ghorbel,Kobayashi,Yang-2016, Zhang&Elia-2016)
- Massive MIMO broadcast channels (Yang, Ngo, Kobayashi-2016)

Summary

- PDAs useful to find good caching schemes with low subpacketization
- Can construction combinational PDAs from standard PDAs $\rightarrow$ good coding schemes for combination networks with low subpacketization levels
- PDA scheme optimal for combination networks with large cache sizes

- Delivery over noisy networks requires joint cache-channel coding
- Adapt cache allocation to channel strengths $\rightarrow$ additional coding opportunities

Literature

- C.-Y. Wang, S. Saeedi Bidokhti, and M. Wigger, "Improved Converses and Gap-Results for Coded Caching," *IT-Trans 2018*
- Q. Yan, M. Wigger, and S. Yang, "Placement Delivery Array Design for Combination Networks with Edge Caching," *ISIT 2018*
- S. Saeedi Bidokhti, M. Wigger, and R. Timo, "Noisy Broadcast Networks with Receiver Caching," *IT-Trans 2018*
- S. Saeedi Bidokhti, M. Wigger, and A. Yener, "Benefits of Cache Assignment on Degraded Broadcast Channels," *ArXiv: 1702.08044*