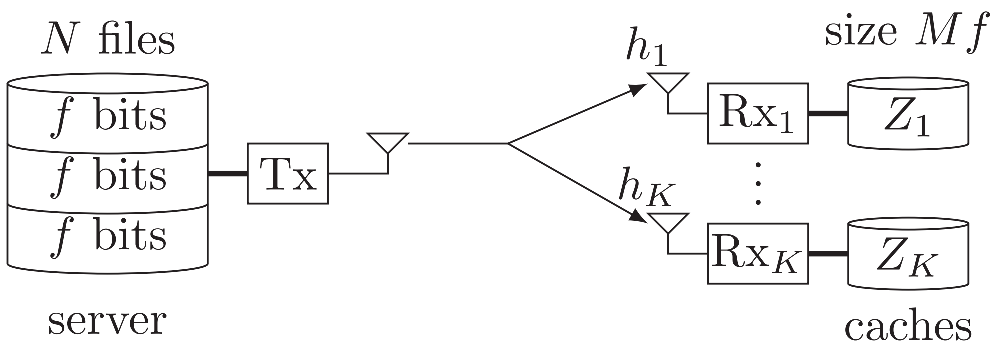
Feedback-boosted coded caching

Jingjing Zhang and Petros Elia

eurecom

Sophia Antipolis - France

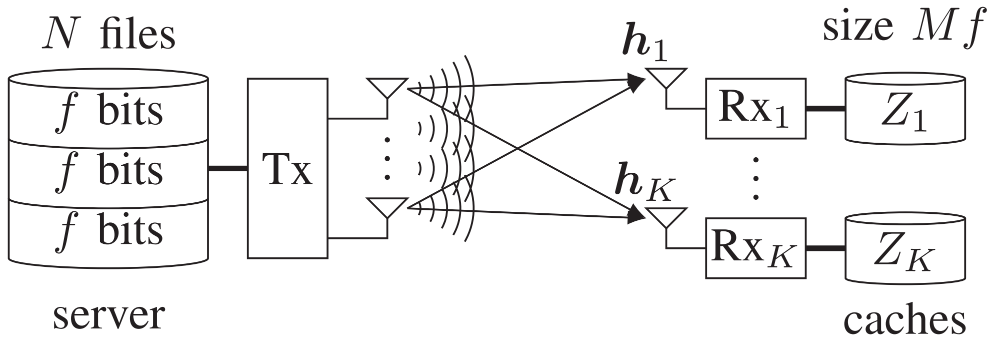
Coded caching



- Coded caching (Maddah-Ali and Niesen)
  - ⋆ by pre-filling the caches $Z_1, Z_2, \ldots, Z_K$
  - ⋆ then encoding over content from different users
  - ⋆ thus increasing multicast opportunities (one tx useful to many)

- Substantial increase in throughput (network load during peak hours)

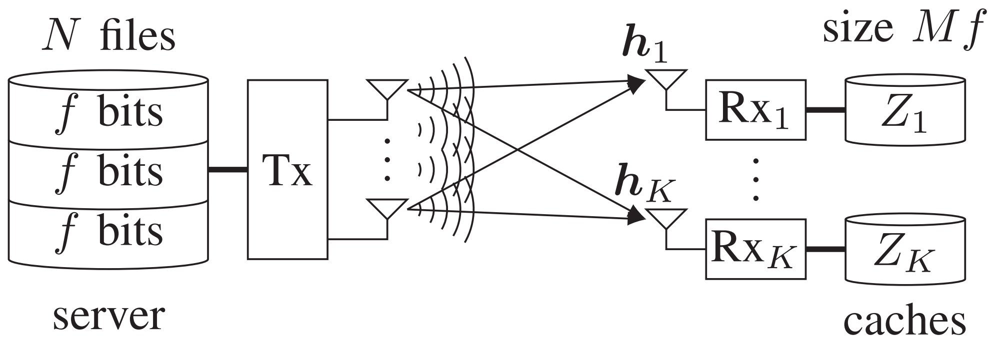
Coding caching in BC with random fading and CSIT



- We explore coded-caching in multi-antenna BC with random fading
  - ⋆ brings to the fore the element of CSIT-type feedback
    - ∗ CSIT is crucial in handling interference
    - ∗ CSIT is hard to get (consider variable quality)
    - ∗ CSIT has ‘intuitive’ connections to coded caching

- Interesting questions arise:
  - ⋆ How to alleviate the real-time feedback bottlenecks?
  - ⋆ How does coded-caching break the linear barrier jointly with feed-back?

Cache-aided K-user MISO BC



- At the transmitter: N distinct files $W_1, \ldots, W_N$, each of size f bits;
- At the receiver, each user $k = 1, \ldots, K$ has a cache Z_k of size Mf bits.
- Placement phase (caching) and delivery phase (commun. after statement of requests)

- Received signal at receiver k

$$y_k = \boldsymbol{h}_k^T \boldsymbol{x} + z_k, \quad k = 1, \ldots, K$$

Measure of performance

- Measure of performance: the duration T of the delivery phase
  - ⋆ per file, per user
  - ⋆ T is a worst-case measure (guarantee any combination of file requests)
  - ⋆ high SNR setting, with $f = \log SNR$ (now T as in Maddah-Ali and Niesen)

- Equivalent measure: Throughput — cache-aided degrees of freedom

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

  - ⋆ R is the throughput of each user
  - ⋆ capture the synergistic effect of feedback and coded caching

General expression

Theorem 1 *In the cache-aided* K*-user MISO BC, with non-real time CSIT, with* $N \geq K$ *files of size* f*, and with caches of size* $M \in \{\frac{N}{K}, \frac{2N}{K}, \cdots, N\}$*, an achievable* T *is characterized as*

$$T = H_K - H_\Gamma,$$

where $H_K = \sum_{i=1}^{K} \frac{1}{i}$*, and* $\Gamma = \frac{KM}{N} = K\gamma$*.*

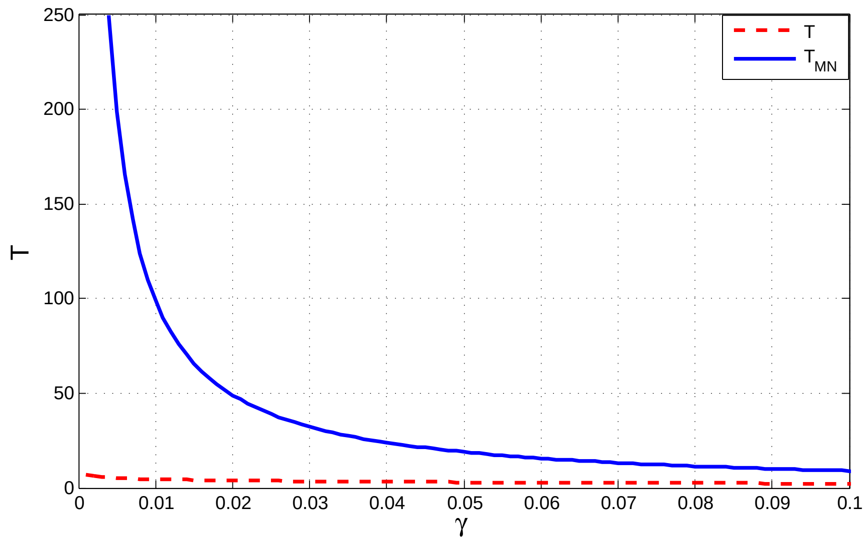
Under the logarithmic approximation, or in the large K *regime, the above* T *takes the form*

$$T \approx \log(\frac{1}{\gamma}) \tag{1}$$

Thus, the corresponding throughput R *for large* K *takes the form*

$$R \approx \frac{1}{\log(\frac{1}{\gamma})} \tag{2}$$

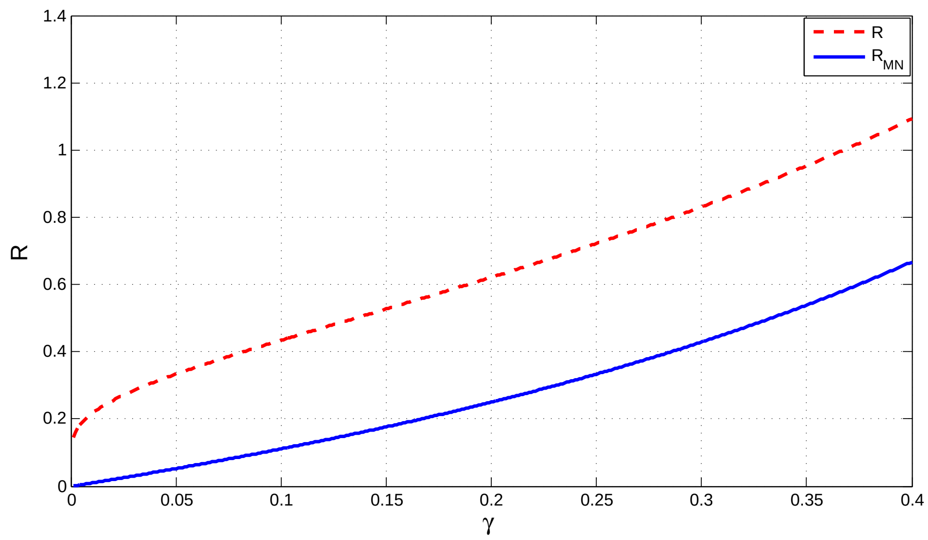
Linear barrier breakthrough



- In real systems, the operational value of γ will be relatively small
  - ⋆ e.g., when $\gamma = 10^{-6}$, $T \approx \log(1/\gamma) \approx 14$
- For the large K and reduced γ regime,

$$T \approx \log(\frac{1}{\gamma}) \quad \text{v.s} \quad T_{MN} = \frac{K(1-\gamma)}{1+K\gamma} \approx \frac{1-\gamma}{\gamma} \approx \frac{1}{\gamma} \tag{3}$$

Linear barrier breakthrough[1]



- For the large K and reduced γ regime,

$$R \approx \frac{1}{\log(\frac{1}{\gamma})} \quad \text{v.s} \quad R_{MN} \approx \gamma \tag{4}$$

- The linear barrier is broken by joint treatment of coded caching and retrospective communications.

Linear barrier breakthrough$_2$

- Without caching for BC, the optimal achievable throughput[1]

$$R = \frac{1}{\log K} \to 0 \tag{5}$$

- A microscopic $\gamma = e^{-G}$ could yield a very satisfactory

$$R(\gamma = e^{-G}) \approx \frac{1}{G} \tag{6}$$

  - ⋆ only a factor G from the interference free optimal $R = 1$.
    - ∗ e.g., $G = 7, \gamma \approx e^{-7} \approx 10^{-3}$, each cache can be one thousand times smaller than the library size.
  - ⋆ $T = G$: any linear decrease in the required performance allows for an exponential reduction in the required cache sizes.

[1]Optimality by Maddah-Ali and Tse 2012

Example

EXAMPLE: $N = K = 3, M = 1$ (i.e., $\gamma = \frac{M}{N} = \frac{1}{3}$)

- Placement: files A, B and C are equally split into 3 subfiles respectively, e.g.,

$$A = (\underbrace{A_1}_{\frac{f}{3}\text{bits}}, A_2, A_3)$$

  - ⋆ set caches $Z_1 = (A_1, B_1, C_1)$, $Z_2 = (A_2, B_2, C_2)$, $Z_3 = (A_3, B_3, C_3)$

- Delivery. Now you know the requests: $W_1 = A, W_2 = B, W_3 = C$.
  - ⋆ Wish to deliver

$$\underbrace{A_2 \oplus B_1}_{\frac{f}{3}\text{bits}}, A_3 \oplus C_1, B_3 \oplus C_2 \tag{7}$$

    - ∗ For simplification, we use AB to denote $A_2 \oplus B_1$, it is the same with others.

Example₁

- Retrospective transmission: two phases.
  - ⋆ Phase one: XORs are sent sequentially by vectors, e.g., $AB = (\underbrace{AB_1}_{\frac{1}{6}\text{bits}}, AB_2)$

$$\boldsymbol{x}_1 = \begin{bmatrix} AB_1 \\ AB_2 \\ 0 \end{bmatrix}, \boldsymbol{x}_2 = \begin{bmatrix} AC_1 \\ AC_2 \\ 0 \end{bmatrix}, \boldsymbol{x}_3 = \begin{bmatrix} BC_1 \\ BC_2 \\ 0 \end{bmatrix}$$

    - ∗ received signals

$$\begin{aligned} \text{User 1}: & f_1(AB_1, AB_2), f_2(AC_1, AC_2), f_3(BC_1, BC_2) \\ \text{User 2}: & f_4(AB_1, AB_2), f_5(AC_1, AC_2), f_6(BC_1, BC_2) \\ \text{User 3}: & f_7(AB_1, AB_2), f_8(AC_1, AC_2), f_9(BC_1, BC_2) \end{aligned}$$

  - ⋆ Phase two: common messages are sent

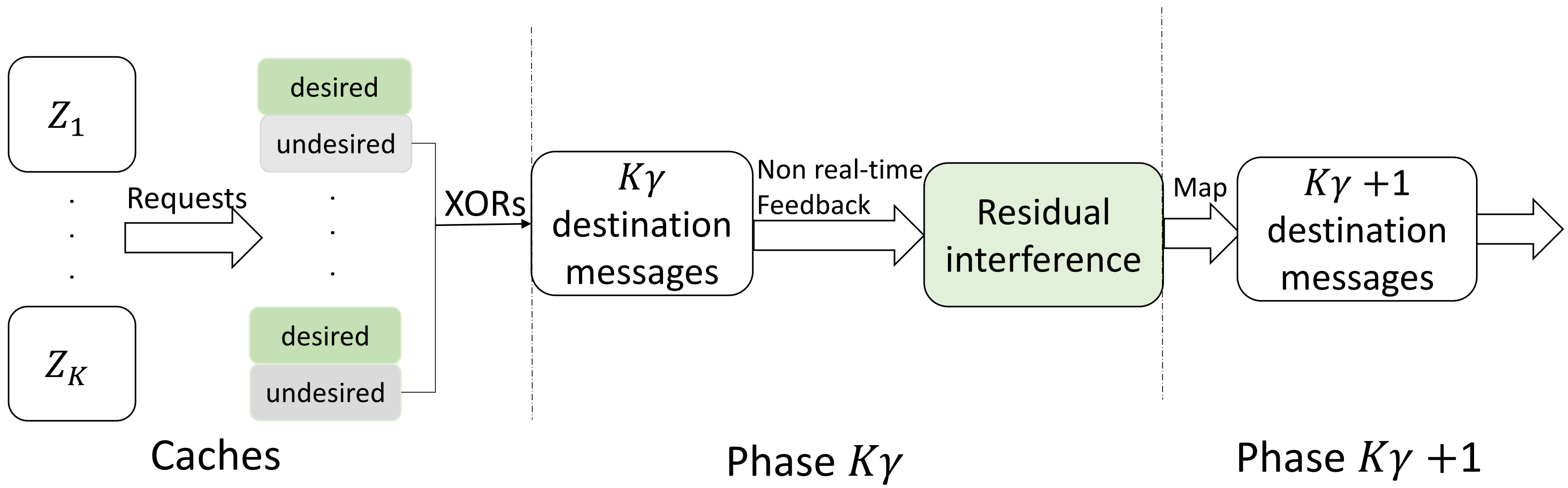
$$\boldsymbol{x}_4 = \alpha_1 f_3(BC_1, BC_2) + \alpha_2 f_5(AC_1, AC_2) + \alpha_3 f_7(AB_1, AB_2) \quad (8)$$
$$\boldsymbol{x}_5 = \beta_1 f_3(BC_1, BC_2) + \beta_2 f_5(AC_1, AC_2) + \beta_3 f_7(AB_1, AB_2) \quad (9)$$

    - ∗ $\alpha_i, \beta_i, i = 1, 2, 3$ are shared with the receivers.

Example$_2$

- Decoding
  - ⋆ Backwards from the received signals:
    - ∗ User 1 can decode AB_1, AB_2 and AC_1, AC_2;
    - ∗ User 2 can decode AB_1, AB_2 and BC_1, BC_2;
    - ∗ User 3 can decode AC_1, AC_2 and AC_1, AC_2;
  - ⋆ Recover $A_2 \oplus B_1, A_3 \oplus C_1, B_3 \oplus C_2$;
  - ⋆ With Z_k, user k reconstruct $W_{F_k}, k = 1, 2, 3$

Example$_3$


Z_1
Z_K
Requests
desired
undesired
desired
undesired
XORs
$K\gamma$
destination
messages
Non real-time
Feedback
Residual
interference
Map
$K\gamma$ +1
destination
messages
Caches
Phase $K\gamma$
Phase $K\gamma$ +1

Fundamental interplay with caching and feedback

Theorem 2 *The optimal* T^* *for the* (K, M, N) *cache-aided* K*-user MISO BC with delayed CSIT, is lower bounded as*

$$T^* \geq \max_{s \in \{1,\ldots,min\{\lfloor \frac{N}{M} \rfloor, K\}\}} \frac{s}{d_s^*(\gamma = 0)}\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right)$$
$$= \max_{s \in \{1,\ldots,min\{\lfloor \frac{N}{M} \rfloor, K\}\}} H_s\left(1 - \frac{M}{\lfloor \frac{N}{s} \rfloor}\right) \quad (10)$$

where $d_s^*(\gamma = 0) = \frac{s}{H_s}$ *is the optimal sum-DoF for the corresponding* s*-user MISO BC.*

Fundamental interplay with caching and feedback[1]

Theorem 3 *The achievable* $T = H_K - H_\Gamma$ *has a gap from the optimal*

$$\frac{T}{T*} < 2 \tag{11}$$

that is less that 2 for all K.

Conclusions

- A exploration of the fundamental limits of cache-aided BC with non-real time CSIT
  - ⋆ the optimal cache-aided DoF within a multiplicative factor of 2.
- Offer insight on the largely unexplored interplay between coded-caching and CSIT
- Our scheme exploited the interesting connections between
  - ⋆ retrospective transmission schemes;
  - ⋆ coded caching schemes.

THANKS