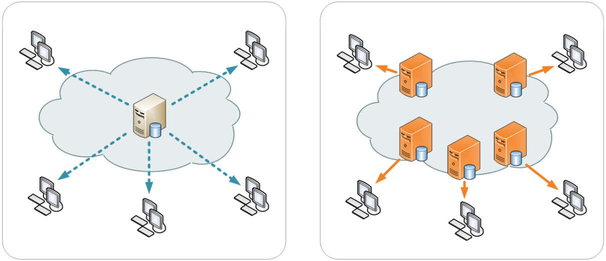
Cache Allocations, Coding Schemes, and Converses for Cellular Networks with Caching

Michèle Wigger

Joint work with Shirin Saeedi Bidokhti, Shlomo Shamai (Shitz), and Roy Timo, Aylin Yener

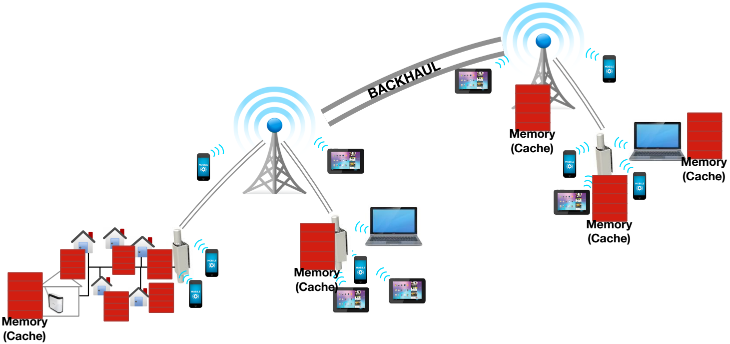
Ben Gurion University, Beer Sheva, Israel, 14 November 2016

Content Delivery Networks



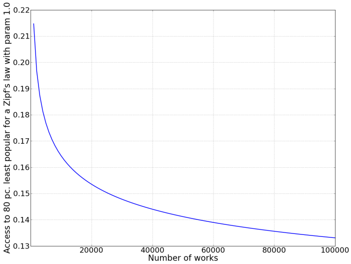
- Store contents in caches before file demands even known
- Reduce network load and latency during high-congestion periods
- Idea useful if certain files very popular and known in advance

Distributed Caches: Promising Solution for Cellular Networks



- Can cache at main BSs, picoBSs, femtoBSs, or directly at end users

File Popularities

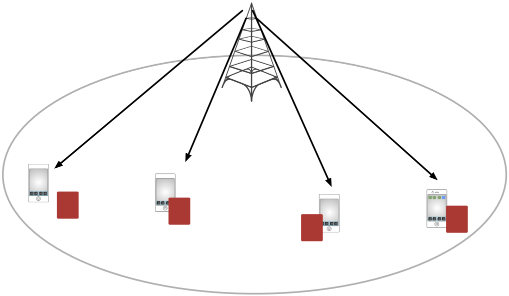


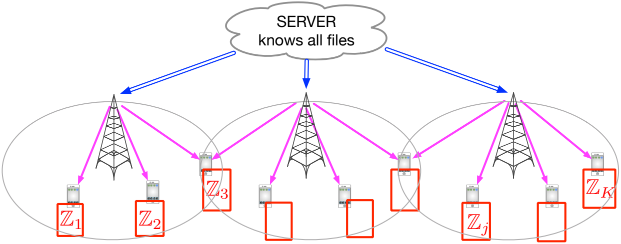
- Static file popularity follows a Zipf distribution $P(x) = Cx^{-\alpha}$

- Evolution of file popularities (youtube videos) can also be predicted

> Use pro-active caching to improve cellular systems!

Cellular Scenarios



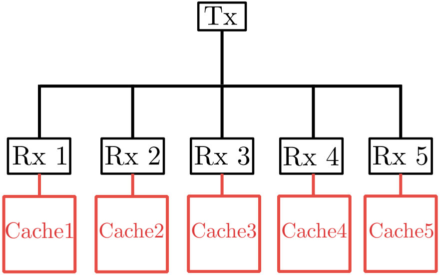


Other Assumptions

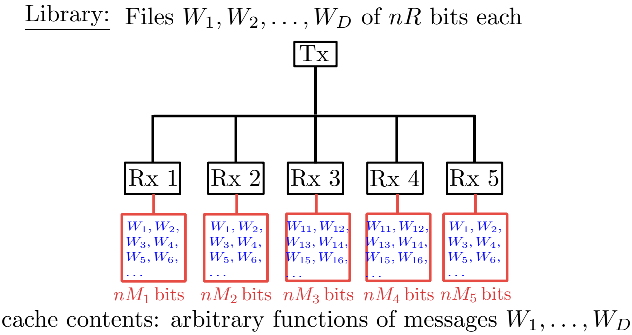
- All files equally popular $\rightarrow$ interested in worst-case performance
- Centralized protocol on how to fill caches
- Caches filled during nights when demands not yet known

Maddah-Ali & Niesen Source Coding Setup

Library: Files $W_1, W_2, \ldots, W_D$ of nR bits each (no popularities)



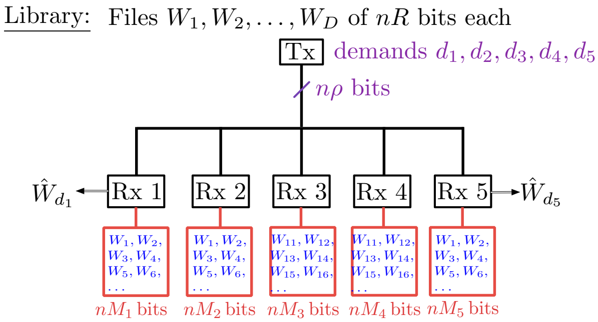
Maddah-Ali & Niesen Source Coding Setup



Communication in two phases:

- Caching phase: Tx fills caches without knowing demands $d_1, \ldots, d_5$

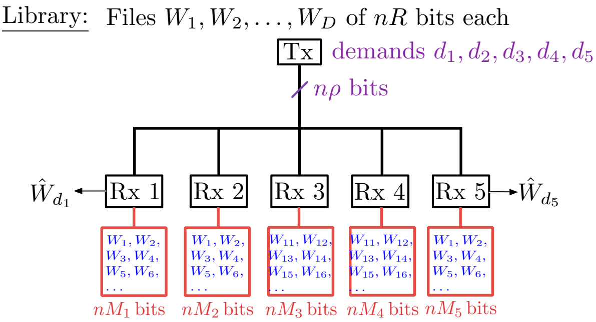
Maddah-Ali & Niesen Source Coding Setup



Communication in two phases:

- Caching phase: Tx fills caches without knowing demands $d_1, \ldots, d_5$
- Delivery phase: Tx describes $W_{d_1}, \ldots, W_{d_5}$ to Rxs $1, \ldots, 5$, respectively, through $n\rho$ common bits

Maddah-Ali & Niesen Source Coding Setup

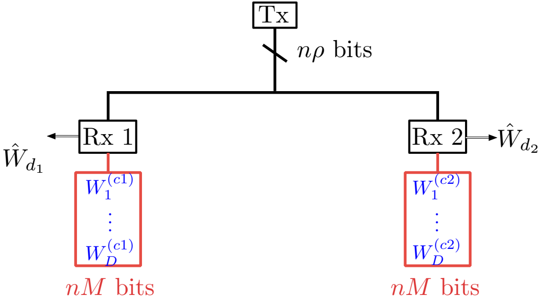


Rates-Memories Tradeoff

For which $(\rho, R, M_1, \ldots, M_K)$ is error-free data transmission possible?

Coded caching for $K = 2$ Receivers [Maddah-Ali&Niesen 2013]

Library: Files $W_1, W_2, \ldots, W_D$ of nR bits each

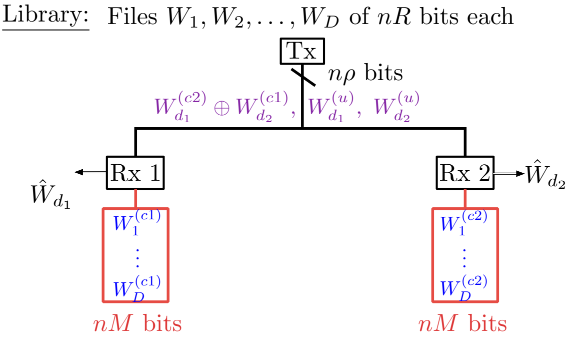


- Split $W_d = \left(W_d^{(c1)}, W_d^{(c2)}, W_d^{(u)}\right)$ of rates $\frac{M}{D}$, $\frac{M}{D}$, and $R - 2\frac{M}{D}$

Rates-Memory Trade-Off for moderate M:

Reconstruction possible, if $R \leq \frac{1}{2}\rho + \frac{3}{2}\frac{M}{D}$

Coded caching for $K = 2$ Receivers [Maddah-Ali&Niesen 2013]



- Split $W_d = \left(W_d^{(c1)}, W_d^{(c2)}, W_d^{(u)}\right)$ of rates $\frac{M}{D}$, $\frac{M}{D}$, and $R - 2\frac{M}{D}$

Rates-Memory Trade-Off for moderate M:

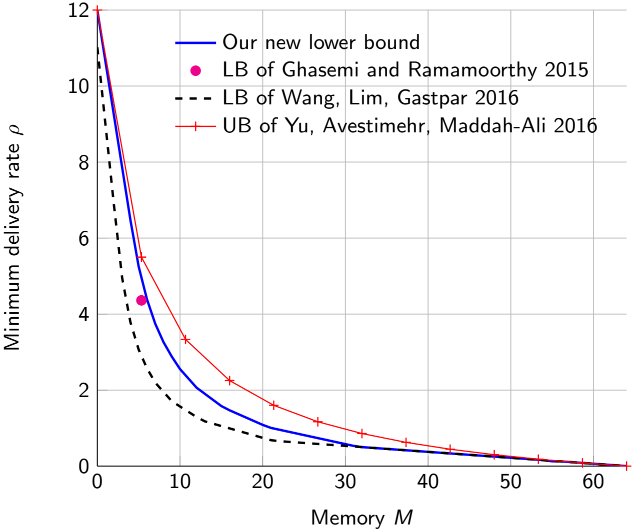
Reconstruction possible, if $R \leq \frac{1}{2}\rho + \frac{3}{2}\frac{M}{D}$

Reconstruction not possible, if $R \geq \frac{1}{2}\rho + \frac{M}{D}\left(1 + \frac{D+1}{2(D-1)}\right)$

(Saeedi,Wigger,Yener-2016)

Current Bounds on Minimum Delivery Rate ρ when $R = 1$

- $K = 12$ users, $D = 64$ files



Extensions

- Decentralized caching

  [M. A. Maddah-Ali, U. Niesen, “Decentralized coded caching attains order-optimal memory-rate tradeoff”]

- Nonuniform or random demands

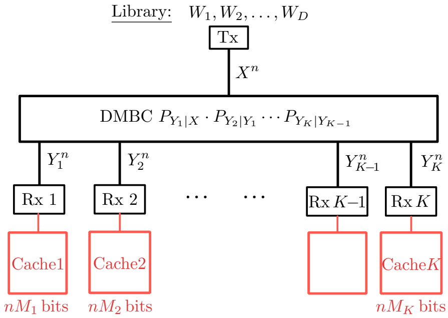
  [U. Niesen and M. A. Maddah-Ali, “Coded caching with nonuniform demands”]
  [Ji, Tulino, Llorca, and Caire, “Order-optimal rate of caching and coded multicasting with random demands”]

- Online caching phase

  [R. Pedarsani, M. A. Maddah-Ali and U. Niesen, “Online coded caching”]

Delivery over Noisy Broadcast Channel (BC)

[Saeedi, Timo, Wigger 2016, Saeedi, Wigger, Yener, 2016]



Capacity-Memory Tradeoff

$C(M_1, \ldots, M_K)$: supremum of achievable message rate R

- New achievability: joint cache-channel schemes based on piggyback coding
- New converse for degraded BCs

Converse for Degraded BCs with Arbitrary Caches

Theorem (Saeedi, Timo, Wigger 16)

$$C(M_1, \ldots, M_K) \leq \min_{\mathcal{S} \subseteq \{1,\ldots,K\}} \left(R_{\text{sym},\mathcal{S}}(M_1, \ldots, M_K) + \sum_{j=1}^{|\mathcal{S}|} \frac{M_{s_j}}{D} \right),$$

$R_{\text{sym},\mathcal{S}}$ and $M_{\mathcal{S}}$: symmetric capacity and total cache at receivers in $\mathcal{S}$

Theorem (Saeedi, Wigger, Yener 16)

$$C(M_1, \ldots, M_K) \leq I(U_1; Y_1) + \frac{M_1}{D},$$

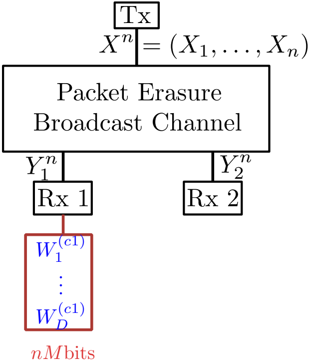
$$C(M_1, \ldots, M_K) \leq I(U_k; Y_k | U_{k-1}) + \frac{\sum_{i=1}^{k} M_i}{D - k + 1}, \qquad \forall k \in \{2, \ldots, K\}$$

where $U_1 - U_2 - \ldots - U_K - X - Y_K - \ldots - Y_1$.

- Suggests that cache-memory M_1 more useful than cache memory M_K
  $\rightarrow$ Each M_k serves for users $k, \ldots, K$

A Simple Example

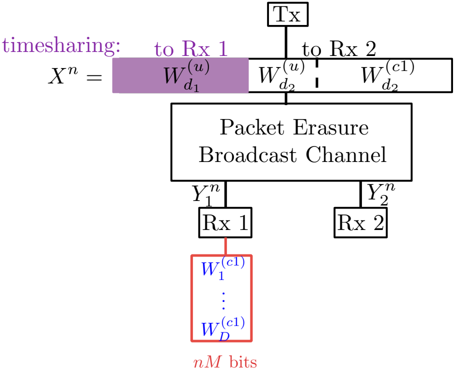
Library: Files $W_1, W_2, \ldots, W_D$ of nR bits each



- $W_d = (W_d^{(c1)},\ W_d^{(u)})$ of rates $(\frac{M}{D},\ R - \frac{M}{D})$

A Simple Example

Library: Files $W_1, W_2, \ldots, W_D$ of nR bits each



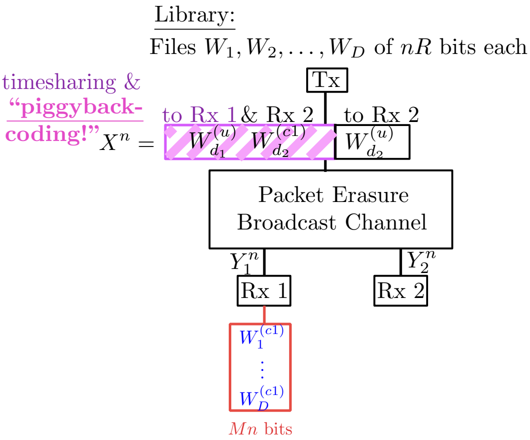
- $W_d = (W_d^{(c1)}, W_d^{(u)})$ of rates $(\frac{M}{D}, R - \frac{M}{D})$

Separate Cache-Channel Coding → No Global Caching Gain

$p(\text{error}) \to 0$ if: $$\frac{R - \frac{M}{D}}{F(1-\delta_1)} + \frac{R}{F(1-\delta_2)} \le 1$$

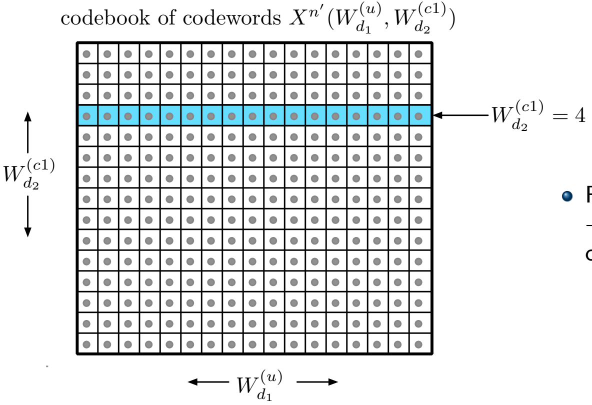
Standard Erasure BC: $p(error)$ if : $$\frac{R_1}{F(1-\delta_1)} + \frac{R_2}{F(1-\delta_2)} \le 1$$

Our Joint Cache-Channel Scheme for this Example



- $W_d = (W_d^{(c1)},\ W_d^{(u)})$ of sub-rates $(\frac{\mathsf{M}}{D},\ \rho - \frac{\mathsf{M}}{D})$

Piggyback Coding to Send $(W_{d_1}^{(u)}, W_{d_2}^{(c1)})$ to Both Rxs

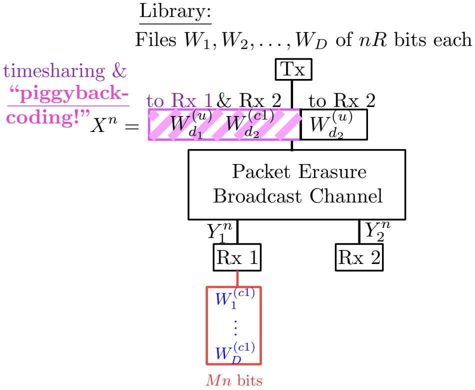


- Rx 1 knows $W_{d_2}^{(c1)}$ $\rightarrow$ restrict decoding to corresponding row

Transmission of $W_{d_2}^{(c1)}$ not affecting Rx 1 at all!

$$p(error) \rightarrow 0 \text{ as } n \rightarrow \infty : \qquad \max\left\{ \frac{R - \frac{M}{D}}{F(1-\delta_1)}, \frac{R}{F(1-\delta_2)} \right\} \leq \frac{n'}{n}$$

Performance of Joint Cache-Channel Scheme for Example

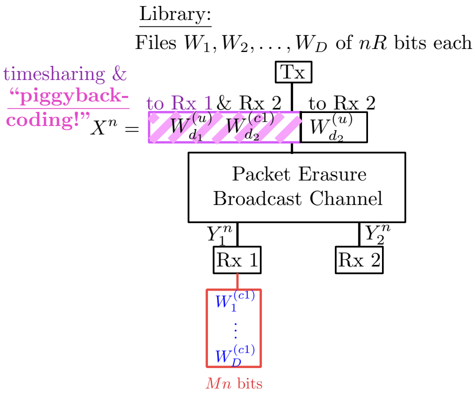


- $W_d = (W_d^{(c1)}, \ W_d^{(u)})$ of sub-rates $(\frac{M}{D}, \ R - \frac{M}{D})$

Joint Cache-Channel Coding → Global Caching Gain!

$$p(error) \to 0 \quad \text{if :} \qquad \underbrace{\max\left\{\frac{R-\frac{M}{D}}{F(1-\delta_1)}, \ \frac{R}{F(1-\delta_2)}\right\}}_{\text{piggyback coding}} + \frac{R-\frac{M}{D}}{F(1-\delta_2)} \leq 1$$

Performance of Joint Cache-Channel Scheme for Example



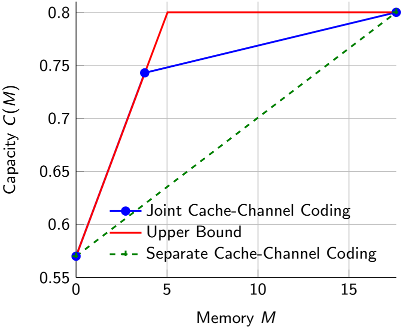
- $W_d = (W_d^{(c1)},\ W_d^{(u)})$ of sub-rates $(\frac{M}{D},\ R - \frac{M}{D})$

Joint Cache-Channel Coding → Global Caching Gain!

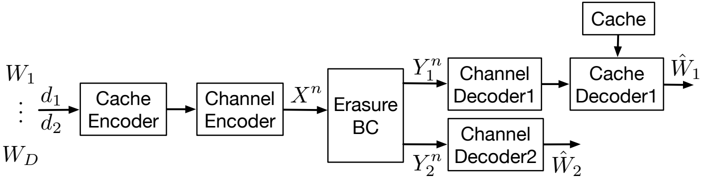
$$p(error) \to 0 \quad \text{if} : R \leq \min \left\{ \underbrace{\frac{F(1-\delta_1)(1-\delta_2)}{(1-\delta_1)+(1-\delta_2)}}_{C_{\text{Sym}}} + \frac{M}{D},\quad \frac{1}{2}F(1-\delta_2) + \frac{M}{2D} \right\}$$

K-User Erasure BC with Cache only at Weakest Receiver

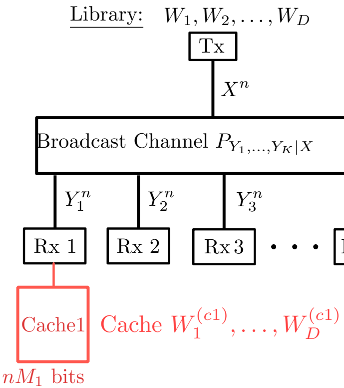
$\delta_1 = 0.8,$
$\delta_2 = \ldots = \delta_K = 0.2,$
$K = 11, \quad D = 22,$
$F = 10$



- Separate Cache-Channel Architecture



Single Cache and $K-1$ stronger receivers



$$W_d = (W_d^{(c1)},\ W_d^{(u)})$$

of rates $\left(\frac{M_1}{D},\ R - \frac{M_1}{D}\right)$

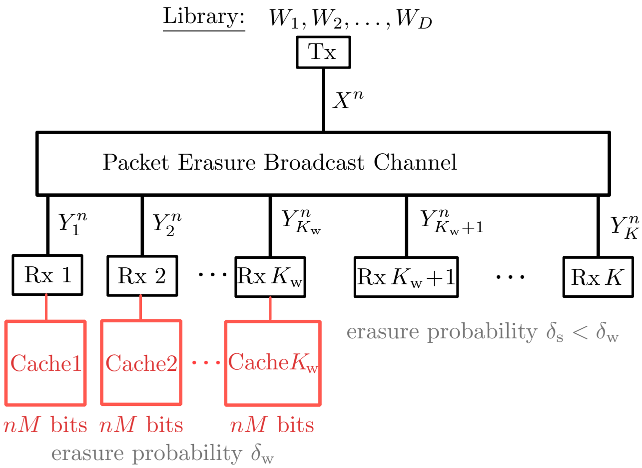
Optimal Coding Scheme for Single Cache and low Cache Memory

- Superposition code achieving C_{Sym} with piggyback cloud center
- Piggyback cloud center encodes $W_{d_1}^{(u)}, W_{d_2}^{(c_1)}, \ldots, W_{d_K}^{(c_1)}$
- Receiver 1 performs restricted decoding
- Receivers $2 \ldots, K$ can decode if $(K-1)\frac{M_1}{D} \leq I(U_1^\star; Y_2) - I(U_1^\star; Y_1)$

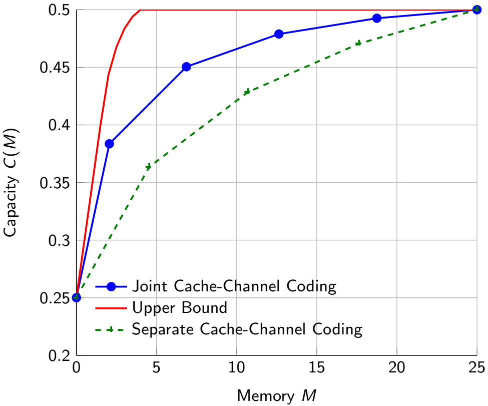
Performance

$$C(M_1, M_2 = 0, \ldots, M_K = 0) = C_{\text{Sym}} + \frac{M_1}{D}, \qquad M_1 \leq D \cdot \frac{I(U_1^\star; Y_2) - I(U_1^\star; Y_1)}{K-1}$$

K_w Weak Receivers with Caches and K_s Strong Receivers Without

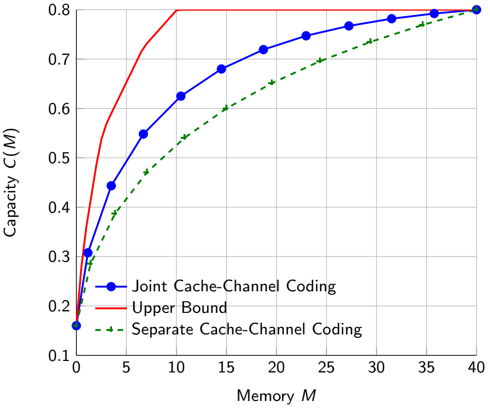


Weak and Strong Receivers: Example I



$$\delta_w = 0.8, \ \ \delta_s = 0.2, \ \ K_w = 4, \ K_s = 16, \ \ D = 50, \ \ F = 10$$

Weak and Strong Receivers: Example II



$$\delta_{\mathrm{w}} = 0.8, \;\; \delta_{\mathrm{s}} = 0.2, \;\; K_{\mathrm{w}} = 10, \; K_{\mathrm{s}} = 10, \;\; D = 50, \;\; F = 10$$

Most Significant Gains for Low Cache Sizes $M/D \leq \Gamma_1$

For $M/D \leq F \frac{(1-\delta_\mathrm{s})}{K_\mathrm{s}} \frac{(\delta_w - \delta_\mathrm{s})}{(K_\mathrm{s}(1-\delta_w)+(1-\delta_\mathrm{s}))}$:

$$C(M) \geq C_{\mathrm{Sym}} + \frac{M}{D} \cdot \gamma_{\mathrm{local}} \cdot \gamma_{\mathrm{global,sep}} \cdot \gamma_{\mathrm{global,joint}},$$

with

$$\gamma_{\mathrm{local}} := \frac{K_\mathrm{w}(1-\delta_\mathrm{s})}{K_\mathrm{w}(1-\delta_\mathrm{s}) + K_\mathrm{s}(1-\delta_\mathrm{w})},$$
$$\gamma_{\mathrm{global,sep}} := \frac{1+K_\mathrm{w}}{2},$$
$$\gamma_{\mathrm{global,joint}} := 1 + \frac{2K_\mathrm{w}}{1+K_\mathrm{w}} \cdot \frac{K_\mathrm{s}(1-\delta_\mathrm{w})}{K_\mathrm{w}(1-\delta_\mathrm{s})}.$$

All Receivers with Cache Memories

- $W_d = (W_d^{(1)}, \ldots, W_d^{(K)})$ or rates $I(X; Y_1), \ldots, I(X; Y_K)$
- Rate $R = \sum_{k=1}^{K} I(X; Y_k)$
- Receiver k caches $W_d^{(1)}, \ldots, W_d^{(k-1)}, W_d^{(k+1)}, \ldots, W_d^{(K)}$

  for all $d = 1, \ldots, D$
- For transmission, use a K-dimensional product codebook

  $\rightarrow$ Receiver k knows $K-1$ dimensions; can decode last dimension of rate $I(X; Y_k)$
- Total cache $(K-1) \sum_{k=1}^{K} I(X; Y_k)$

Capacity-Memory Tradeoff under Global Memory Constraint

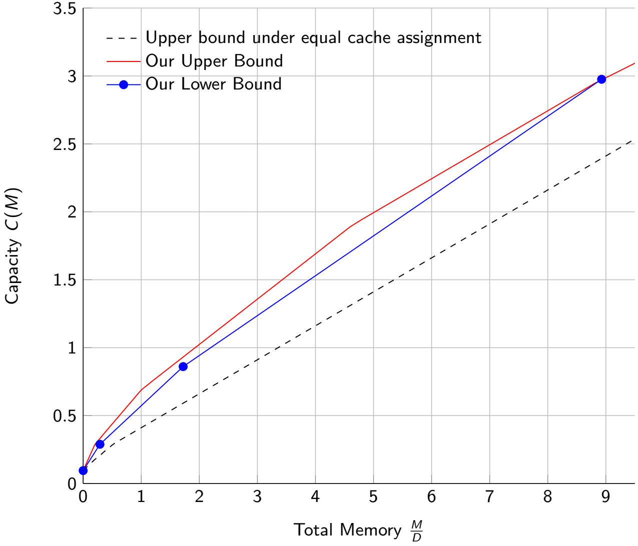
$$C^{\star}(M) \leq \max C(M_1, \ldots, M_K),$$

where max over $(M_1, \ldots, M_K)$: $\sum_{k=1}^{K} M_k \leq M$

‘

- Have upper and lower bounds from previous our results
- They are tight for small and large cache memories M
- Equa cache assignment highly suboptimal for non-symmetric BCs

Global Capacity-Memory Tradeoff for a Gaussian-BC Example



Insights and Intuition

- Important to consider noisy communication channel:
  - Joint cache-channel coding (piggyback coding)
  - Caching gains combine with feedback gains

    [A. Ghorbel, M. Kobayashi, S. Yang, “Cache-enabled broadcast packet erasure channels with state feedback”]
  - Interplay between caching gains and CSI gains

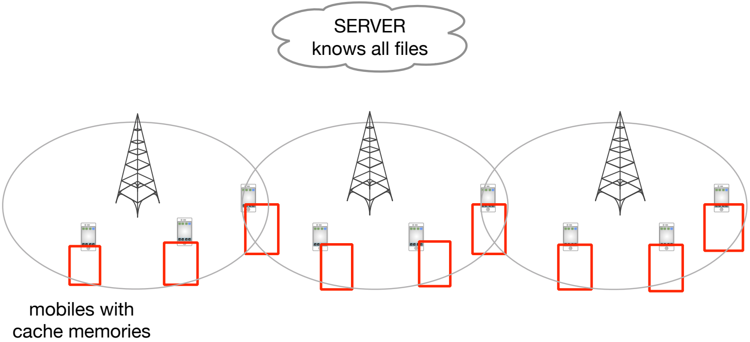
    [J. Zhang and P. Elia, “Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback”]

- Gains with cache allocation $\rightarrow$ larger caches to weak receivers

- Cache allocation gains boosted through joint cache-channel coding

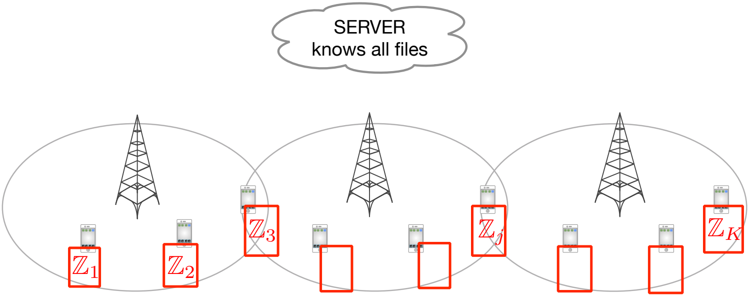
- Piggyback coding useful whenever info for strong Rx in cache of weak Rx!

Multi-Cell Model with 3 Communication Phases



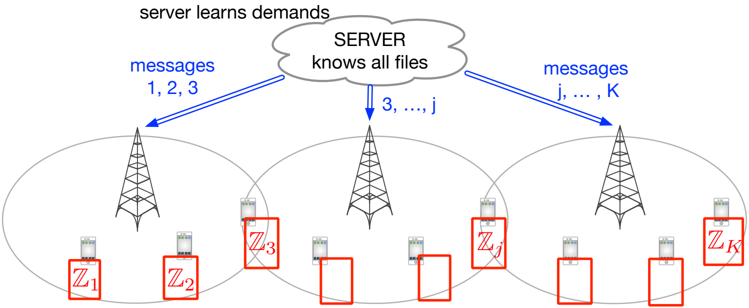
1. Caching Phase: Server fills caches while ignoring demands $d_1, \ldots, d_K$
2. Download from Server: BSs download messages of connected rxs
3. Delivery to Users: BSs communicate messages to rxs over network

Multi-Cell Model with 3 Communication Phases



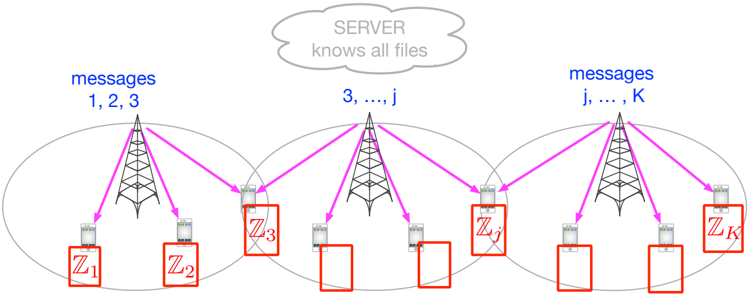
1. Caching Phase: Server fills caches while ignoring demands $d_1, \ldots, d_K$
2. Download from Server: BSs download messages of connected rxs
3. Delivery to Users: BSs communicate messages to rxs over network

Multi-Cell Model with 3 Communication Phases



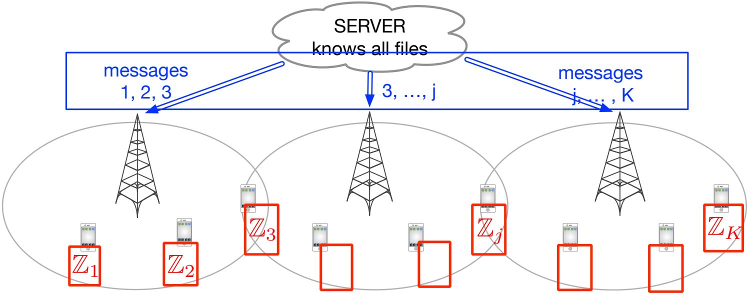
1. Caching Phase: Server fills caches while ignoring demands $d_1, \ldots, d_K$
2. Download from Server: BSs download messages of connected rxs
3. Delivery to Users: BSs communicate messages to rxs over network

Multi-Cell Model with 3 Communication Phases



1. Caching Phase: Server fills caches while ignoring demands $d_1, \ldots, d_K$
2. Download from Server: BSs download messages of connected rxs
3. Delivery to Users: BSs communicate messages to rxs over network

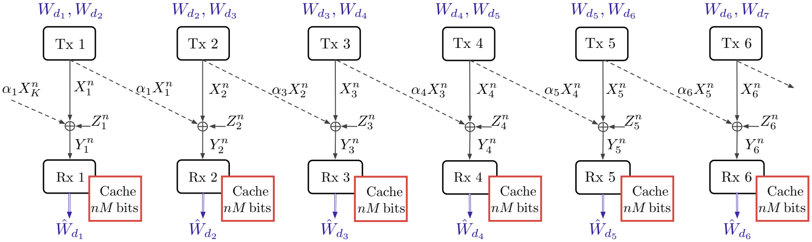
Multi-Cell Model with 3 Communication Phases



- All files revealed to all BSs → BC
- Caches also at BSs & rate-limited pipes from servers to BSs: Karamchandani, Niesen, Maddah-Ali, Diggavi-2014.
- BSs download files only in caching phase (tx&rx caches): Maddah Ali&Niesen-2015; Pooya-Abolfazl-Hossein-2015; Naderializadeh-Maddah Ali-Avestimehr-2016; Hachem-Niesen-Diggavi-2016

Caching in Wyner's Cellular Networks

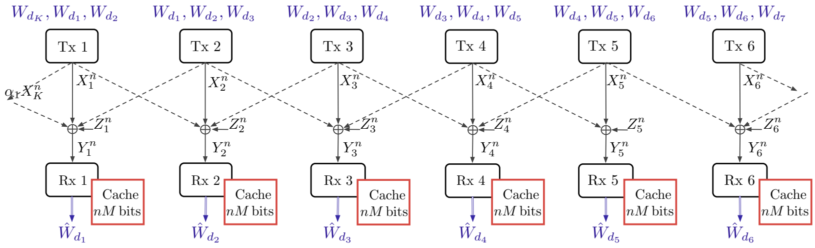
- Delivery over *Wyner's Asymmetric Network*



- Transmitters constrained to power P

Caching in Wyner's Cellular Networks

- Delivery over *Wyner's Symmetric Network*



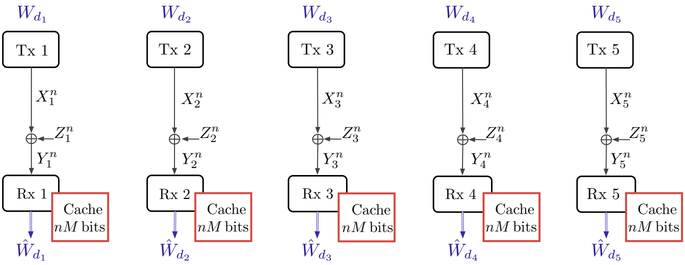
- Transmitters constrained to power P

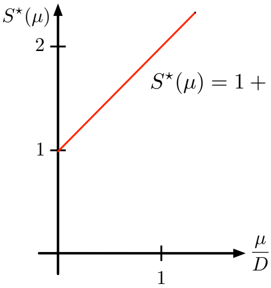
DoF Rate-Memory Tradeoff

- Power $P \gg 1$
- Rate $R = S \cdot \frac{1}{2}\log(1+P)$
- Memory $M = \mu \cdot \frac{1}{2}\log(1+P)$

$$S^\star(\mu) := \max\left\{S : \quad (S,\mu) \text{ achievable}\right\}$$

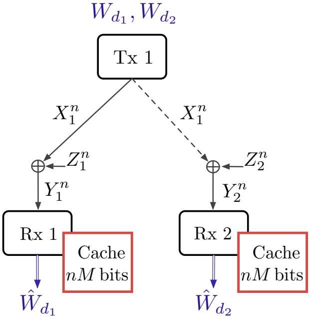
DoF Rate-Memory Tradeoff for Network with Parallel Links

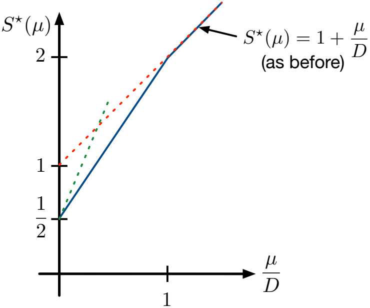




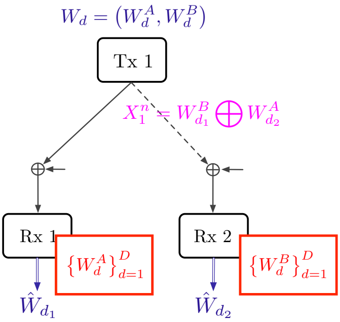
$$S^\star(\mu) = 1 + \frac{\mu}{D} : \quad \text{local caching gain!}$$

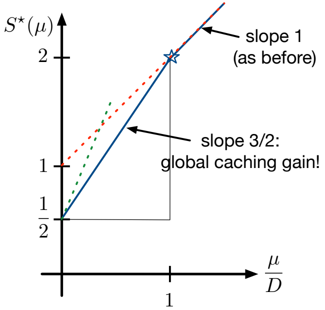
DoF Rate-Memory Tradeoff of Two-User BC





DoF Rate-Memory Tradeoff of Two-User BC

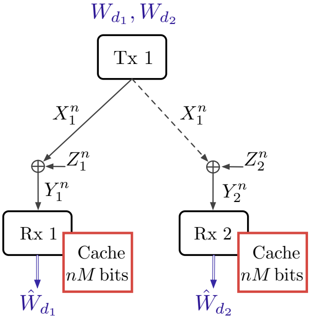


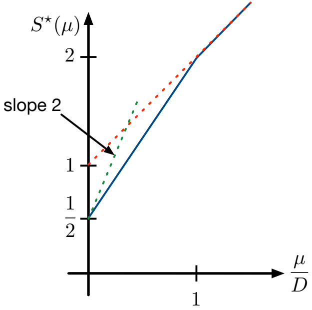


Maddah-Ali&Niesen Coded Caching:

- Cache half of each message at Rx1 and the other at Rx2 $\rightarrow \mu = 1$
- Send XOR of missing halves! $\rightarrow S = 2\mu/D = 2$

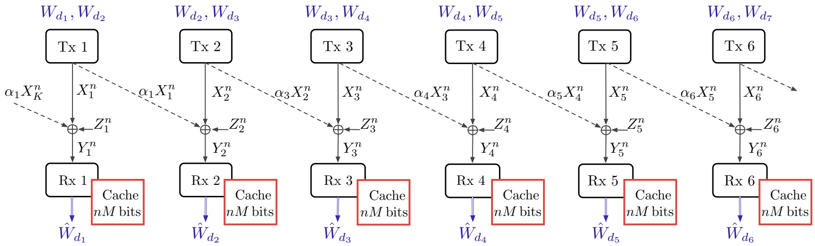
DoF Rate-Memory Tradeoff of Two-User BC





- Slope 2: each receiver can fully profit from both caches (Upper bound from [Saeedi-W-Timo 2016])

Back to Asymmetric Circular Wyner Model

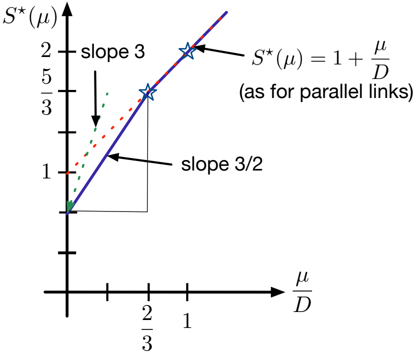


- Without cache memories: $S^\star(\mu = 0) = 2/3$

Our Bounds on DoF Rate-Memory Tradeoff for Asymmetric Model

Theorem

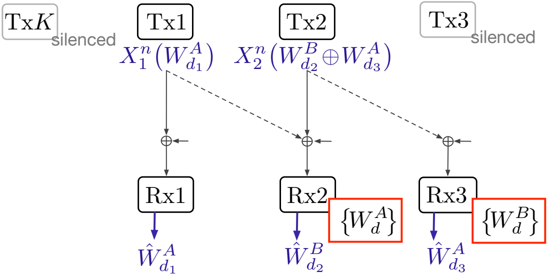
- *If* $\frac{\mu}{D} \leq \frac{2}{3}$: $\quad \frac{2}{3} + \frac{3}{2}\frac{\mu}{D} \leq S^\star \leq \min\left\{\frac{2}{3} + 3\frac{\mu}{D}\ ,\ 1 + \frac{\mu}{D}\right\}$
- *If* $\frac{\mu}{D} \geq \frac{2}{3}$: $\quad S^\star = 1 + \frac{\mu}{D}$



Scheme achieving $\mu/D = 2/3$ and $S = 5/3$ for Asymmetric Network

- Split $W_d = (W_d^A, W_d^B, W_d^C, W_d^D, W_d^E)$ each of DoF 1
- 3 delivery phases: in each we silence every third transmitter $\rightarrow$ subnets

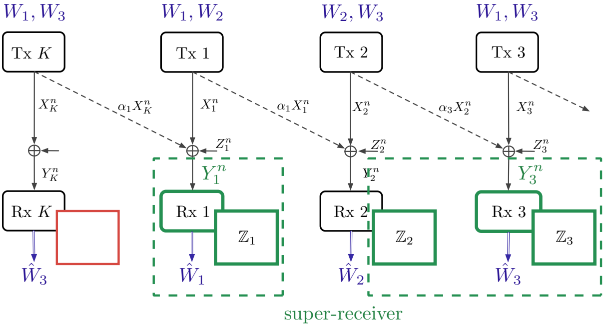
Phase 1 Subnet:



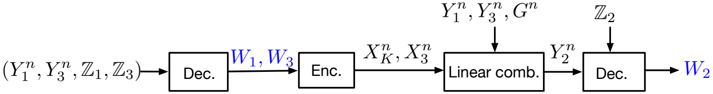
- Rx 3: Decodes $W_{d_2}^B \bigoplus W_{d_3}^A$; gets $W_{d_3}^A$;
- Rx 2: Forms $Y_2^n - X_1^n(W_{d_1}^A)$; decodes $W_{d_2}^B \bigoplus W_{d_3}^A$; gets $W_{d_2}^B$

Converse for Asymmetric Network

- Choose a demand vector (worst-case error)

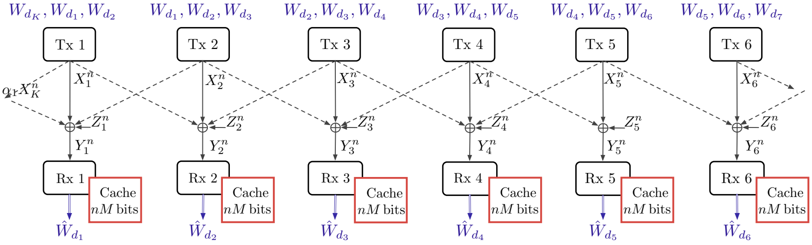


- MAC- bound: If Rx 1–3 decode correctly, super receiver can do the same:



- Average over demand vectors! $\rightarrow$ factor $1/D$

Symmetric Circular Wyner Model

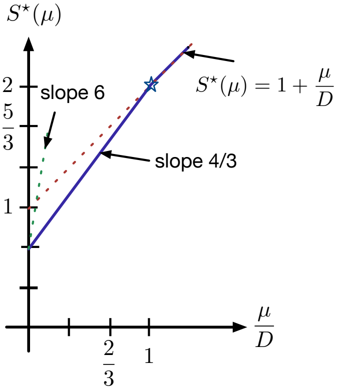


- Without cache memories: $S^{\star}(\mu = 0) = 2/3$

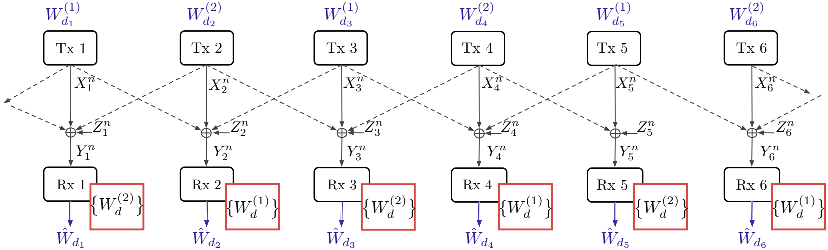
Our Bounds on DoF Rate-Memory Tradeoff for Symmetric Model

Theorem

- *If* $\frac{\mu}{D} \leq 1$: $\quad \frac{2}{3} + \frac{4}{3}\frac{\mu}{D} \leq S^\star \leq \min\left\{\frac{2}{3} + 6\frac{\mu}{D}\ ,\ 1 + \frac{\mu}{D}\right\}$
- *If* $\frac{\mu}{D} \geq 1$: $\quad S^\star = 1 + \frac{\mu}{D}$



Coding Scheme for Symmetric Network



- Split $W_d = (W_d^A, W_d^B)$ each of DoF 1
- Cache contents allow receivers to pre-cancel interference

Summary

- Noisy channels $\rightarrow$ joint cache-channel coding (piggyback coding)

- Cache allocation: more caches to weaker receivers

  $\rightarrow$ joint cache-channel codes boost cache allocation gains

- Caching allows to completely cancel interference in cellular networks
  $\rightarrow$ smart code design significantly reduces required cache sizes

- Outer bounds for degraded BCs and sparse interference networks with receiver caching