# The Interplay between Index Coding, Caching, and Beamforming for Fog Radio Access Networks

[1]Salwa Mostafa,[1]Chi Wan Sung,[2]Terence H. Chan,[3]Guangping Xu

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong.
[2]Institute for Telecommunication Research, University of South Australia, Australia.
[3]School of Computer and Communication Engineering, Tianjin University of Technology, China

*smostafa3-c@my.cityu.edu.hk*

GLOBECOM 2020

# Outline

1. Introduction
2. System Model
3. Cache Placement Schemes
4. Cache Delivery Schemes
5. Simulation Results
6. Conclusion

# Motivation

- **Problem**: Fronthaul link being the bottleneck.
- **Solution**: Edge caching.

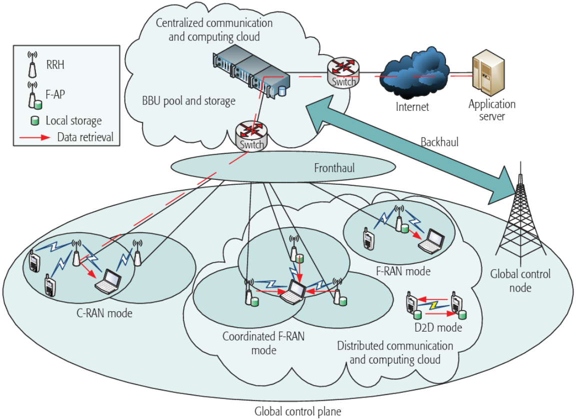


Figure 1: Fog-Radio Access Network (F-RAN) Architecture.[1]

# Objective and Contribution

- **Our Objective**
  - Joint design the cache placement and delivery to minimize the fronthaul traffic.
  - Consider two delivery schemes under random network connectivity.
    - Direct
    - Beamforming
- **Main Contribution**
  - Incorporate distributed beamforming.
  - Investigate the interplay between index coding, caching, and beamforming.
  - Beamforming can be exploited to reduce fronthaul traffic by more than 30%.

# System Model

Consider a F-RAN consists of

- A cloud server, $M$ cache-enabled F-APs, and $N$ users.
- A library of $F$ popular files, each of which has a size of $B$ bits.
- Each F-AP has a cache space of $C$ bits and a peak power constraint of $P$.
- F-AP $m$ is connected to user $n$ via a time-invariant Gaussian channel with amplitude gain $h_{nm}$.
- A target signal-to-noise ratio (SNR), $\gamma$, has to be met.
- If $|h_{nm}|^2 \geq \frac{\gamma}{P}$, the link is said to be *strong* and information can be successfully delivered.
- If $\frac{\gamma}{4P} \leq |h_{nm}|^2 < \frac{\gamma}{P}$, the link is said to be *weak*.
- Beamforming to transmit identical bits to user $n$

$$(|h_{nm}|^2 + |h_{nm'}|^2 + 2|h_{nm}||h_{nm'}|)P \geq \gamma.$$

# Cache Placement Schemes

We consider three caching schemes

- *Uncoded Caching* ($k = M$): Each F-AP $m$ stores the subfile $W_m^{(f)}$, for $m \in \mathcal{M}$.
- *Repetition Caching* ($k = \frac{M}{2}$): Each F-AP $m$ stores the subfile $W_{(m \bmod k)+1}^{(f)}$, for $m \in \mathcal{M}$.
- *MDS-Coded Caching* ($k \leq M$): The $k$ subfiles are encoded using an $(M + k, k)$ MDS code to obtain $M + k$ coded packets. The $M$ are placed in the F-APs and the remaining $k$, denoted by $\mathcal{Z}$, are stored only in the cloud.

Caching strategy *Most Popular First* (MPF).

# Cache Delivery Schemes

- Transmission modes over the access channel
  - Direct.
  - Beamforming.
- The connectivity is represented by a ternary association matrix,

$$\boldsymbol{A}_{N\times M} \triangleq [a_{nm}] = \begin{cases} 0, & missing \\ 1, & weak \\ 2, & strong \end{cases}$$

  - Fully Connected Networks :- if each user is associated, either weakly or strongly, to all F-APs.
  - Partially Connected Networks :- if each user is associated, either weakly or strongly, to some F-APs.

# Design Index Coding for Fully Connected Networks

- **Repetition caching**, each subfile of $W^{(f)}$ is stored twice. Thus, no fronthaul traffic.
- **Uncoded caching**, each user needs the subfiles of $W^{(f)}$ cached on all F-APs.
  - If an F-AP connects to all users via strong links, its cached subfile can be obtained by all users.
  - If an F-AP connects to some users via weak links, its subfile needs to be sent over the fronthaul.
    - Let $\mathcal{M}' \triangleq \{m \in \mathcal{M} \mid a_{nm} = 1 \text{ for some } n \in \mathcal{N}\}$.
    - For any distinct $i, j \in \mathcal{M}'$, if $W_i^{(f)} \oplus W_j^{(f)}$ is transmitted, all users can obtain both $W_i^{(f)}$ and $W_j^{(f)}$ via two packet transmissions either over one strong link or beamforming on two weak links.
    - The packets can be paired up arbitrarily to form XOR packets. If the number of those packets is odd, the unpaired one is sent uncoded.
  - The minimum number of packets need to deliver $W^{(f)}$ to all users is $\lceil |\mathcal{M}'|/2 \rceil$.

# Design Index Coding for Fully Connected Networks

- **MDS-coded caching**, to determine which packets to deliver, binary linear programming (LP) can be used:

$$\begin{aligned} \text{min} \quad & \sum_{m=1}^{M} x_m \\ \text{subject to} \quad & \boldsymbol{P}\boldsymbol{x} \geq \boldsymbol{r}, \end{aligned} \tag{1}$$

where

$$\boldsymbol{P}_{N\times M} \triangleq [P_{nm}] = \begin{cases} 1, & \text{missing or weak} \\ 0, & \textit{strong} \end{cases}$$

$$\boldsymbol{x} \triangleq [x_1, x_2, \ldots, x_m] = \begin{cases} 1, & \text{send } W_m^{(f)} \\ 0, & \text{otherwise} \end{cases}$$

$\boldsymbol{r} = \max(\boldsymbol{k} - \boldsymbol{s}, \boldsymbol{0})$ where $\boldsymbol{s} \triangleq (s_1, s_2, \ldots, s_N)$, user $n$ has $s_n$ strong links.

  - After an optimal vector $\boldsymbol{x}$ is obtained, the corresponding packets are paired up for XOR transmissions.
  - If the weight of $\boldsymbol{x}$ is odd, the last packet is transmitted without index coding.

# Fronthaul Traffic Analysis

We analyze the expected fronthaul traffic for each caching scheme.

## Theorem

*Consider a fully connected networks with $F = 2$ and $MC = 2B$. The two files are requested by a user with probability $p_1$ and $p_2 \triangleq 1 - p_1$, where $p_1 \geq 0.5$ and each link is strong with probability $q$ and weak with probability $1 - q$.*

- *For repetition caching, $E[\Lambda]$ is given by*

$$(1 - p_1^N)B.$$

- *For uncoded caching, $E[\Lambda]$ is given by*

$$\sum_{n=0}^{N} b_{N,p_1}(n) \sum_{j=0}^{M} \left[ b_{M,1-q^n}(j) + b_{M,1-q^{N-n}}(j) \right] \left\lceil \frac{j}{2} \right\rceil \frac{B}{M},$$

  *where $b_{N,p}(i) \triangleq \binom{N}{i} p^i (1-p)^{N-i}$.*

- *For MDS-coded caching with $k = \lceil Mq \rceil$, $E[\Lambda]$ is bounded below by*

$$(1 - p_1^N) \left( 1 - \frac{\lceil Mq \rceil}{M} \right) 2B, \;\; \text{for } q \geq 0.5.$$

  *Moreover, the lower bound is asymptotically tight when $M$ goes to infinity.*

# Design Index Coding for Partially Connected Networks

We first show that the problem with repetition caching can be reduced to that with uncoded caching.

- **Repetition caching**, every pair of F-APs that cache the same subfile can be combined into one single F-AP, so the network can be transformed into one that has $M/2$ F-APs with a new $N \times M/2$ association matrix $\boldsymbol{A}'$, whose entries are defined by $a'_{n,m} = \min(a_{n,m} + a_{n,m+M/2}, 2)$, for all $n \in \mathcal{N}$ and $m \in \mathcal{M}$.

**Example**

*Consider 4 F-APs and 2 users, a file* $W^{(f)} = \begin{bmatrix} W_1^{(f)} & W_2^{(f)} & W_1^{(f)} & W_2^{(f)} \end{bmatrix}$

$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 1 & 1 & 2 & 0 \end{bmatrix}$. *Then,* $\boldsymbol{A}' = \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix}$

- It suffices to design algorithms for uncoded caching and MDS-coded caching only.

# Optimal Index Coding for Uncoded Caching

- A pair of distinct subfiles, $i$ and $j$, denoted by $(i, j)$, is said to be a potential coded group, if the sum of each row of $A[i, j]$ is greater than or equal to two.
- It has the property that if $W_i \oplus W_j$ is transmitted over the fronthaul, all users must have both $W_i$ and $W_j$.

## Example

*Consider the file*

$$W^{(f)} = \begin{bmatrix} W_1^{(f)} & W_2^{(f)} & W_3^{(f)} \end{bmatrix}$$

*and the association matrix*

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

.

$W_1^{(f)} \oplus W_3^{(f)}$ *and* $W_2^{(f)} \oplus W_3^{(f)}$ *are potential coded groups while* $W_1^{(f)} \oplus W_2^{(f)}$ *is not.*

# Optimal Index Coding for Uncoded Caching

**Algorithm 1** Index Coding for Uncoded Caching in Partially Connected Networks

**Input :** A set of F-APs $\mathcal{M}$, a set of users $\mathcal{N}$, an association matrix $\boldsymbol{A}$.
**Output:** A set of packets $\mathcal{I}$.

1: Let $\mathcal{V} := \mathcal{M} \setminus \{m \in \mathcal{M} \mid a_{nm} = 2 \; \forall n \in \mathcal{N}\}$;
2: Construct a graph $G(\mathcal{V}, \mathcal{E})$, where $(i,j) \in \mathcal{E}$ if $(i,j) \in \mathcal{V}^2$ is a potential coded group;
3: Find a maximum matching $\mathcal{I}$ for $G$;
4: Add all unmatched vertices in $\mathcal{V}$ to $\mathcal{I}$;
5: **return** $\mathcal{I}$;

- The overall time complexity of Algorithm 1 is $O(NM^{2.5})$.
- Algorithm 1 is optimal.

# Heuristic Index Coding for MDS-Coded Caching

**Algorithm 2** Index Coding for MDS-Coded Caching in Partially Connected Networks

**Input :** A set of F-APs $\mathcal{M}$, a set of users $\mathcal{N}$, an association matrix $\boldsymbol{A}$, a set of MDS coded packets $\mathcal{Z}$.

**Output:** A set of packets $\mathcal{I}$.

```
Let r_n be the extra number of packets required by user n for n ∈ 𝒩;
Let 𝒱 := ℳ \ {m ∈ ℳ | a_nm = 2 ∀n ∈ 𝒩};
Construct a graph G(𝒱, ℰ), where (i, j) ∈ ℰ if (i, j) ∈ 𝒱^2 is a potential coded group;
Find a maximum matching 𝒫 for G;
while r_n > 0 for some n do
    if 𝒫 is non-empty then
        Move an arbitrary element p from 𝒫 to ℐ;
        Update r_n for all n, assuming p is broadcast;
    else
        Move max_n r_n elements from 𝒵 to ℐ;
        Let r_n := 0 for all n;
    end if
end while
return ℐ;
```

- The overall time complexity of Algorithm 1 is $O(NM^{2.5})$.

# Simulation Parameters

- Single cell of radius $R$ with a cloud server located at its center.
- The F-APs and the users are randomly distributed according to a homogeneous poisson point process.
- The F-APs are restricted to an inner concentric circle with radius $R/2$ while the users are distributed over the whole cell.

Table 1: Parameters for Partially Connected Networks

| Parameters | Value |
|---|---|
| Cell radius ($R$) | 500 m |
| Number of F-APs ($M$) | 10 F-APs |
| Number of Users ($N$) | $5-55$ users |
| F-APs Peak Power ($P$) | 2 W |
| Target SNR ($\gamma$) | $6-14$ dB |
| Path loss at distance $d$ Km | $140.7+36.7\log_{10} d$, dB |
| Noise Power ($\sigma^2$) (10 MHz bandwidth) | $-102$ dBm |
| Number of Files ($F$) | 10 files |
| Distribution Skewness ($\alpha$) | 1.5 |
| File Size ($B$) | 100 Mbits |
| Cache Size ($C$) | 100 Mbits |

# Outage Probability

- A user is said to be in outage if he is unable to obtain his requested file.

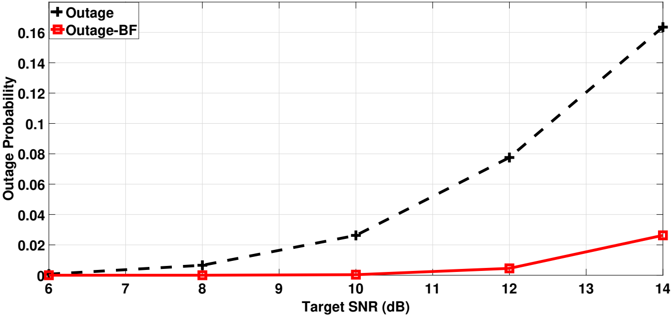


Figure 2: Outage probability for partially connected networks.

Fig. 2 shows that beamforming reduces outage probability significantly for high target SNR.

# Fronthaul Traffic Load

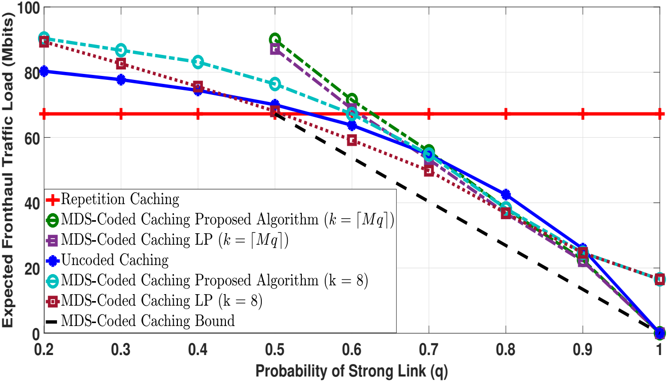


Figure 3: Expected fronthaul traffic load with beamforming for fully connected networks where $N = 5$, $M = 10$, $F = 2$ and $p_1 = 0.8$.

# Fronthaul Traffic Load

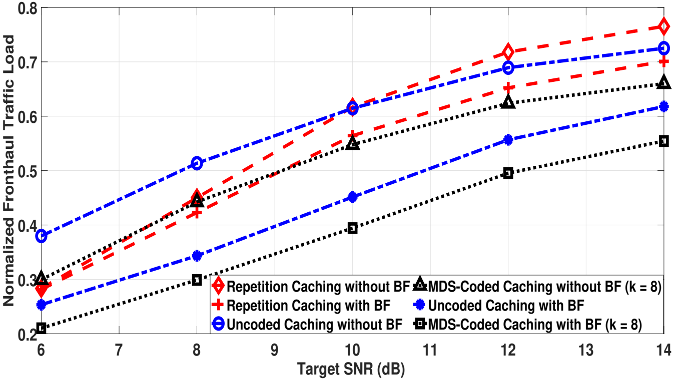


Figure 4: Normalized fronthaul traffic load for partially connected network where $N = 15$ and $\alpha = 1.5$.

# Fronthaul Traffic Load

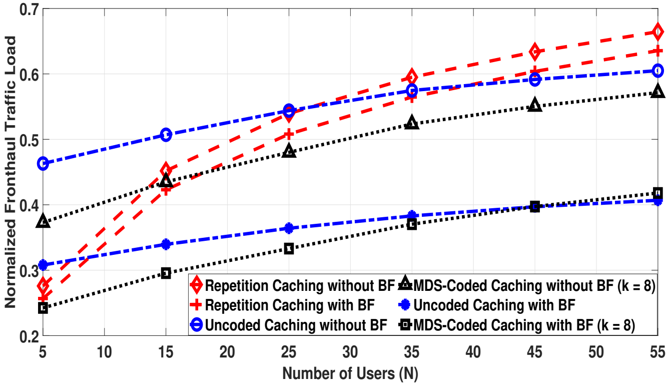


Figure 5: Normalized fronthaul traffic load for partially connected network where $\gamma = 8$ dB and $\alpha = 1.5$.

# Conclusion

- Distributed beamforming is a promising physical-layer technique to increase cell coverage and boast received SNR.
- Distributed beamforming can lower the outage probability and the fronthaul traffic load of a F-RAN with cache-enabled F-APs.
- MDS-coded caching, in general, outperforms the uncoded and repetition caching schemes, except only in more extreme cases.

# References


Y.-J. Ku, D.-Y. Lin, C.-F. Lee, P.-J. Hsieh, H.-Y. Wei, C.-T. Chou, and A.-C. Pang, “5g radio access network design with the fog paradigm: Confluence of communications and computing,” *IEEE Communications Magazine*, vol. 55, no. 4, pp. 46–52, 2017.

M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

K. Zhang and C. Tian, “Fundamental limits of coded caching: From uncoded prefetching to coded prefetching,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1153–1164, 2018.

X. Wu, Q. Li, V. C. Leung, and P. Ching, “Joint fronthaul multicast and cooperative beamforming for cache-enabled cloud-based small cell networks: An MDS codes-aided approach,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4970–4982, 2019.

R. Sun, Y. Wang, N. Cheng, L. Lyu, S. Zhang, H. Zhou, and X. Shen, “QoE-driven transmission-aware cache placement and cooperative beamforming design in cloud-RANs,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 636–650, 2019.

M.-M. Zhao, Y. Cai, M.-J. Zhao, B. Champagne, and T. A. Tsiftsis, “Improving caching efficiency in content-aware C-RAN-based cooperative beamforming: A joint design approach,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4125–4140, 2020.

M. Tao, E. Chen, H. Zhou, and W. Yu, “Content-centric sparse multicast beamforming for cache-enabled cloud RAN,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6118–6131, 2016.

S. Mostafa, C. W. Sung, and G. Xu, “Code rate maximization of cooperative caching in ultra-dense networks,” in *IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, 2019.