This work was supported in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 690893.

Decentralized Coded Caching in Wireless Networks: Trade-off between Storage and Latency

Antonious M. Girgis[1], Ozgur Ercetin[2], Mohammed Nafie[1,3] and Tamer ElBatt[1,3]

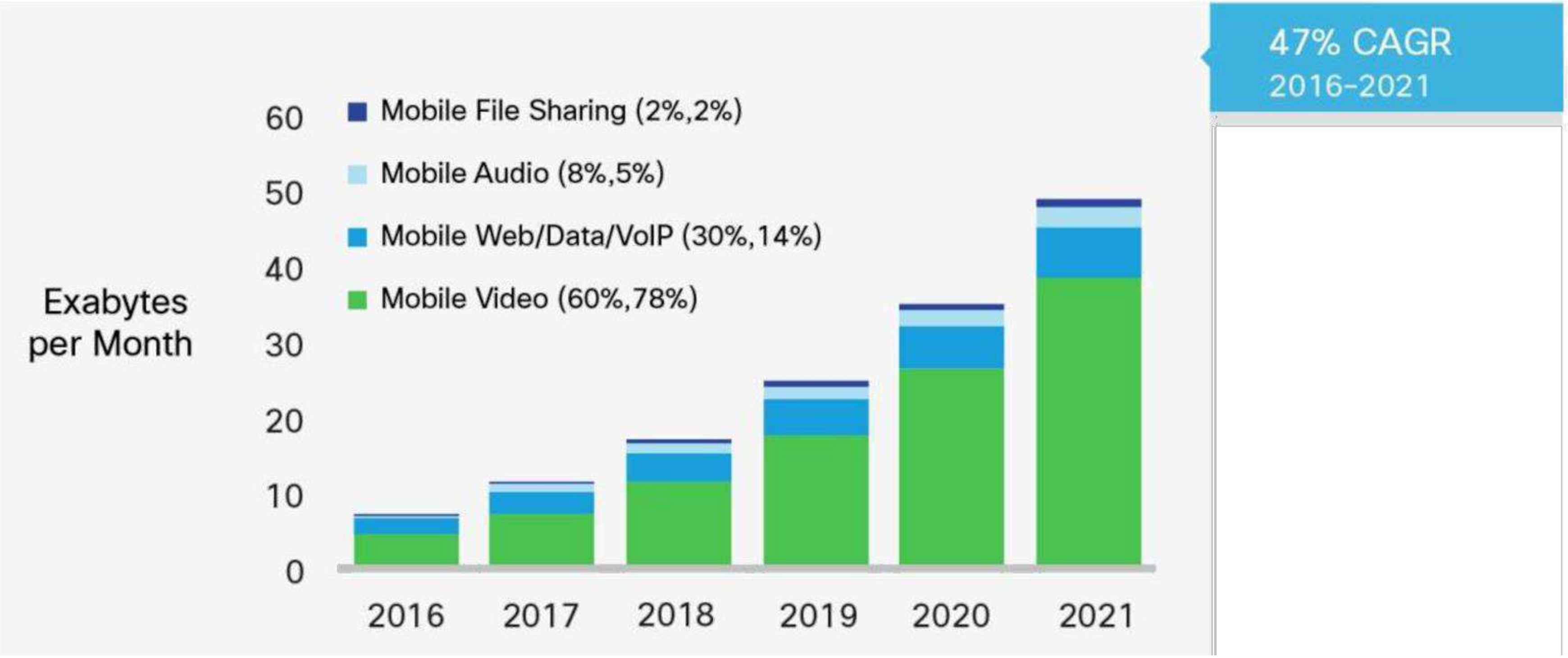
[1]WIRELESS Intelligent Network Center (WINC), Nile University, Cairo, Egypt

[2]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey

[3]Electronics and Communications Dept., Faculty of Engineering, Cairo University, Giza, Egypt

Context and motivation

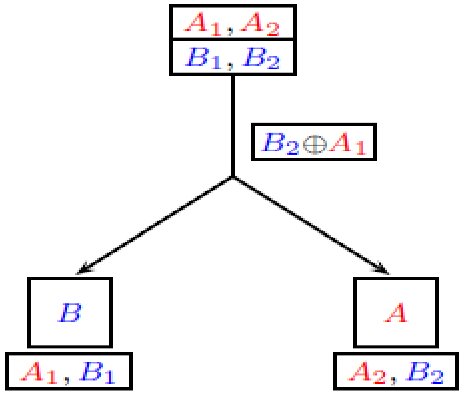
- The increasing number of wireless devices is leading to rapid evolution of the traffic load.
- Mobile video will generate much of the mobile traffic growth through 2021 [1].



- Caching most popular contents close to user terminals is a promising solution.

[1] Cisco visual networking index: Global mobile data traffic forecast white paper, Feb. 2017.

Context and motivation



- An information theoretic view for caching systems [1].

  A novel centralized caching scheme for an error-free broadcast channel

- Decentralized placement [2].

  The caches of users are filled independently for each other.

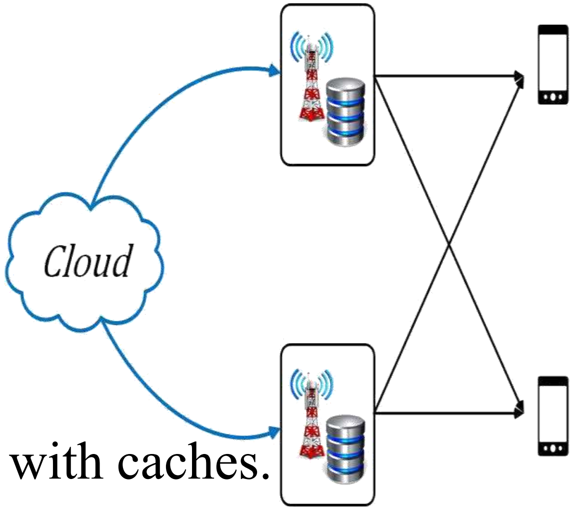
- Coded caching concept for interference networks[3]

  Characterization of DoF for 3×3 interference channel with caches at transmitters.

- Coded Caching in Fog-Radio Access Networks (F-RAN)[4]

  Centralized baseband processing at the cloud $+$ Edge processing at ENs equipped with caches.

  Introducing Normalized Delivery Time (NDT) as a performance metric.

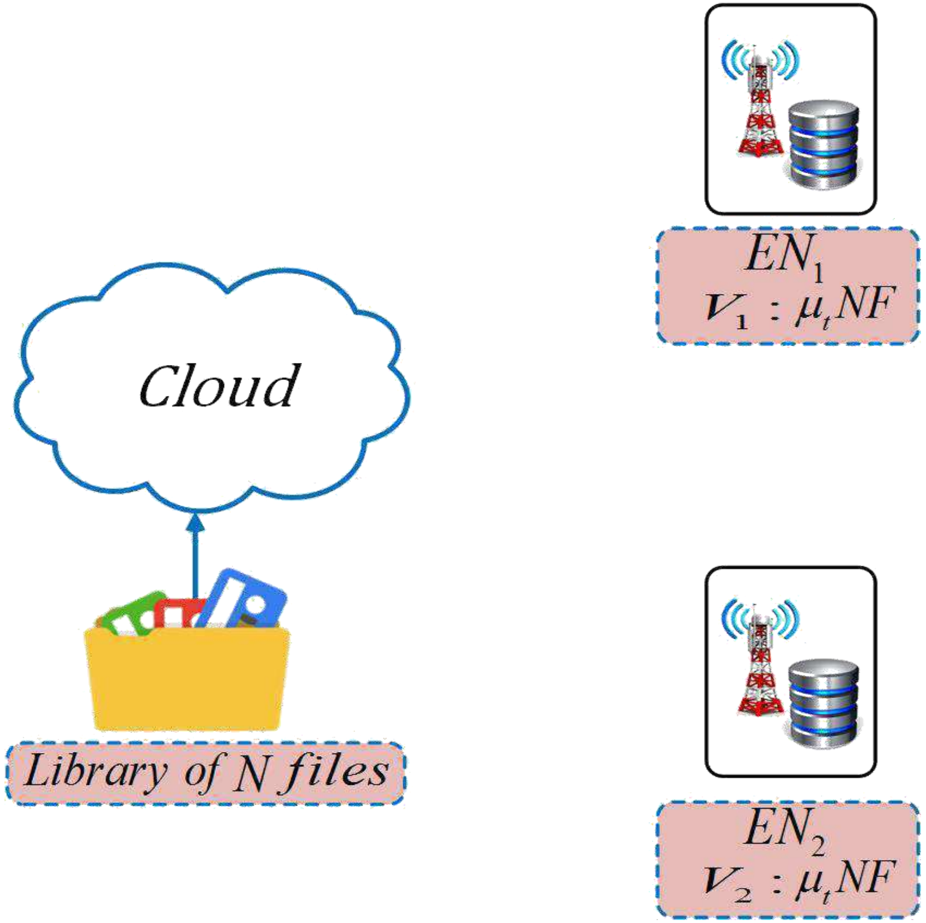


Decentralized coded caching problem for $2 \times K_r$ F-RAN architecture.

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," IEEE Trans. Inf. Theory, 2014.
[2] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," IEEE/ACM Trans. Netwo., 2015.
[3] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in IEEE ISIT, 2015.
[4] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in IEEE ISIT, 2016.

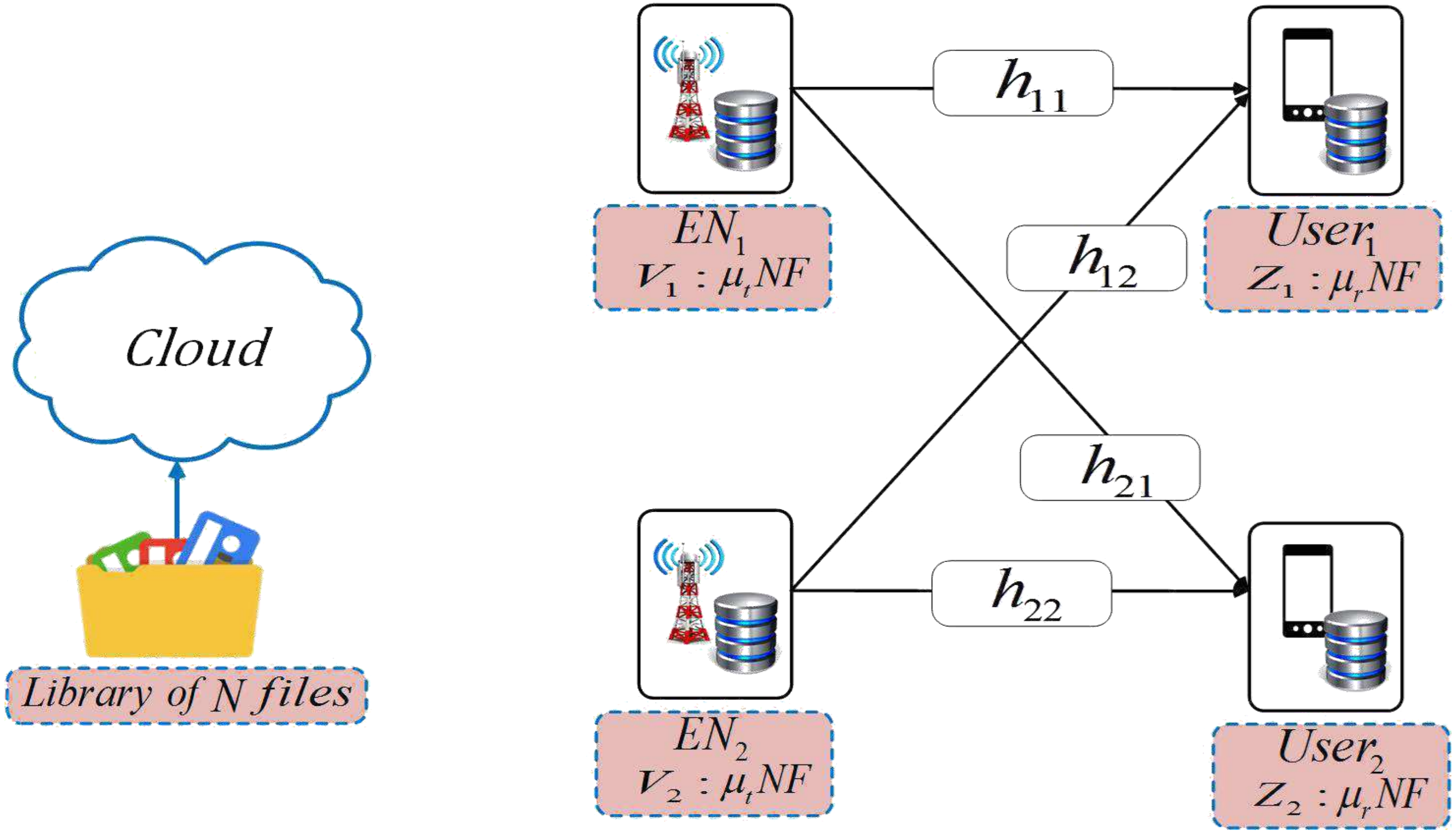
System model

- A cloud server has a library of N files $\mathcal{W} \triangleq \{W_1, \cdots, W_N\}$ each of size F bits.
- A set of K_t Edge nodes, $EN_1, \cdots, EN_{K_t}$, each is equipped with a cache memory V_i of size $\mu_t NF$ bits.



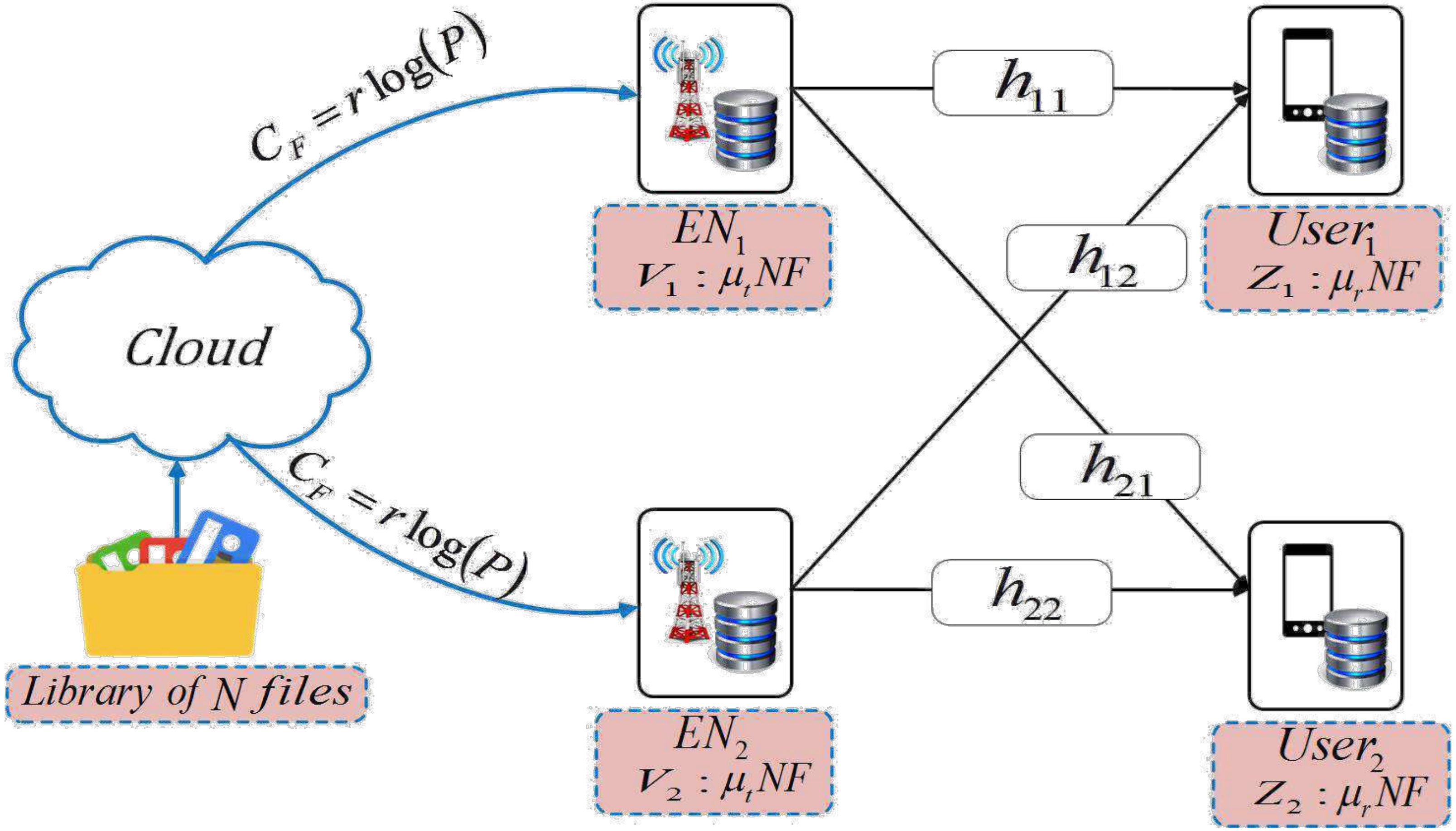
System model

- ENs are ready to serve requests of K_r users through a Gaussian interference channel.
- User k is equipped with a cache memory Z_k of size $\mu_r NF$ bits.



System model

- The cloud is connected to each EN via a fronthaul link of capacity $C_F = r\log(P)$ bits per channel use, where r measures to the multiplexing gain of the fronthaul link.



System model

- **Placement Phase:** Fill the cache memory of each node at the off-peak hours.
- **Delivery Phase:** Users' requests are revealed
  - Fronthaul Transmission: The cloud sends a message U_i of block length T_F to EN_i over the fronthaul link.
  - Edge Transmission: Each EN_i transmits a message X_i of block length T_E over the wireless channel.
  - Decoding function: Each user k can decode the requested file from its cache contents Z_k and the received signals $Y_k(1), \cdots, Y_k(T_E)$.

$$Y_k(t) = \sum_{i=1}^{K_t} h_{ki} X_i(t) + Z_k(t)$$

  - Performance metric: *Normalized Delivery time (NDT)*

$$\delta = \lim_{P\to\infty} \lim_{F\to\infty} sup \frac{T}{F/\log(P)}$$

T refers to end-to-end latency of transmission.

$F/\log(P)$ is the delivery time for interference-free, unlimited cache system.

Transmission types

- **Serial transmission:** Fronthaul and edge transmissions occur *consecutively.*
  - End-to-end Latency $T = T_F + T_E$.

$$\delta_S = \delta_F + \delta_E$$

- **Pipelined transmission:** Fronthaul and edge transmissions occur *simultaneously*.
  - End-to-end latency $T = \max(T_F, T_E)$.

$$\delta_P = \max(\delta_F, \delta_E).$$

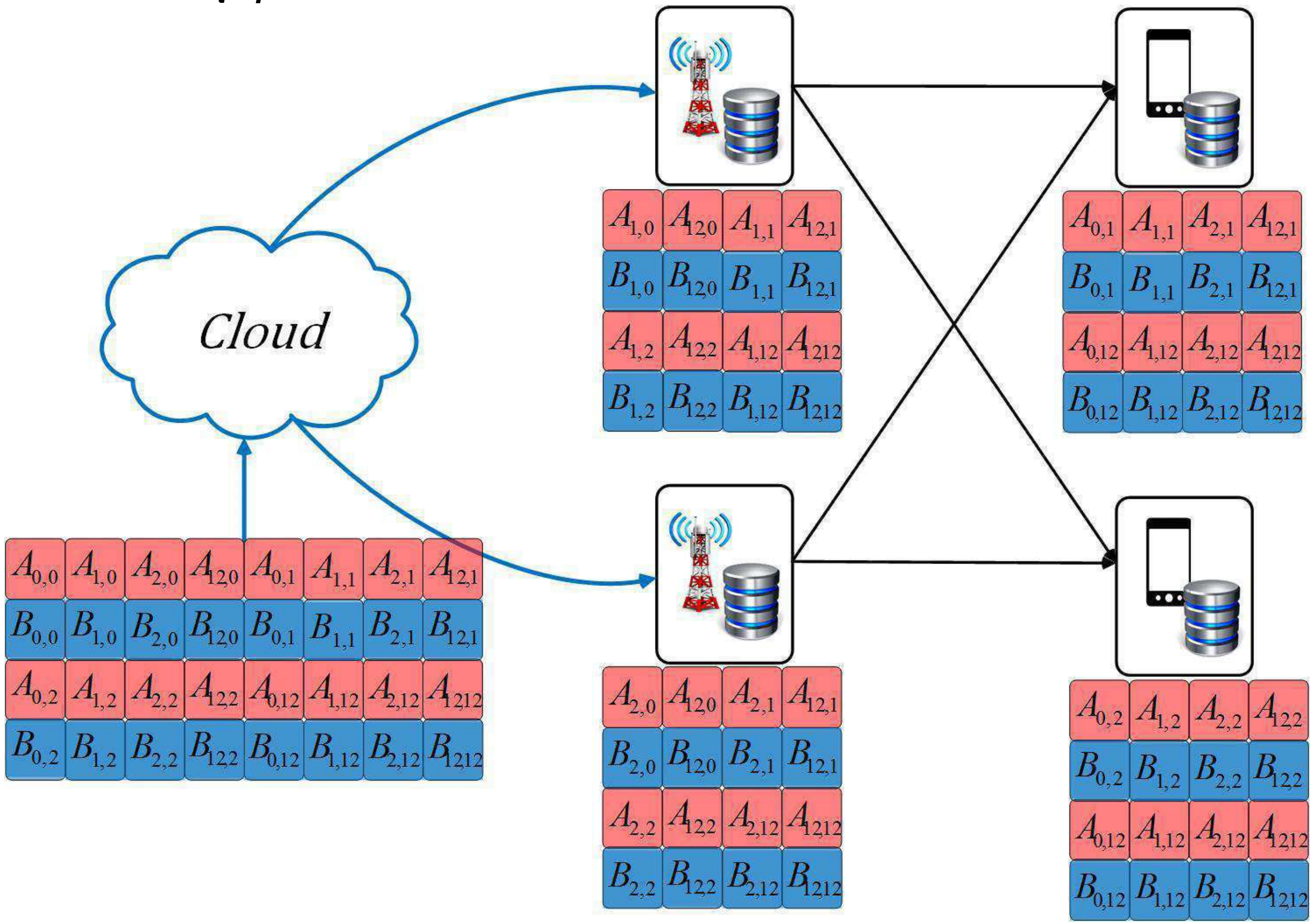
Decentralized placement

- Each EN stores independently at random $\mu_t F$ bits from each file.
- Each user stores independently at random $\mu_r F$ bits from each file.

Each file is split into $2^{K_t+K_r}$ fragments.

$$W_{j,S_t,S_r}$$

Fragment of file j stored exclusively at $S_t \subset [K_t]$ ENs and $S_r \subset [K_r]$ users.



Coded delivery

- Let user one request file A and user two request file B.
- User one wants fragments $A_{0,0}, A_{1,0}, A_{2,0}, A_{12,0}, A_{0,2}, A_{1,2}, A_{12,2}$
- User two wants fragments $B_{0,0}, B_{1,0}, B_{2,0}, B_{12,0}, B_{0,1}, B_{1,1}, B_{12,1}$
- The delivery scheme is divided into five stages.

Coded delivery

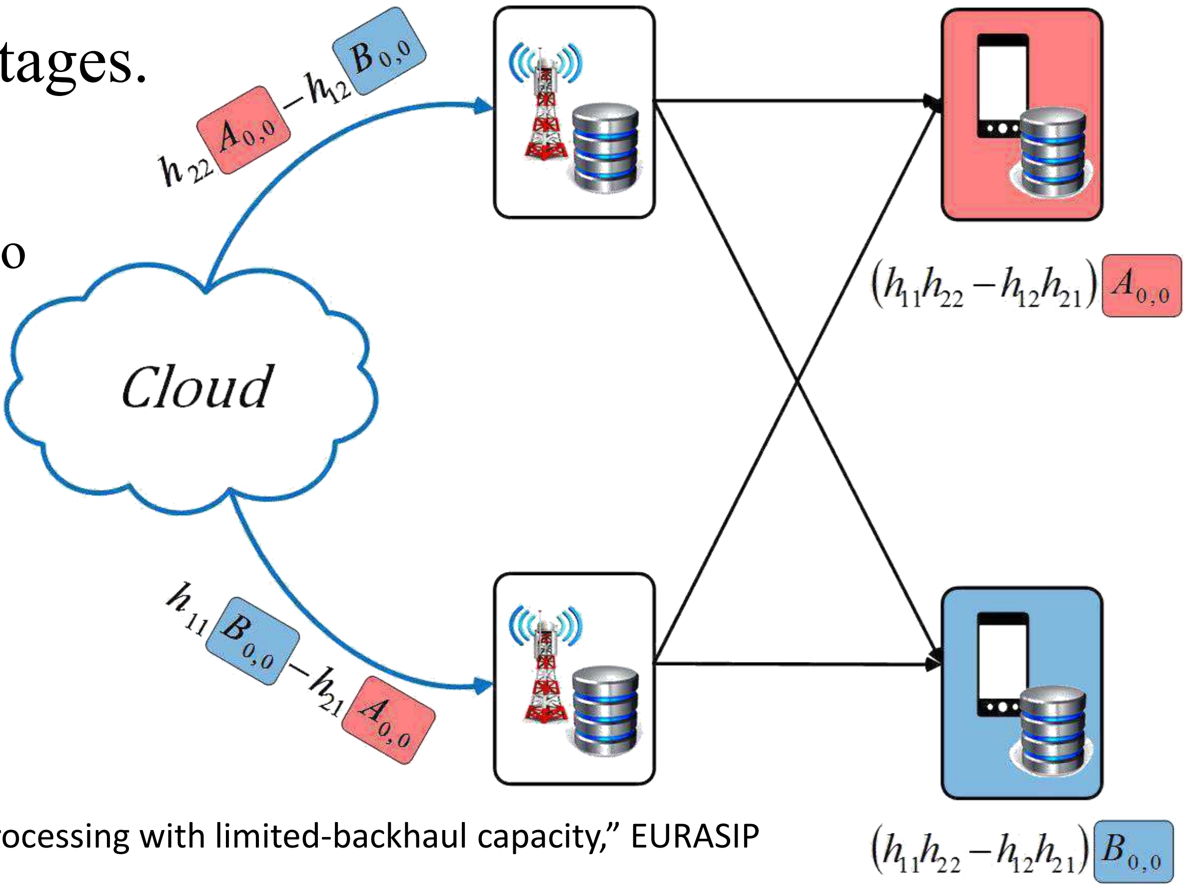
- Let user one request file A and user two request file B.
- User one wants fragments $A_{0,0}, A_{1,0}, A_{2,0}, A_{12,0}, A_{0,2}, A_{1,2}, A_{12,2}$
- User two wants fragments $B_{0,0}, B_{1,0}, B_{2,0}, B_{12,0}, B_{0,1}, B_{1,1}, B_{12,1}$
- The delivery scheme is divided into five stages.

   1. Delivery of fragments $A_{0,0}, B_{0,0}$

   - The cloud implements Zero-forcing beamforming to be transmitted to ENs over the fronthaul links [1].
   - ENs deliver the cloud messages to users over the wireless channel.

$$\delta_F^{(1)} = \frac{|A_{0,0}|}{r}$$

$$\delta_E^{(1)} = |A_{0,0}|$$



[1] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," EURASIP Journal on Advances in Signal Processing, 2009.

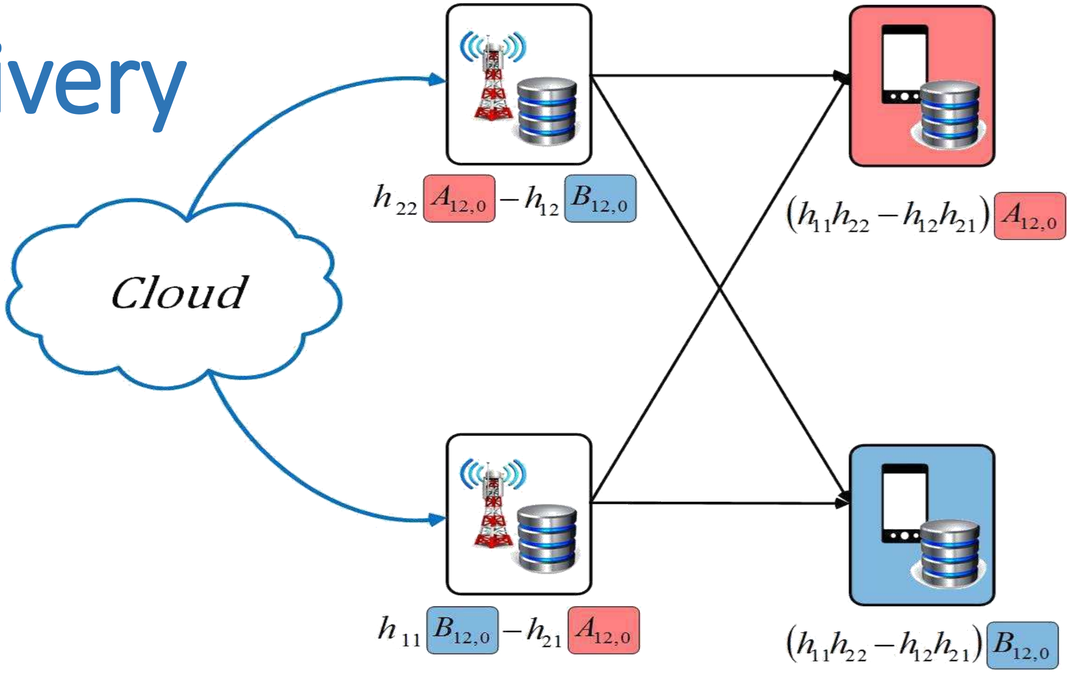
Coded delivery

2. Delivery of fragments $A_{12,0}$, $B_{12,0}$

- ENs apply ZF-beamforming over interference channel

$$\delta_F^{(2)} = 0$$

$$\delta_E^{(2)} = |A_{12,0}|$$



Coded delivery

2. Delivery of fragments $A_{12,0}, B_{12,0}$

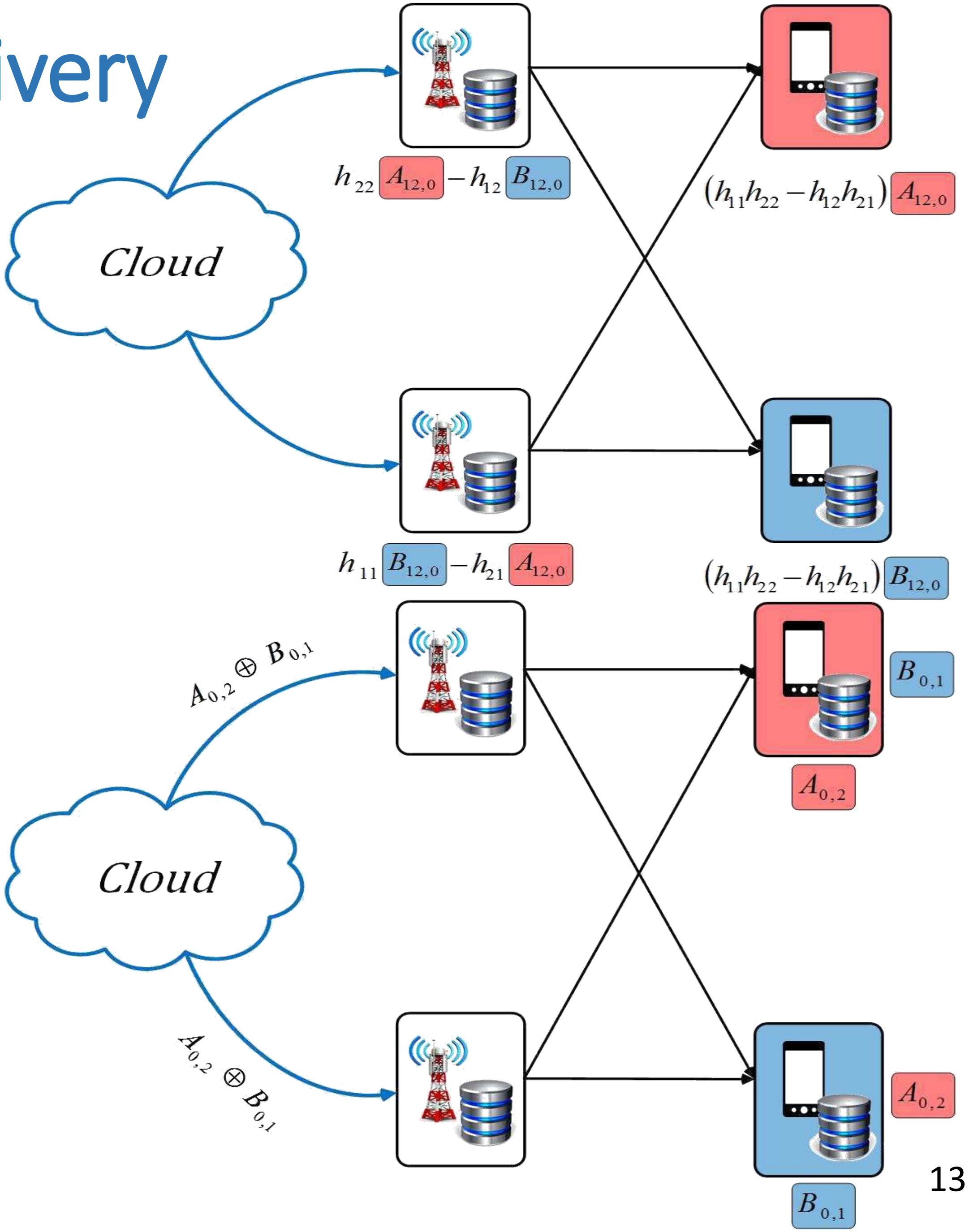
- ENs apply ZF-beamforming over interference channel

$$\delta_F^{(2)} = 0$$
$$\delta_E^{(2)} = |A_{12,0}|$$

3. Delivery of fragments $A_{0,2}, B_{0,1}$

- The cloud sends $A_{0,2} \oplus \mathrm{B}_{0,1}$.
- ENs deliver the cloud messages to users over the wireless channel.

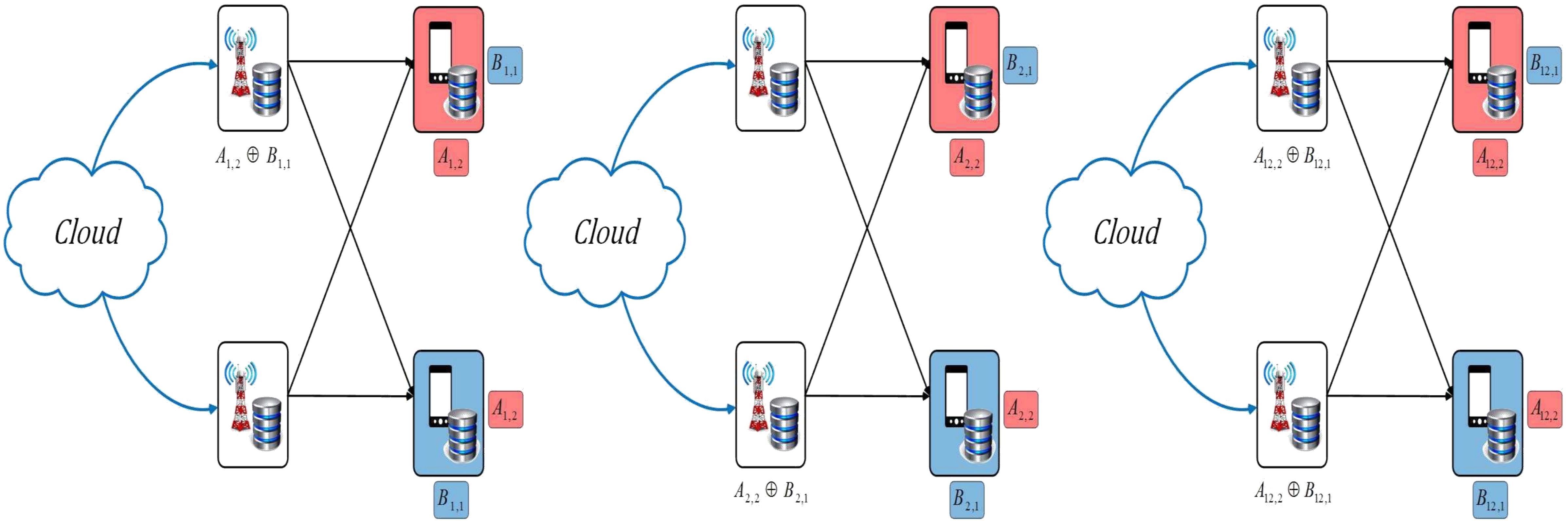
$$\delta_F^{(3)} = \frac{|A_{0,2}|}{r}$$
$$\delta_E^{(3)} = |A_{0,2}|$$



Coded delivery

4. Delivery of fragments $A_{1,2}, B_{1,1}, A_{2,2}, B_{2,1}, A_{12,2}, B_{12,1}$

- EN_1 sends $A_{1,2} \oplus \mathrm{B}_{1,1}$.
- EN_2 send $A_{2,2} \oplus \mathrm{B}_{2,1}$.
- ENs send $A_{12,2} \oplus \mathrm{B}_{12,1}$.



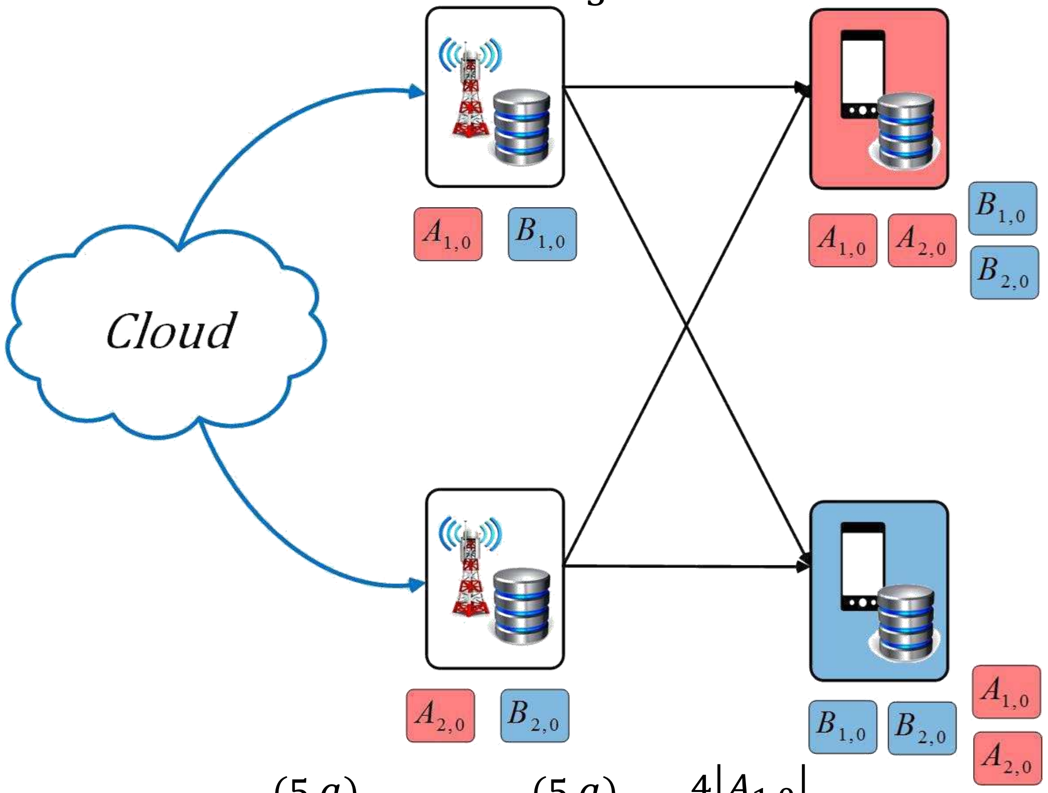
$$\delta_F^{(4)} = 0,\ \delta_E^{(4)} = |A_{1,2}| + |A_{2,2}| + |A_{12,2}|$$

Coded delivery

5. Delivery of fragments $A_{1,0}, B_{1,0}, A_{2,0}, B_{2,0}$

A. Mapping to 2×2 X-channel.

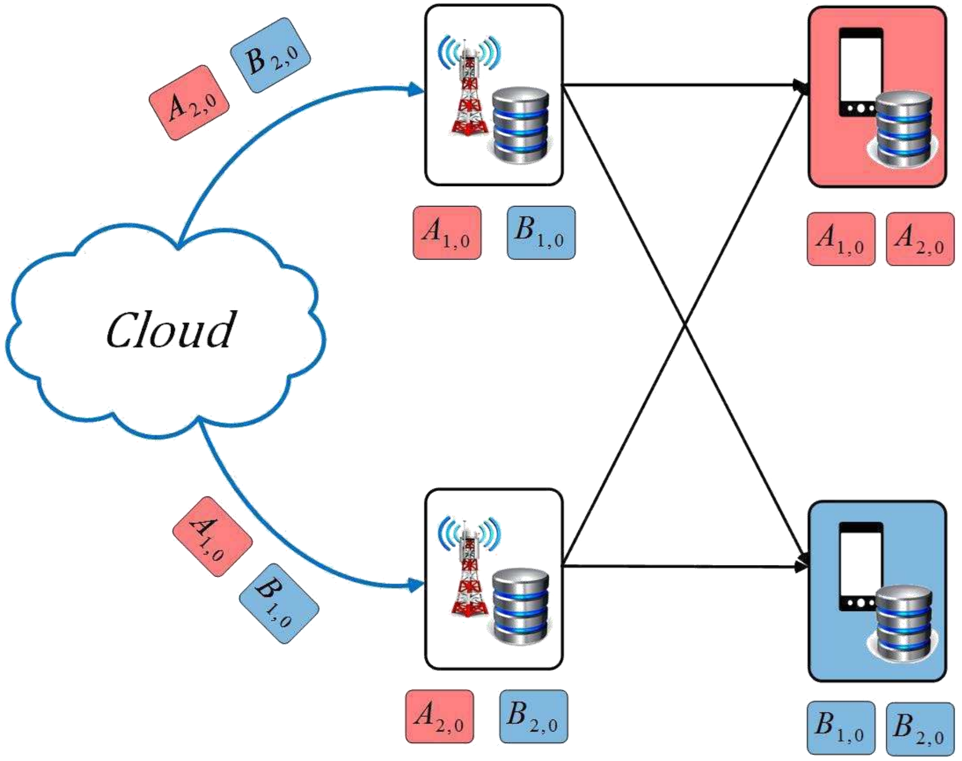
- EN_s apply interference alignment scheme [1] to achieve $DoF = \frac{4}{3}$.



$$\delta_F^{(5,a)} = 0,\ \delta_E^{(5,a)} = \frac{4|A_{1,0}|}{4/3} = 3|A_{1,0}|$$

B. Mapping to MISO broadcast channel.

- Cloud send $A_{1,0}, B_{1,0}$ to EN_2, and $A_{2,0}, B_{2,0}$ to EN_1.
- EN_s apply ZF-beamforming to achieve $DoF = 2$.



$$\delta_F^{(5,b)} = \frac{2|A_{1,0}|}{r},\ \delta_E^{(5,b)} = \frac{4|A_{1,0}|}{2} = 2|A_{1,0}|$$

Achievable Scheme performance

Theorem: For $2 \times K_r$ F-RAN with decentralized caching placement

$$\delta_S^{dec} = \begin{cases} \delta_F^{(a)} + \delta_E^{(a)} & r < K_r \\ \delta_F^{(b)} + \delta_E^{(b)} & r \geq K_r \end{cases}$$

$$\delta_P^{dec} = \min\left\{\max\left(\delta_F^{(a)}, \delta_E^{(a)}\right), \max\left(\delta_F^{(b)}, \delta_E^{(b)}\right)\right\}$$

$$\delta_F^{(a)} = \frac{(1-\mu_t)^2(1-\mu_r)}{r\mu_{\mathrm{r}}}\left[1-(1-\mu_r)^{K_r} - \frac{K_r}{2}\mu_r(1-\mu_r)^{K_r-1}\right]$$

$$\delta_F^{(b)} = \frac{(1-\mu_t)^2(1-\mu_r)}{r\mu_{\mathrm{r}}}\left[1-(1-\mu_r)^{K_r} - \frac{K_r}{2}\mu_r(1-\mu_r)^{K_r-1}\left(\frac{1-3\mu_t}{1-\mu_t}\right)\right]$$

$$\delta_E^{(a)} = \frac{(1-\mu_r)}{\mu_{\mathrm{r}}}\left[1-(1-\mu_r)^{K_r} - \left(\frac{K_r}{2}-\mu_t(1-\mu_t)\right)\mu_r(1-\mu_r)^{K_r-1}\right]$$

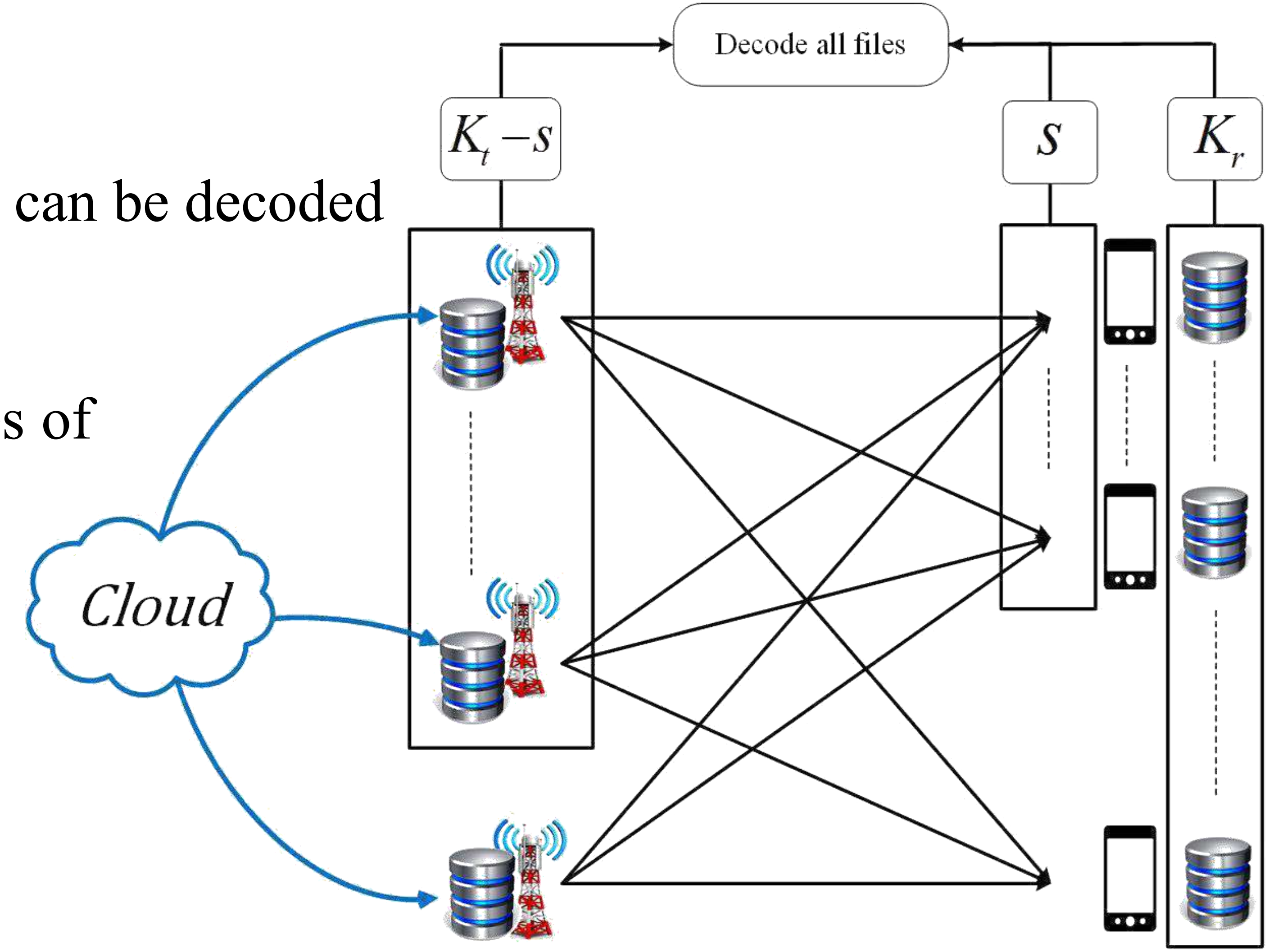
$$\delta_E^{(b)} = \frac{(1-\mu_r)}{\mu_{\mathrm{r}}}\left[1-(1-\mu_r)^{K_r} - \frac{K_r}{2}\mu_r(1-\mu_r)^{K_r-1}\right]$$

Lower bound

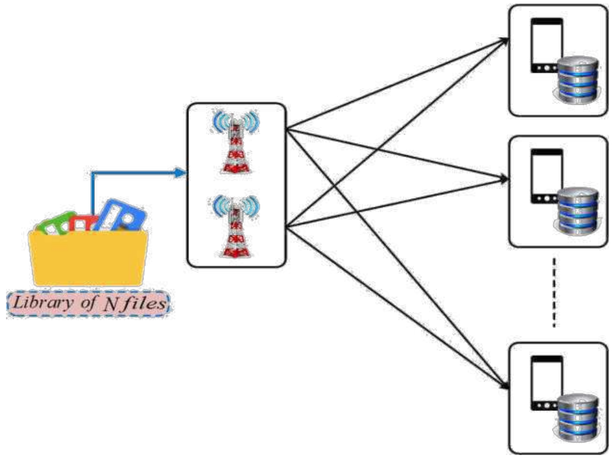
Using cut set argument, K_r requests can be decoded by observing

- s channel output
- cache contents and cloud messages of $(K_t - s)\ EN_s$
- cache contents of K_r users.

For $s \in \{0, \cdots, \min(K_t, K_r)\}$



Special cases



For $\mu_t = 1$, and/or $r = \infty$: two EN_s have all the library.

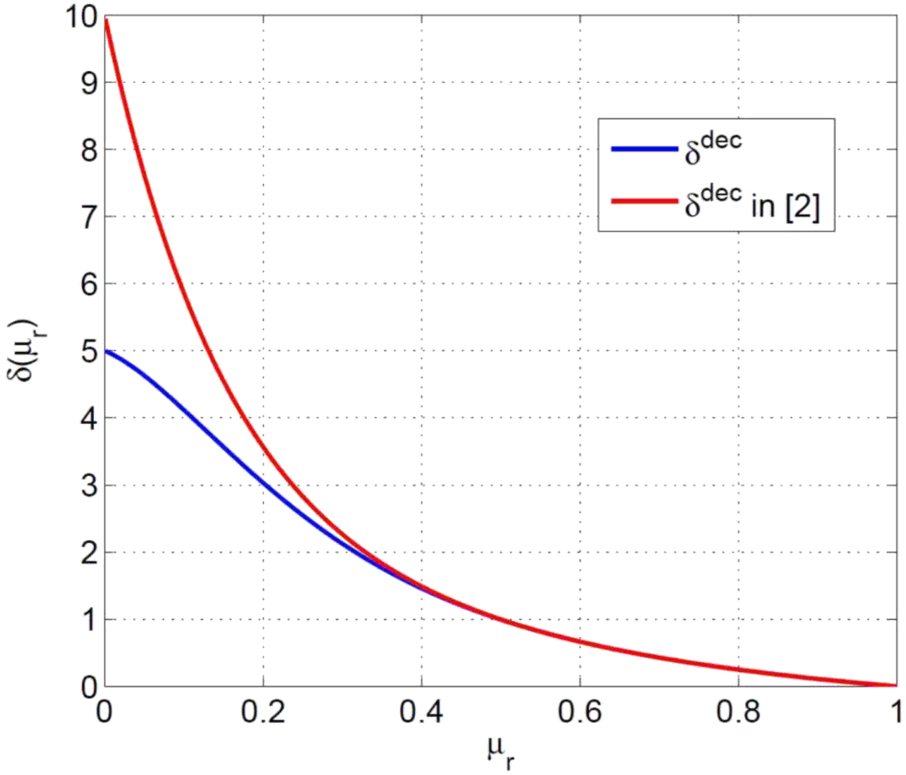
$$\delta = \frac{(1-\mu_r)}{\mu_r}\left[1-(1-\mu_r)^{K_r} - \frac{K_r}{2}\mu_r(1-\mu_r)^{K_r-1}\right]$$

The NDT of broadcast channel with single antenna [2]

$$\delta = \frac{(1-\mu_r)}{\mu_r}\left[1-(1-\mu_r)^{K_r}\right]$$

The achievable NDT is lower than NDT in [2] with term

$$\frac{K_r}{2}(1-\mu_r)^{K_r}$$



[2] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," IEEE/ACM Trans. Netwok., 2015.

Special cases

For $\mu_r = 0$ **:** The caches are available at EN_s only.

Theorem: For pipelined transmission, the decentralized scheme is optimal in region $r \geq (1 - \mu_t^2)$ and $0 < r \leq (1 - \mu_t^2)$, $0 < \mu_t \leq 0.5$.
For serial transmission, the decentralized scheme achieves

$$\frac{\delta_S^{dec}}{\delta_S^*} \leq 3$$

in region $r \geq 1$ and $0 < r < 1, 0 \leq \mu_t \leq \sqrt{2} - 1$

The decentralized scheme is not tight for large cache size $(\mu_t > 0.5)$ due to the random placement.

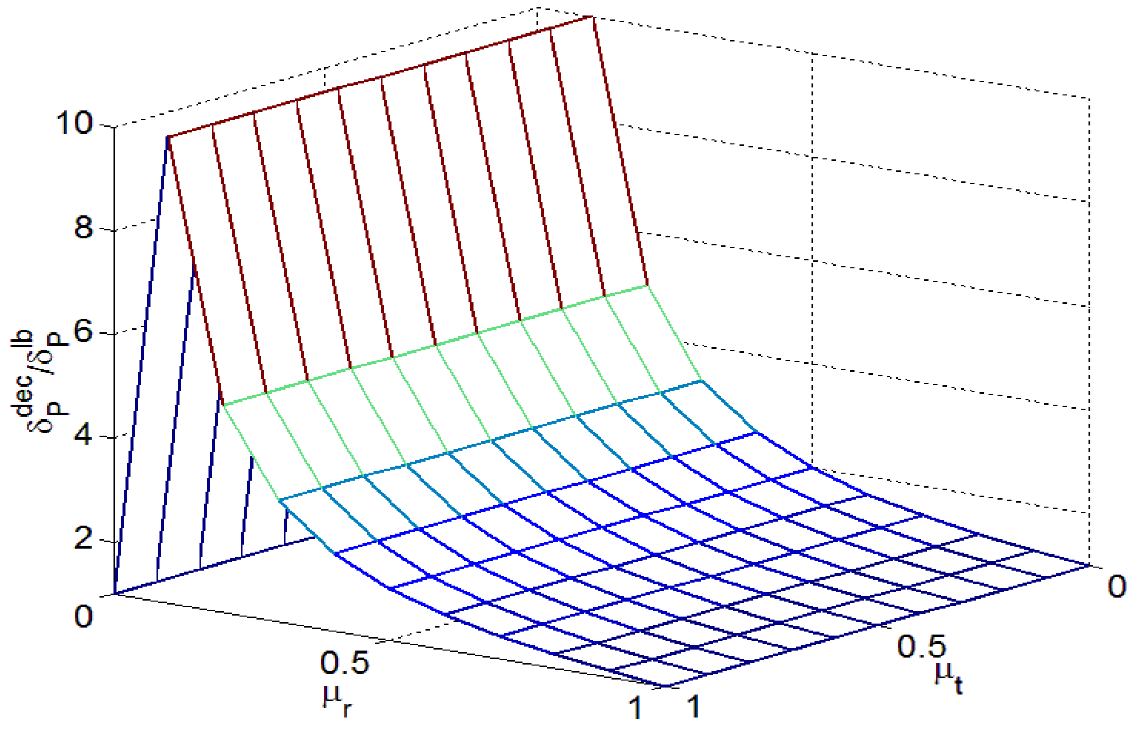
[2] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," IEEE/ACM Trans. Netwok., 2015.

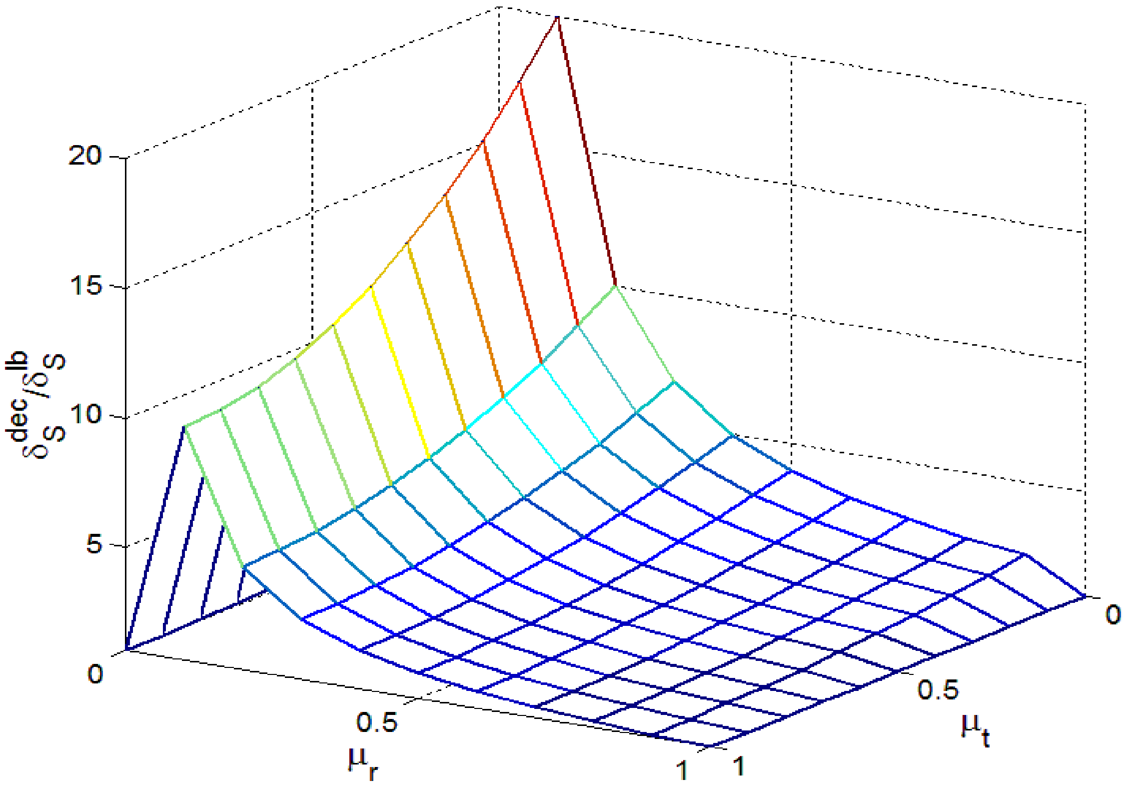
Numerical Results

The performance of the decentralized scheme for the general case

- $K_r = 100.$
- $r = 1.$

- The maximum gap is about 10 for pipelined transmission.
- The maximum gap is about 20 for serial transmission.





[2] M. A. Maddah-Ali and U. Niesen, “Decentralized coded caching attains order-optimal memory-rate tradeoff,” IEEE/ACM Trans. Netwok., 2015.

Conclusion

- Studying the decentralized coded caching for $2 \times K_r$ Fog-Radio Access Networks for pipelined and serial transmissions.
- Deriving lower bound on the minimum NDT for $K_t \times K_r$ Fog-Radio Access Networks for pipelined and serial transmissions.
- Evaluation of the performance of the decentralized scheme.
- Extensions:
  - Characterization of the NDT for $K_t \times K_r$ F-RAN with decentralized placement.
  - Studying the NDT for an F-RAN under nonuniform distribution.