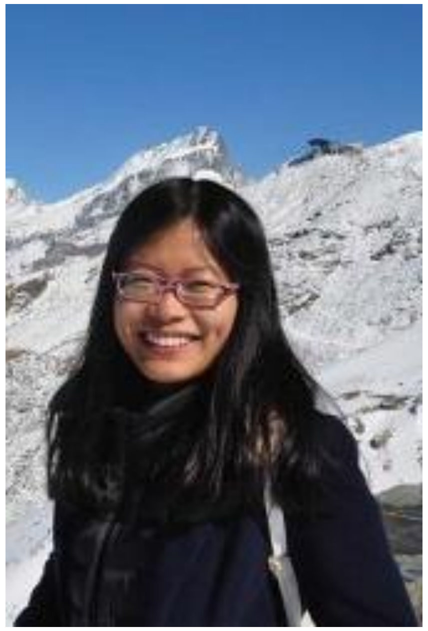
Fog-Aided Wireless Networks: An Information-Theoretic View

Osvaldo Simeone



Jingjing Zhang
King's College London

Ravi Tandon
University of Arizona

Avik Sengupta
Virginia Tech

Mohammadreza Azimi
NJIT

Shlomo Shamai
Technion

Seok-Hwan Park
Chonbuk University

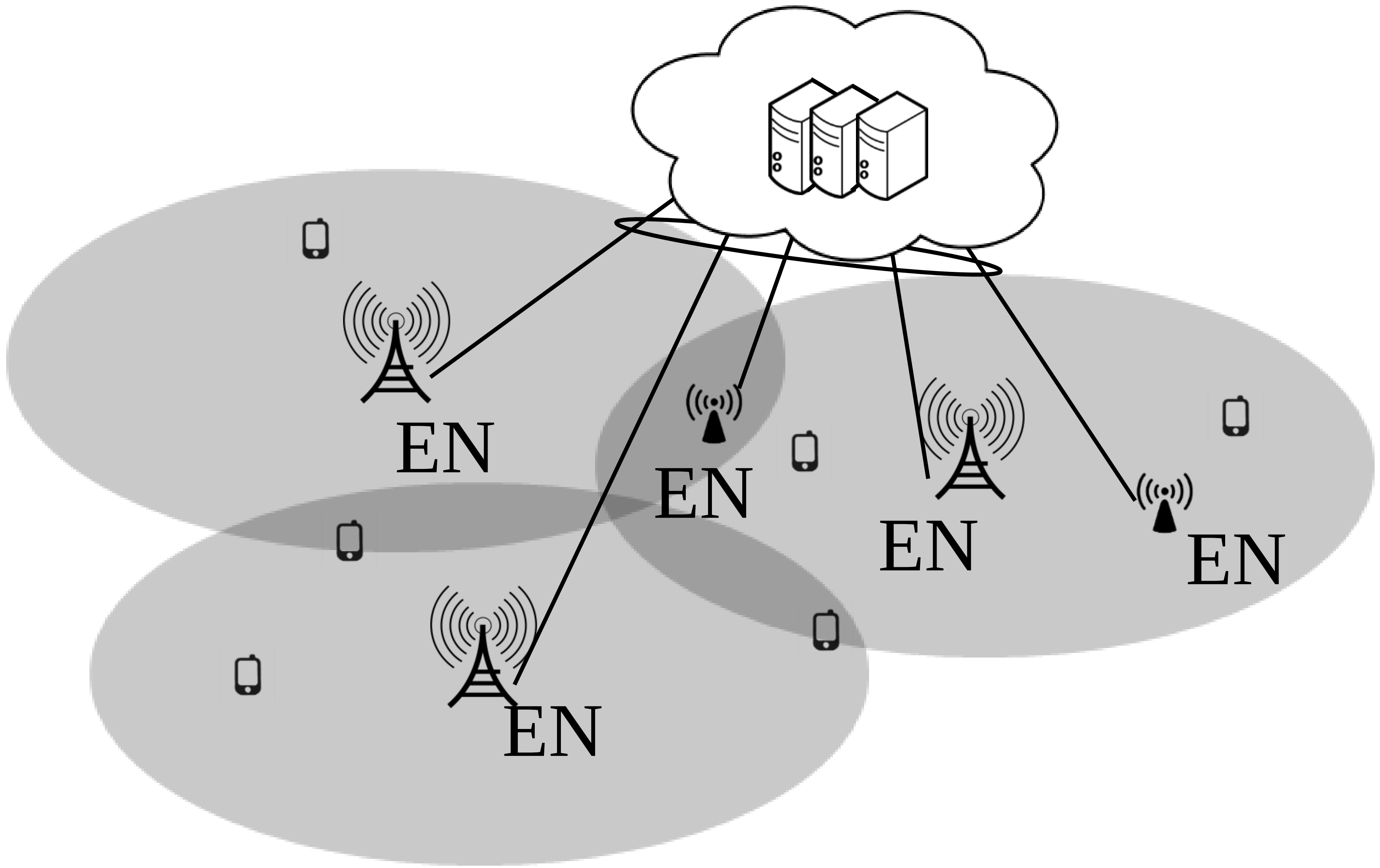
Roy Karasik
Technion

Joonhyuk Kang
KAIST

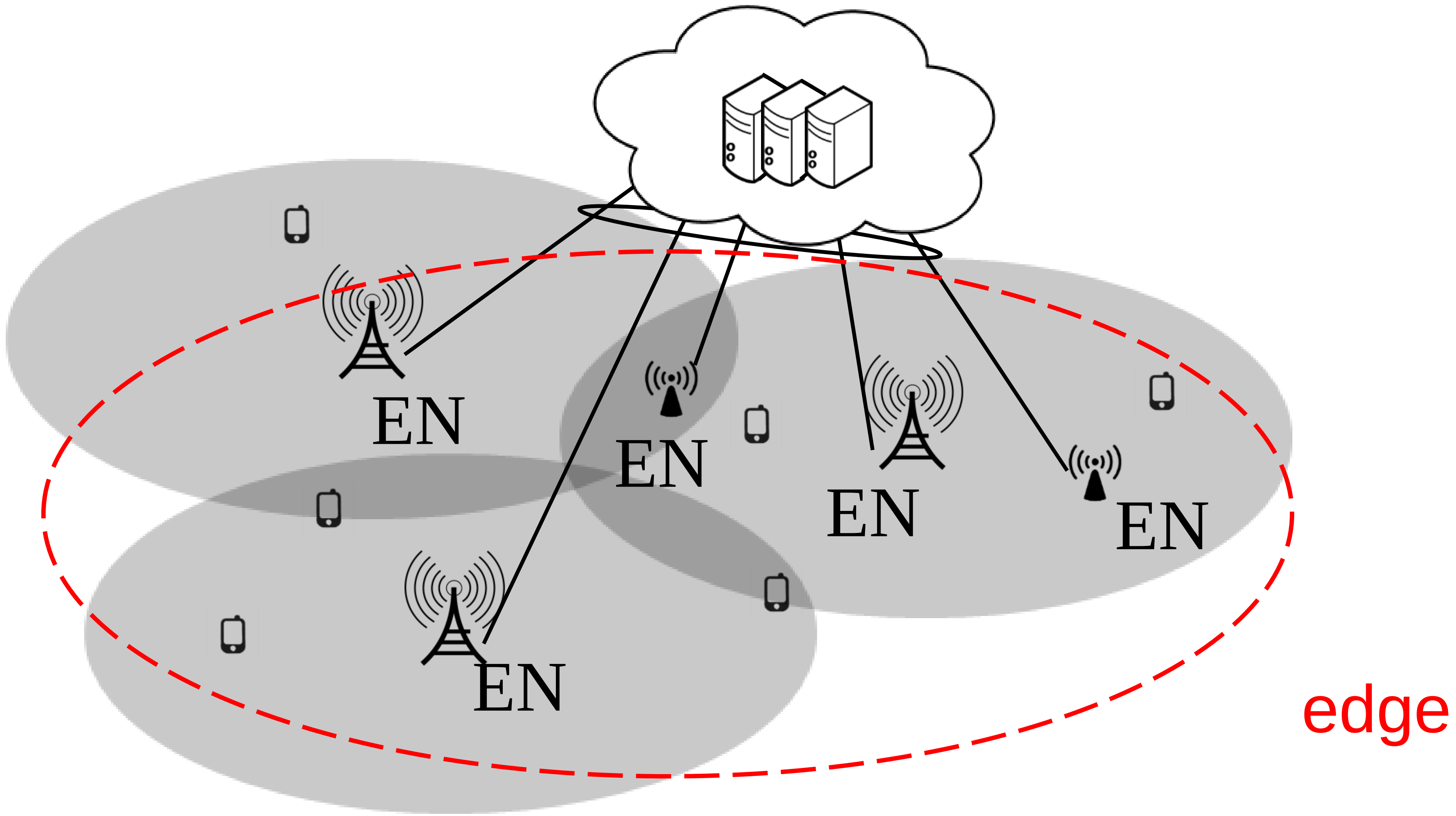
Jasper Goseling
Univ. of Twente

Jeongwan Koh
KAIST

Fog-Radio Access Network (F-RAN)

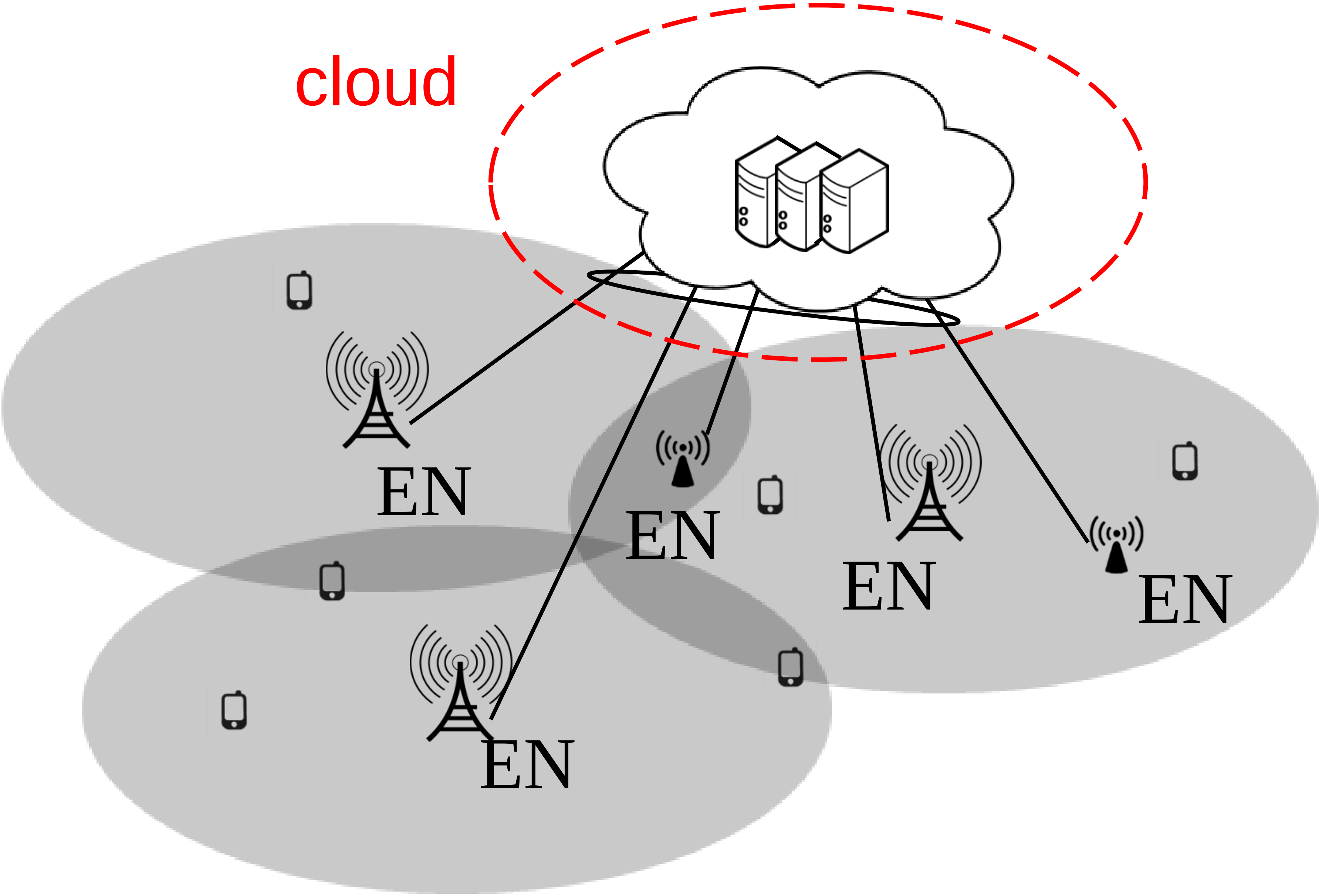


Fog-Radio Access Network (F-RAN)

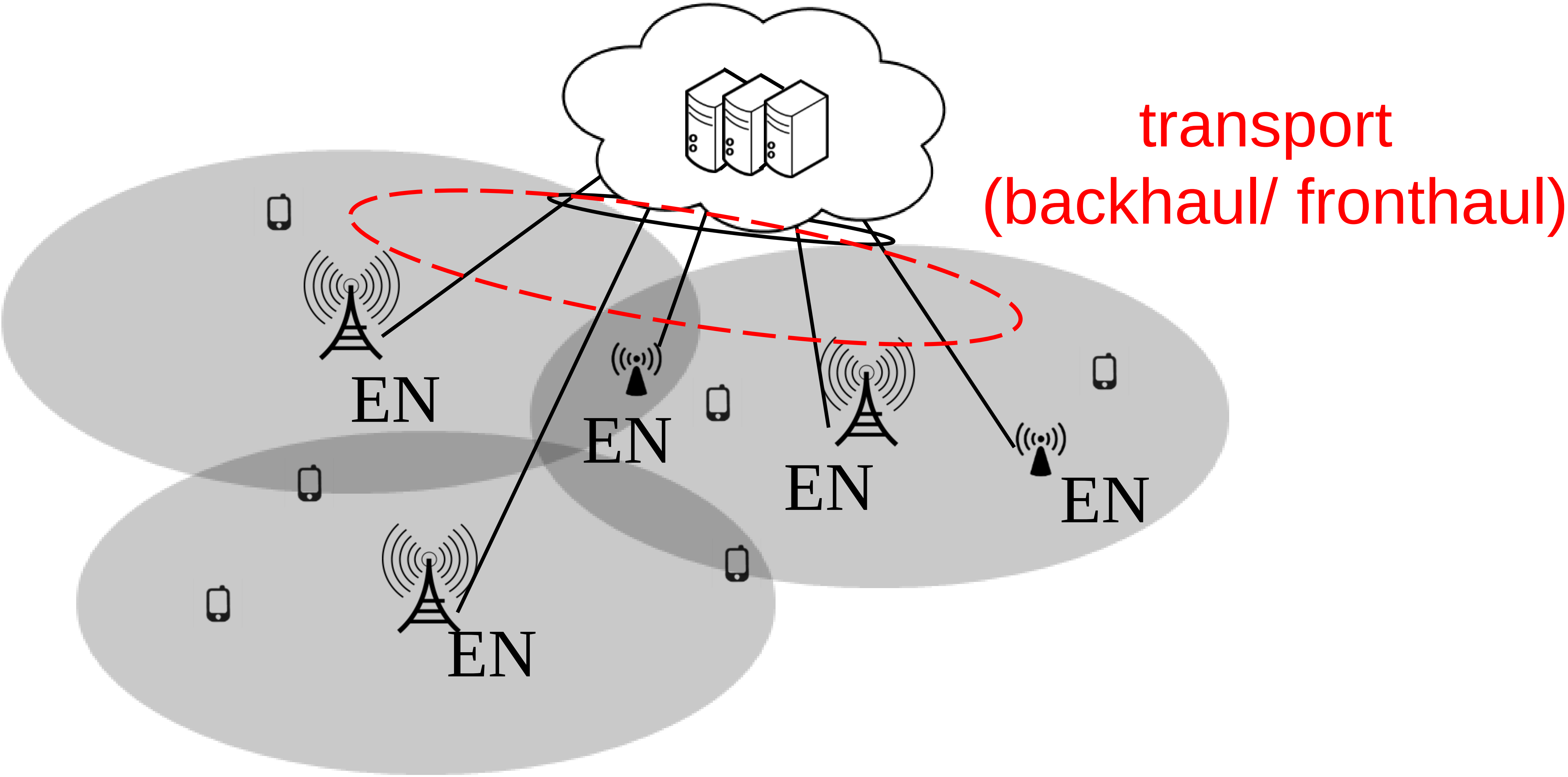


EN = Edge Node

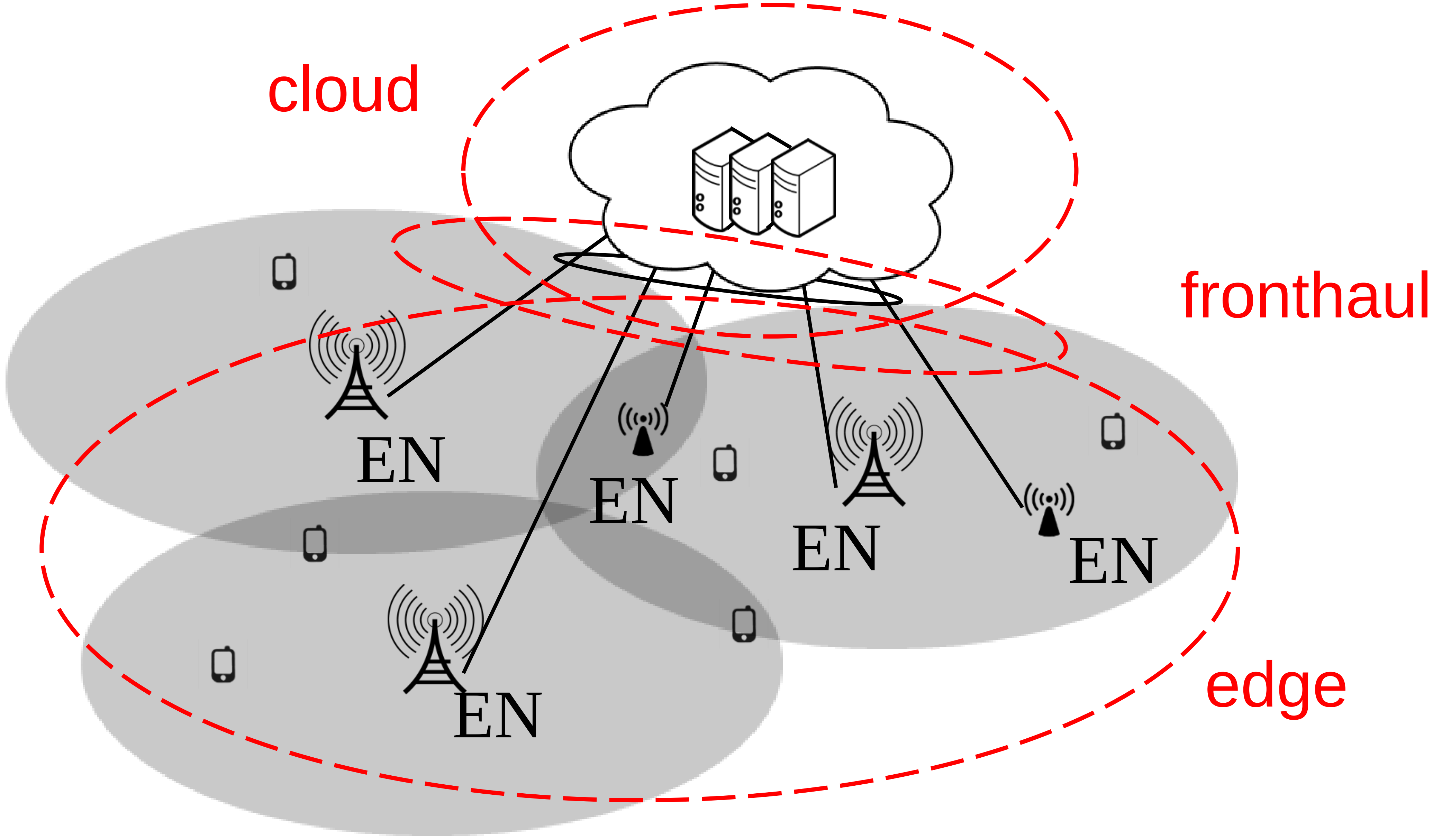
Fog-Radio Access Network (F-RAN)



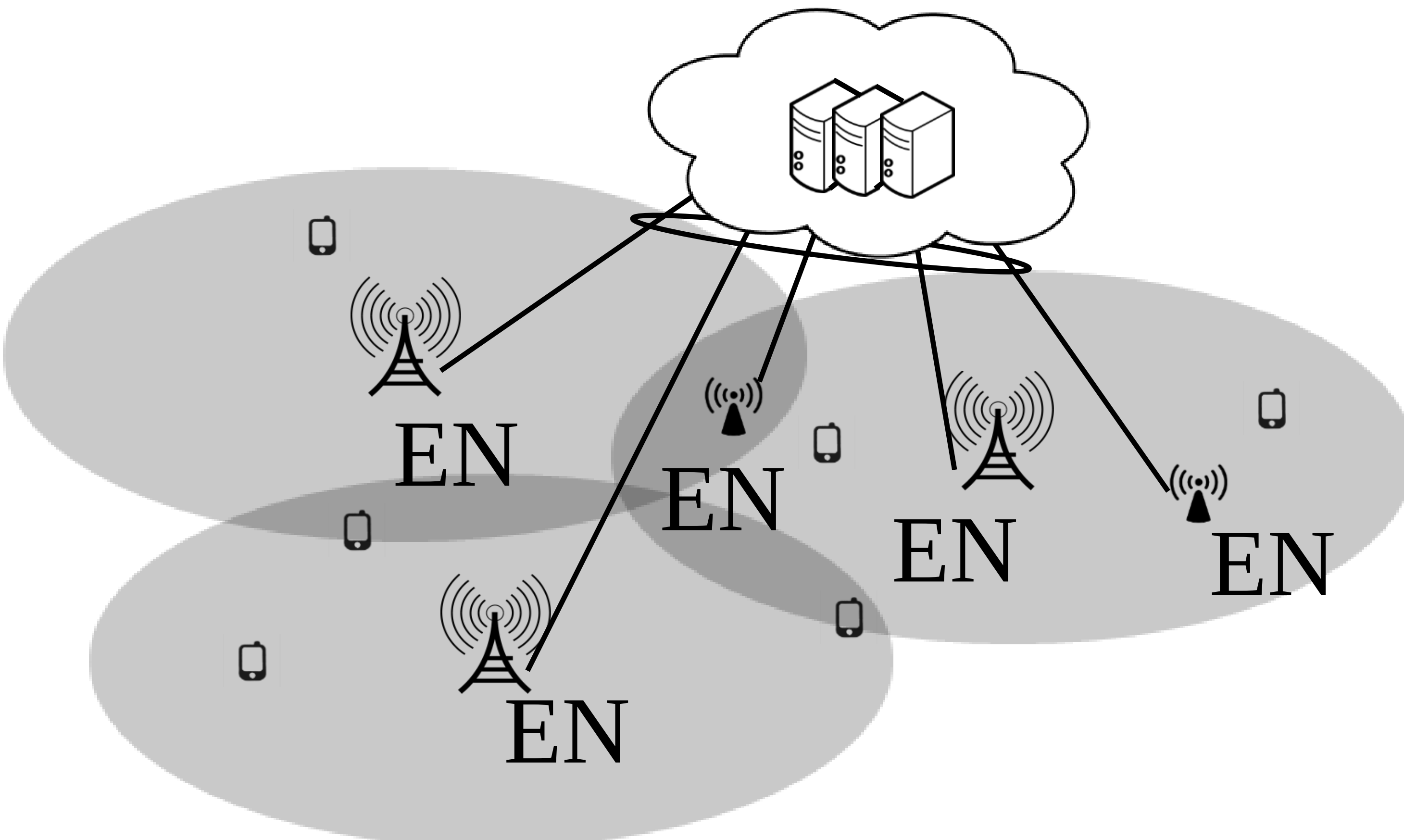
Fog-Radio Access Network (F-RAN)



Fog-Radio Access Network (F-RAN)

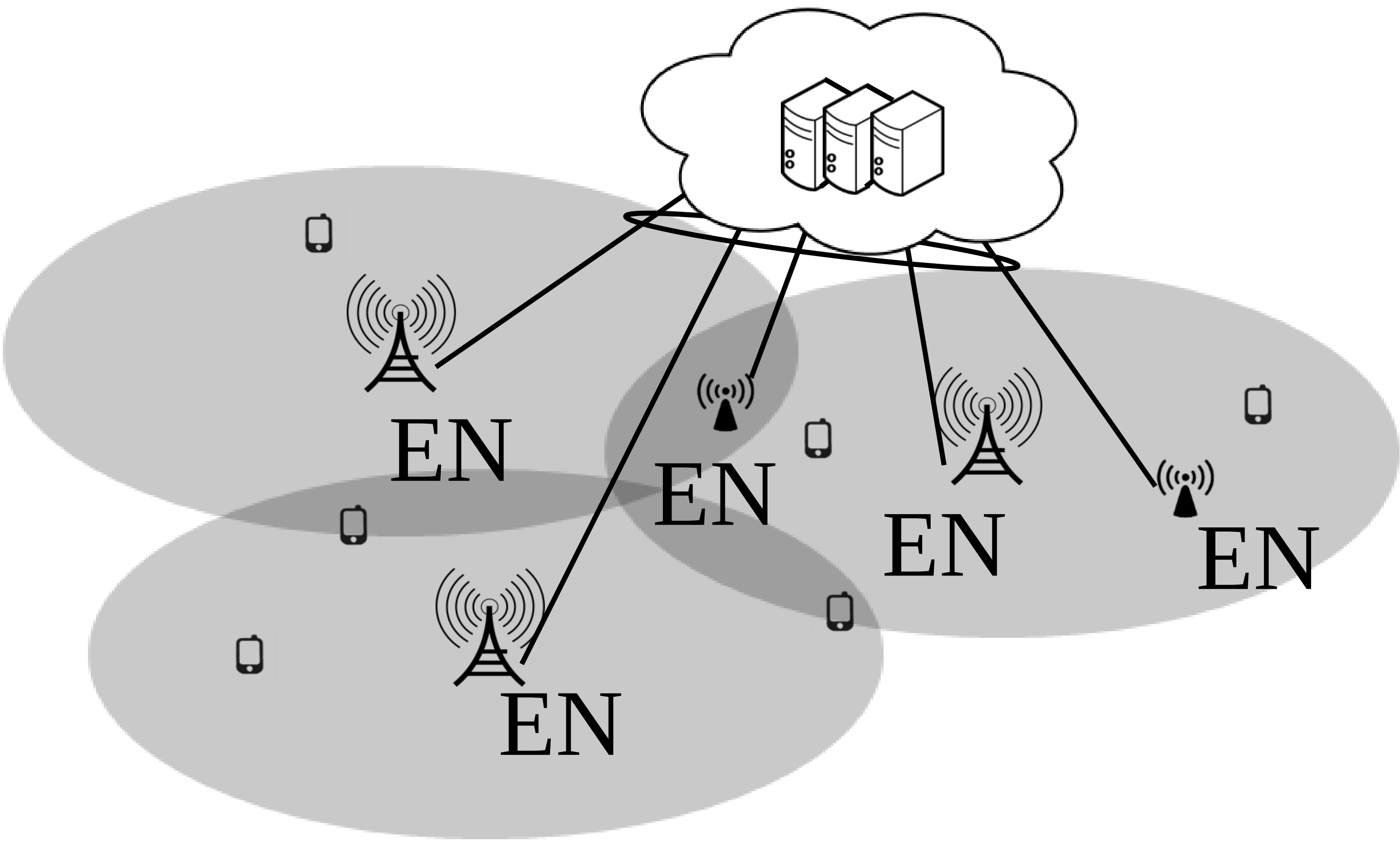


Fog Networking



- Fog networks: computing, storage, and communication functions along the cloud-to-user continuum [Chiang et al '17]

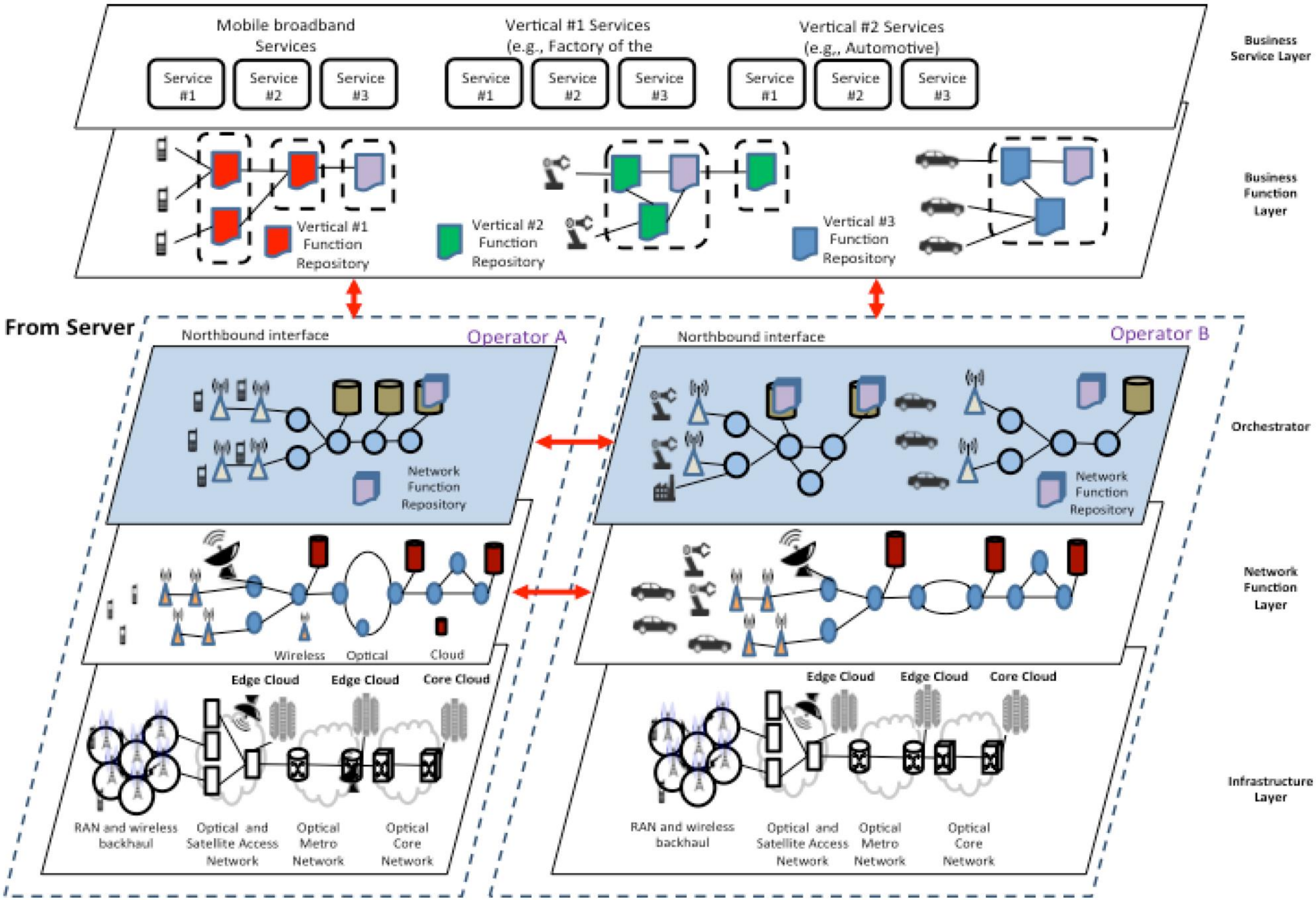
Fog Networking and Softwarization



- Fog networks: computing, storage, and communication functions along the cloud-to-user continuum [Chiang et al '17]

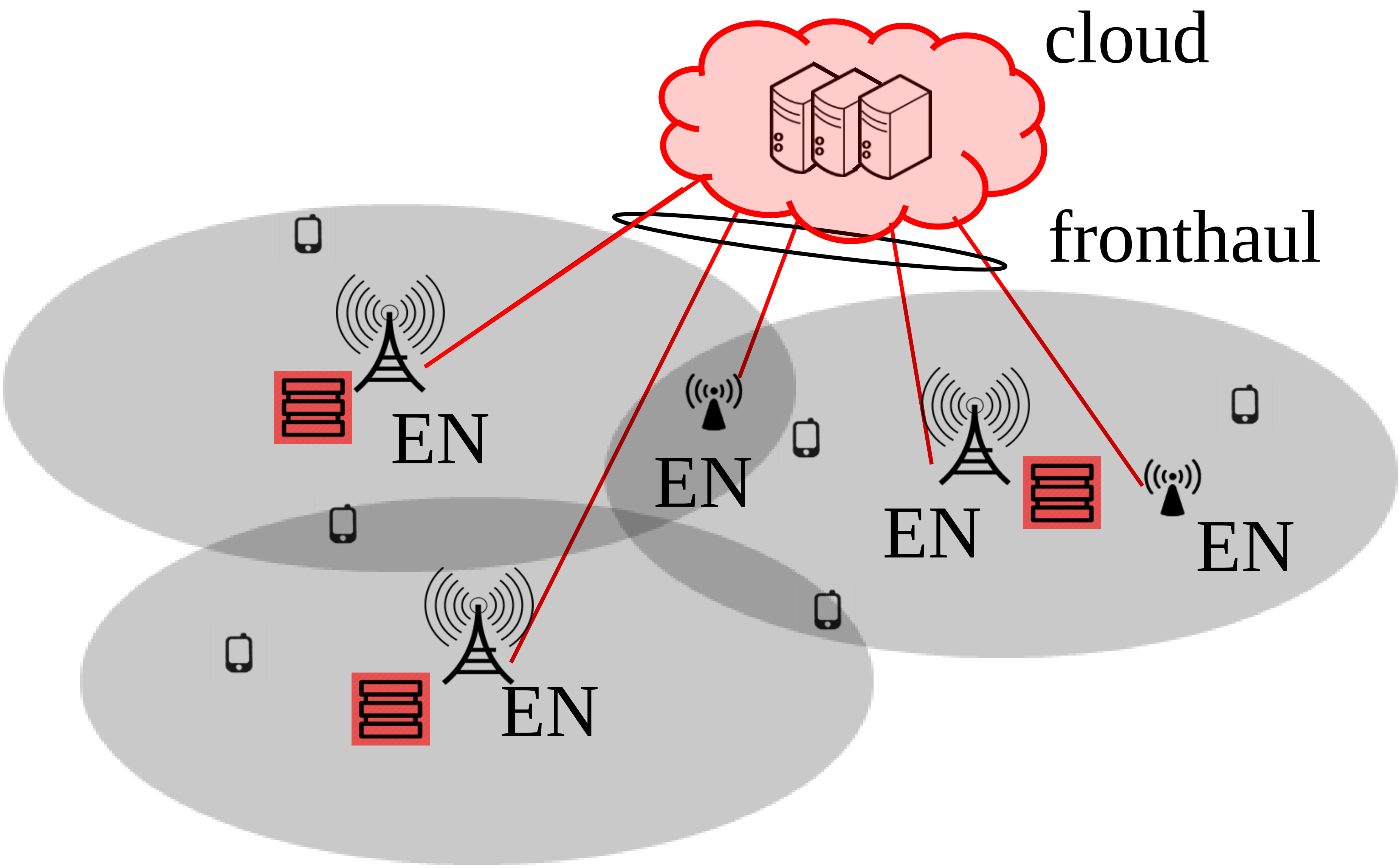
- Network softwarization allows the optimization of cloud vs. edge functional allocation.

F-RAN and 5G



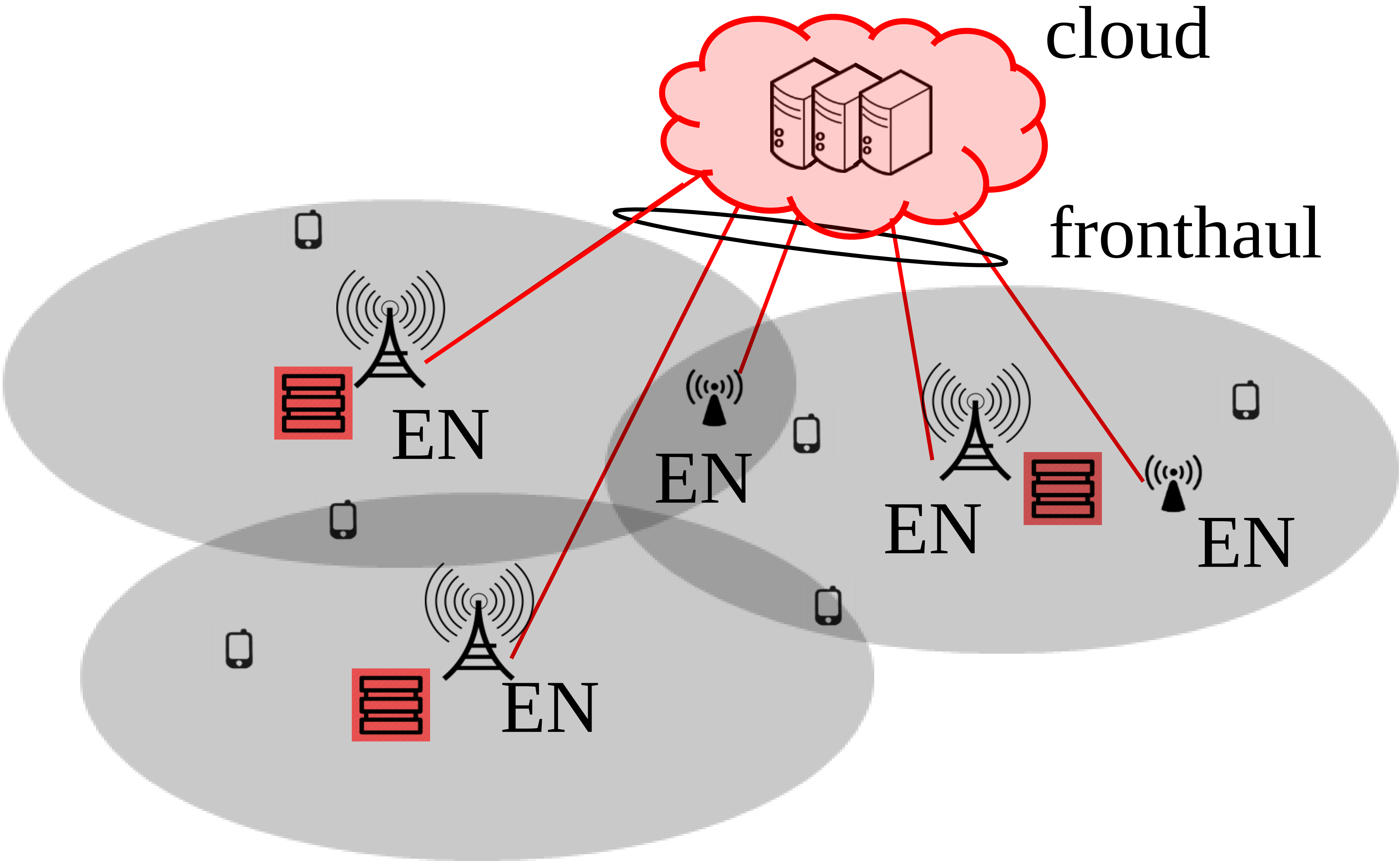
[5G PPP '16]

This Talk



- Content delivery: caching and delivery

This Talk



- Content delivery: caching and delivery
- Information-theoretic approach

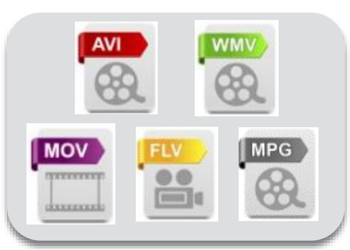
Overview

- Edge Caching
- F-RAN
- Enabling general transport and delivery
- Extensions

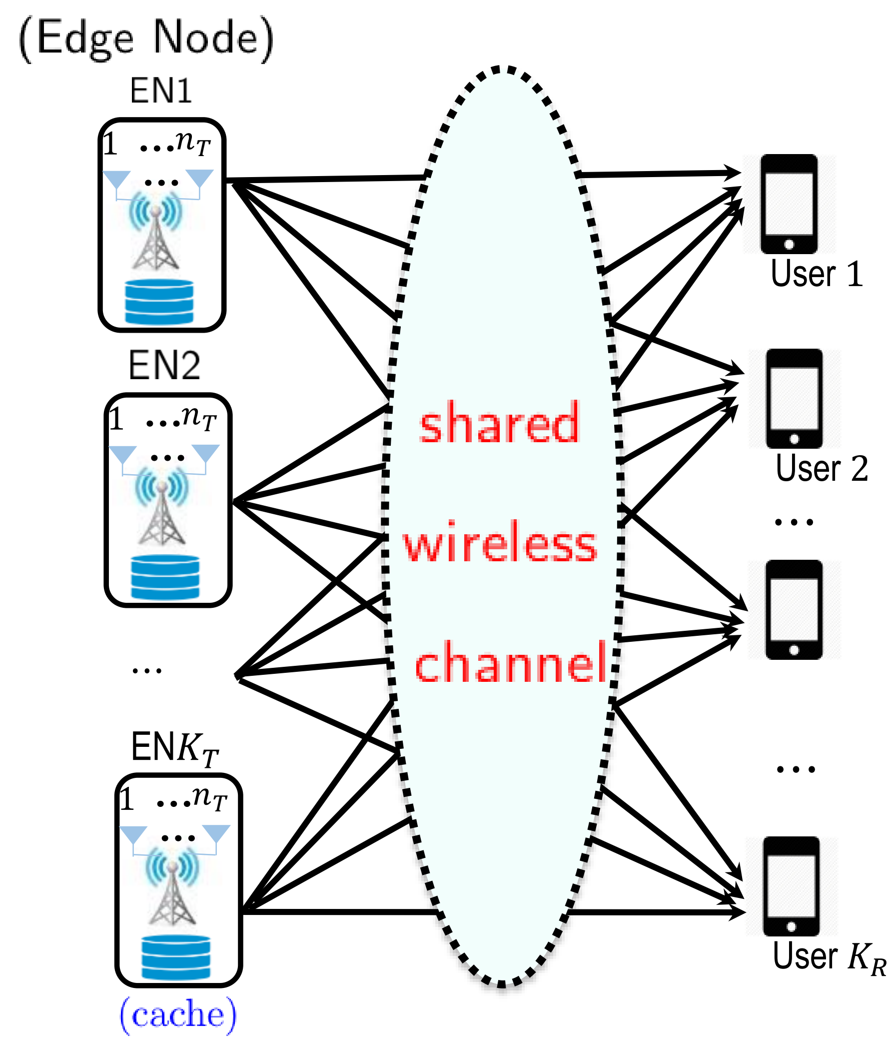
Edge Caching
(under constrained delivery)

J. Zhang and O. Simeone, “Fundamental Limits of Cloud and Cache-Aided Interference Management with Multi-Antenna Base Stations,” arXiv:1712.04266.

Edge Caching

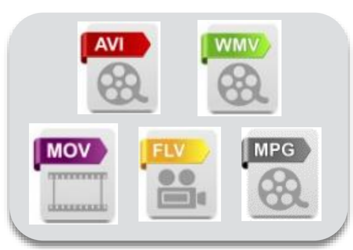


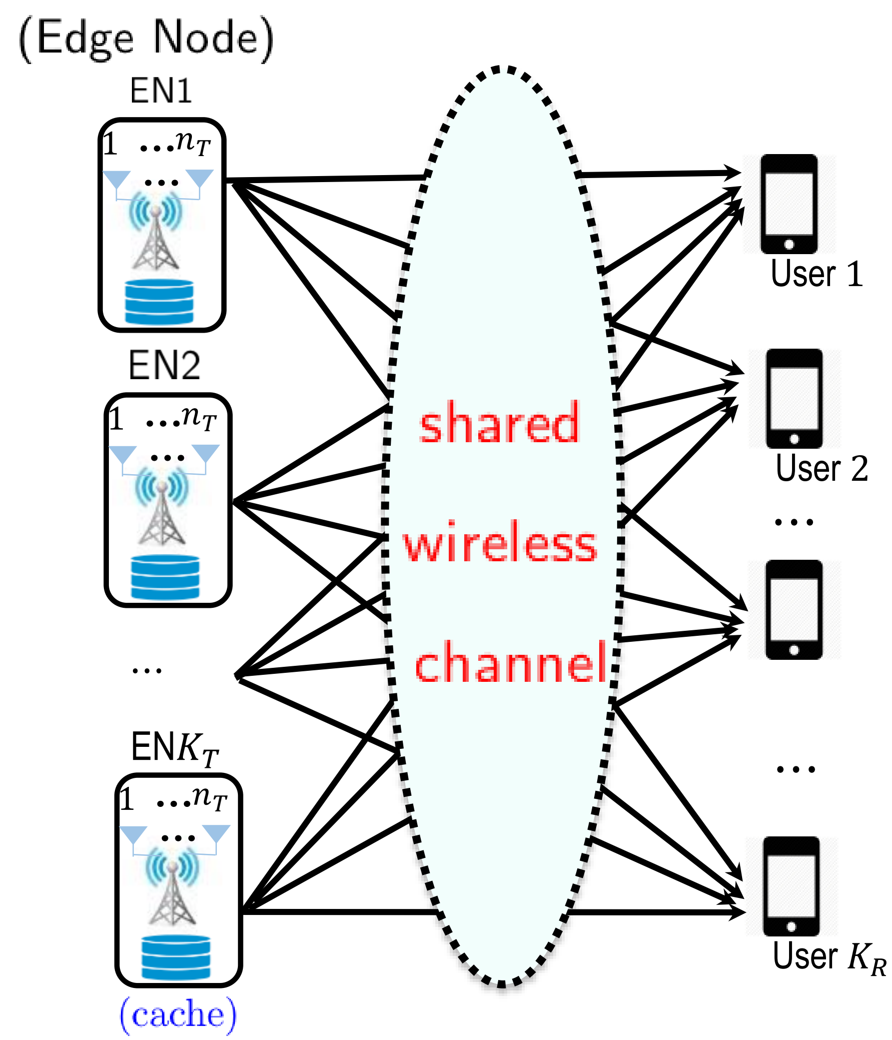
Set of popular files



- Caching at the edge nodes (ENs) can reduce delivery latency and network congestion [Golrezaei et al ‘12]

Edge Caching





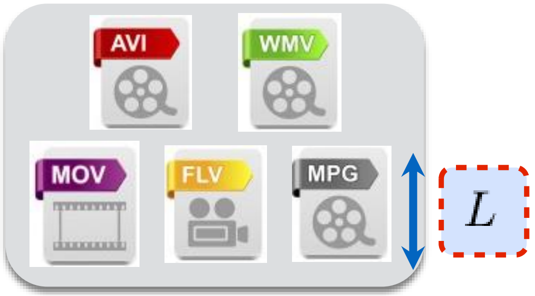
- Edge caching enables coordination and cooperation at the ENs [Maddah-Ali and Niesen ’15] [Hachem et al ‘16] [Xu et al ’16] [Roig et al ‘17] [Girgis et al ‘17]

System Model

- To start, we consider the following constraints:
  - ✓ Uncoded (fractional) caching
  - ✓ One-shot linear precoding

- Extension of [Naderializadeh et al ‘17] to multi-antenna ENs and a more efficient packetization method

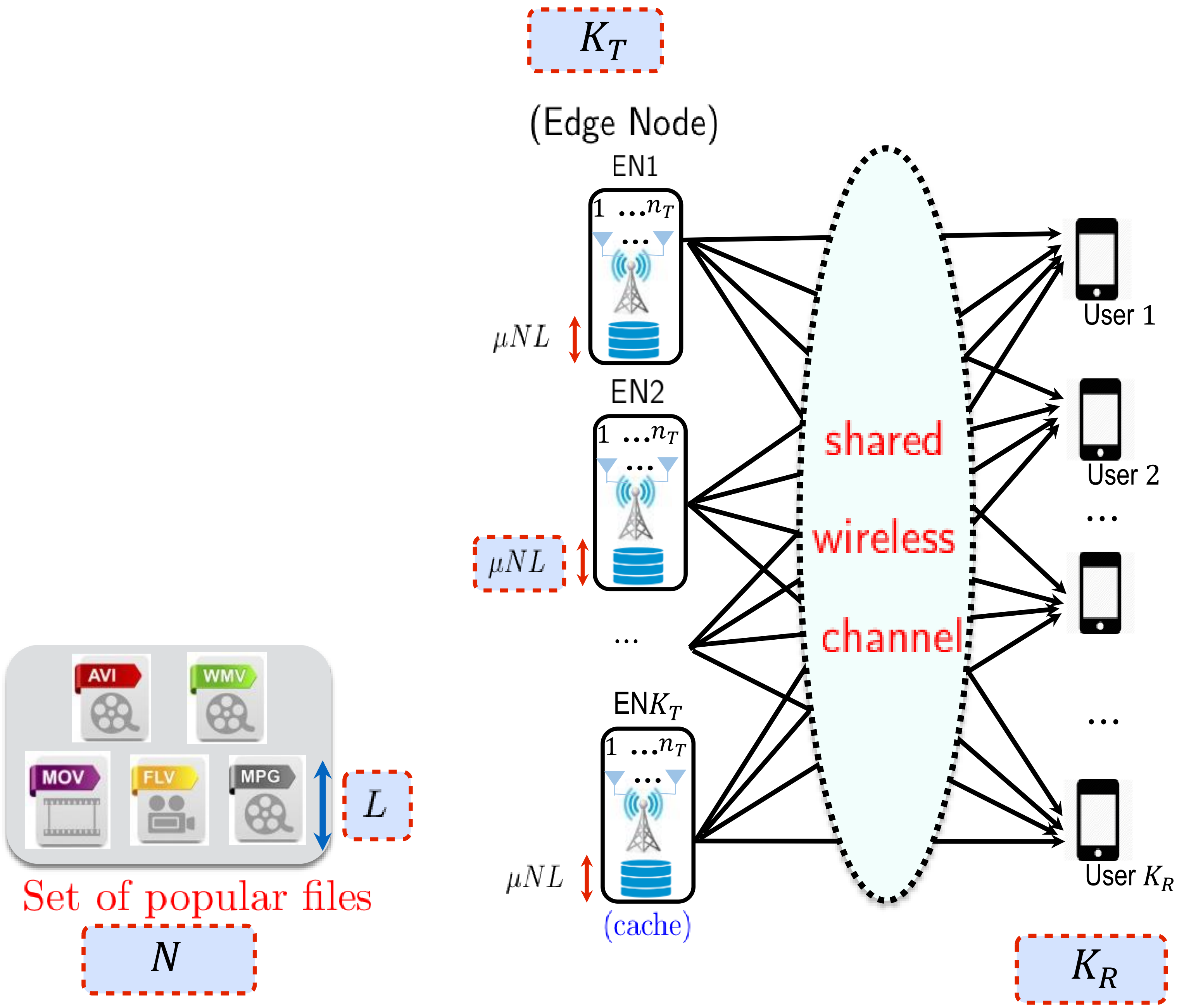
System Model



Set of popular files

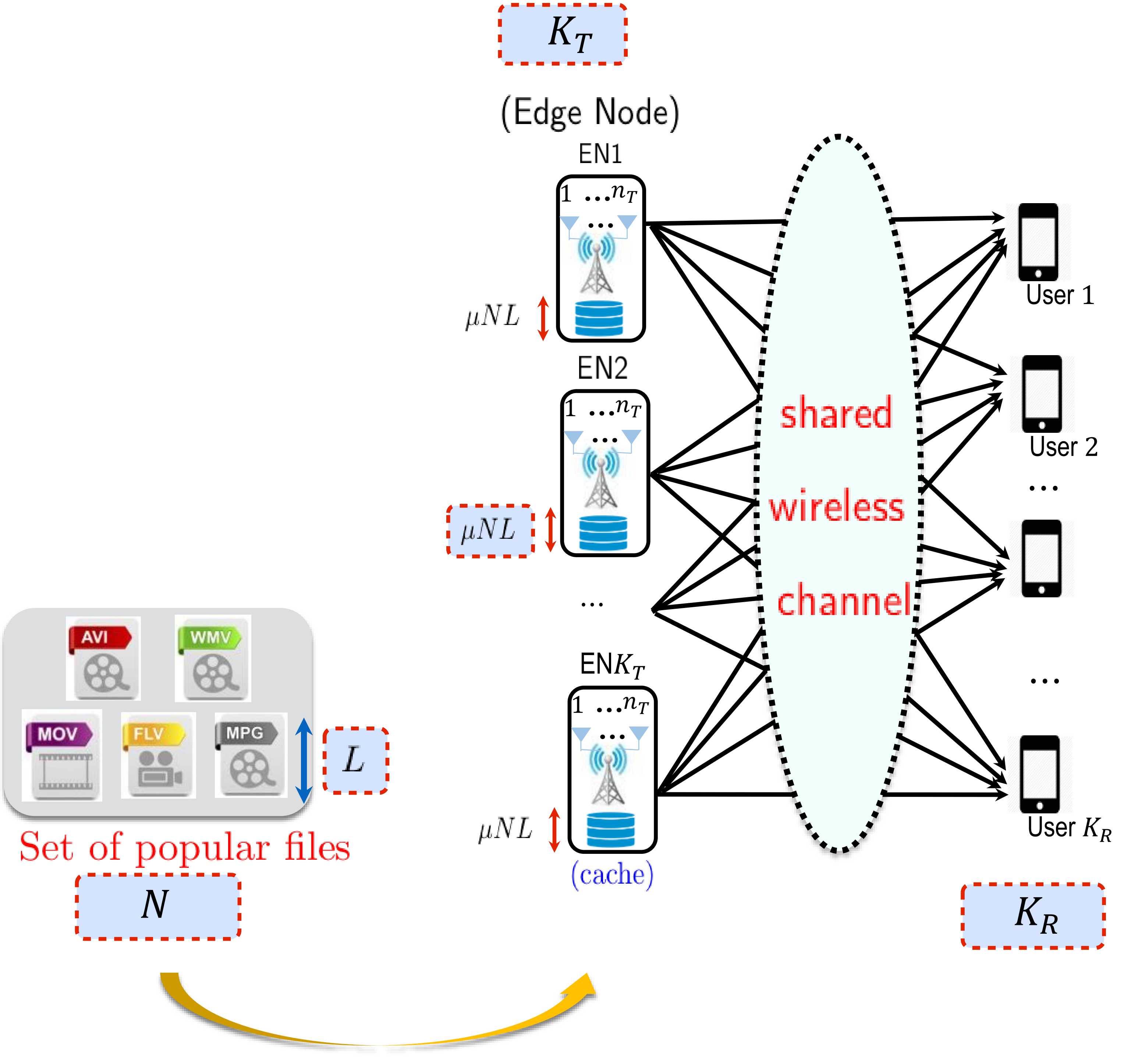
N

System Model



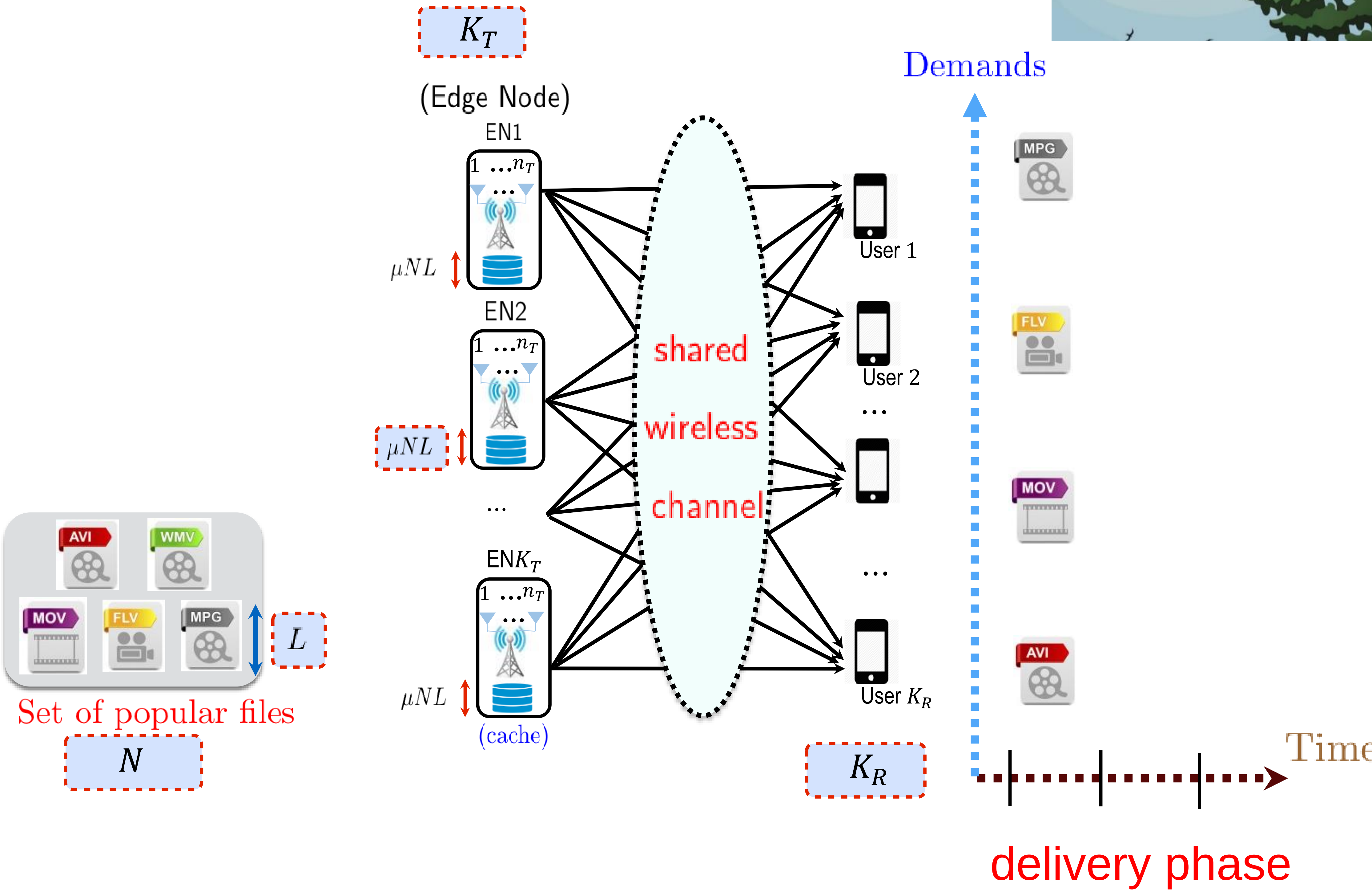
- n_T = number of per-EN transmit antennas
- μ = fractional cache size
- $P(\rightarrow \infty)$ = per-EN transmit power
- Full CSI

System Model

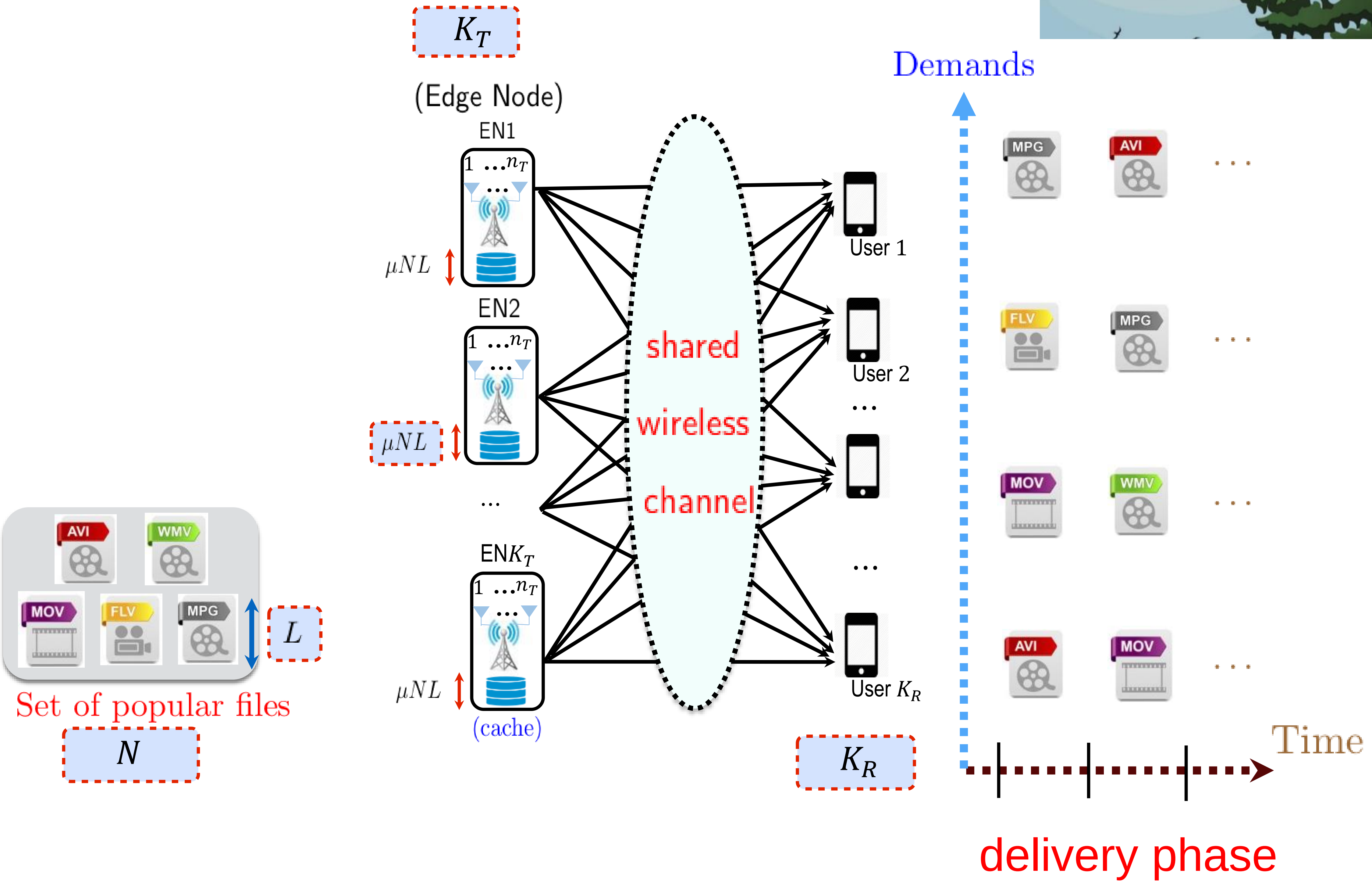


placement phase: uncoded fractional caching

System Model

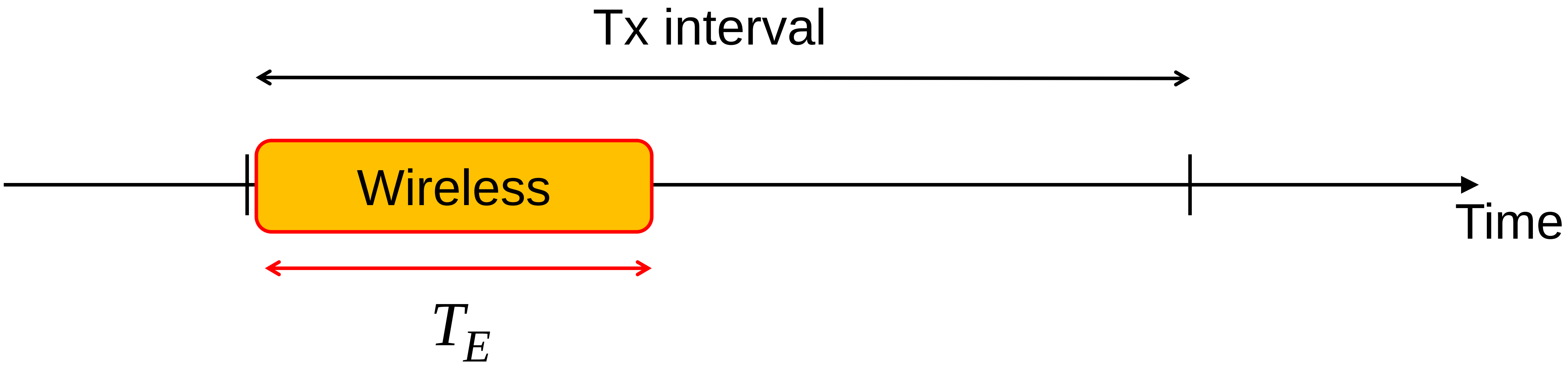


System Model

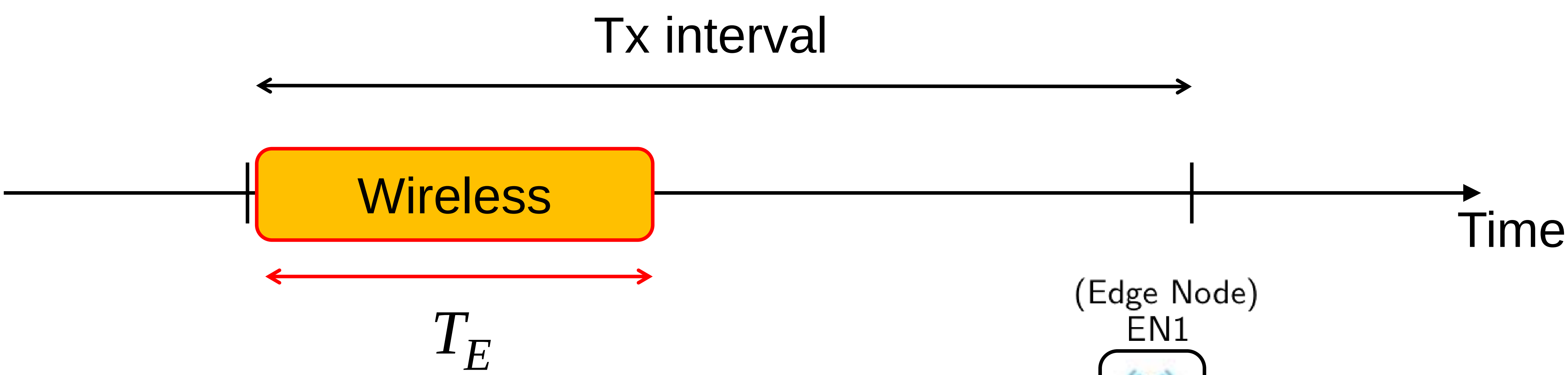


linear one-shot precoding

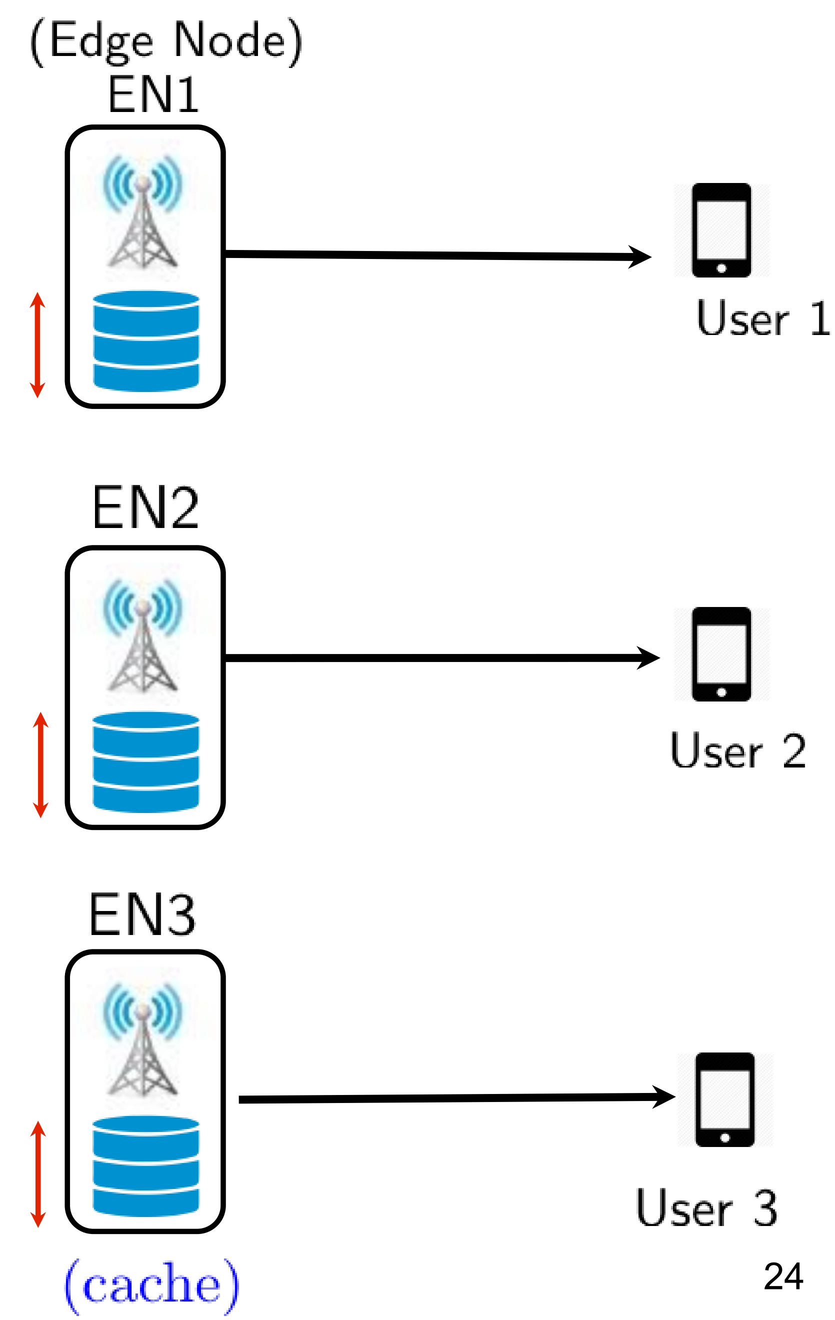
Normalized Delivery Time (NDT)



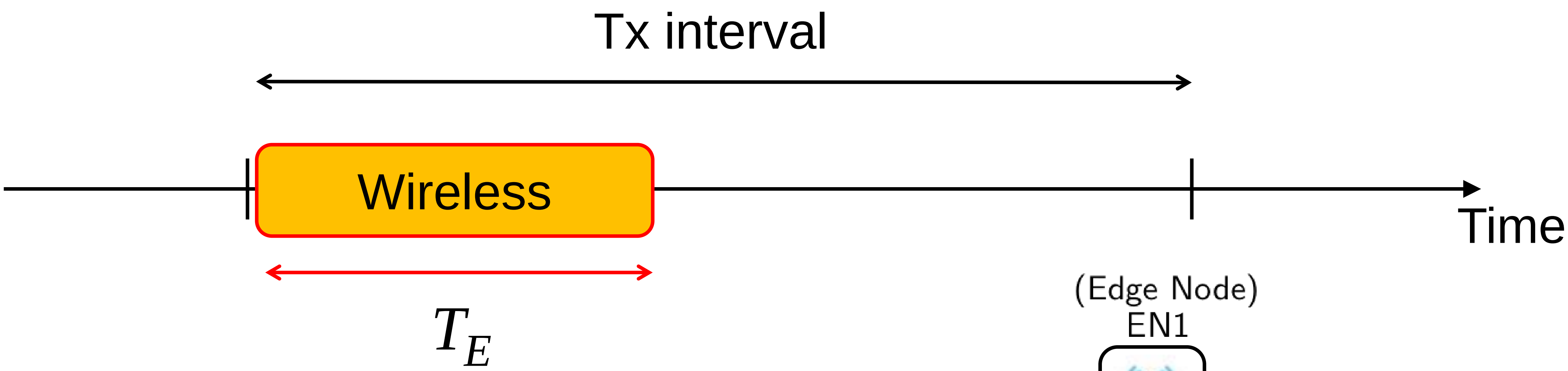
Normalized Delivery Time (NDT)

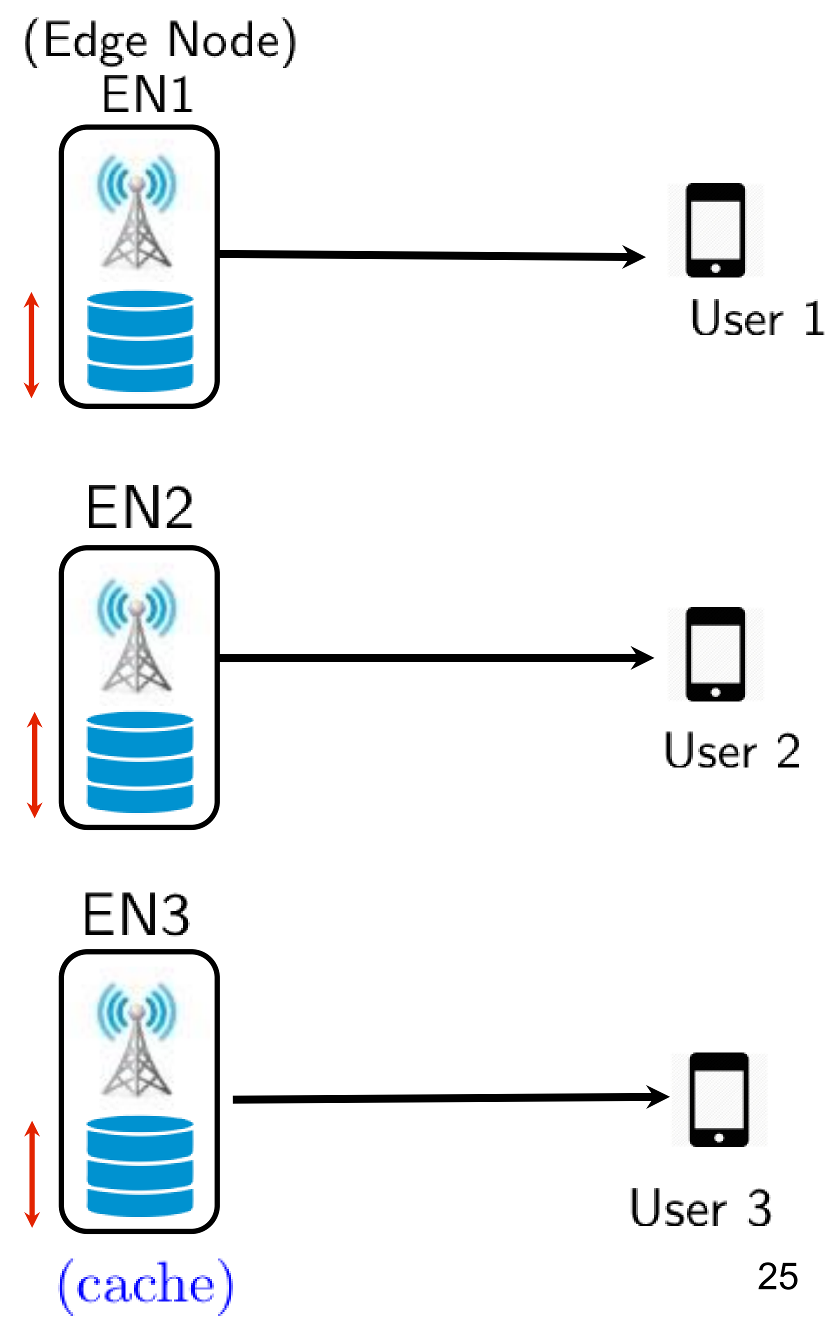


- Delivery time normalized by that of an ideal reference system (right) in which each user receives at maximum log(*P*) rate at all times



Normalized Delivery Time (NDT)





- Edge NDT:

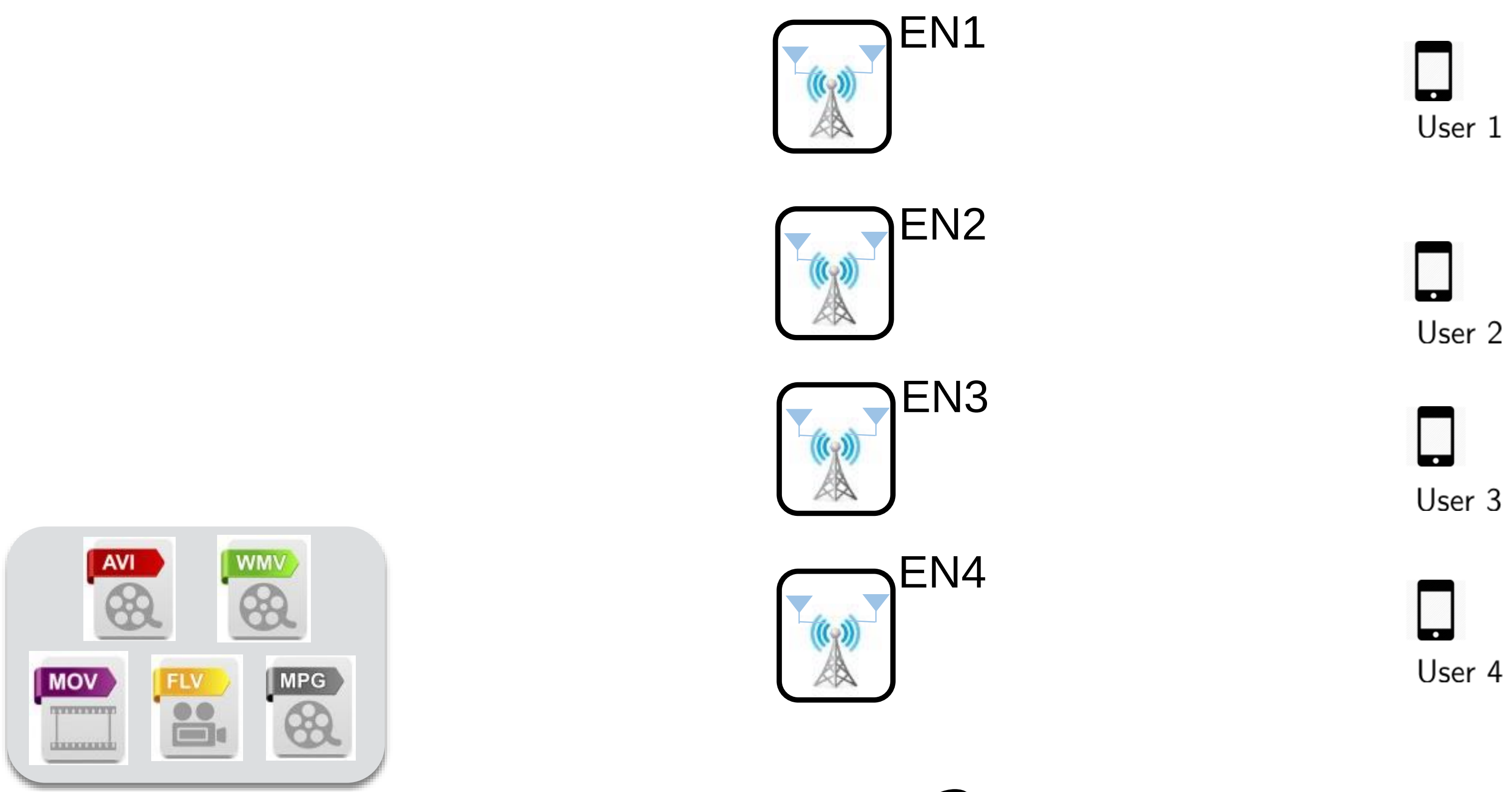
$$\delta_E = \lim_{P\to\infty}\lim_{L\to\infty}\frac{T_E}{L/\log(P)}$$

Multiplicity

- The multiplicity m of a content is the number of times that a content appears across the caches of all the ENs
- $m(\mu) = \mu K_T$ = content multiplicity afforded by edge caching
- The multiplicity determines the number of users $u(m)$ that can be served simultaneously by means of cooperation

Example

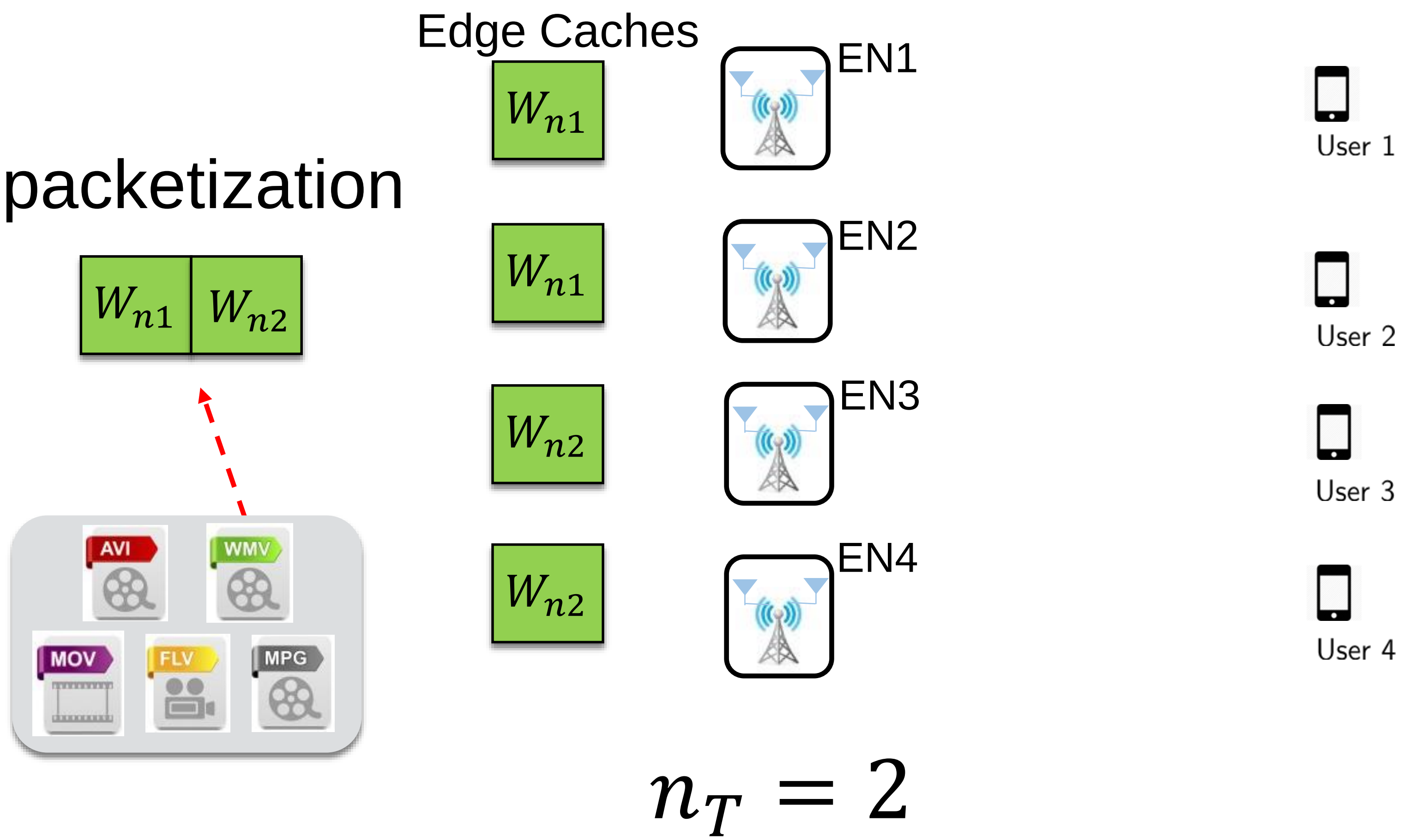
- Ex.: $\mu = 0.5 \Rightarrow m(\mu) = \mu K_T = 2$



$n_T = 2$

Example

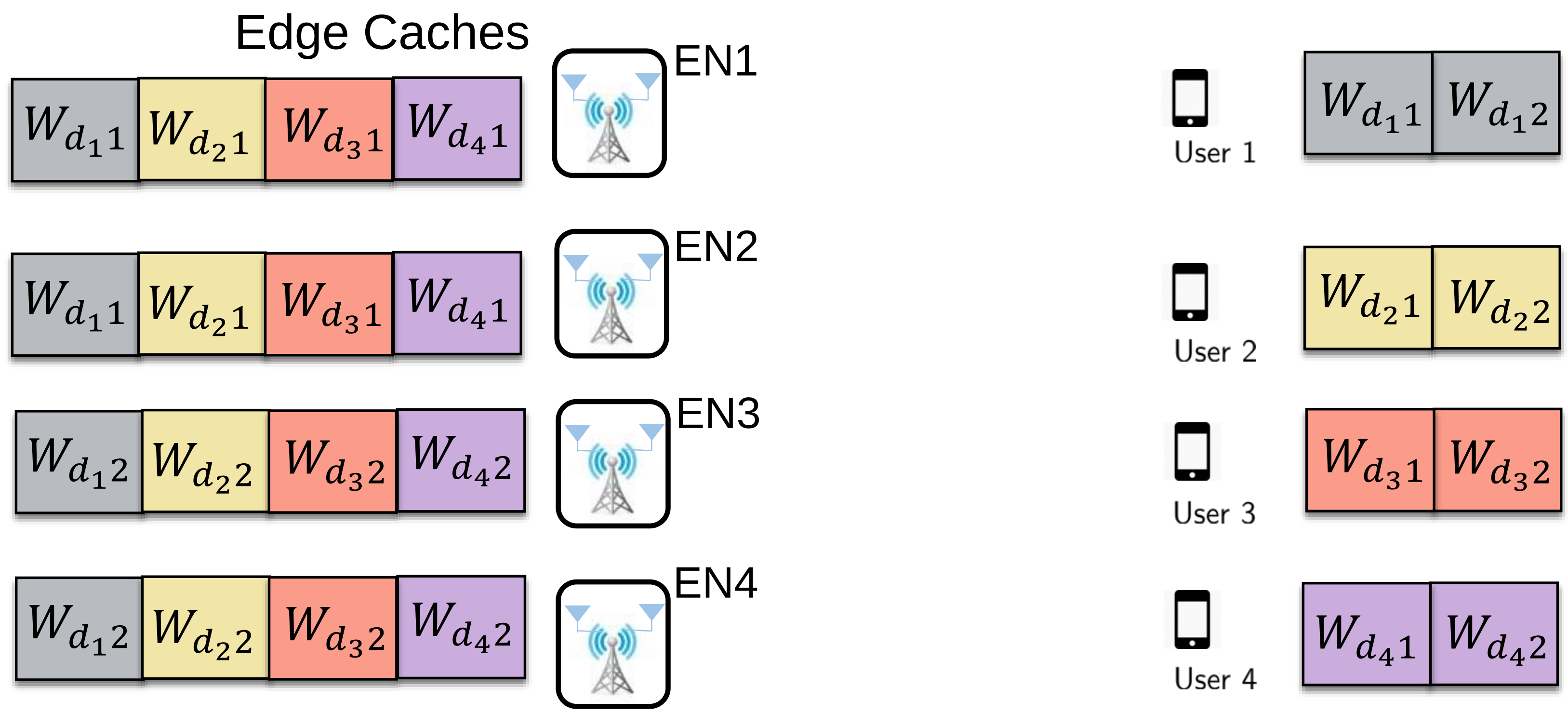
- Ex.: $\mu = 0.5 \Rightarrow m(\mu) = \mu K_T = 2$



placement phase

Example

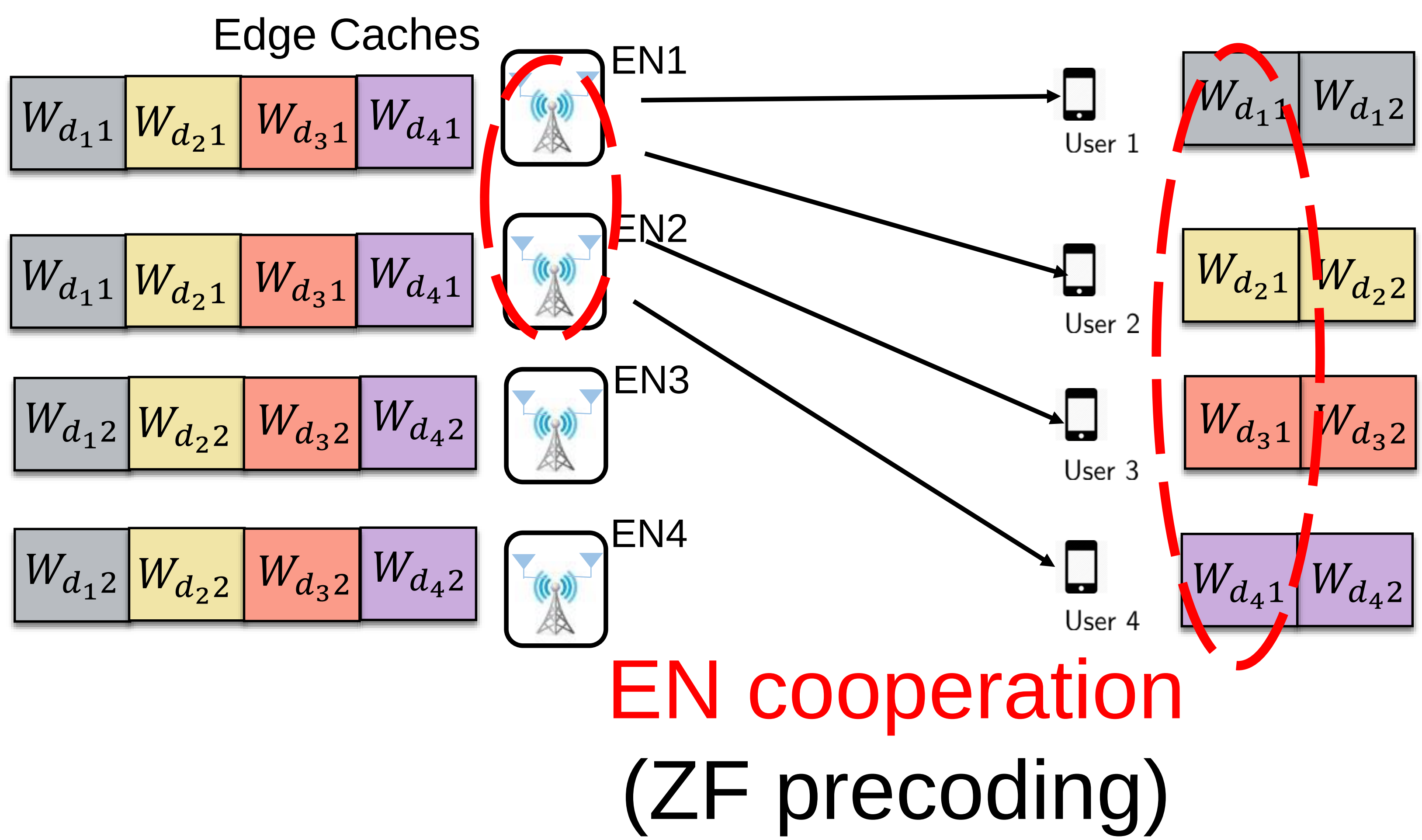
- Ex.: $\mu = 0.5 \Rightarrow m(\mu) = \mu K_T = 2$



delivery phase

Example

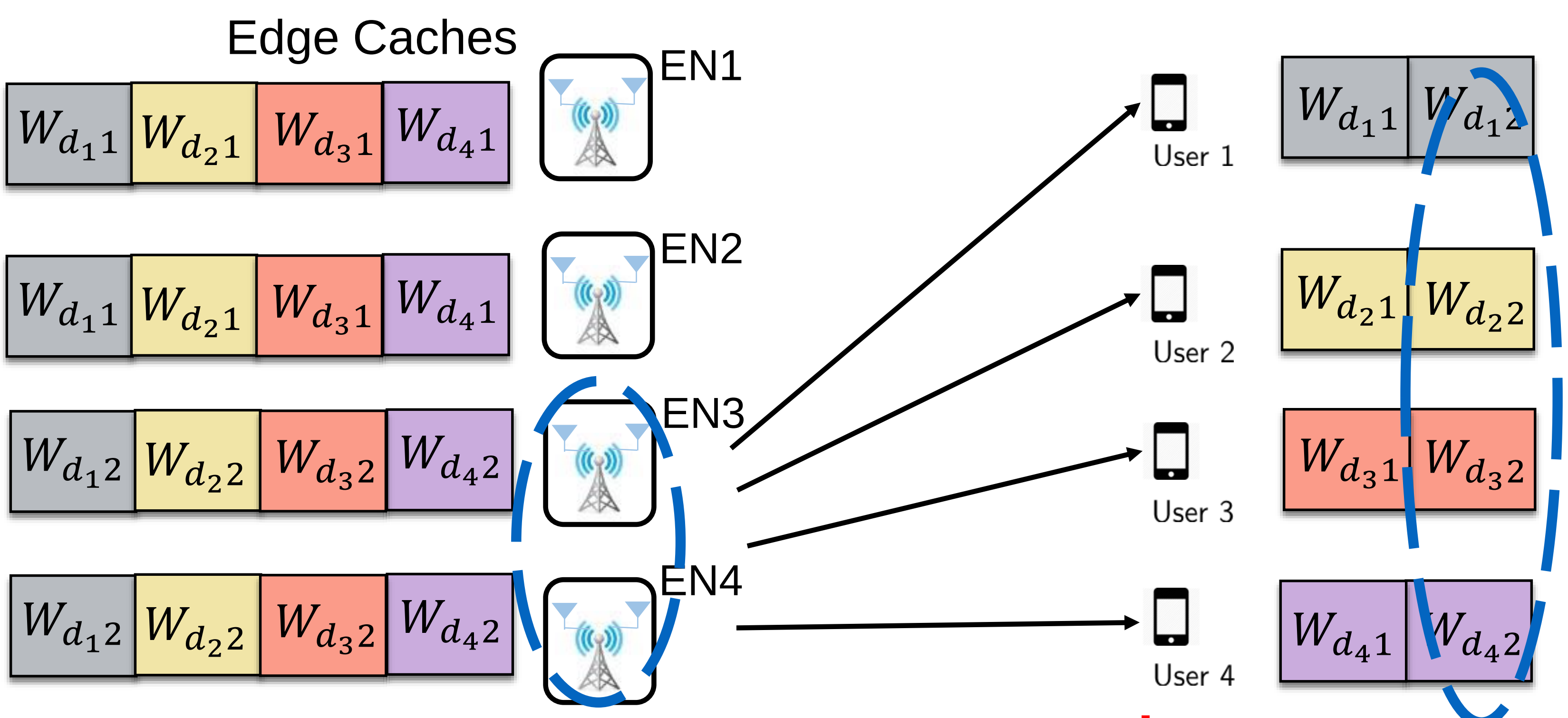
- Ex.: $\mu = 0.5 \Rightarrow m(\mu) = \mu K_T = 2$ and $u(m) = 4$



delivery phase

Example

- Ex.: $\mu = 0.5 \ \Rightarrow m(\mu) = \mu K_T = 2$ and $u(m) = 4$

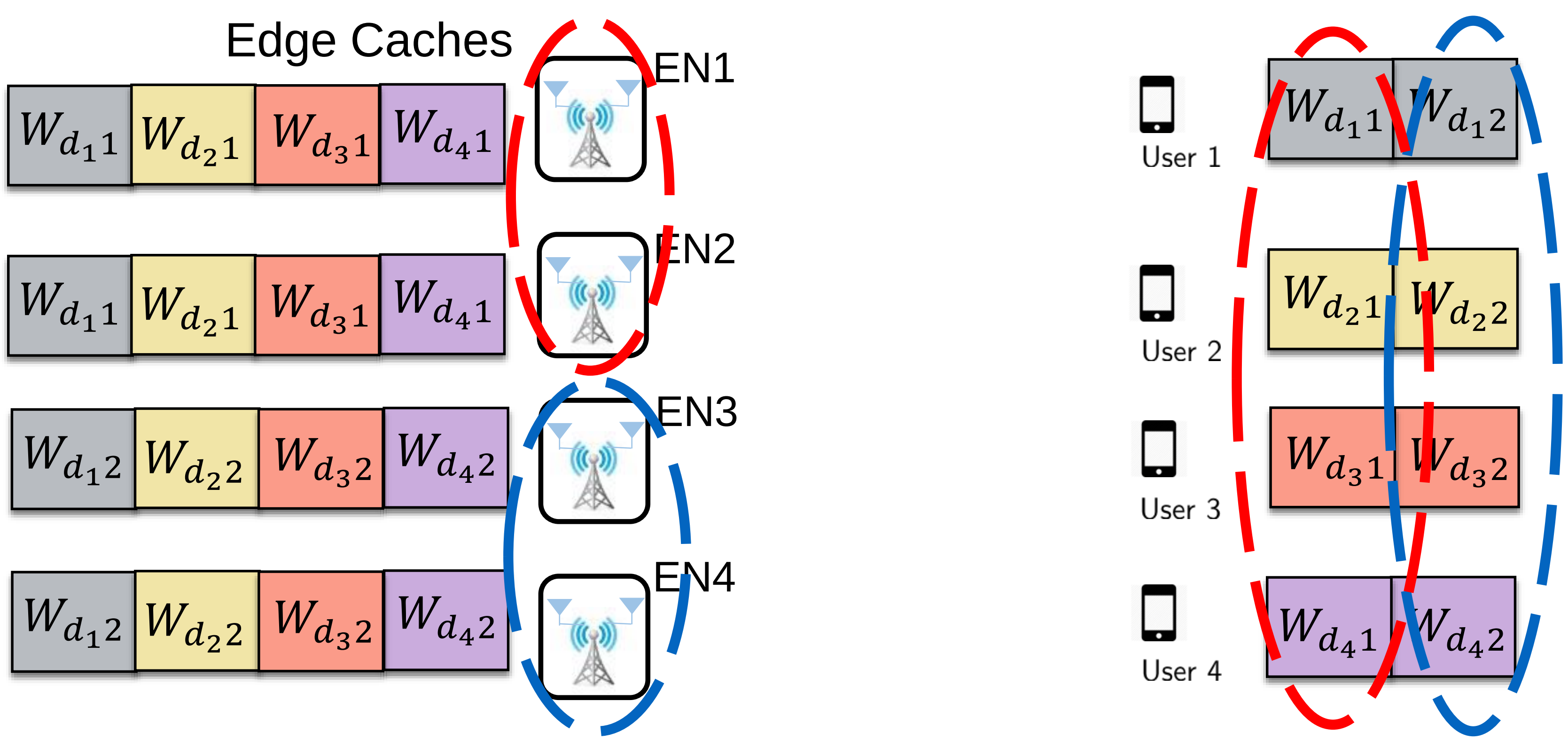


EN cooperation
(ZF precoding)

delivery phase

Example

- Ex.: $\mu = 0.5 \Rightarrow m(\mu) = \mu K_T = 2$ and $u(m) = 4$



- Normalized Delivery Time (NDT)

$$\delta_E = 1$$

Edge NDT via Edge Caching

- Generalizing this example, with multiplicity

$$m(\mu) = \lfloor \mu K_T \rfloor,$$

clustered cooperative EN transmission enables the simultaneous transmission to a number of users equal to

$$u(m) = \max\{m n_T, K_R\}$$

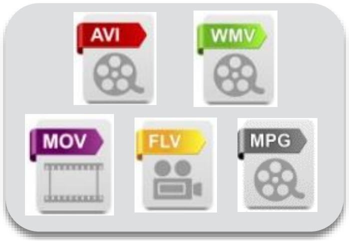
- The resulting edge NDT is

$$\delta_E(m) = \frac{K_R}{u(m)}$$

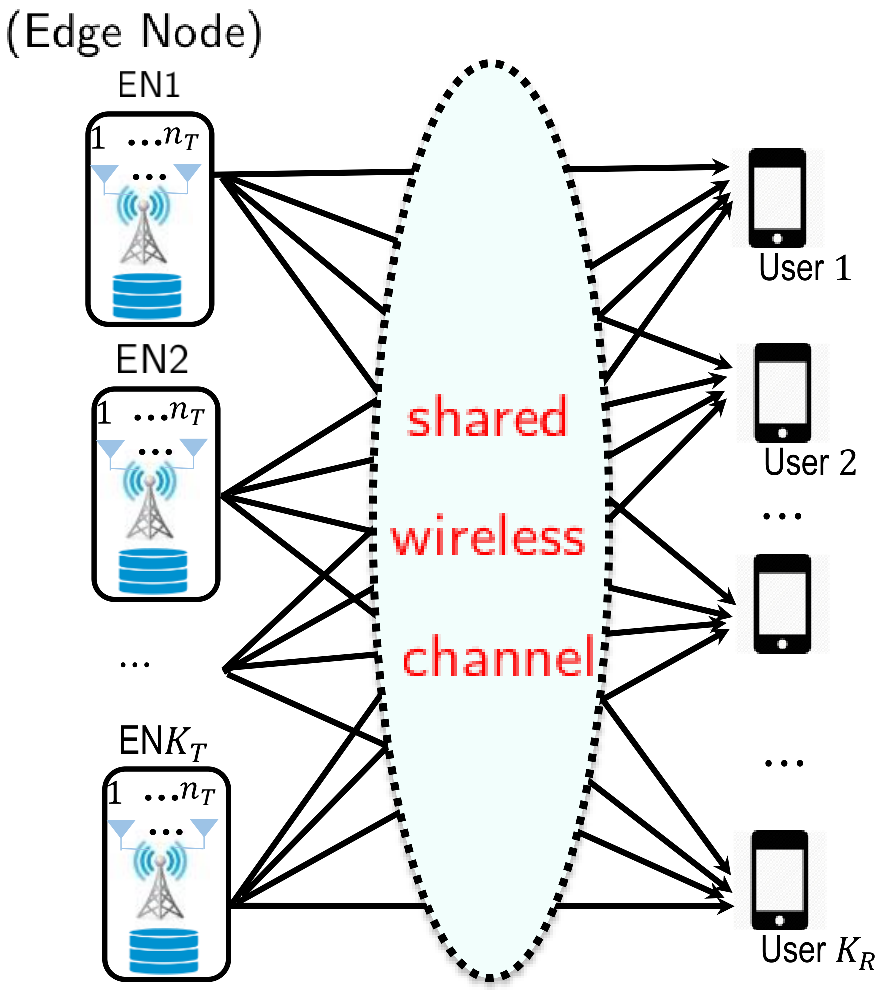
F-RAN
(under constrained delivery)

J. Zhang and O. Simeone, “Fundamental Limits of Cloud and Cache-Aided Interference Management with Multi-Antenna Base Stations,” arXiv:1712.04266.

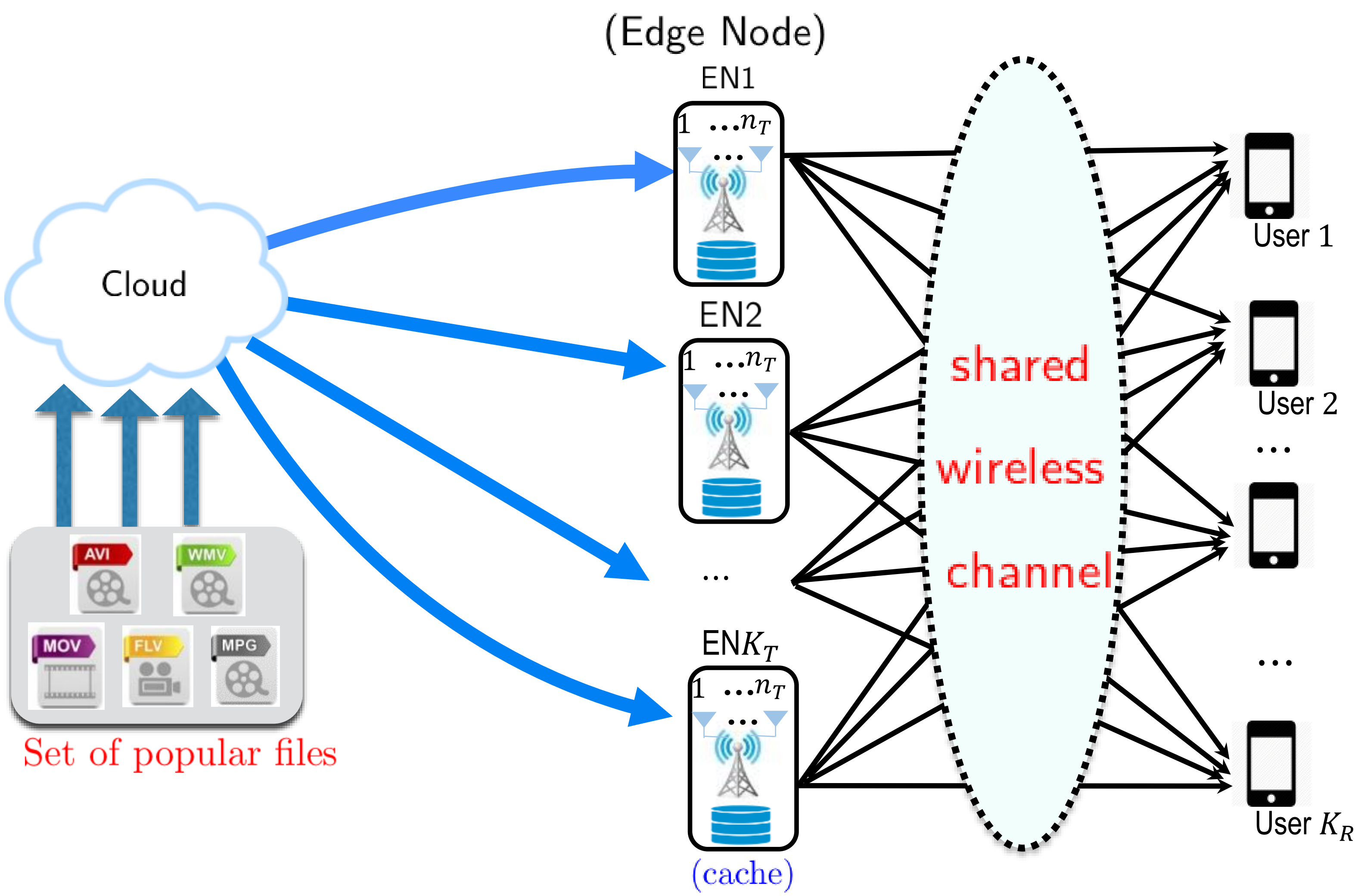
Edge Caching



Set of popular files



F-RAN

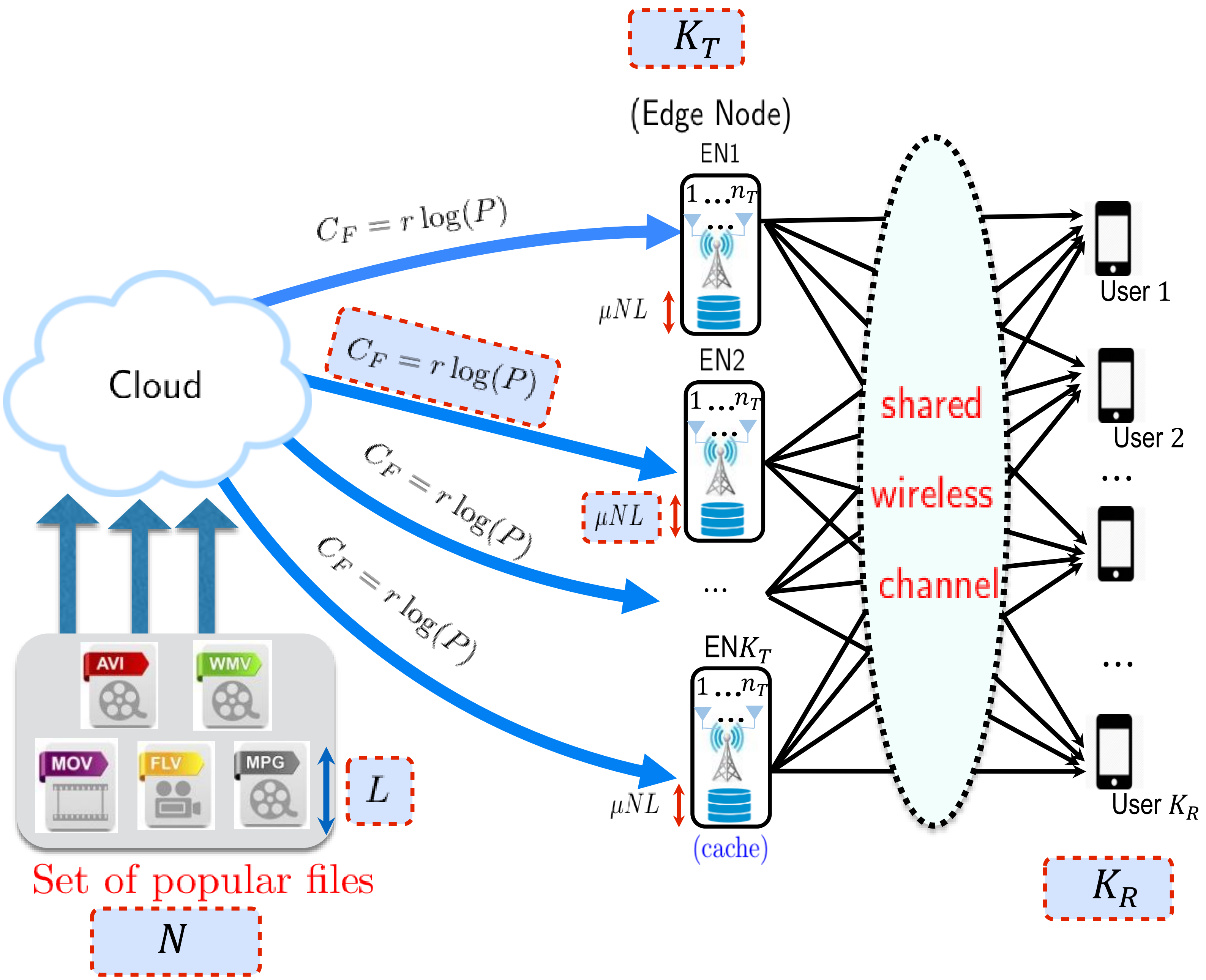


- Fronthaul links can be used to deliver uncached files and/or to enhance interference management capabilities [Sengupta et al ’17] [Azimi et al ‘17] [Kakar et al ’17] [Goseling et al ’17] [Roig et al ‘18]

System Model

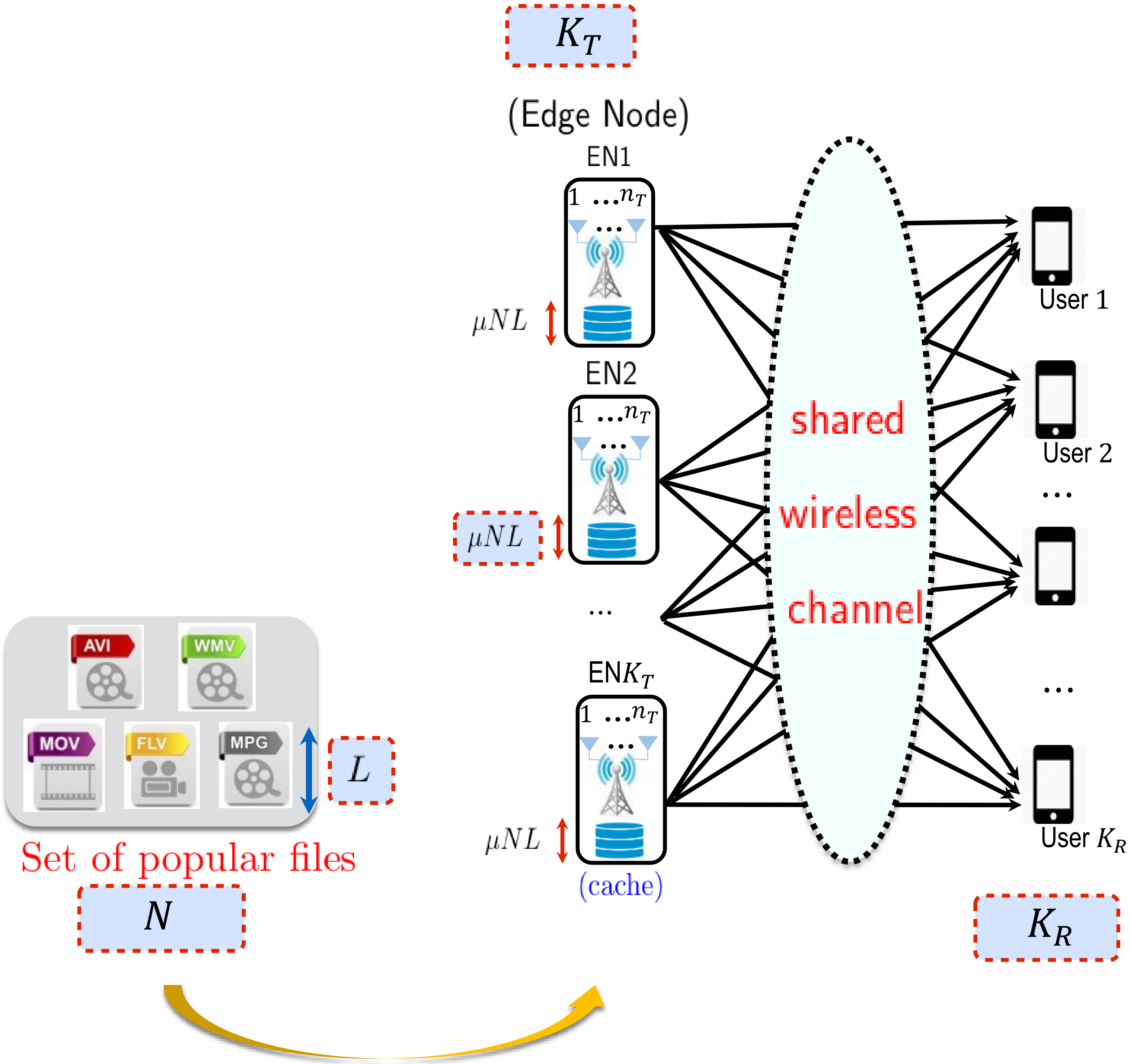
- We consider the simplifying assumptions:
  - ✓ Uncoded (fractional) caching
  - ✓ Transport of uncoded (fractional) contents
  - ✓ One-shot linear precoding

System Model



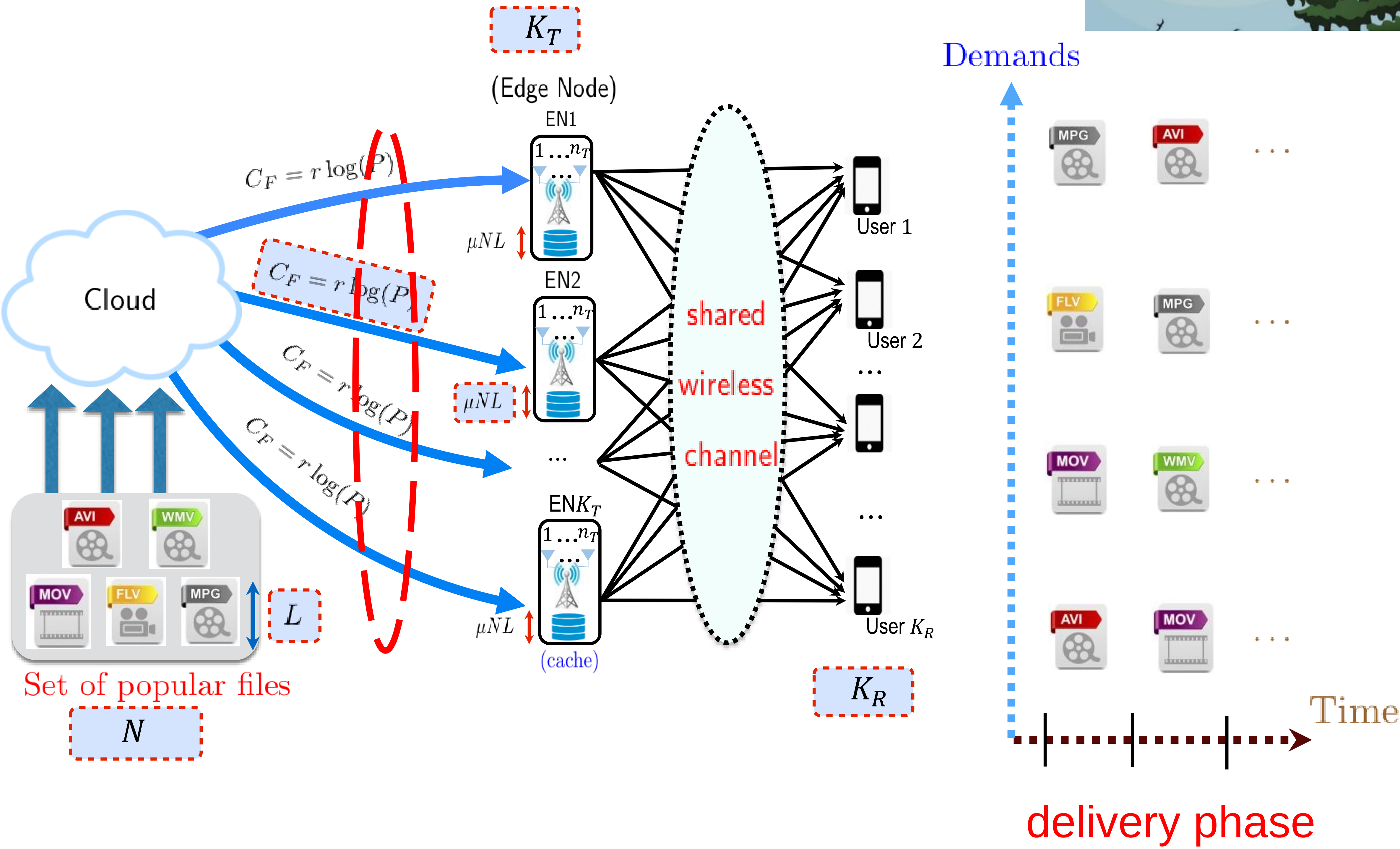
r = fronthaul rate

System Model



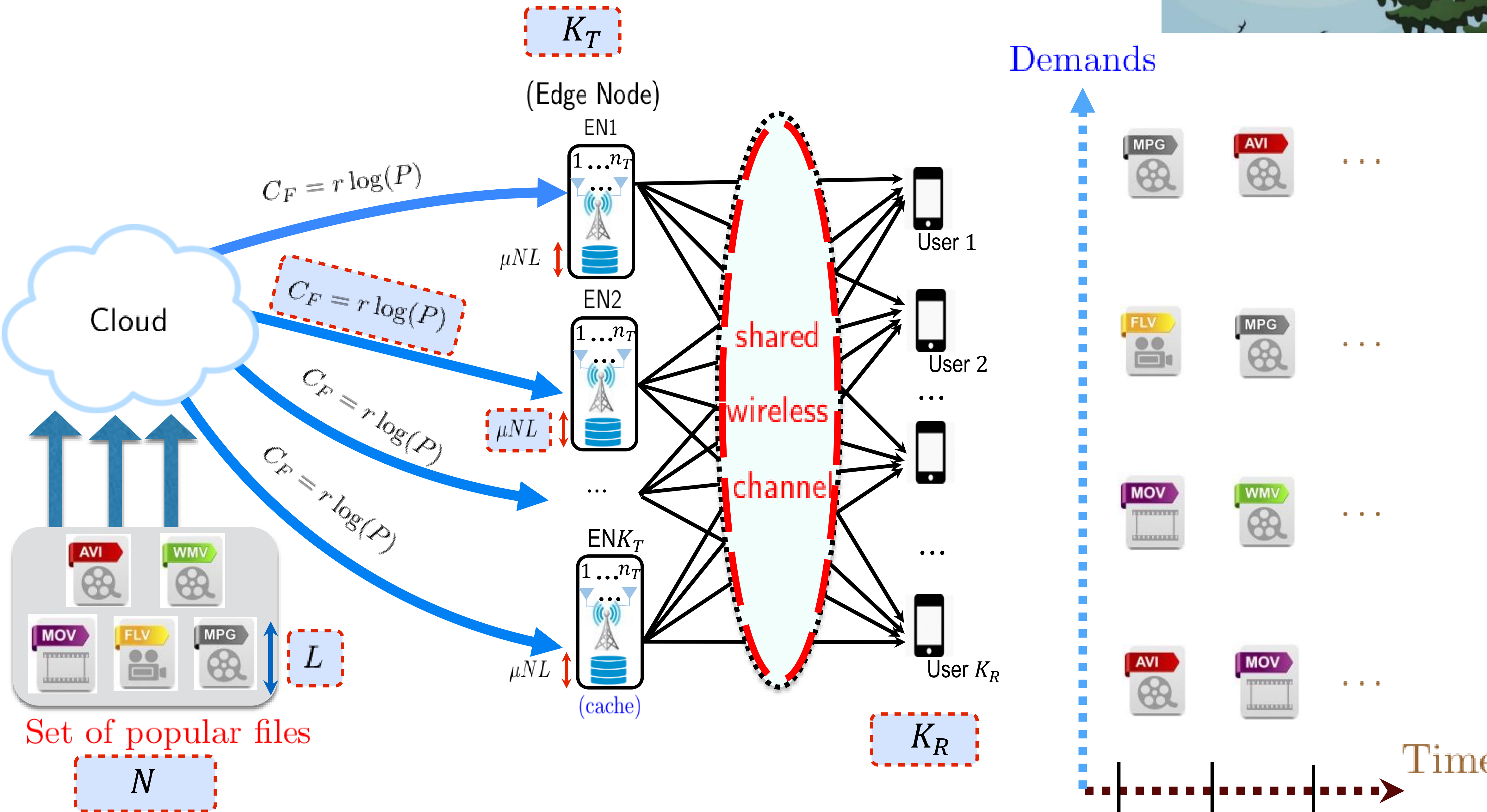
placement phase: uncoded fractional caching

System Model



uncoded fronthaul transmission

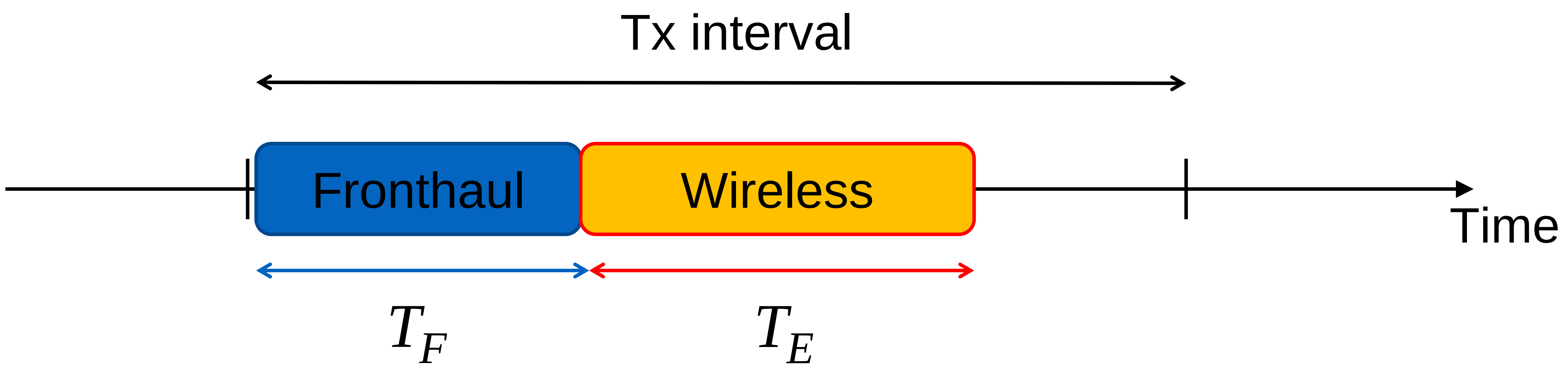
System Model



delivery phase

linear one-shot precoding

Normalized Delivery Time (NDT)



- Fronthaul NDT:

$$\delta_F = \lim_{P\to\infty}\lim_{L\to\infty}\frac{T_F}{L/\log(P)}$$

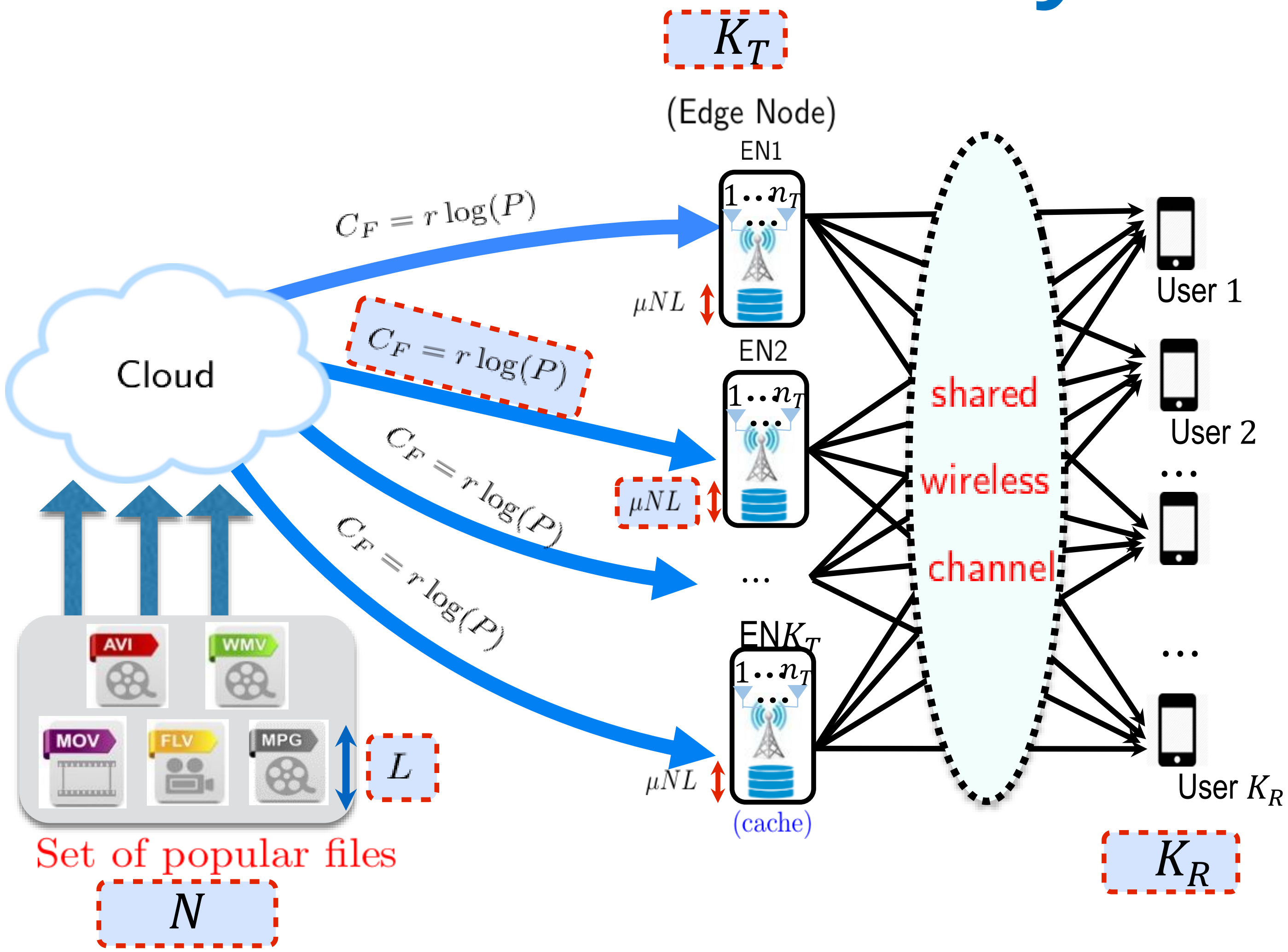
- Edge NDT:

$$\delta_E = \lim_{P\to\infty}\lim_{L\to\infty}\frac{T_E}{L/\log(P)}$$

- NDT:

$$\delta = \delta_F + \delta_E$$

Placement and Delivery Strategy

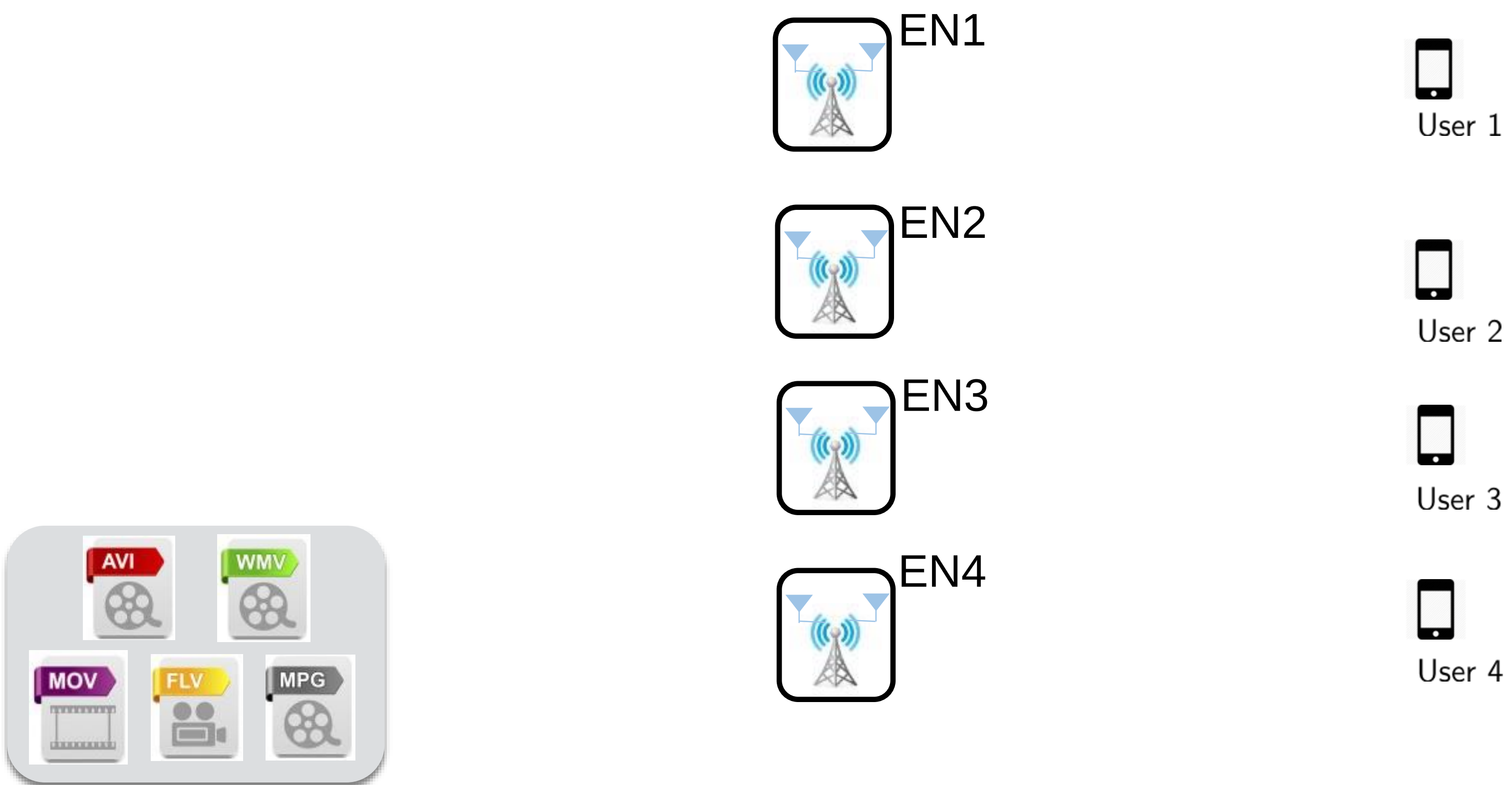


- Ensure a given multiplicity *m* for the requested files via both caching and fronthaul transmission.

- Use clustered EN cooperation to serve $u(m)$ users at a time.

Example

- Ex.: $\mu < 0.5,\ n_T = 2, m = 2$



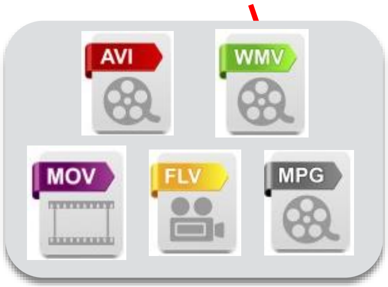
$$n_T = 2$$

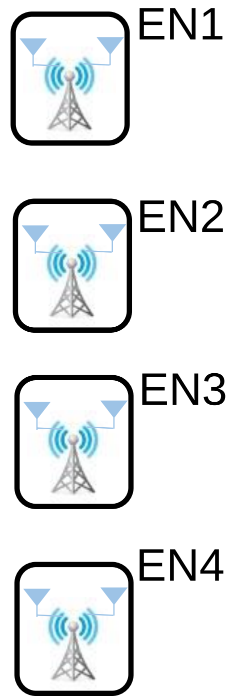
Example

- Ex.: $\mu < 0.5,\ n_T = 2, m = 2$

packetization

W_{n1} | W_{n2}



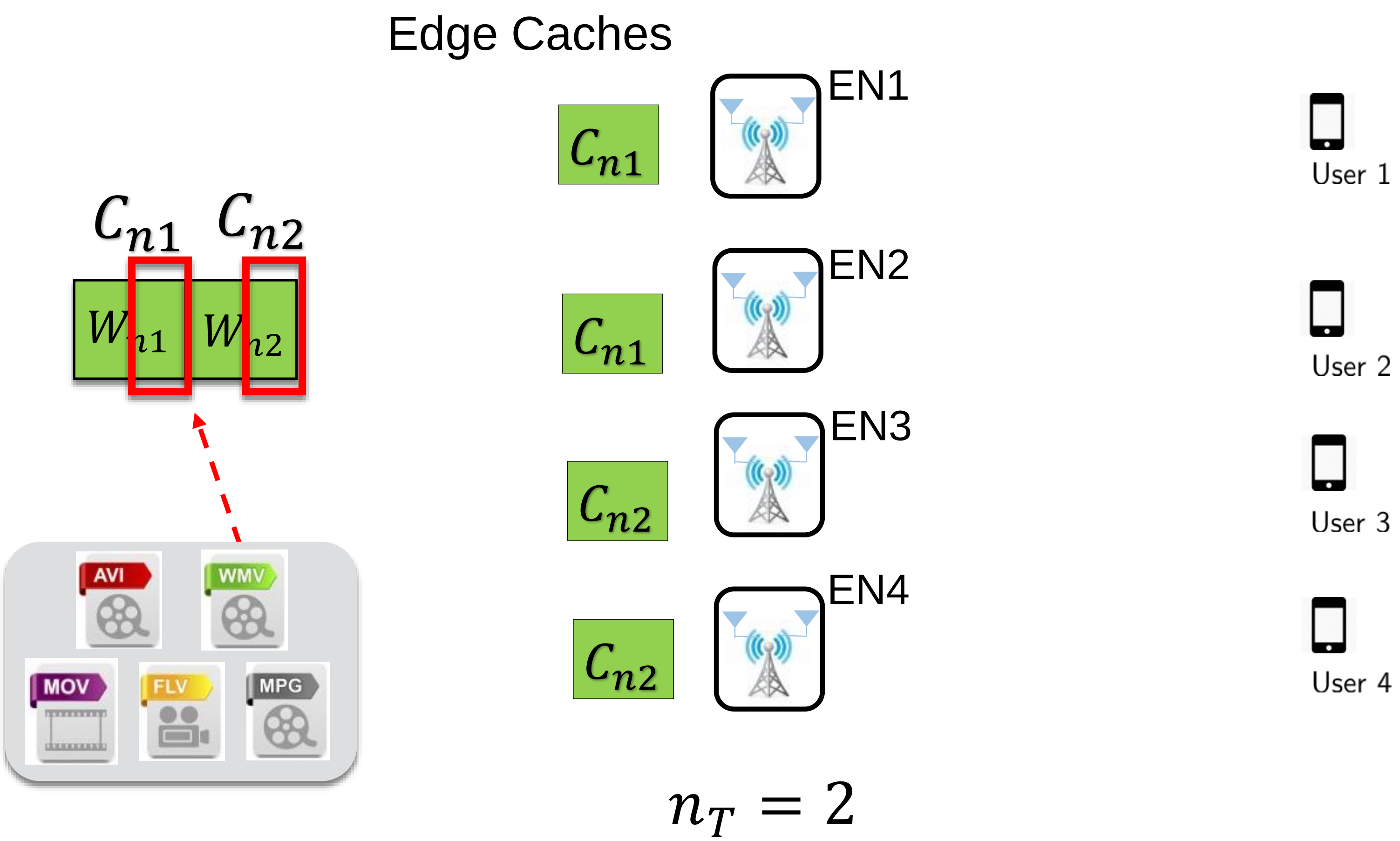


$n_T = 2$

placement phase

Example

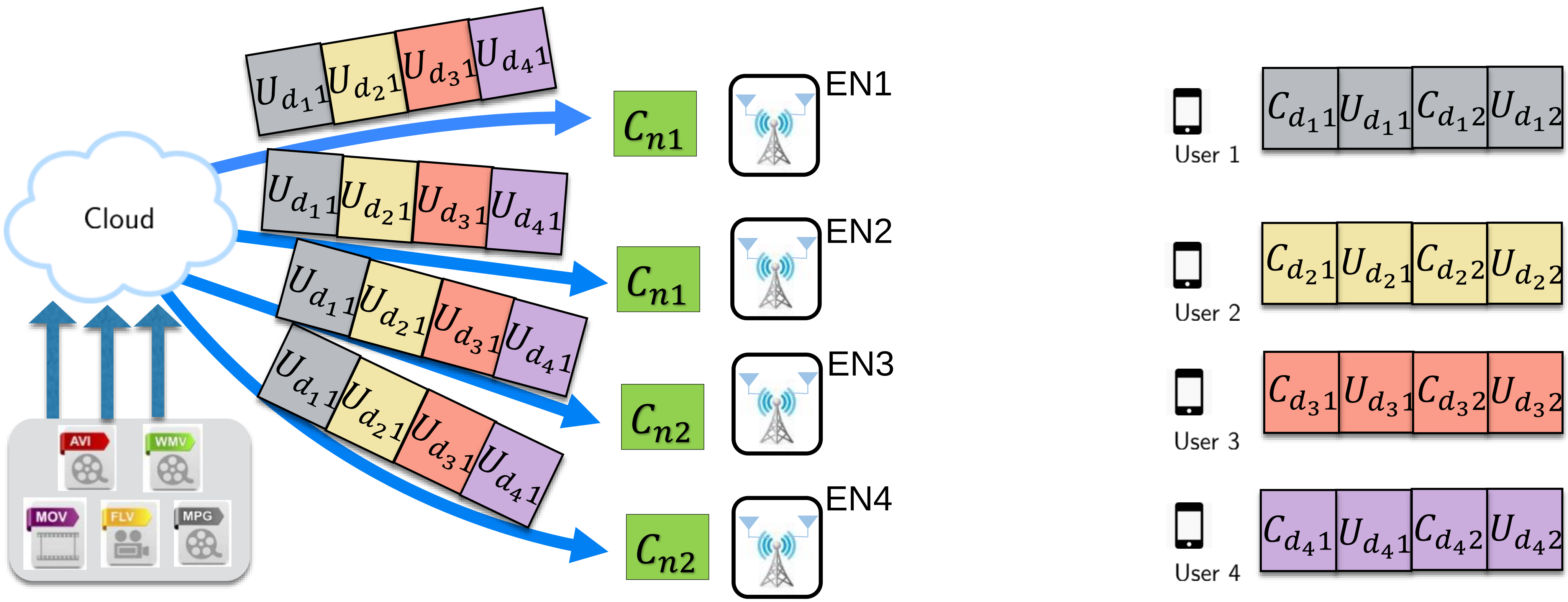
- Ex.: $\mu < 0.5,\ n_T = 2, m = 2$



placement phase

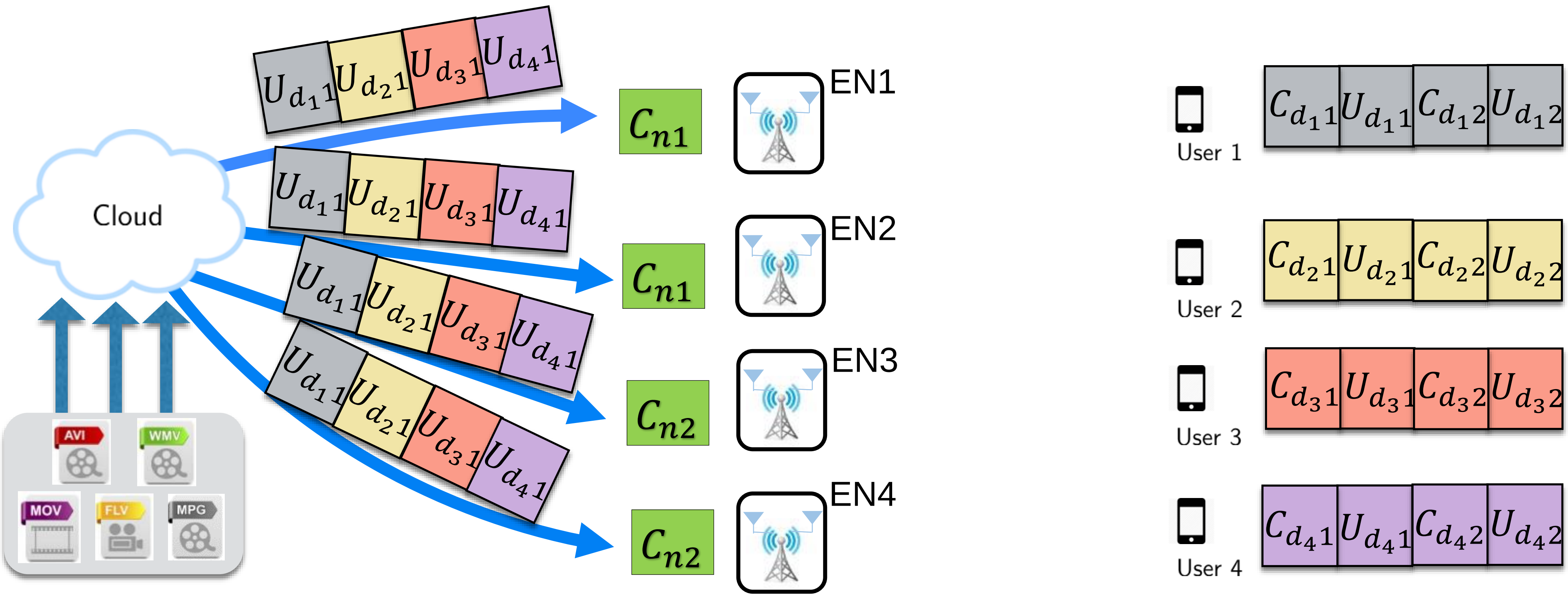
Example

- Ex.: $\mu < 0.5,\ n_T = 2, m = 2$



delivery phase

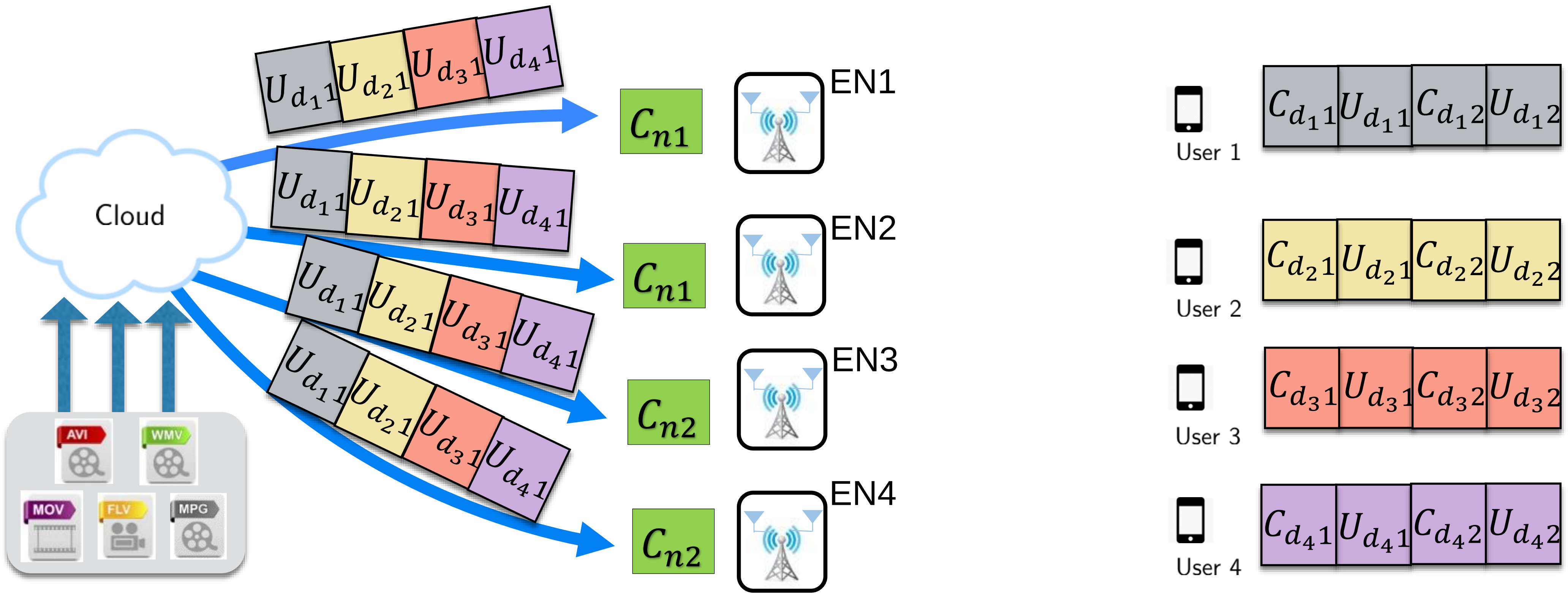
Achievable NDT



- To achieve a multiplicity of m, the required fronthaul NDT is

$$\delta_F(m) = \frac{K_R(m - \mu K_T)}{K_T r}$$

Achievable NDT



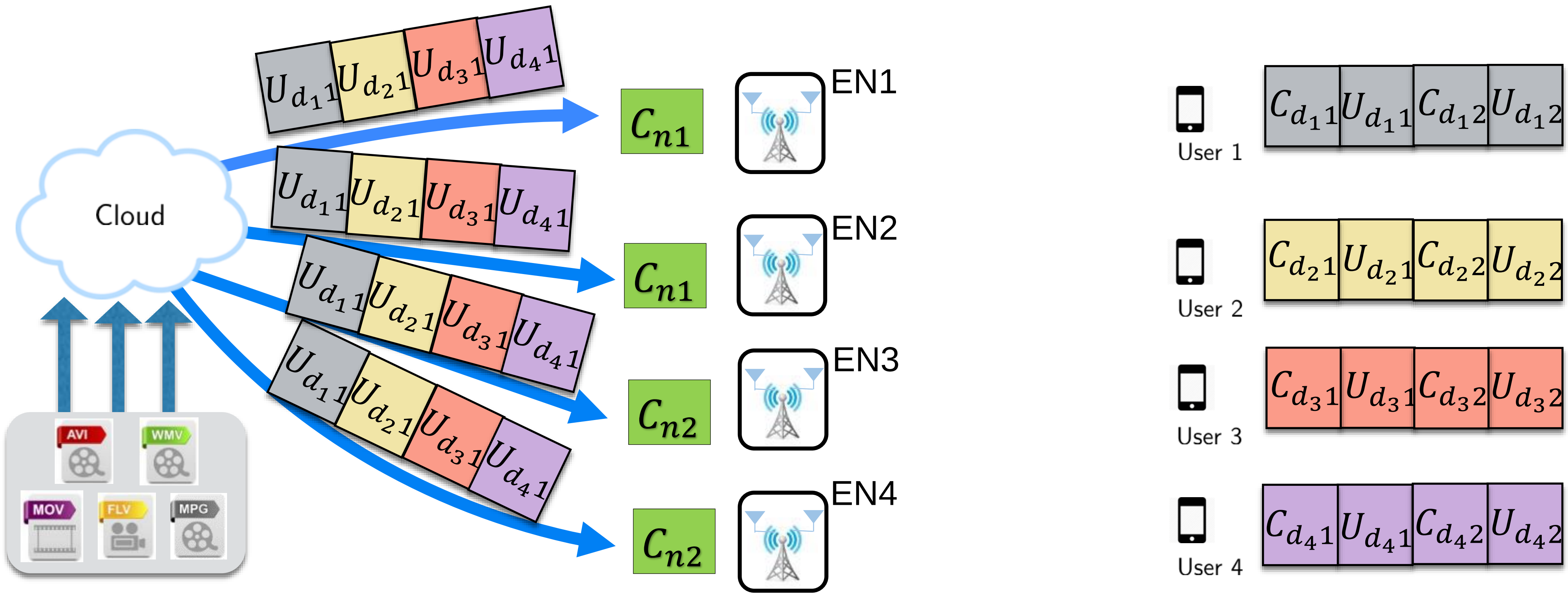
- To achieve a multiplicity of m, the required fronthaul NDT is

$$\delta_F(m) = \frac{K_R(m - \mu K_T)}{K_T r}$$

- And the edge NDT is

$$\delta_E(m) = \frac{K_R}{u(m)}$$

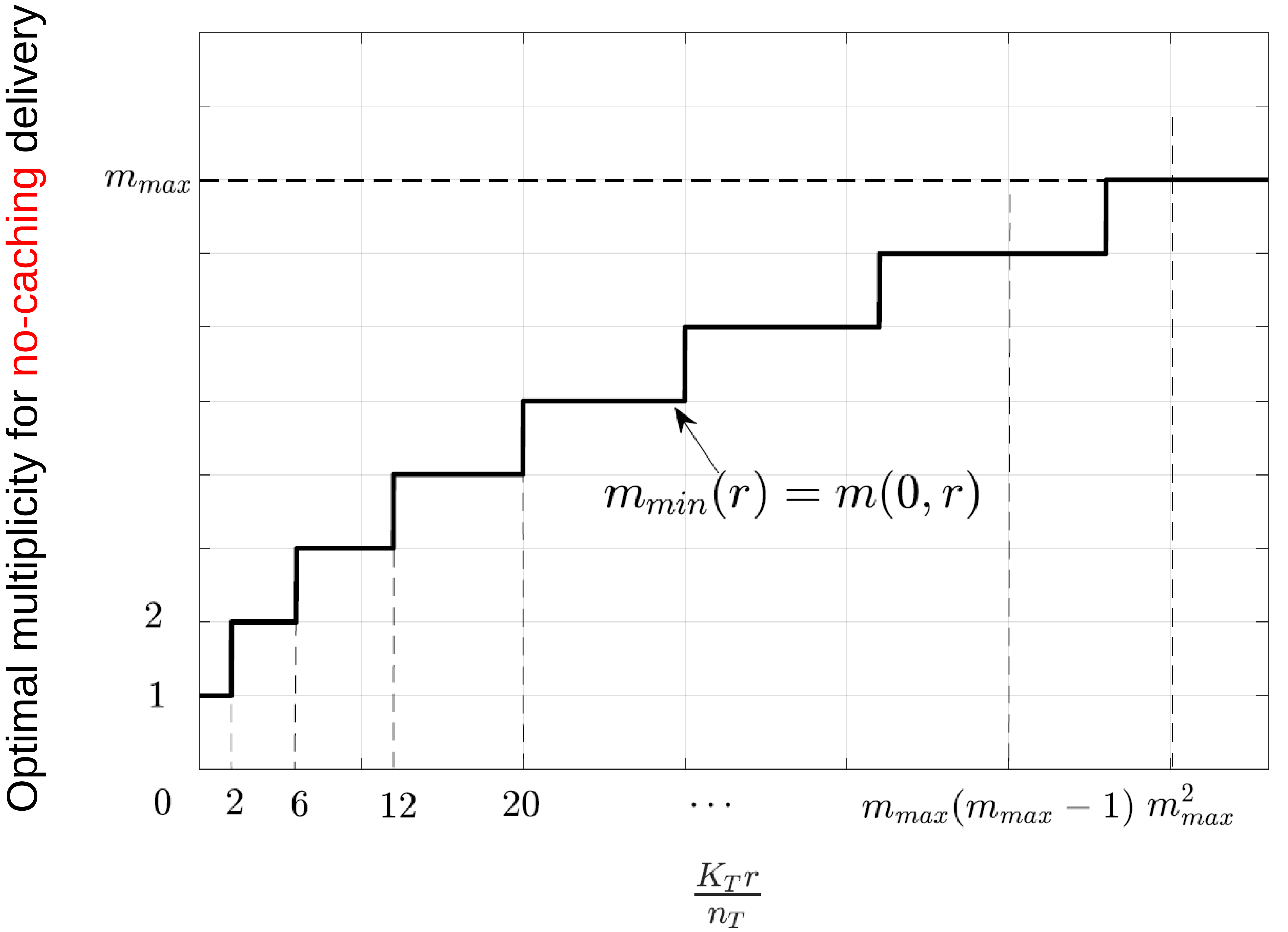
Achievable NDT



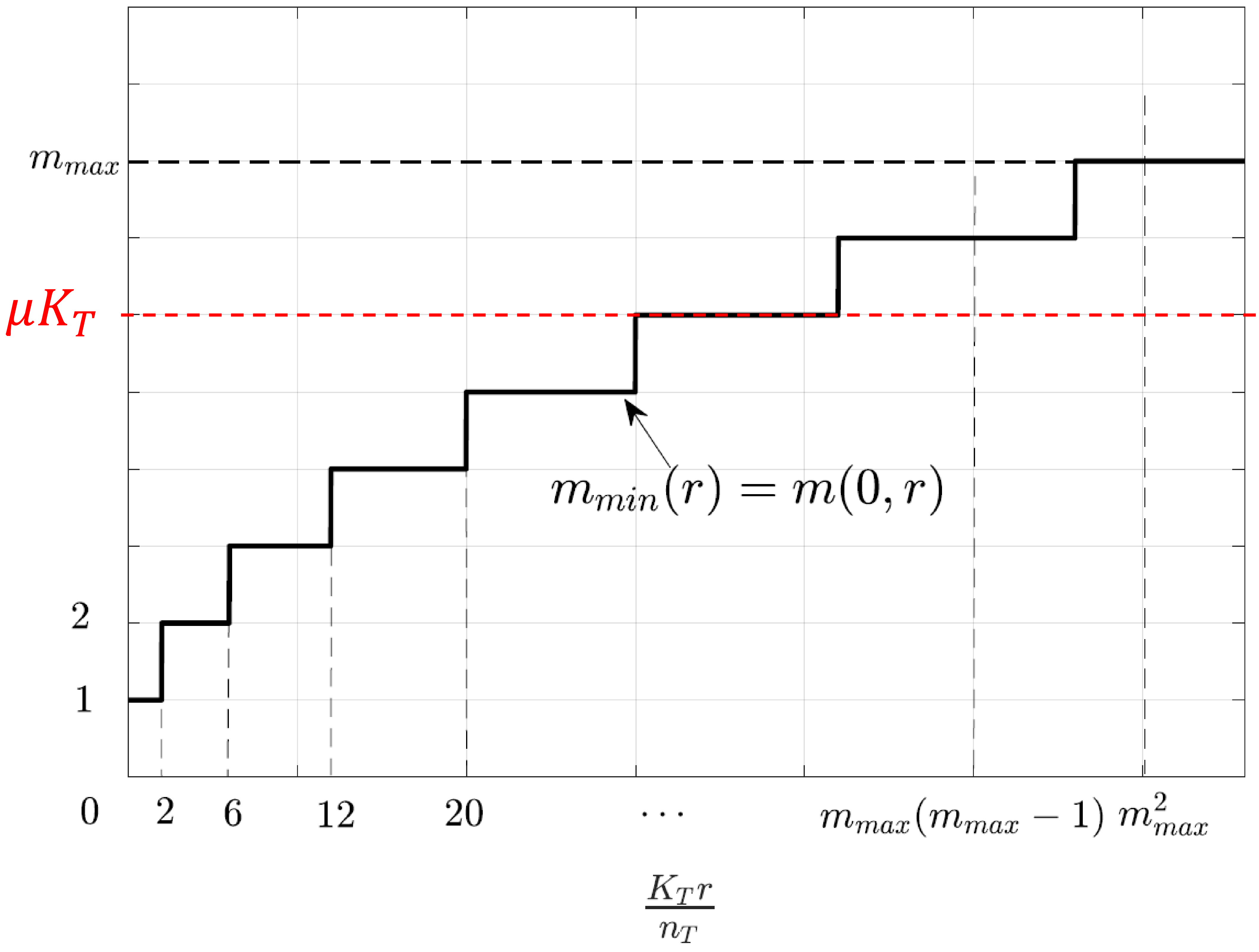
- Choice of the multiplicity

$$\text{minimize } \delta_F(m) + \delta_E(m)$$

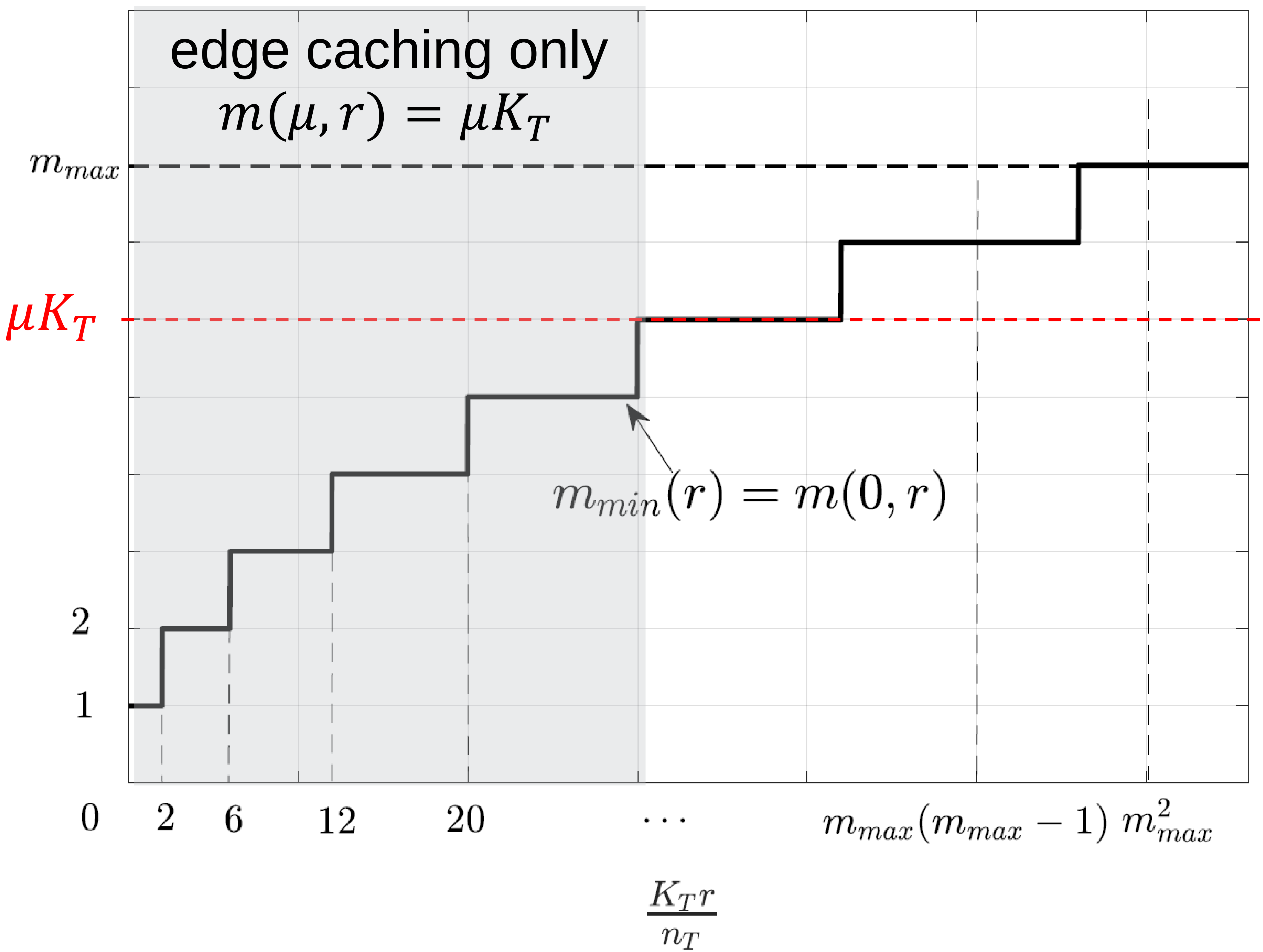
Optimal Multiplicity



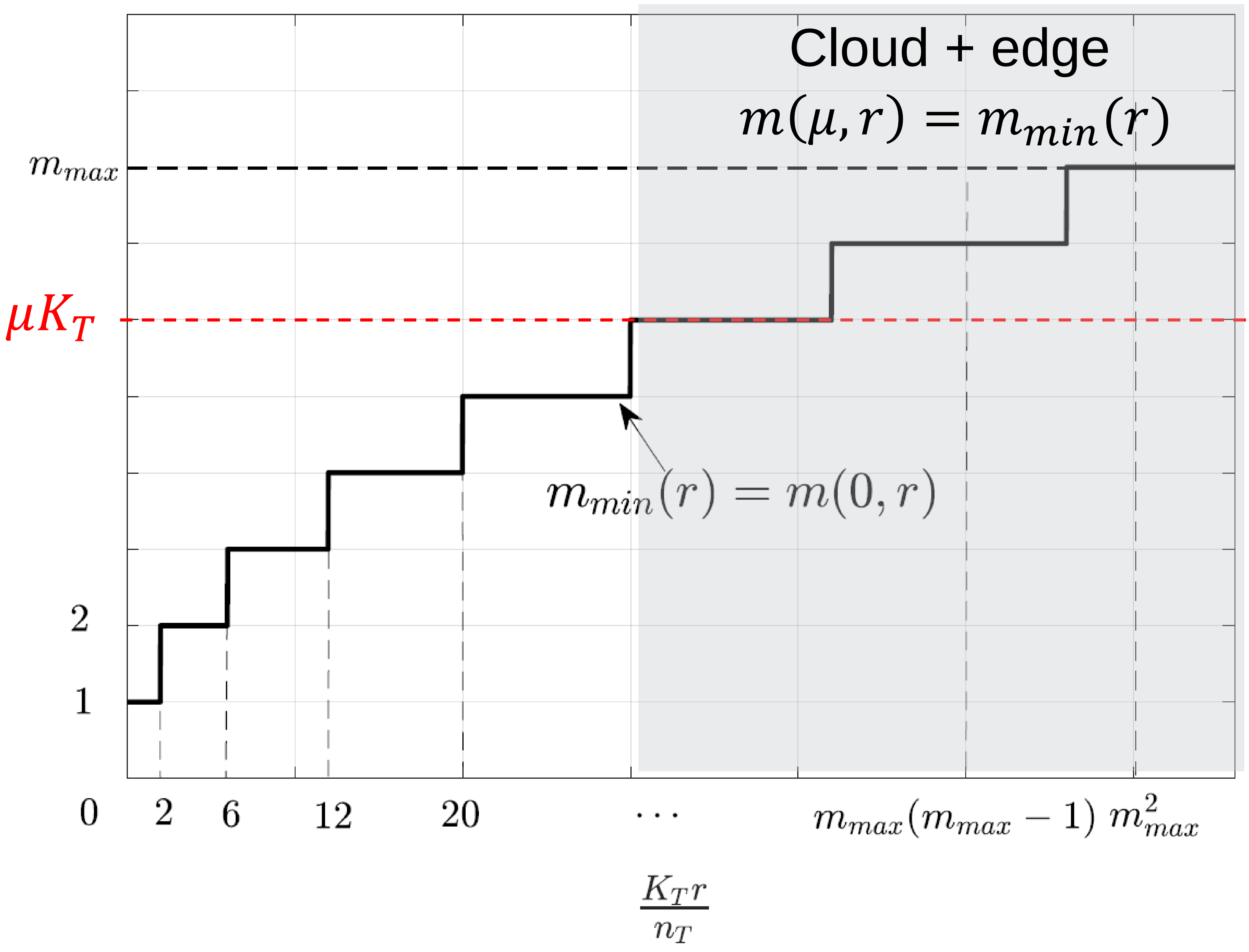
Optimal Multiplicity



Optimal Multiplicity



Optimal Multiplicity



Achievable NDT

Proposition: The following NDT is achievable

$$\delta_{ach}(\mu,r) = \begin{cases} \delta_F(\mu,r) + \delta_E(\mu,r), & \text{for } \mu K_T \leq m_{min}(r), \\ \alpha\delta_E(\mu,r) + (1-\alpha)\delta'_E(\mu,r), & \text{for } \mu K_T \geq m_{min}(r), \end{cases}$$

with the fronthaul NDT

$$\delta_F(\mu,r) = \frac{K_R(m(\mu,r\,) - \mu K_T)}{K_T r}$$

and the edge NDTs

$$\delta_E(\mu,r) = \frac{K_R}{u(\mu,r)}, \delta'_E(\mu,r) = \frac{K_R}{\min(\lceil \mu K_T \rceil n_T, K_R)}$$

and $\alpha = 1 + \lfloor \mu K_T \rfloor - \mu K_T$

Minimum NDT

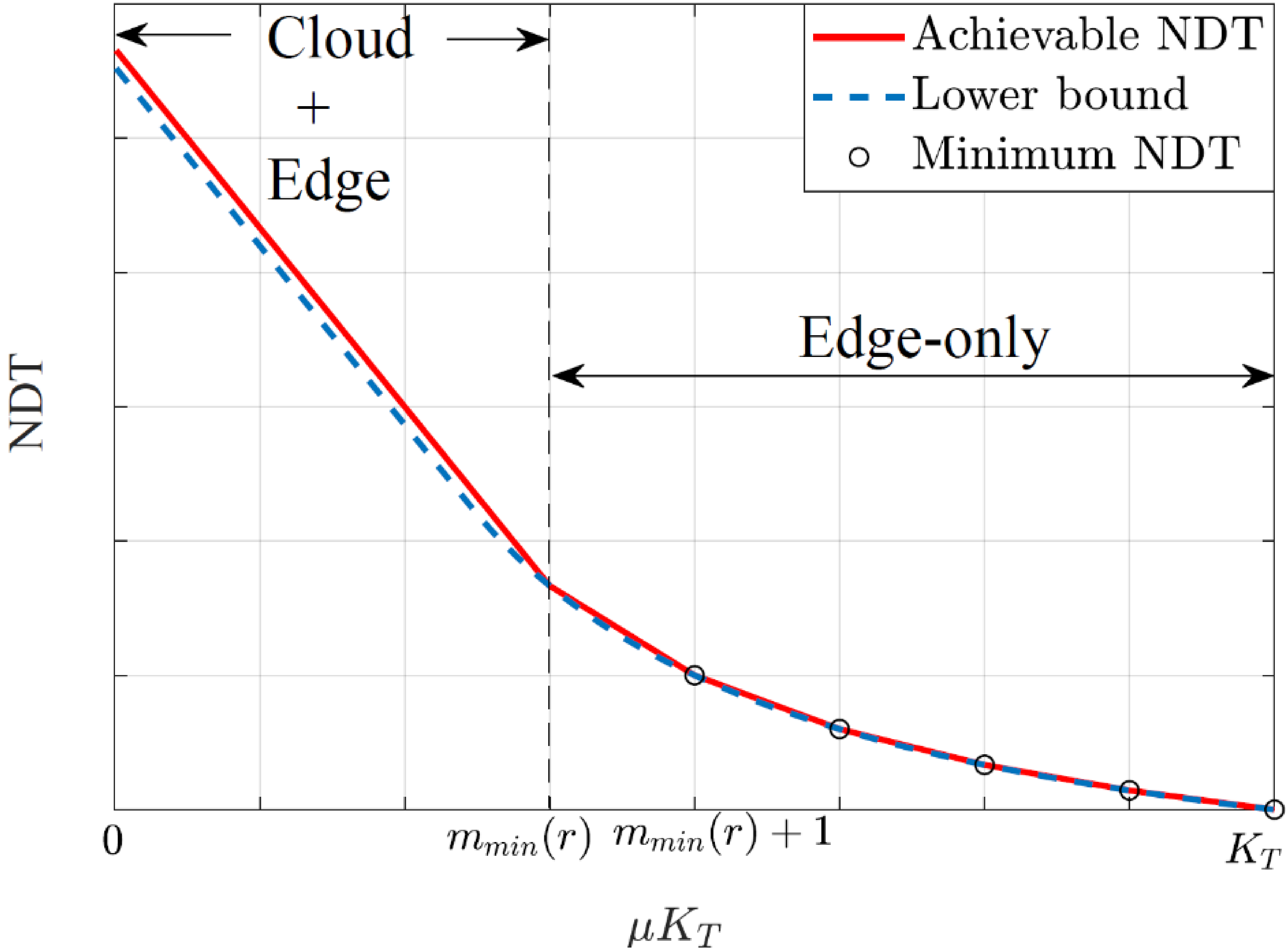
Proposition: For an F-RAN system with n_T antennas at each EN, the minimum NDT $\delta^*(\mu, r)$ is given as

$$\delta^*(\mu, r) = \begin{cases} \max\left\{\frac{K_R(1-\mu K_T)}{K_T r} + \frac{K_R}{n_T}, 1\right\}, & \text{for } \mu K_T \in [0,1] \text{ and } r \in [0, \frac{n_T}{K_T}], \\ \max\left\{\frac{K_R}{\mu K_T n_T}, 1\right\}, & \text{for } \mu K_T \in [m_{min}(r)+1, \cdots, m_{max}] \cup [m_{max}, K_T] \end{cases}$$

Proposition: For an F-RAN system with n_T antennas at each EN, and any value of $\mu \geq 0$ and $r \geq 0$, we have the inequality

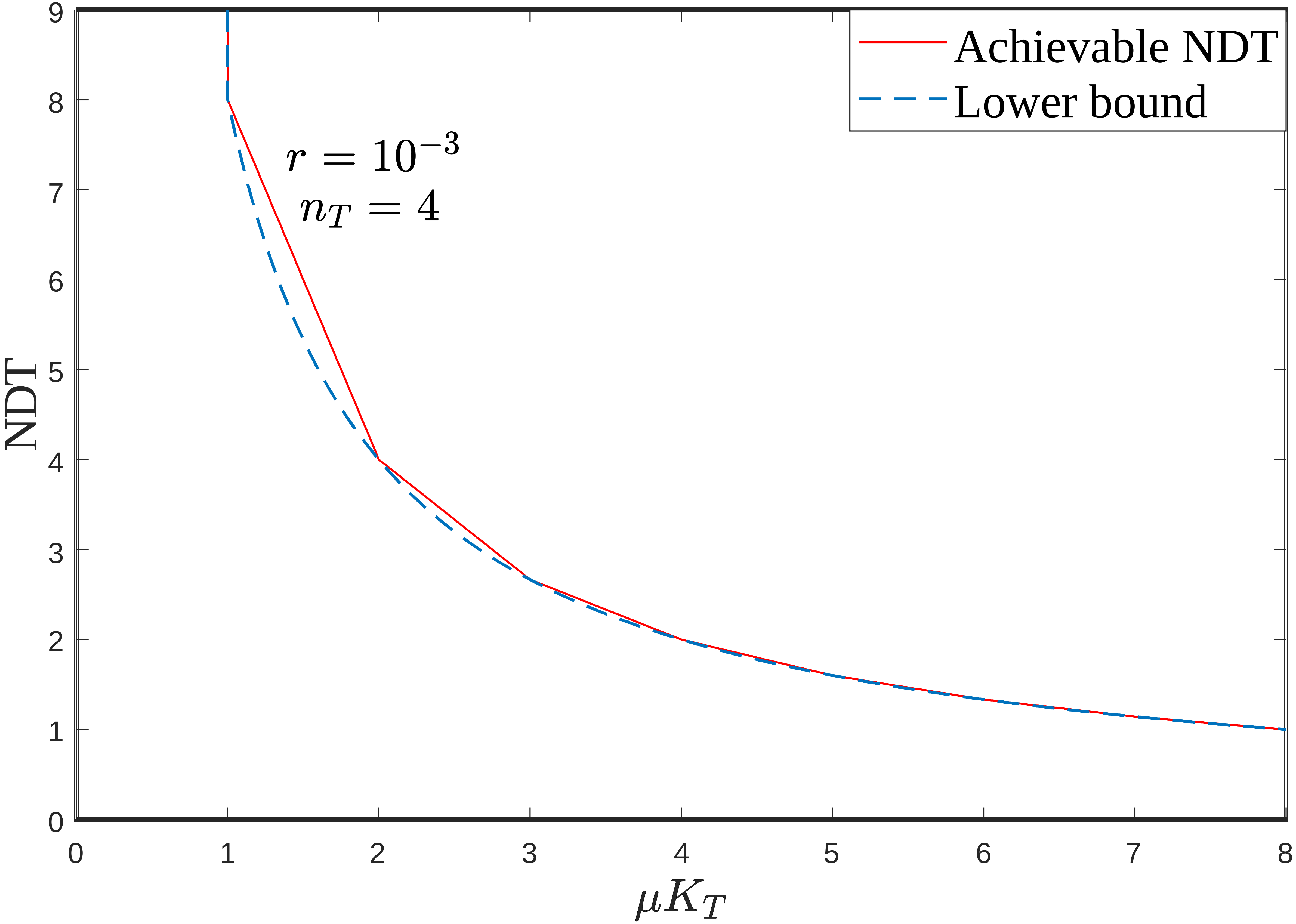
$$\frac{\delta_{ach}(\mu, r)}{\delta^*(\mu, r)} \leq \frac{3}{2}$$

Minimum NDT

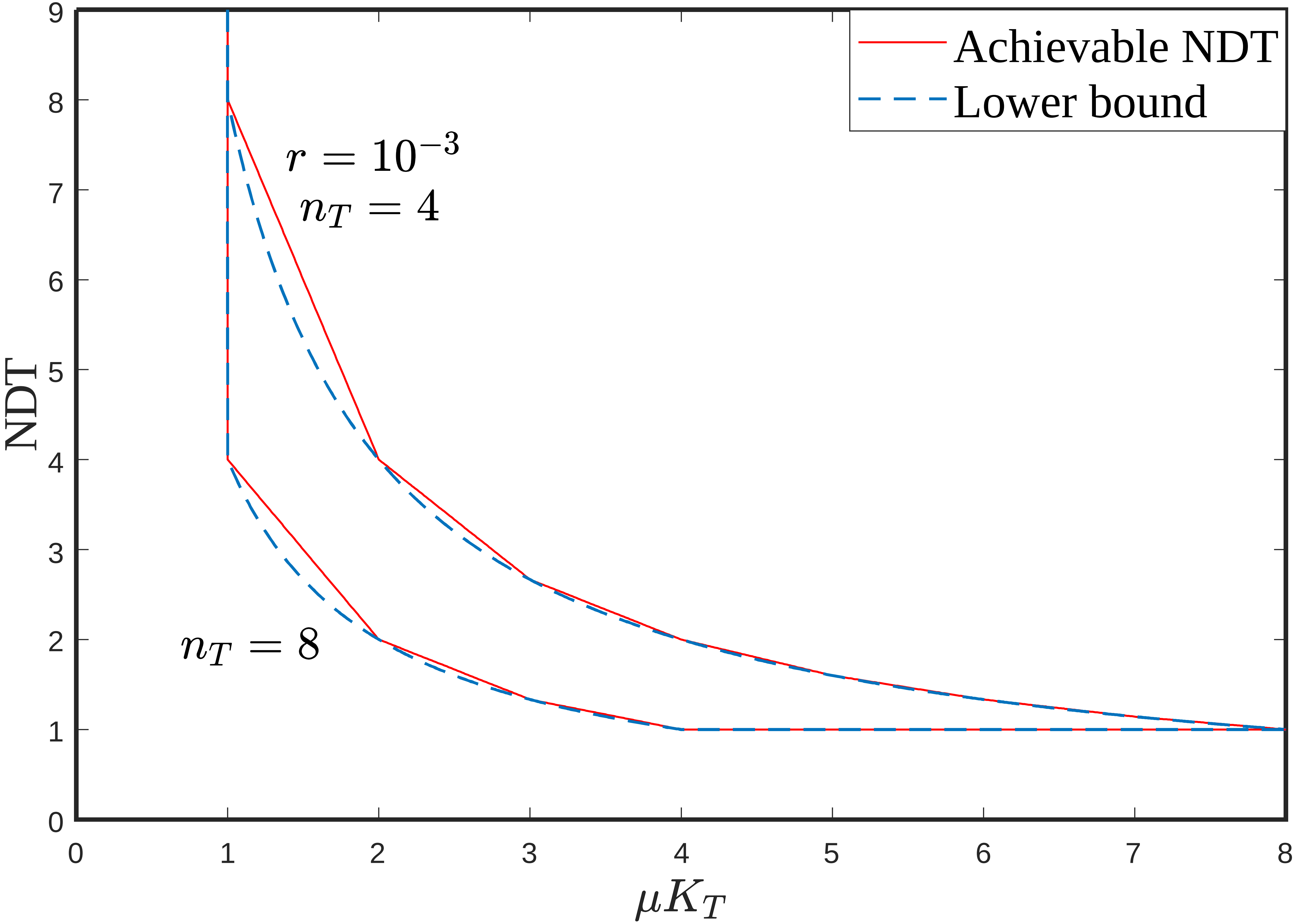


Cloud
+
Edge
Achievable NDT
Lower bound
Minimum NDT
Edge-only
NDT
0
$m_{min}(r)$
$m_{min}(r)+1$
K_T
μK_T

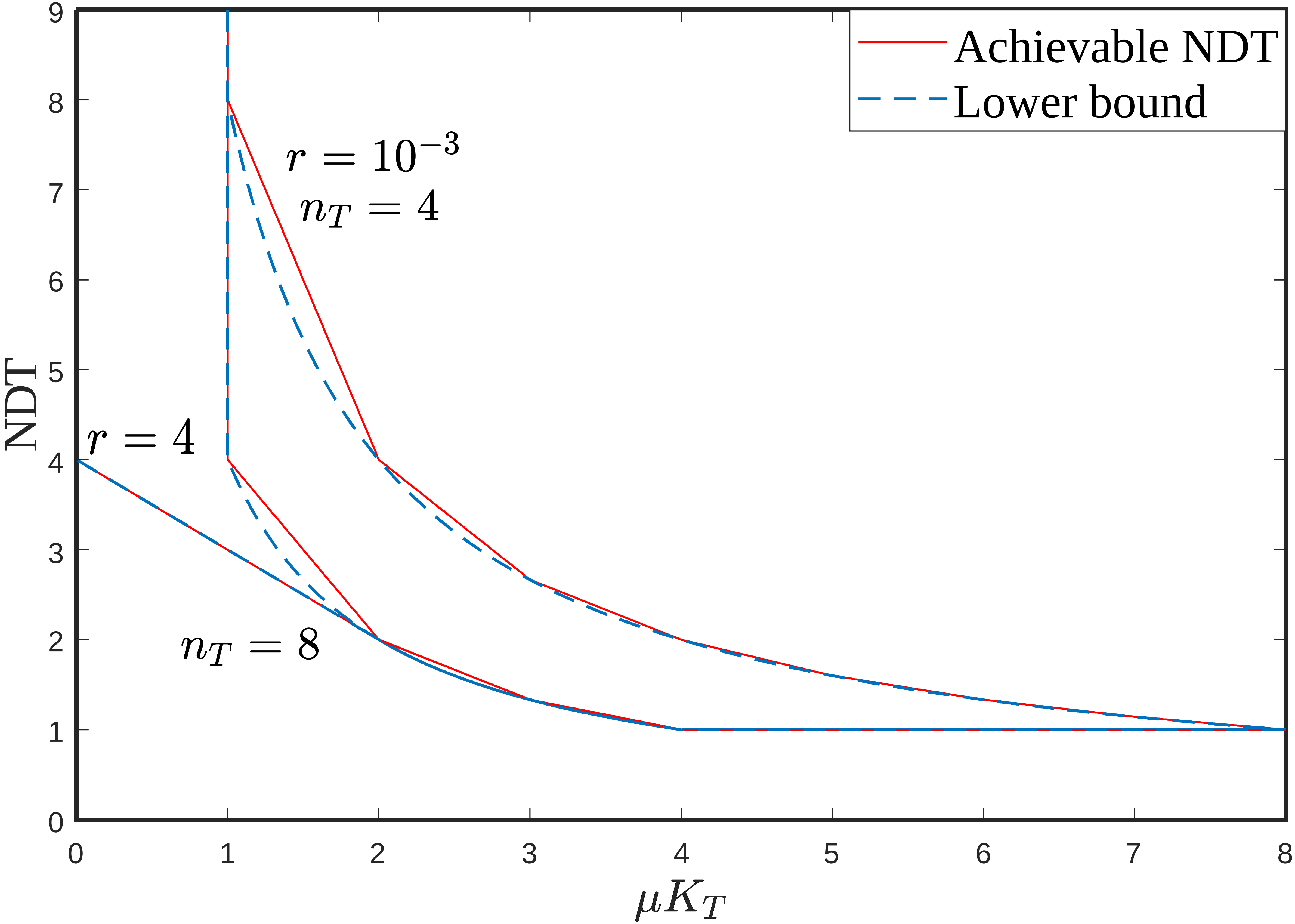
Minimum NDT



Minimum NDT


Achievable NDT
Lower bound
$r = 10^{-3}$
$n_T = 4$
$n_T = 8$
NDT
μK_T
0
1
2
3
4
5
6
7
8
9

Minimum NDT



Achievable NDT
Lower bound
$r = 10^{-3}$
$n_T = 4$
$r = 4$
$n_T = 8$
NDT
μK_T
0
1
2
3
4
5
6
7
8
9

F-RAN
(under general delivery)

A. Sengupta, R. Tandon and O. Simeone, “Fog-Aided Wireless Networks for Content Delivery: Fundamental Latency Trade-Offs,” IEEE Trans. Inf. Theory, Oct. 2018.

Constraints

✓ Uncoded (fractional) caching

✓ Transport of uncoded (fractional) contents

✓ One-shot linear precoding

Removing Constraints

- ✓ ~~Uncoded (fractional) caching~~

  Allow for intra-file coding

- ✓ ~~Transport of uncoded (fractional) contents~~

  General fronthaul strategy

- ✓ ~~One-shot linear precoding~~

  General edge delivery strategy

Removing Constraints

- ✓ ~~Uncoded (fractional) caching~~

  Allow for intra-file coding

- ✓ ~~Transport of uncoded (fractional) contents~~

  General fronthaul strategy

- ✓ ~~One-shot linear precoding~~

  General edge delivery strategy

- Single antenna ENs

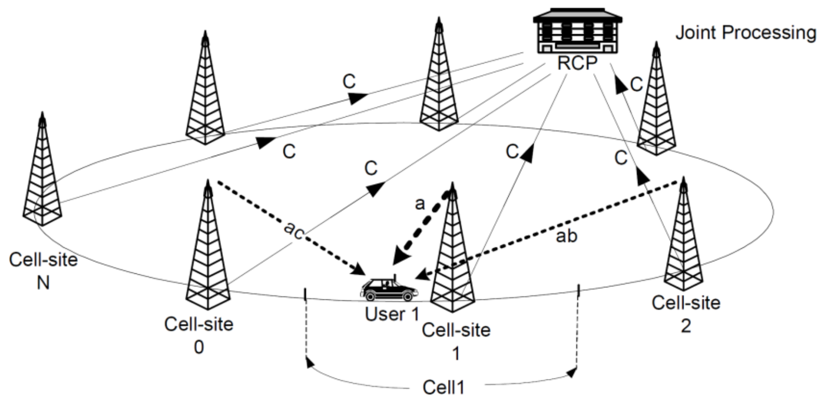
Cloud Radio Access Network

- From paper…

2007

Uplink Macro Diversity with Limited Backhaul Capacity

Amichai Sanderovich*, Oren Somekh†, and Shlomo Shamai (Shitz)*



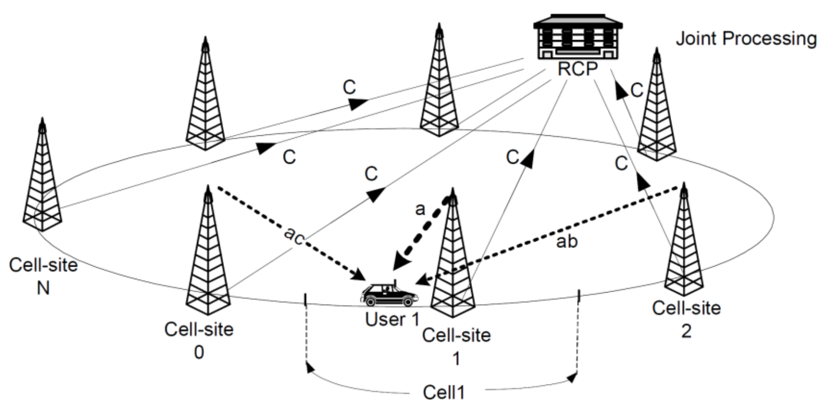
Cloud Radio Access Network

- From paper to industry white paper…

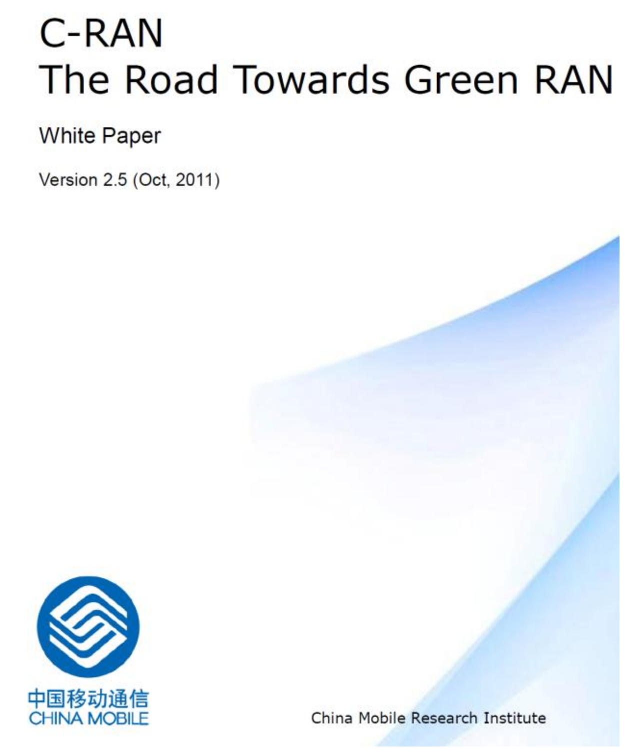
2007

Uplink Macro Diversity with Limited Backhaul Capacity

Amichai Sanderovich*, Oren Somekh†, and Shlomo Shamai (Shitz)*



2011


C-RAN
The Road Towards Green RAN

White Paper

Version 2.5 (Oct, 2011)

中国移动通信 CHINA MOBILE

China Mobile Research Institute

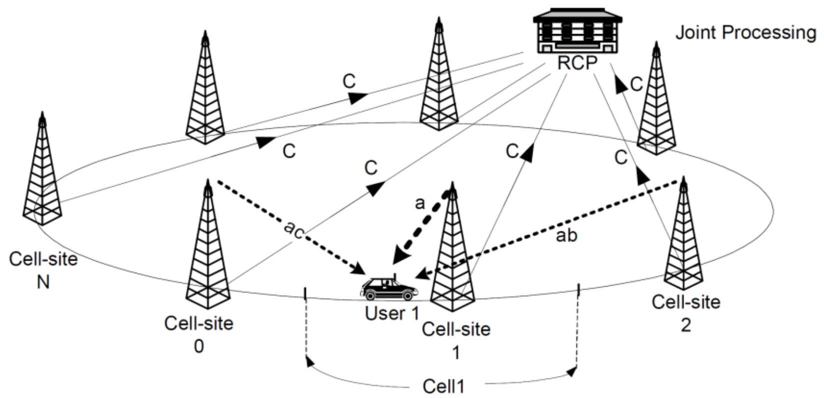
Cloud Radio Access Network

- From paper to industry white paper to deployment…

2007

Uplink Macro Diversity with Limited Backhaul Capacity

Amichai Sanderovich*, Oren Somekh†, and Shlomo Shamai (Shitz)*

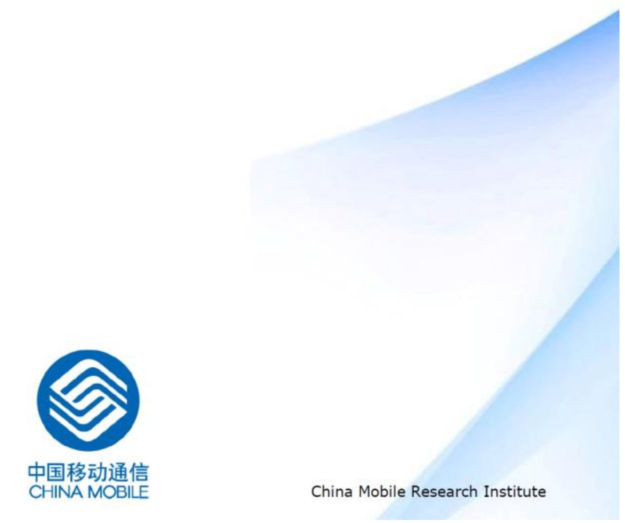


2011

C-RAN
The Road Towards Green RAN

White Paper

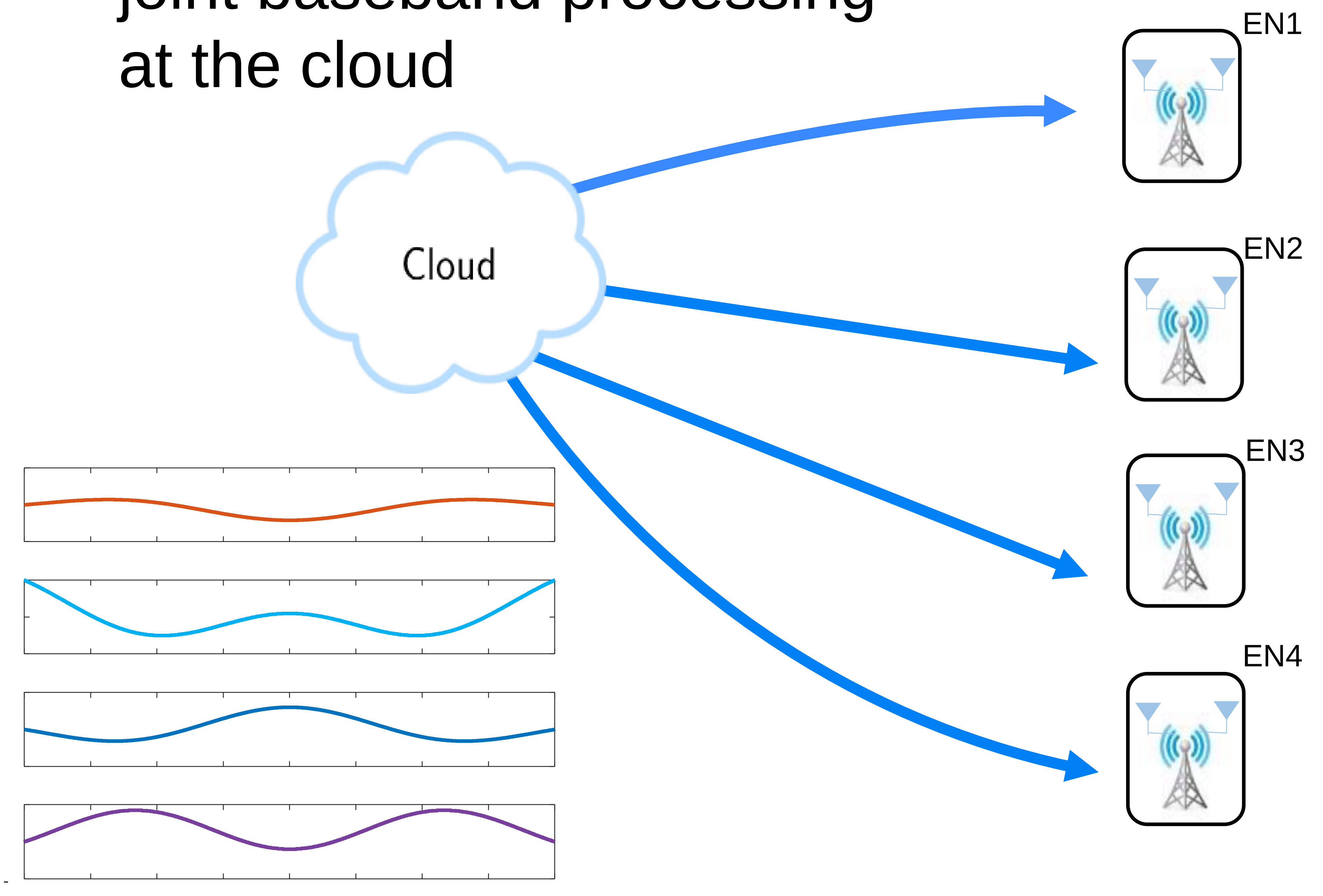
Version 2.5 (Oct, 2011)



2017

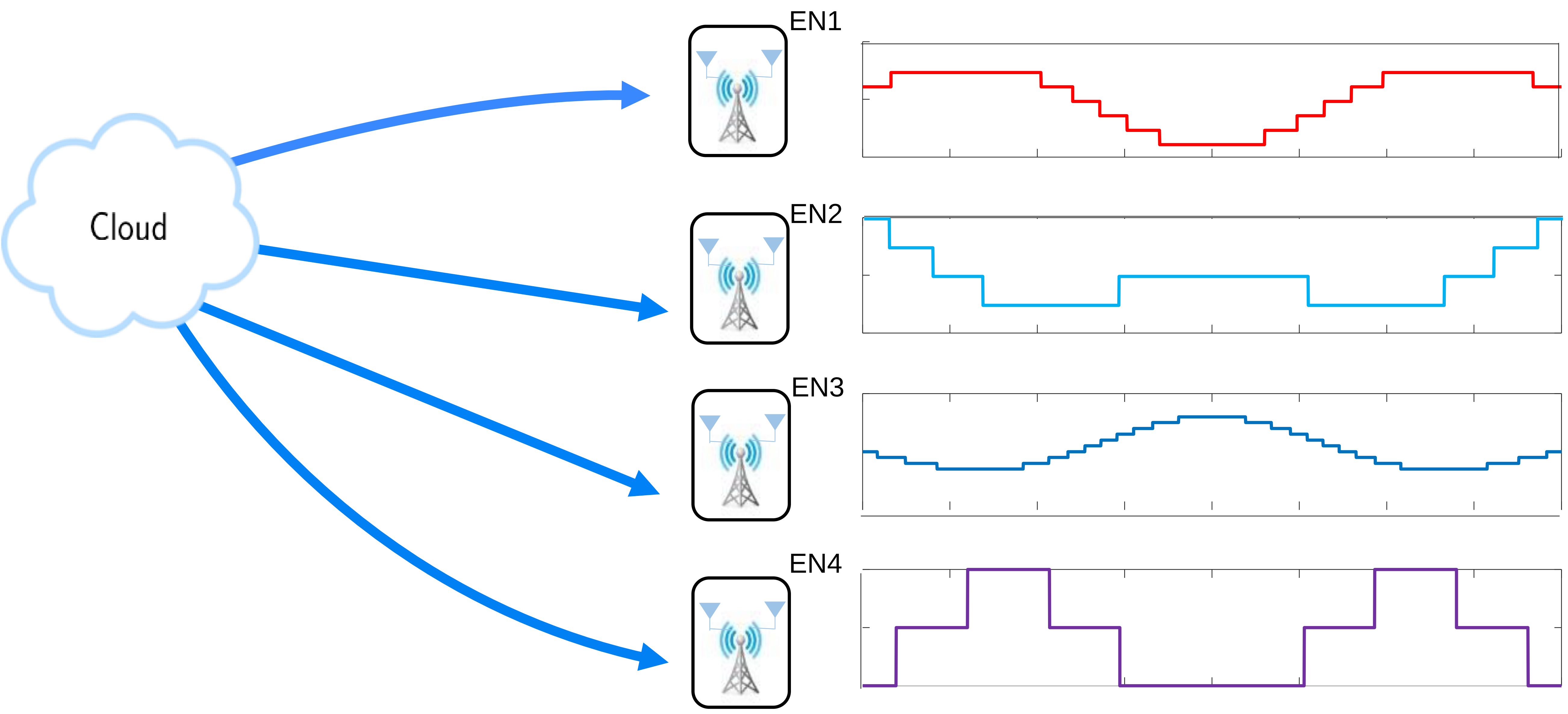
Cloud Radio Access Network

joint baseband processing
at the cloud



Cloud Radio Access Network

fronthaul quantization/ compression and transmission



Uncoded Transmission vs C-RAN

- Consider $K_T = K_R = K$ and $\mu = 0$

- In order to achieve an edge NDT of 1, uncoded transmission requires a fronthaul NDT

$$\delta_F = \frac{m}{r} = \frac{K}{r}$$

Uncoded Transmission vs C-RAN

- Consider $K_T = K_R = K$ and $\mu = 0$
- In order to achieve an edge NDT of 1, uncoded transmission requires a fronthaul NDT

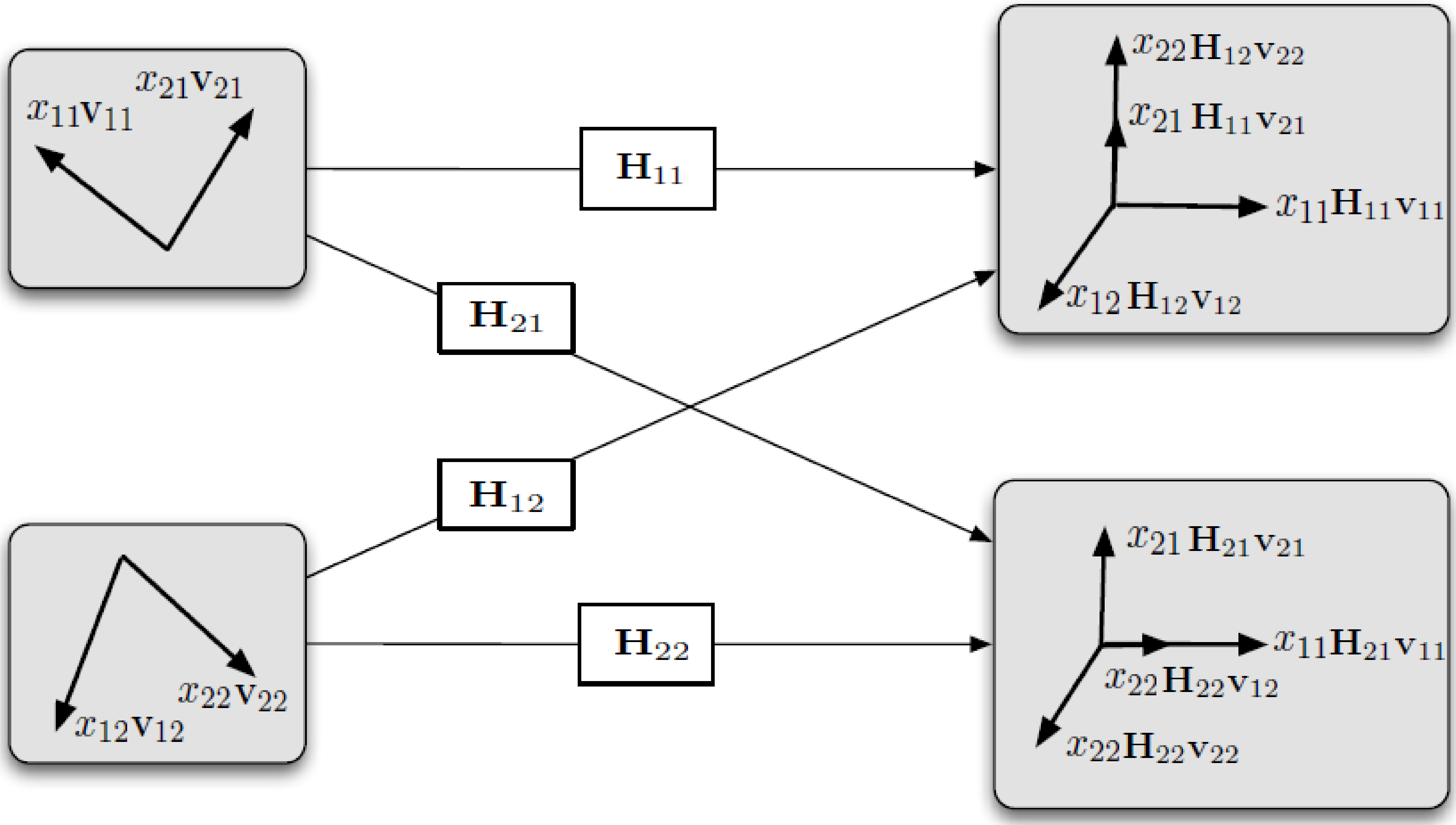
$$\delta_F = \frac{m}{r} = \frac{K}{r}$$

- With a proper choice of the quantization resolution, fronthaul compression requires

$$\delta_F = \frac{1}{r}$$

which does not scale with *K*.

Interference Alignment



- EN coordination based on precoding over linear multiple symbols or non-linear precoding [Cadambe and Jafar ‘09] [Motahari et al ’14].

One-Shot Beamforming vs Interference Alignment

- Consider $K_T = K_R = K,\ \mu = \frac{1}{K}$ and $r = 0$
- With one-shot linear precoding, since $m(\mu) = \mu K = 1$, we have

$$\delta_E = K$$

One-Shot Beamforming vs Interference Alignment

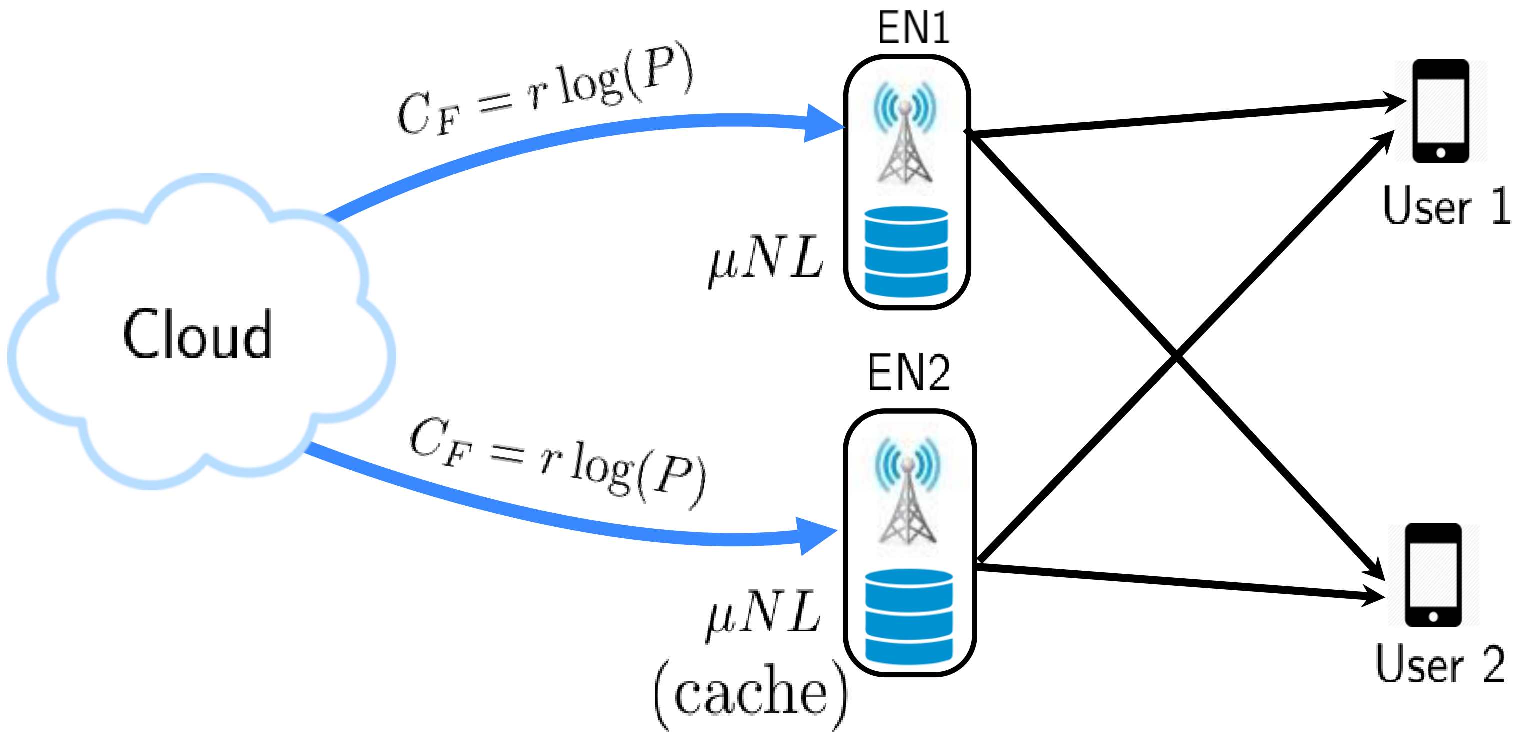
- Consider $K_T = K_R = K,\ \mu = \frac{1}{K}$ and $r = 0$
- With one-shot linear precoding, since $m(\mu) = \mu K = 1$, we have

$$\delta_E = K$$

- With interference alignment (on an X-channel), we have [Cadambe and Jafar '09][Motahari et al '14]

$$\delta_E = 2 - 1/K$$

Example



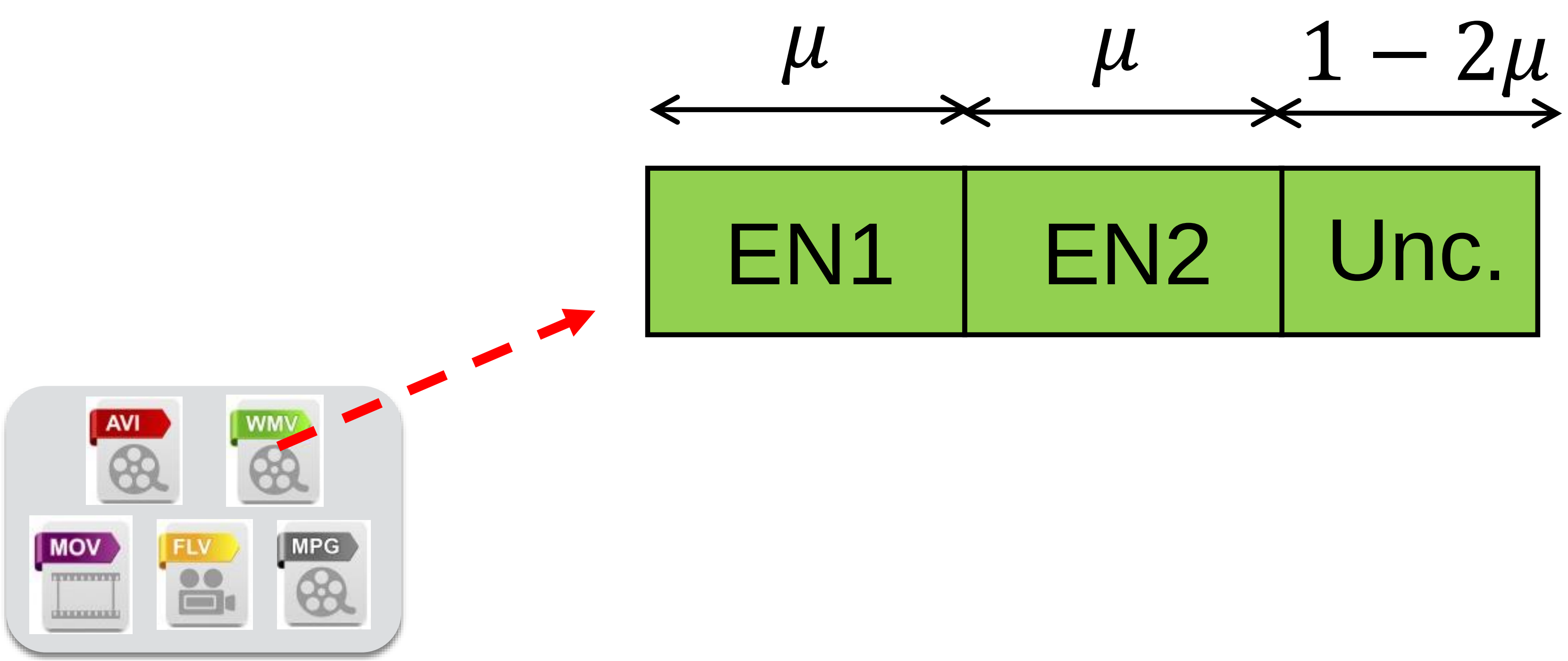
Low fronthaul capacity

$r < 1$

Low cache capacity

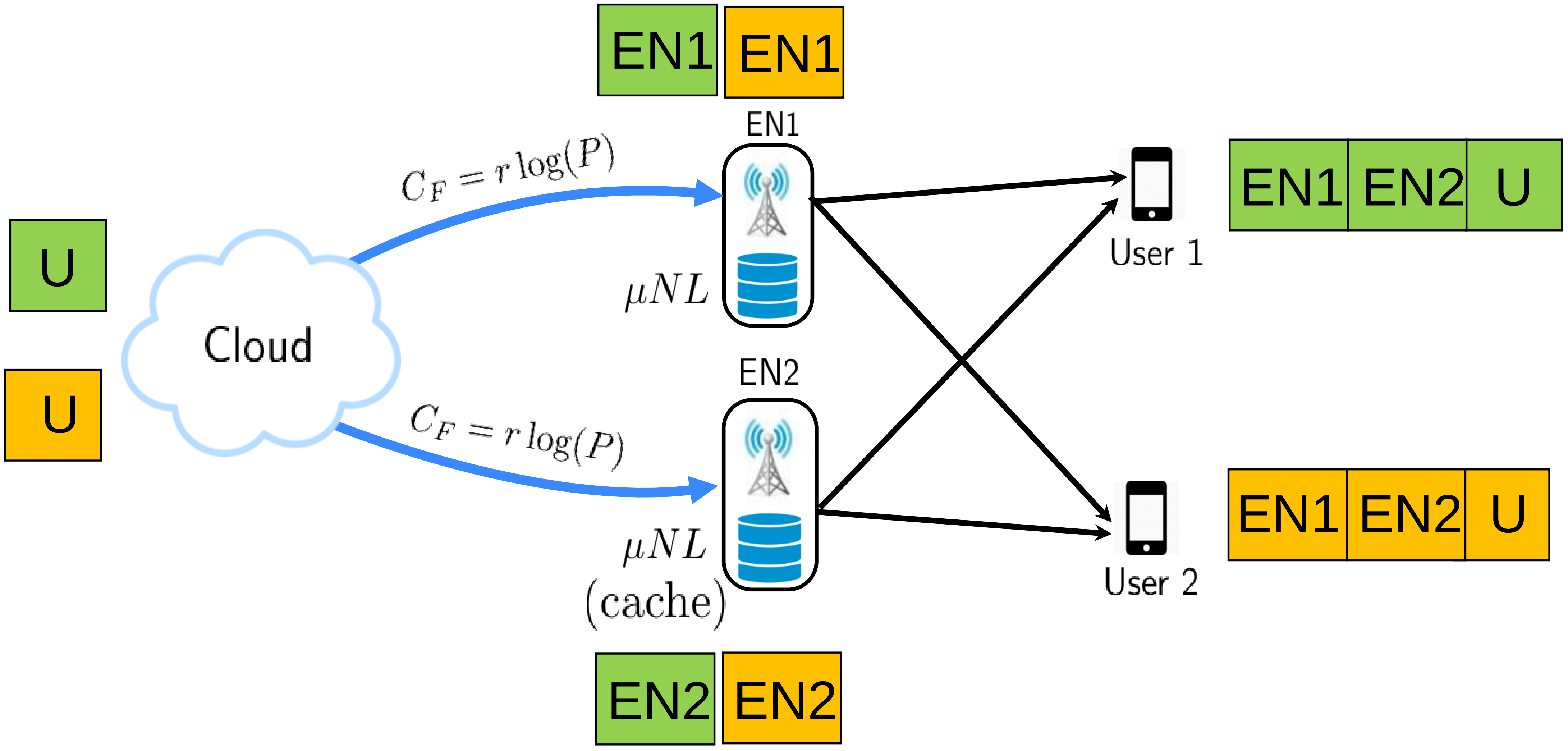
$\mu \leq \frac{1}{2}$

Example



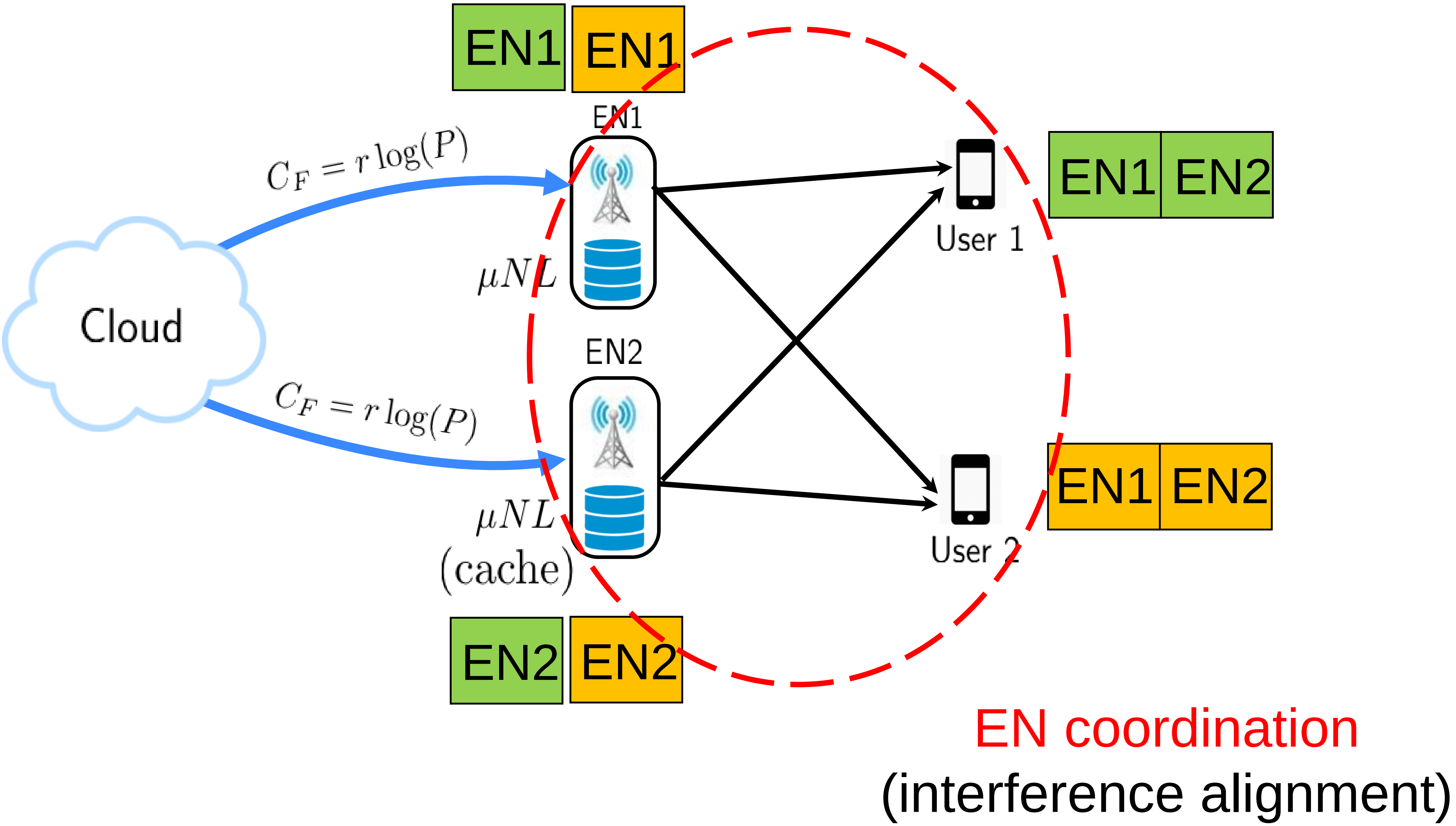
placement phase

Example



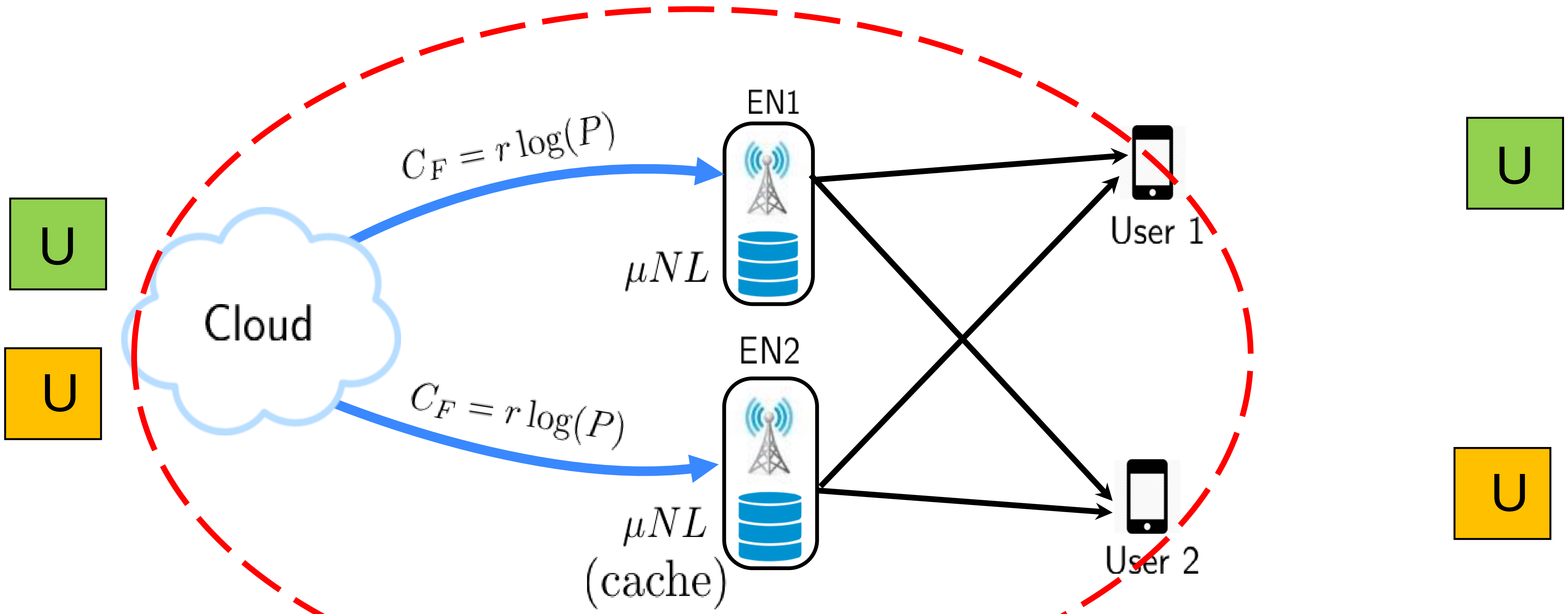
delivery phase

Example



delivery phase

Example



C-RAN transmission
(fronthaul compression)

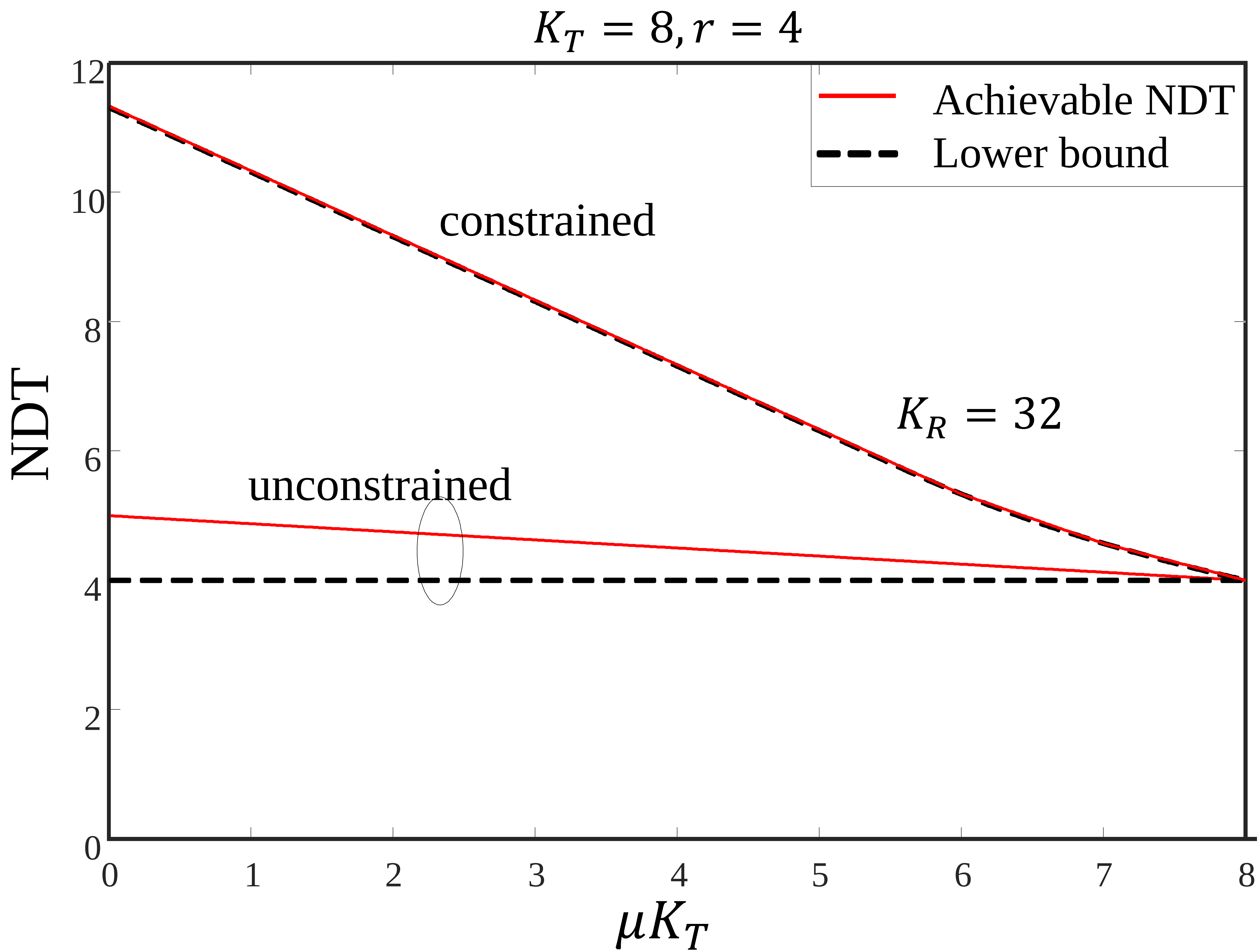
delivery phase

Minimum NDT

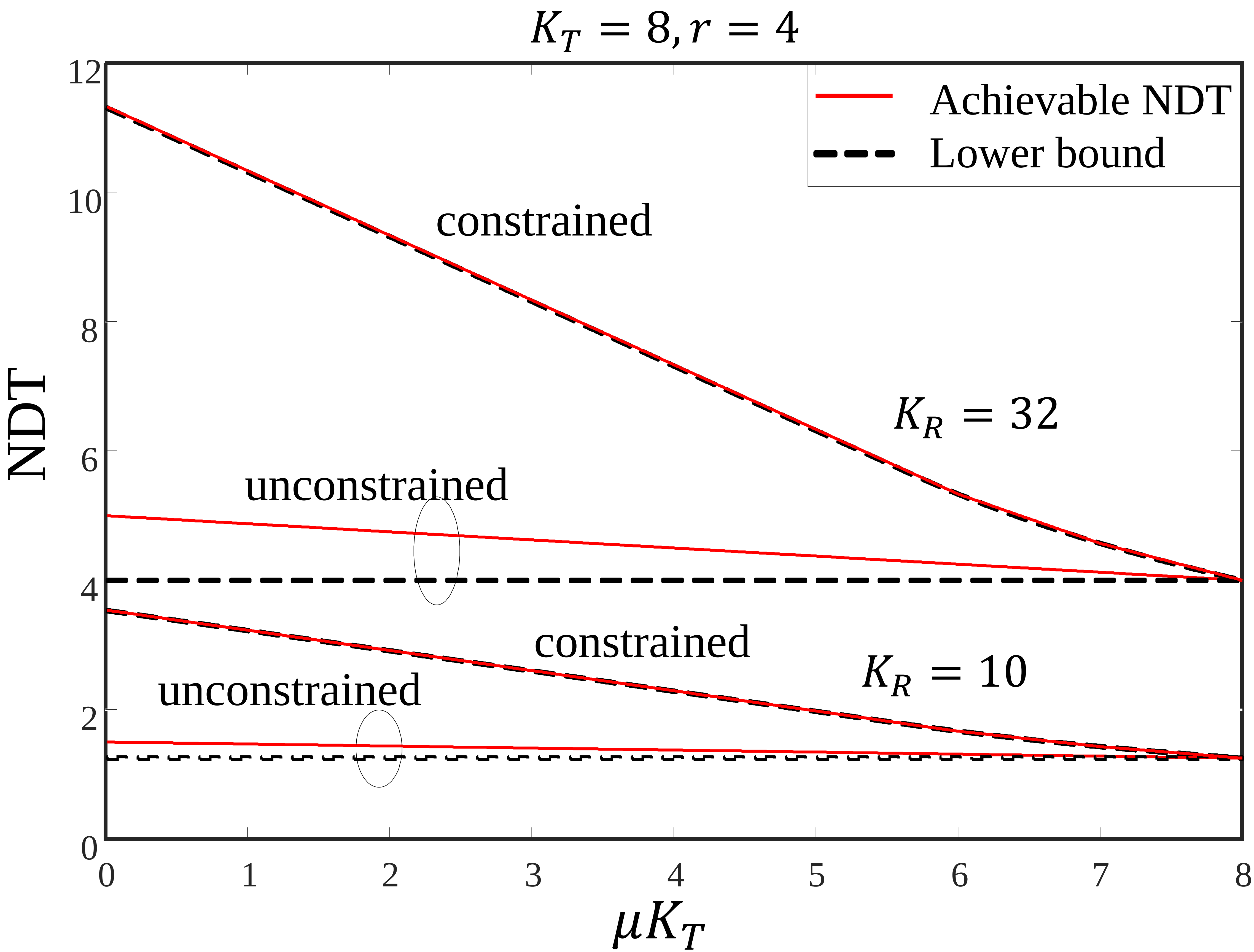
Theorem: Integrating fronthaul compression and interference alignment, the minimum NDT can be achieved within a multiplicative factor of 2 for $N \geq K$, i.e.,

$$\frac{\delta_{\mathrm{off,ach}}(\mu, r)}{\delta^*(\mu, r)} \leq 2$$

Performance Comparison



Performance Comparison



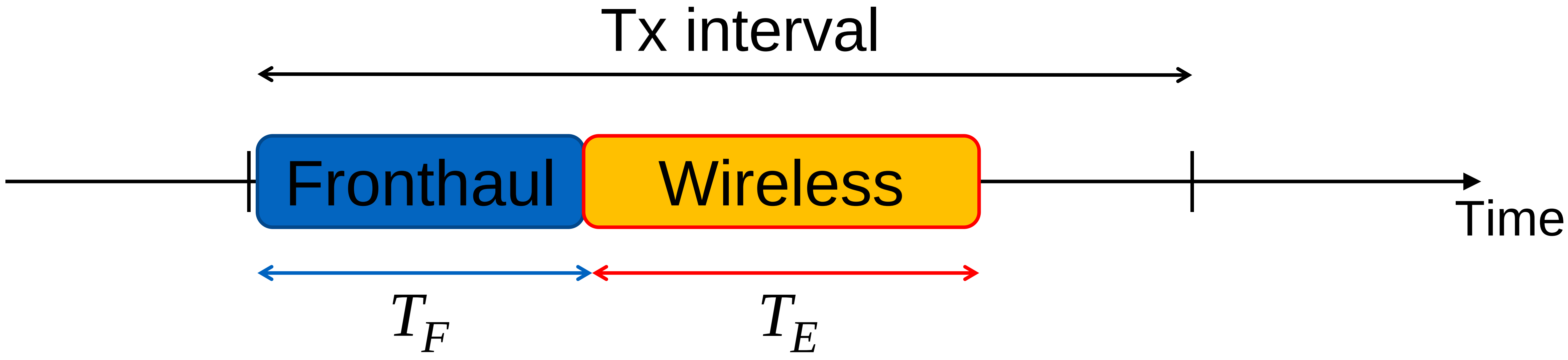
Extensions:
Delivery Protocol

A. Sengupta, R. Tandon and O. Simeone, “Fog-Aided Wireless Networks for Content Delivery: Fundamental Latency Trade-Offs,” IEEE Trans. Inf. Theory, Oct. 2018.

J. Zhang and O. Simeone, “Fundamental Limits of Cloud and Cache-Aided Interference Management with Multi-Antenna Base Stations,” arXiv:1712.04266.

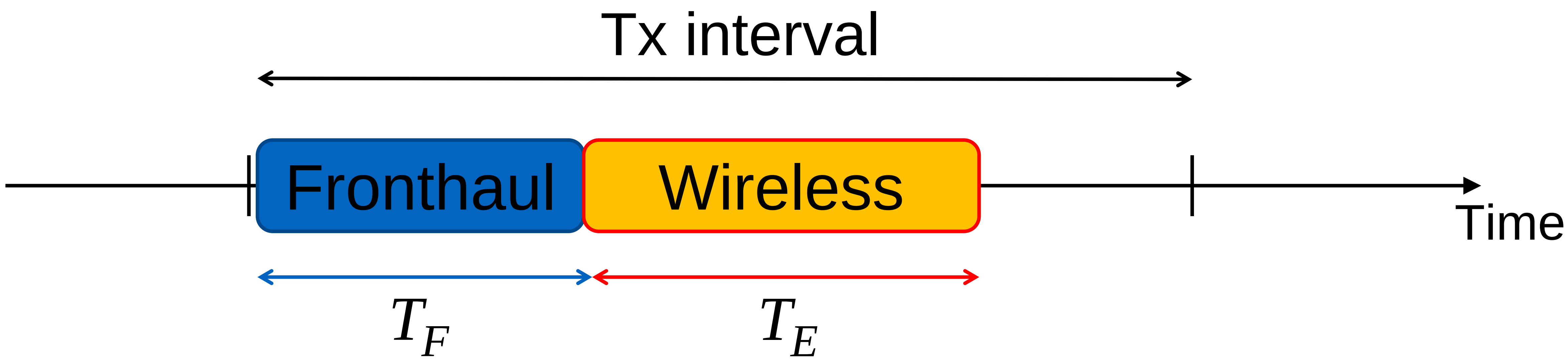
Serial Delivery…

- Serial fronthaul-edge transmission

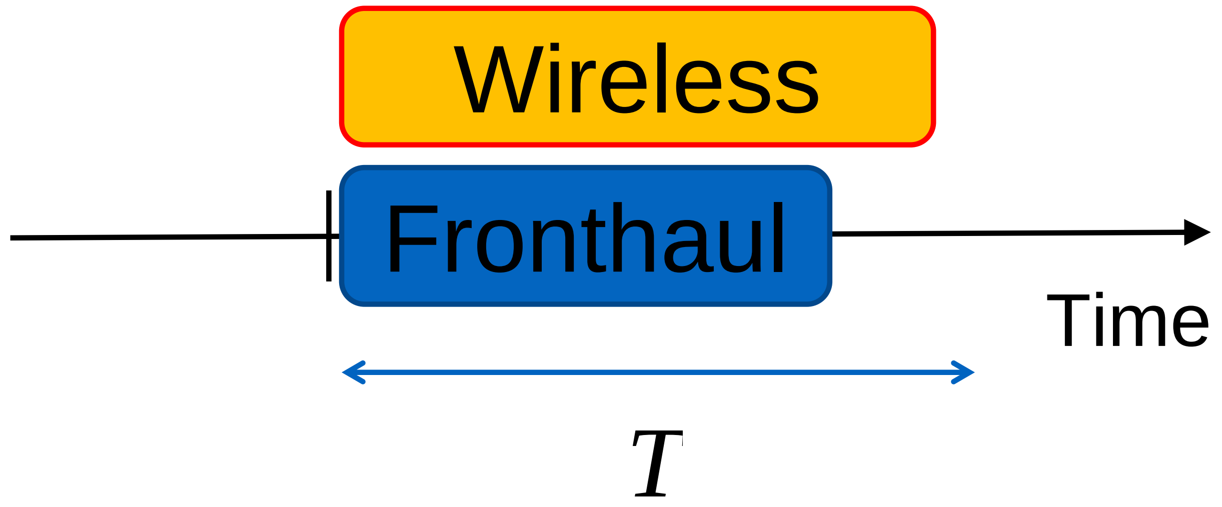


… vs Pipelined Delivery

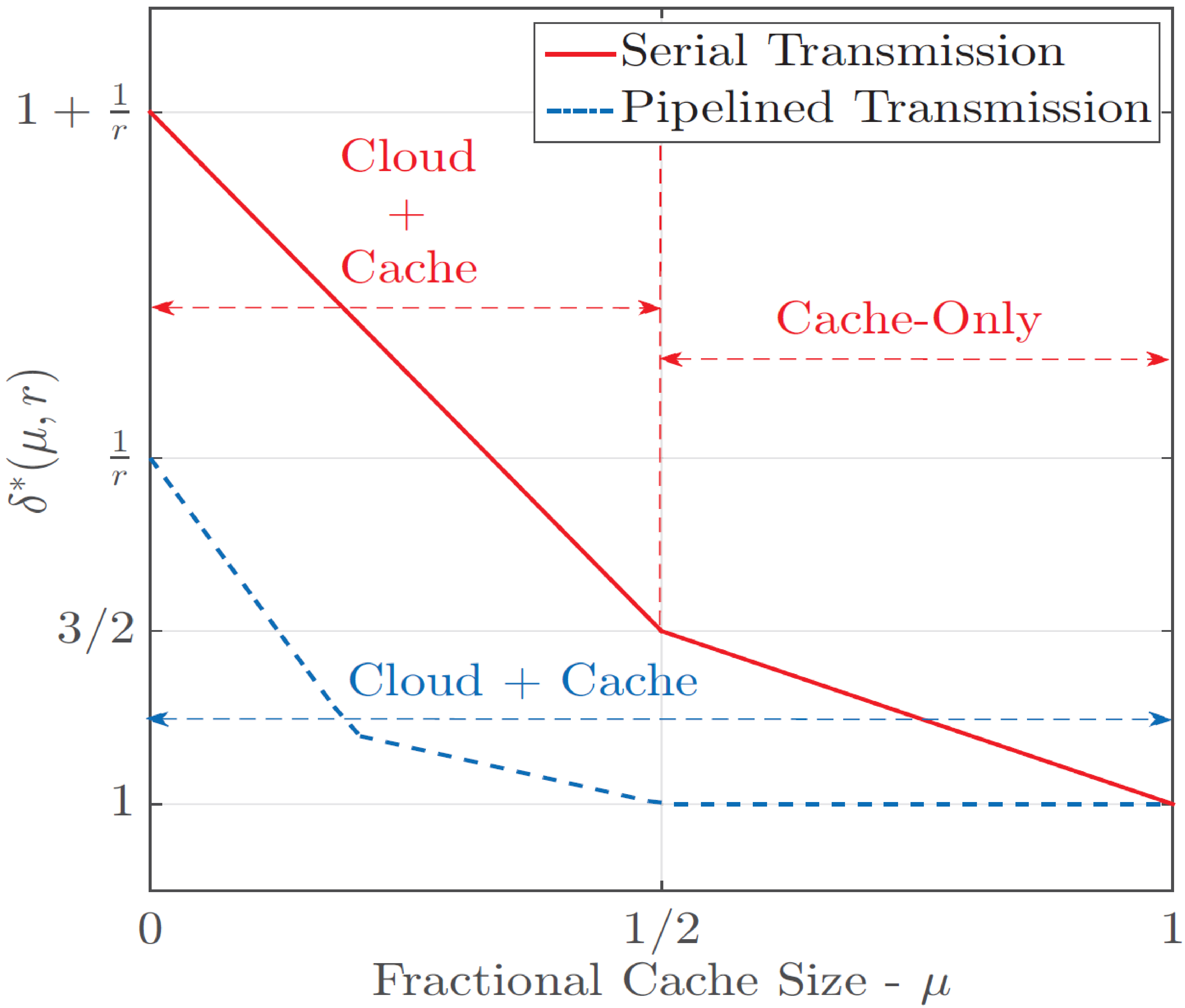
- Serial fronthaul-edge transmission



- Pipelined fronthaul-edge transmission



Minimum NDT : Pipelined Transmission

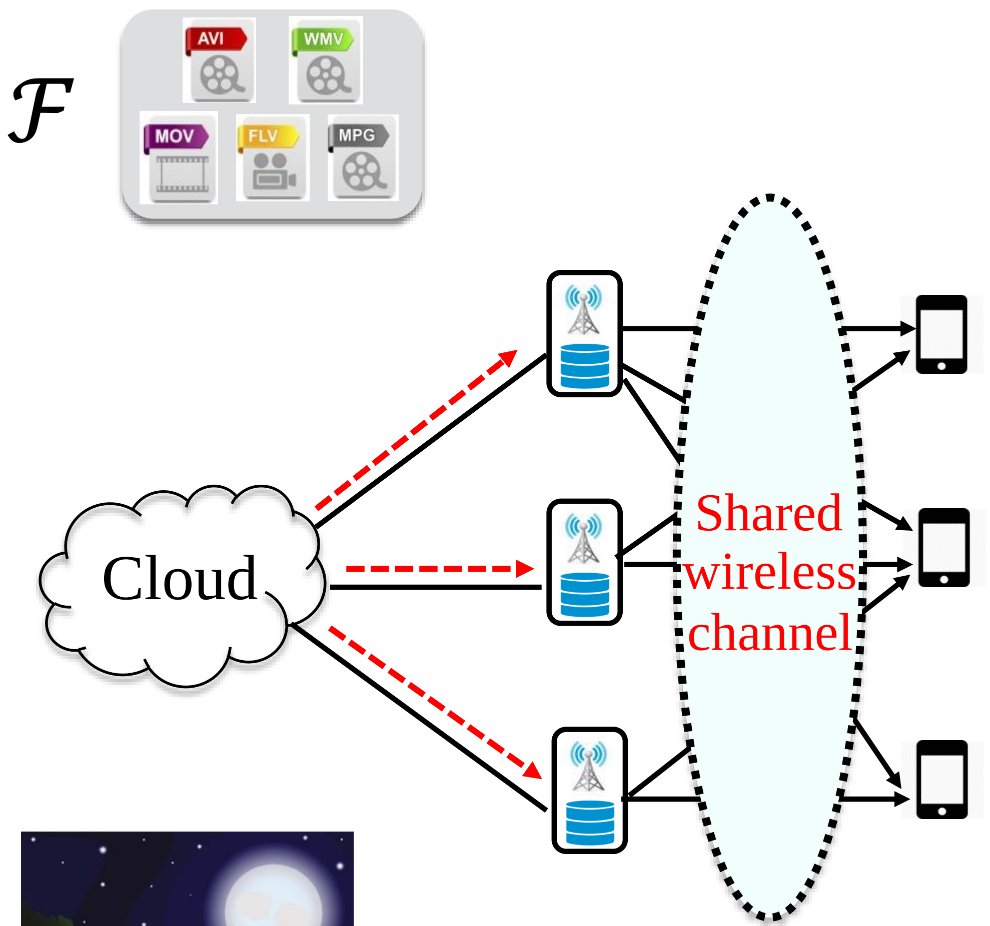


$K = 2$

Extensions:
Caching Protocol

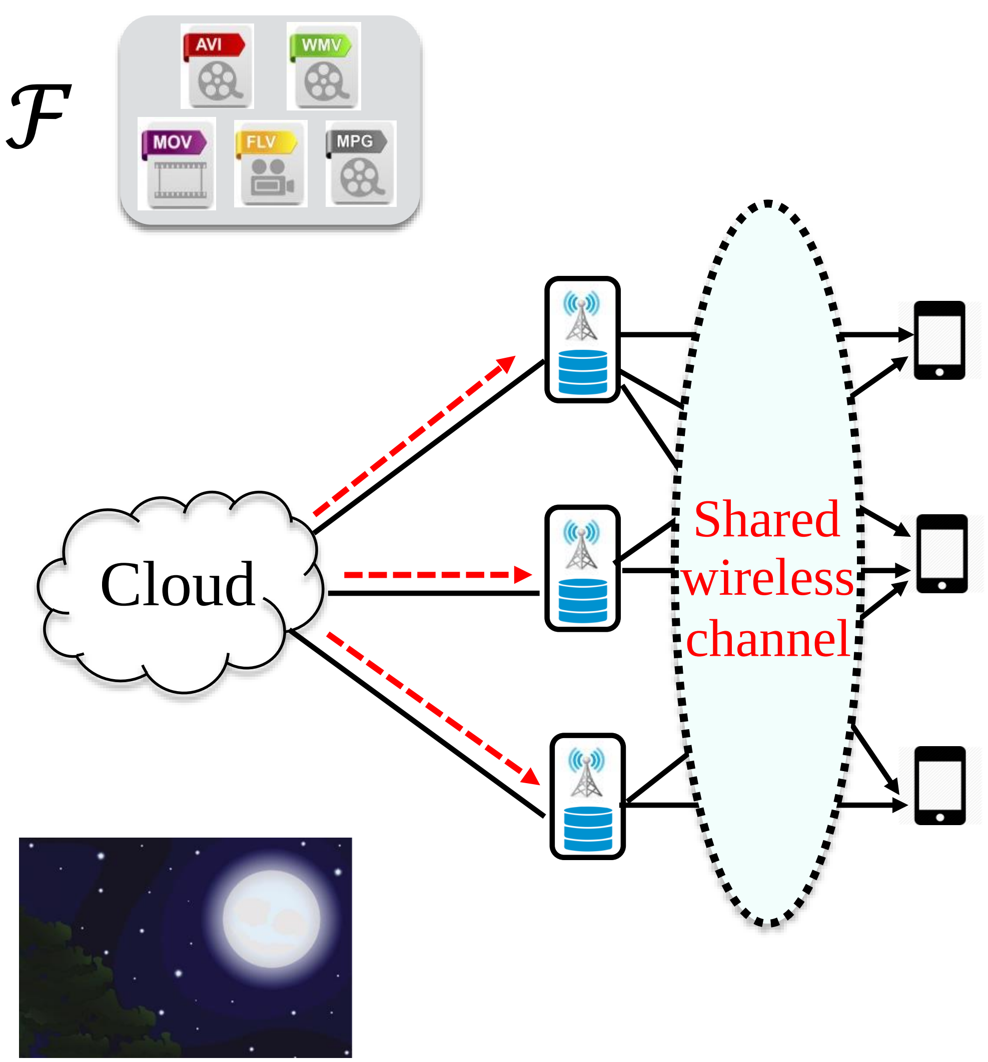
M. Azimi, O. Simeone, A. Sengupta, and R. Tandon, “Online Edge Caching and Wireless Delivery in Fog-Aided Networks with Dynamic Content Popularity”, arXiv:1711.10430

Offline Edge Caching…

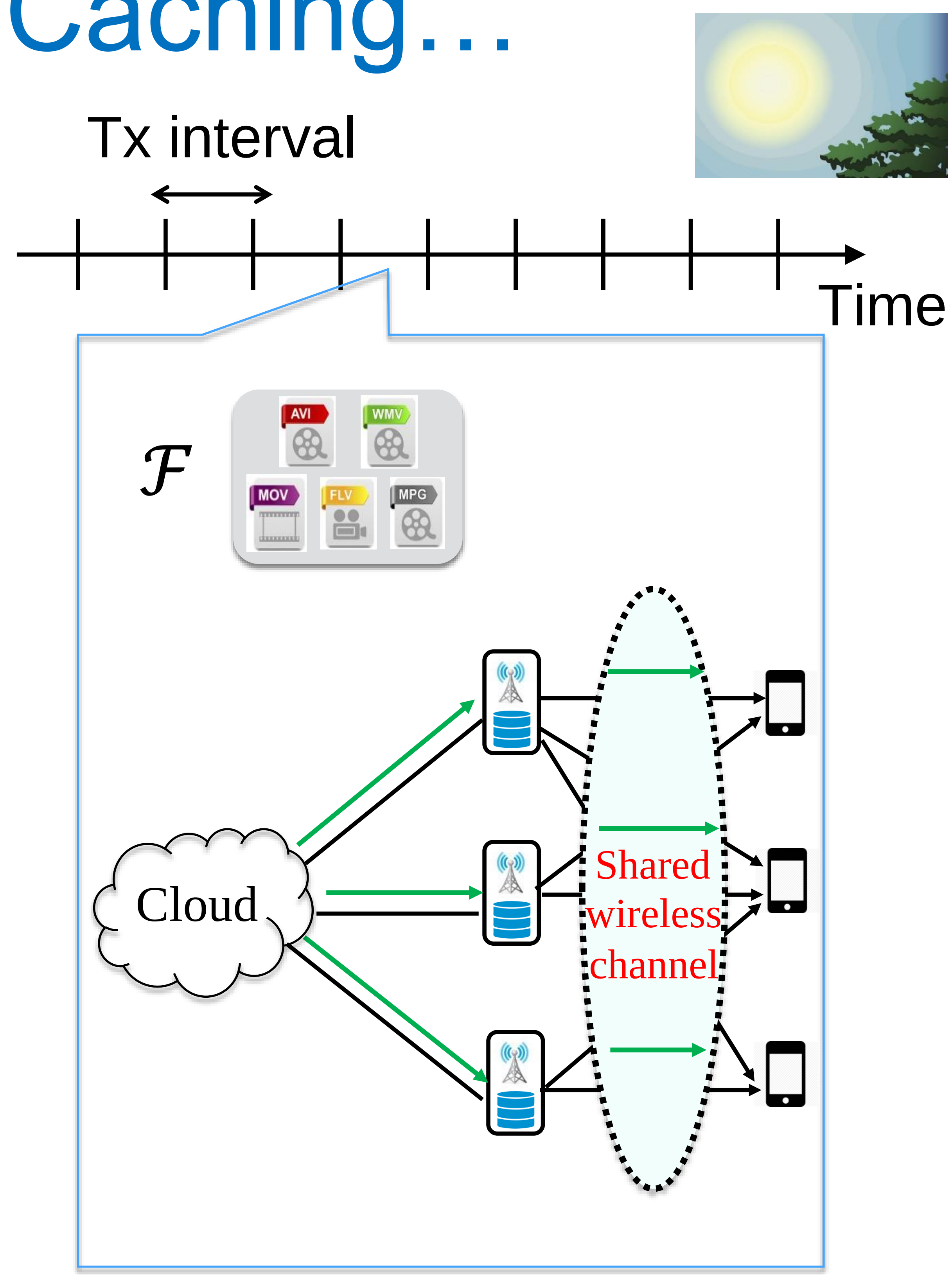


Placement phase

Offline Edge Caching…

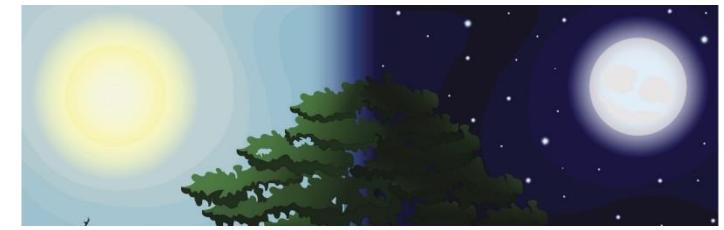


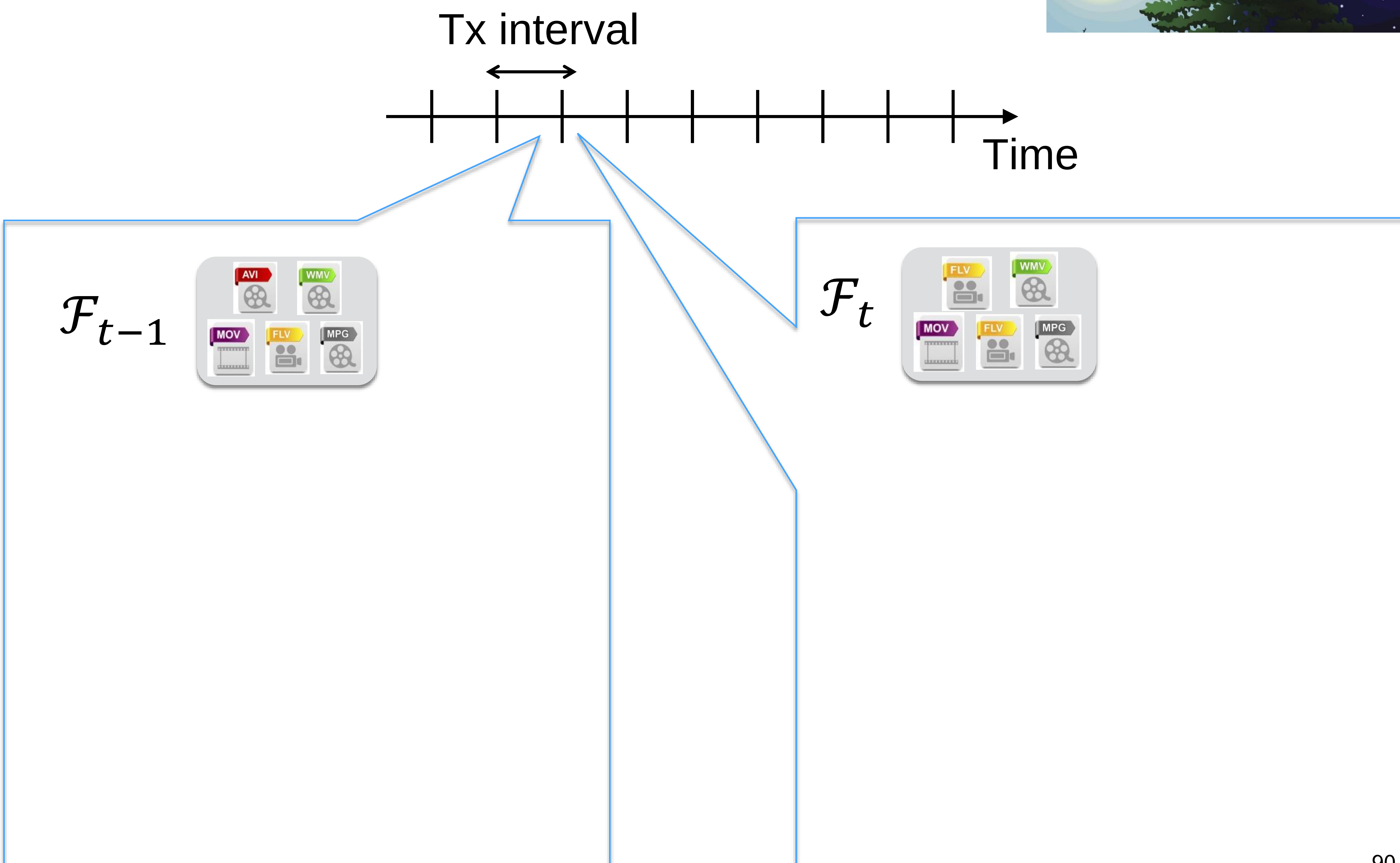
Placement phase



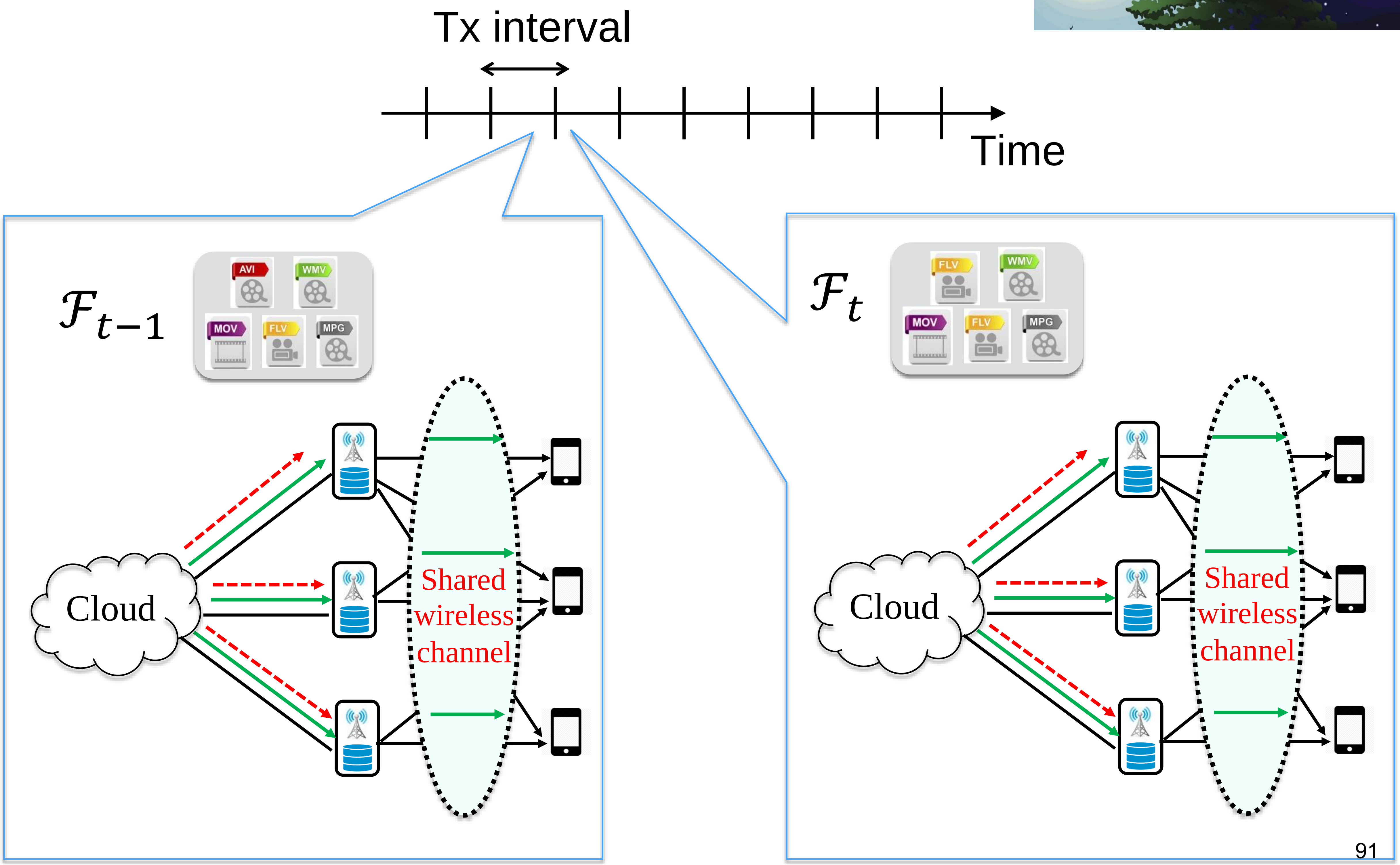
Delivery phase

… vs Online Edge Caching



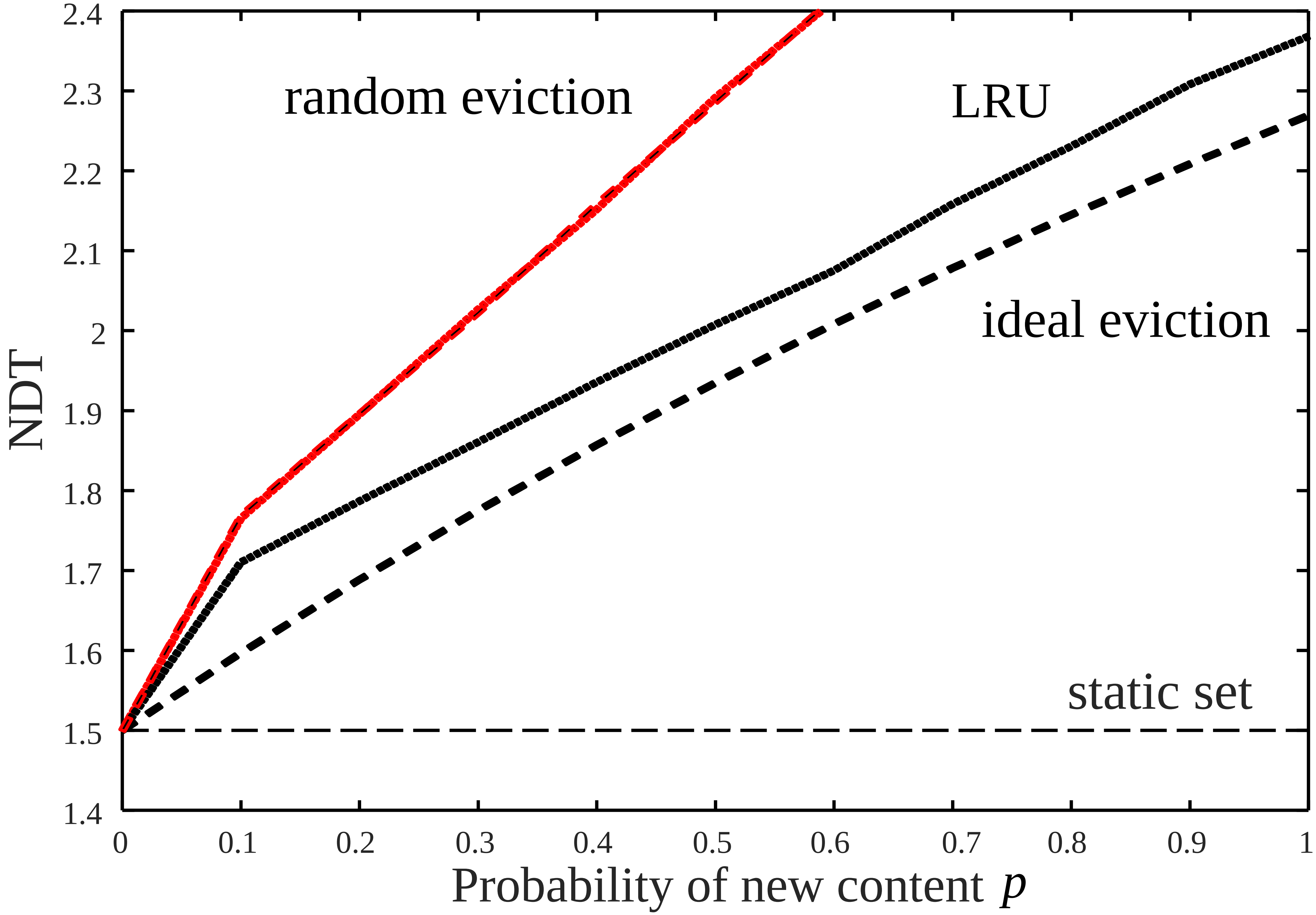


… vs Online Edge Caching



Numerical Example

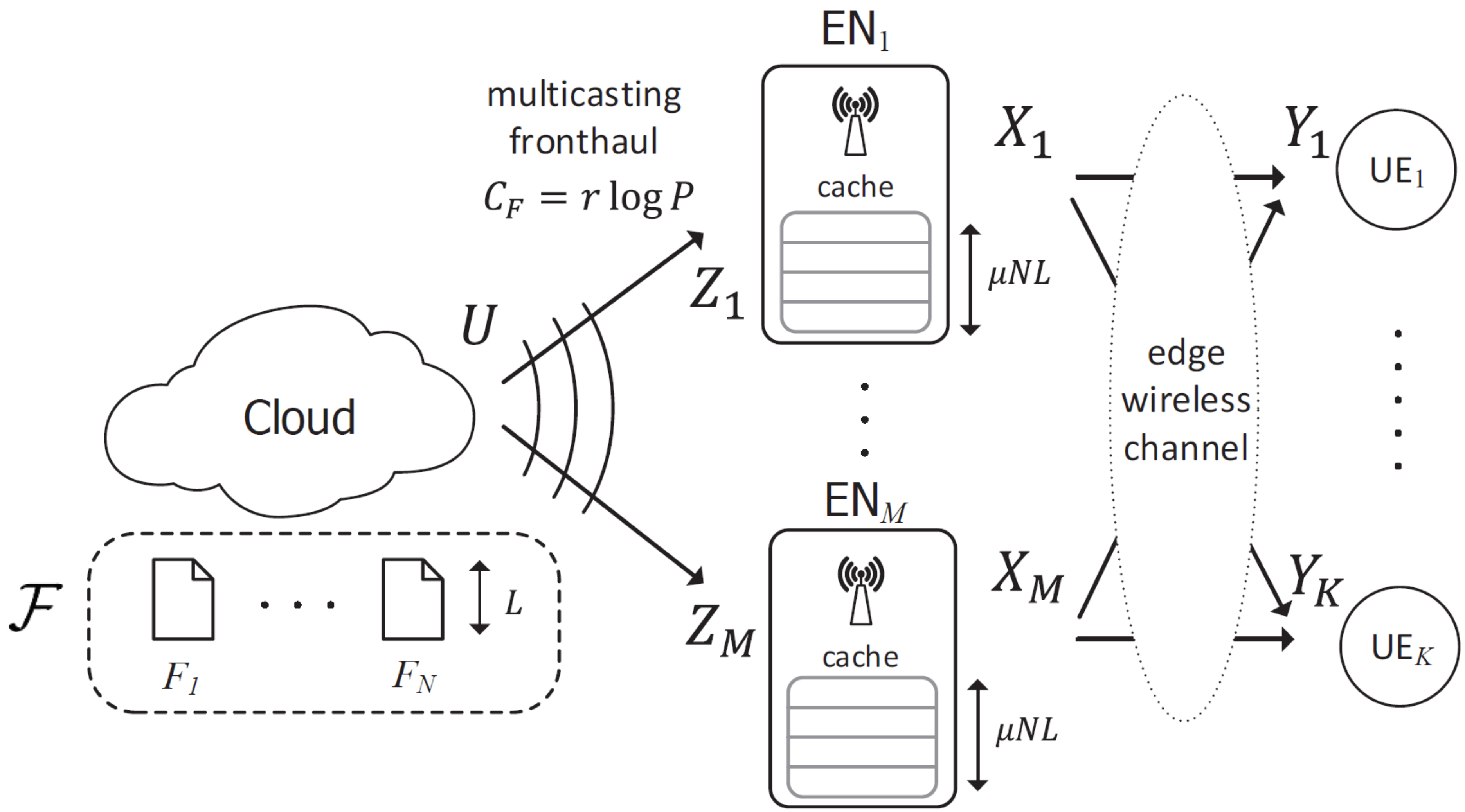
$r = 0.5, K = 5, M = 10, \mu = 0.8, N = 10$



Extensions:
Network Architecture 1

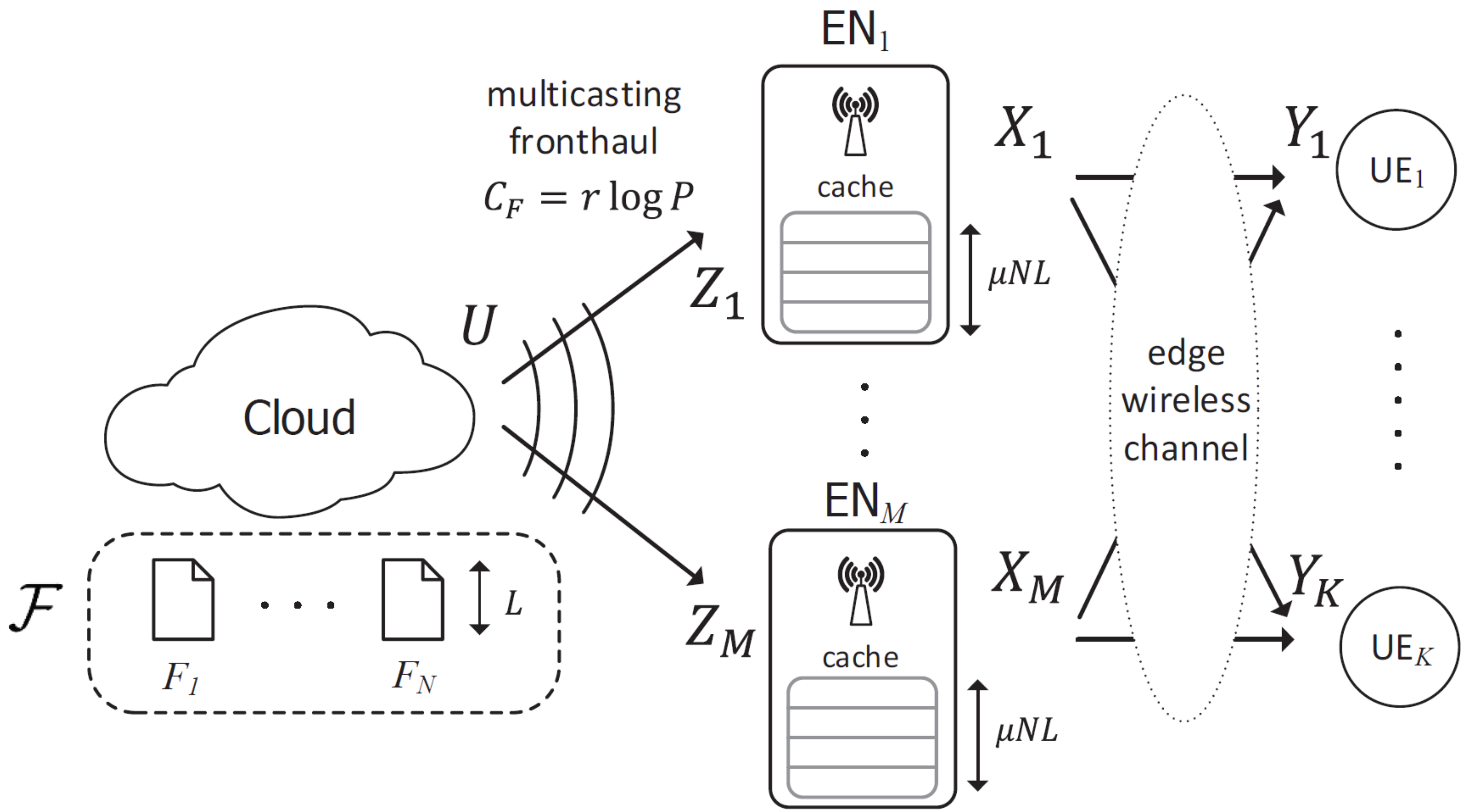
J. Koh, O. Simeone, R. Tandon, and J. Kang, “Cloud-aided edge caching with wireless multicast fronthauling in fog radio access networks,” Proc. WCNC 2017.

Multicast Fronthauling



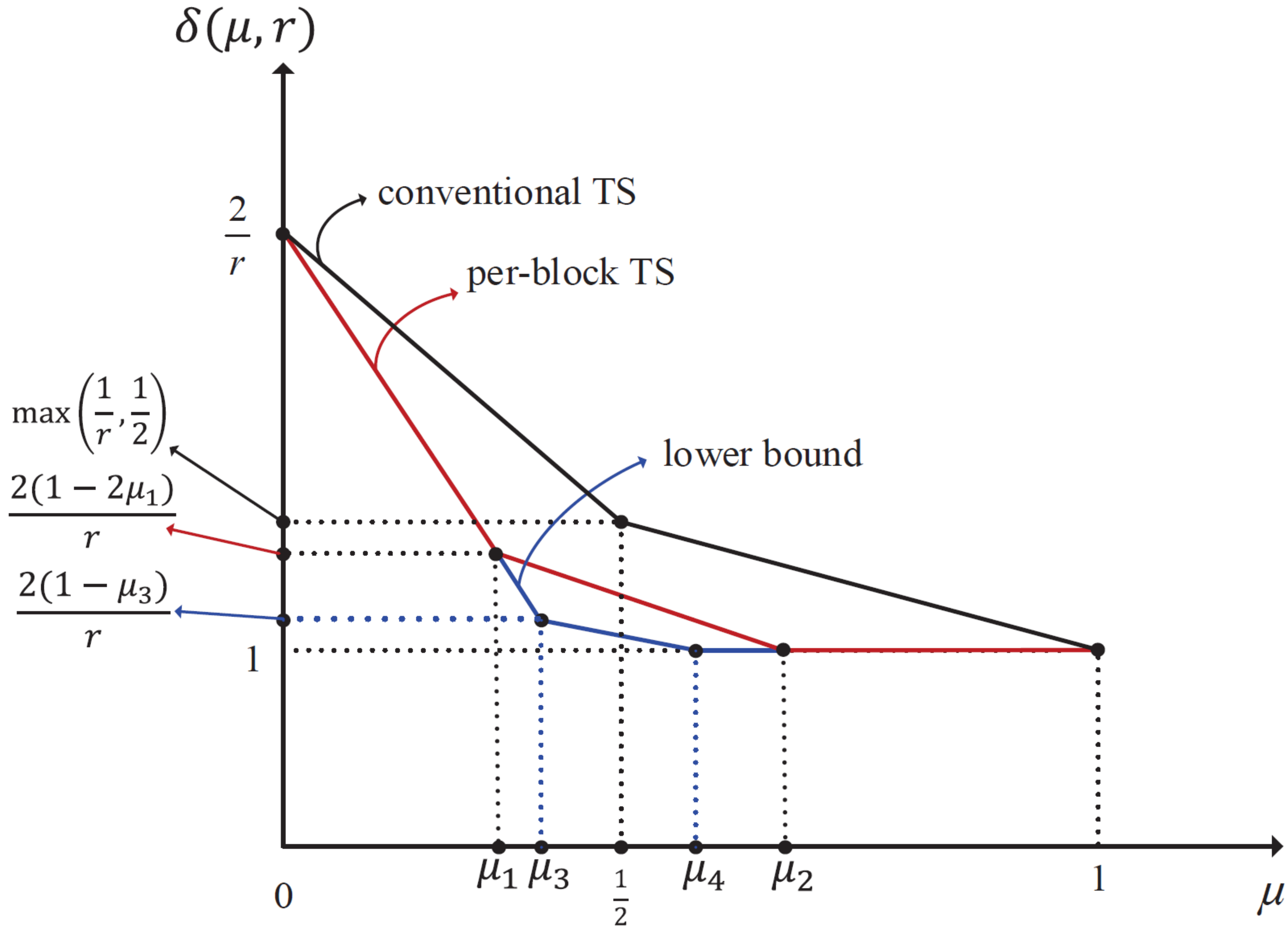
- Can coded multicasting [Maddah-Ali and Niesen '14] be useful?

Multicast Fronthauling



- Can coded multicasting [Maddah-Ali and Niesen '14] be useful?
  - ✓ Caching at the receivers of a multicast link
  - x End users are reachable through multiple ENs

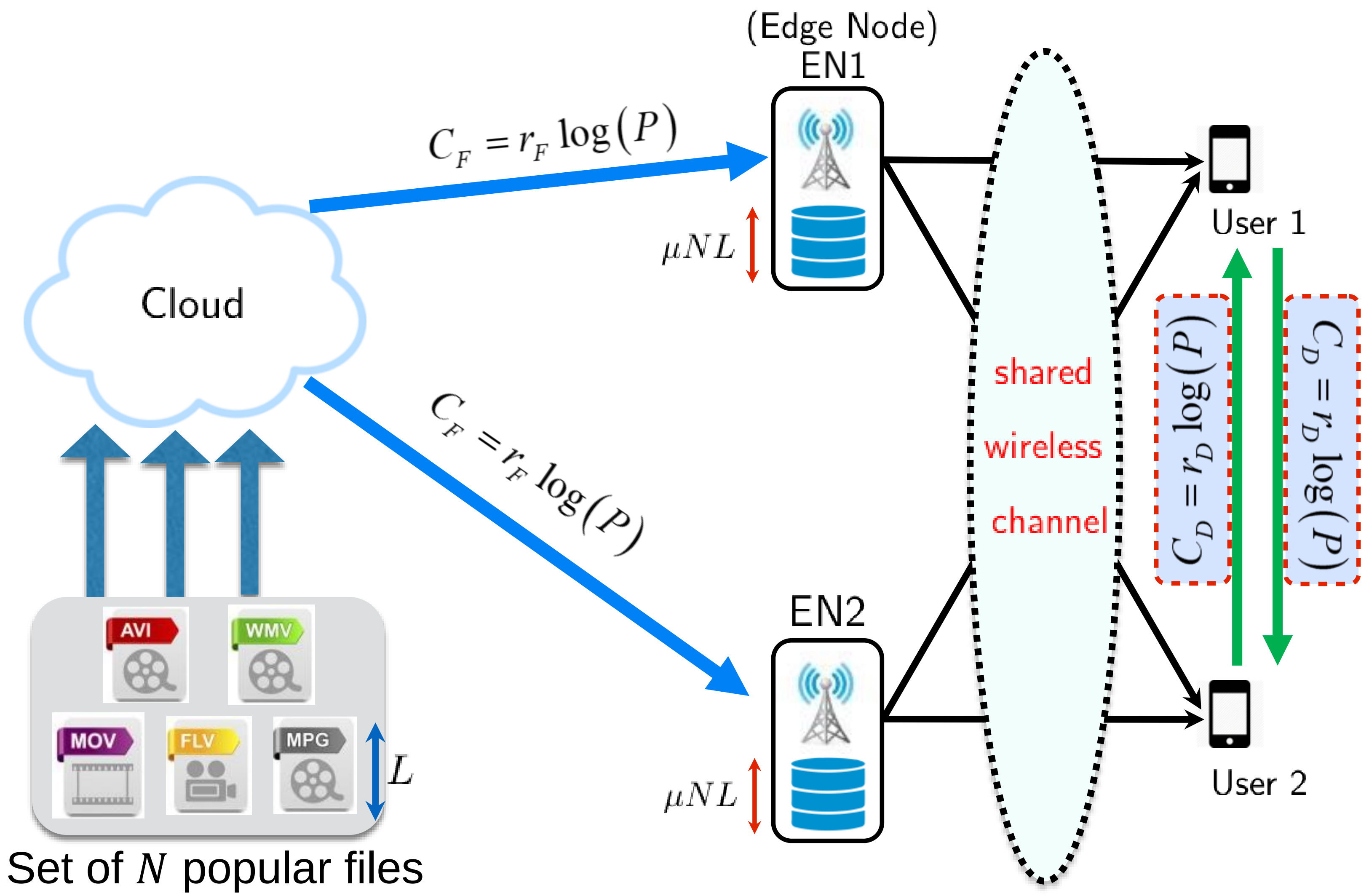
Multicast Fronthauling



Extensions:
Network Architecture 2

R. Karasik, O. Simeone, and S. Shamai (Shitz), “Fundamental Latency Limits for D2D-Aided Content Delivery in Fog Wireless Networks,” arXiv:1801.00754

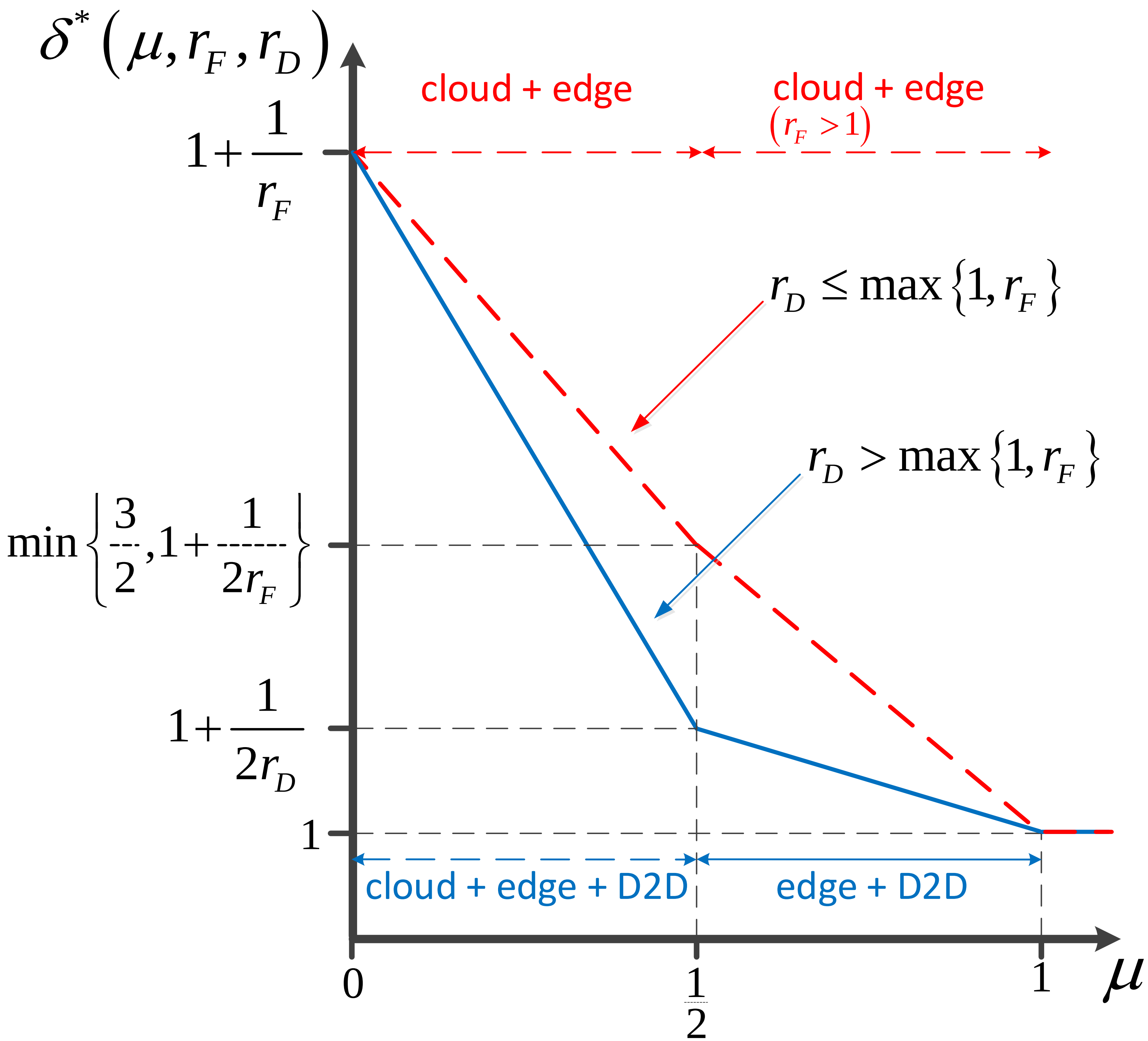
D2D-Aided F-RAN



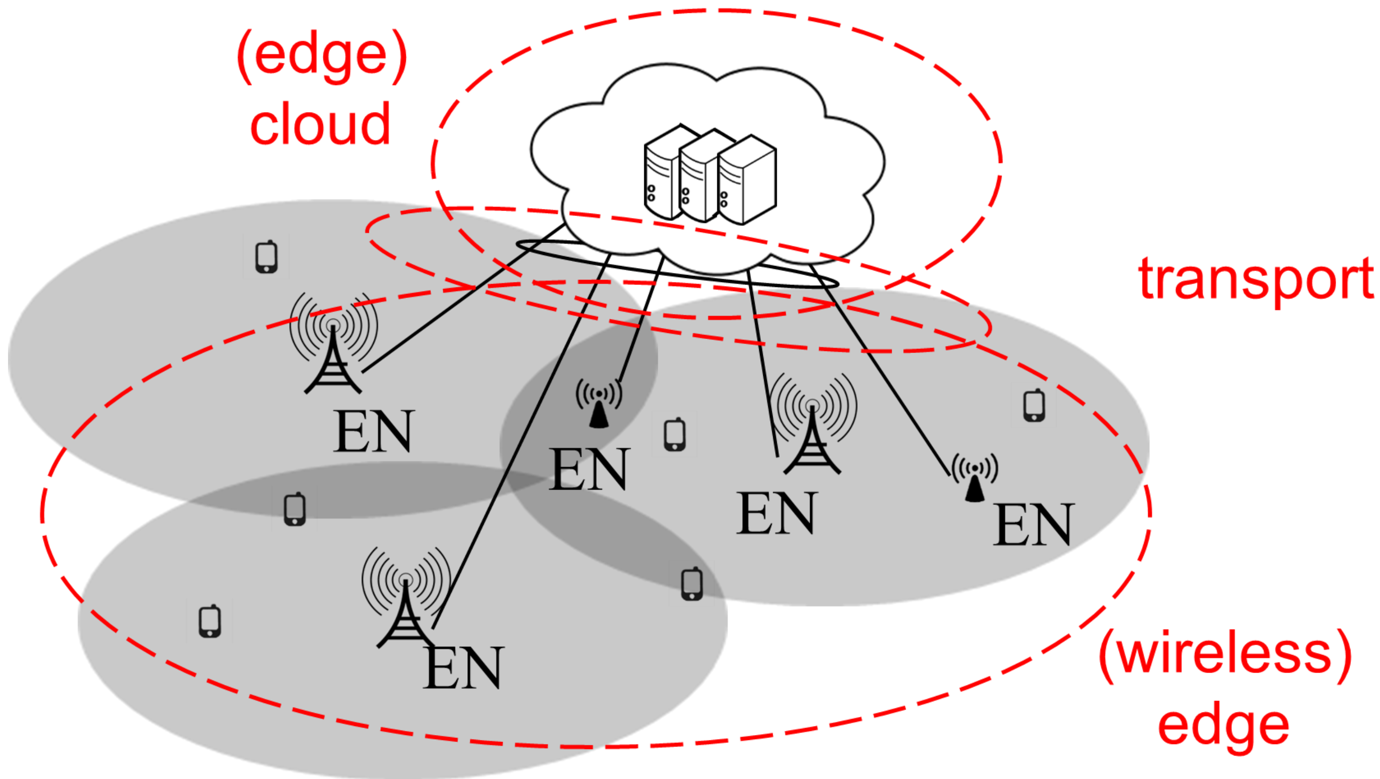
- D2D capacity parametrized as $C_D = r_D \log(P)$
- Serial transmission
- Interactive D2D conferencing

Minimum NDT

- Is D2D useful to reduce delivery latency? It depends --



Conclusions



- Cloud vs. edge processing in F-RAN systems
- Information-theoretic view
- Design insights: EN cooperation/ coordination, fronthaul compression

- Open questions: Limited connectivity? Finite-SNR performance? Full fog architecture? Other services?

101

Extensions:
Operating Regimes

S.-H. Park, O. Simeone and S. Shamai, “Joint Optimization of Cloud and Edge Processing for Fog Radio Access Neworks," IEEE Trans. Wireless Commun., 2016.

Operating Regimes

- Infinite SNR

  - Focus on the effect of interference

- Infinite file length *L*

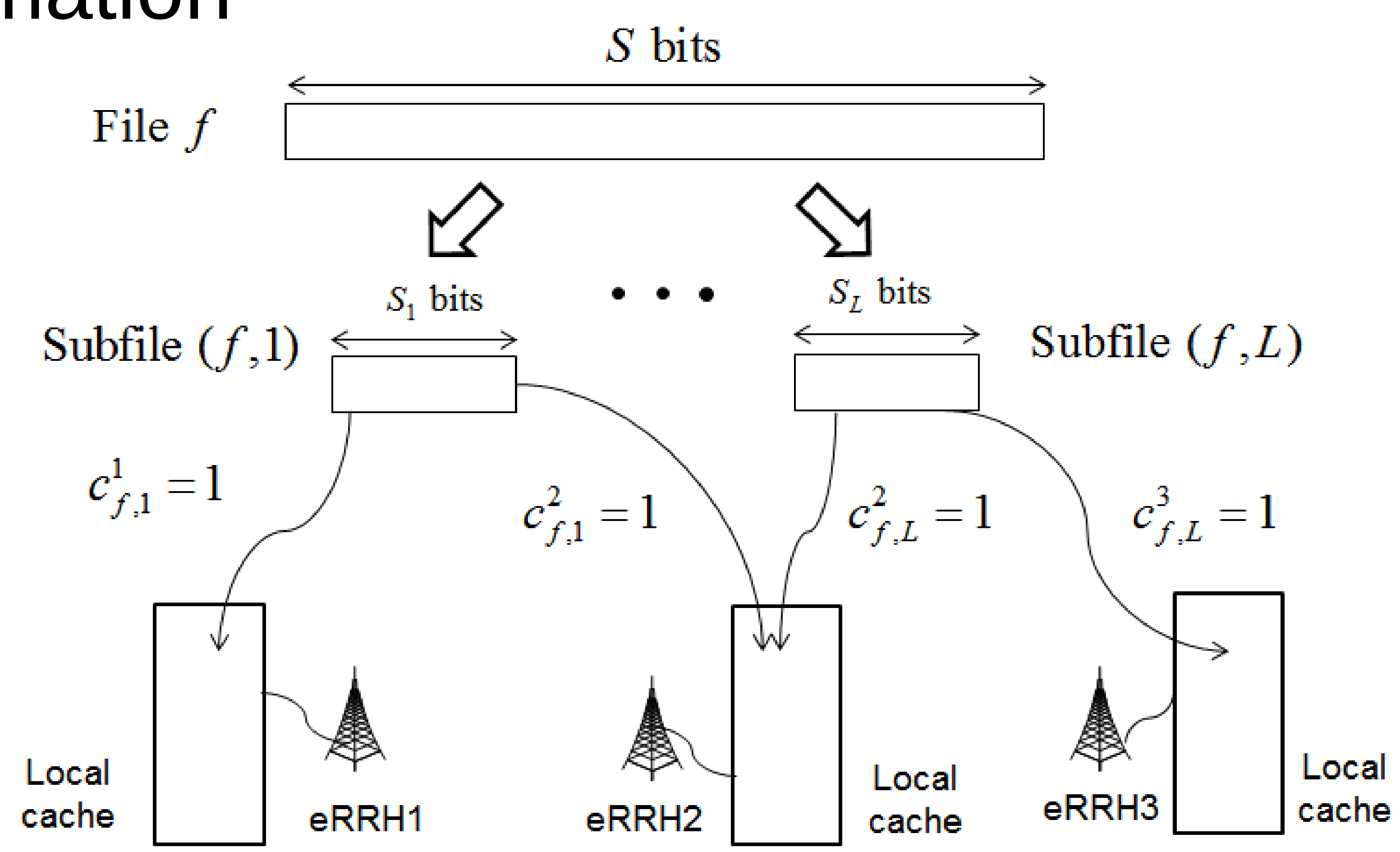
  - Neglect finite-blocklength penalties

- Is the fronthaul compression still optimal when backing off from these asymptotic regimes?

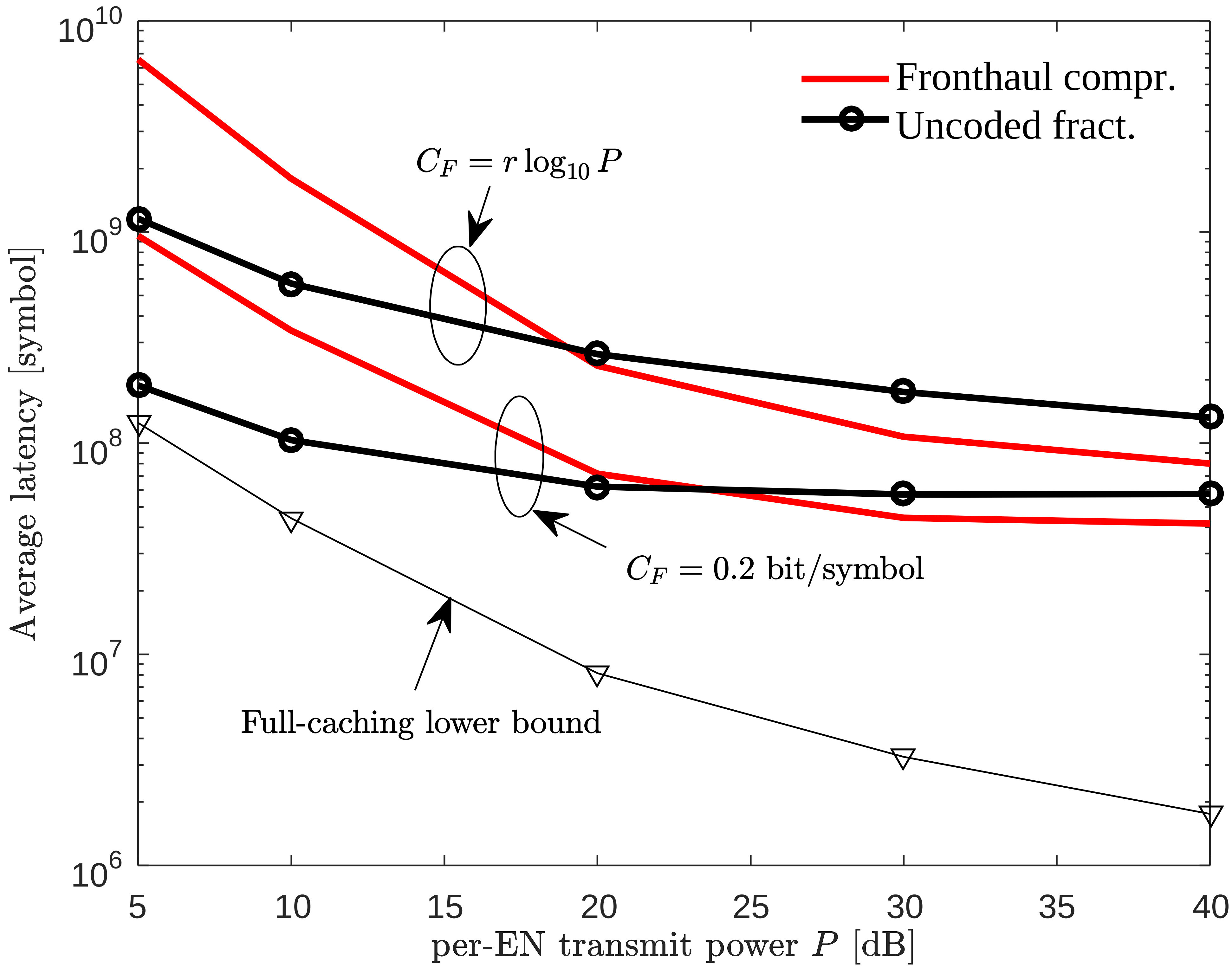
Operating Regimes

- Rayleigh fading, random ENs' and users' placements and fixed caching (randomized fractional caching)

- Linear precoding optimized using Successive Convex Approximation



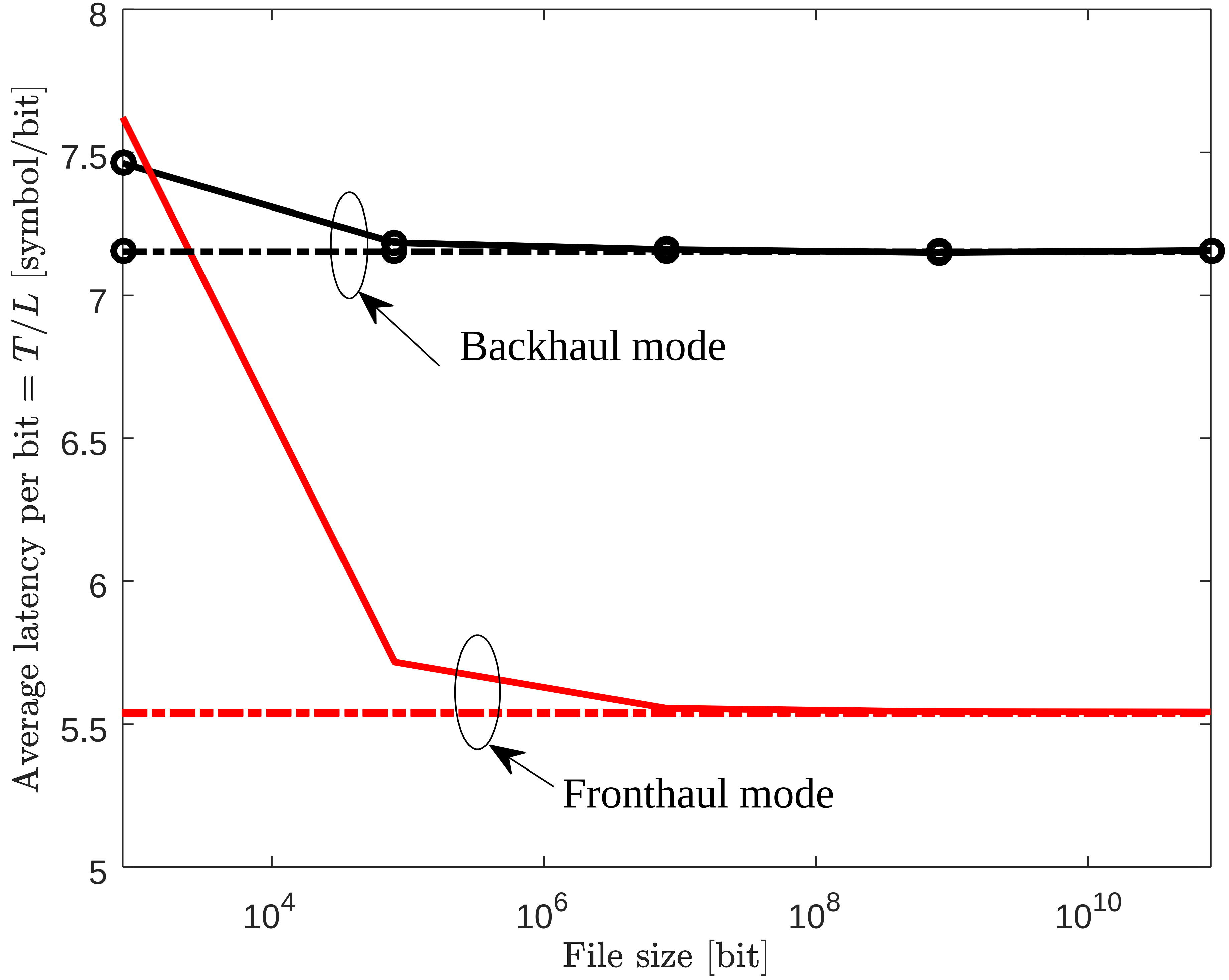
Finite SNR



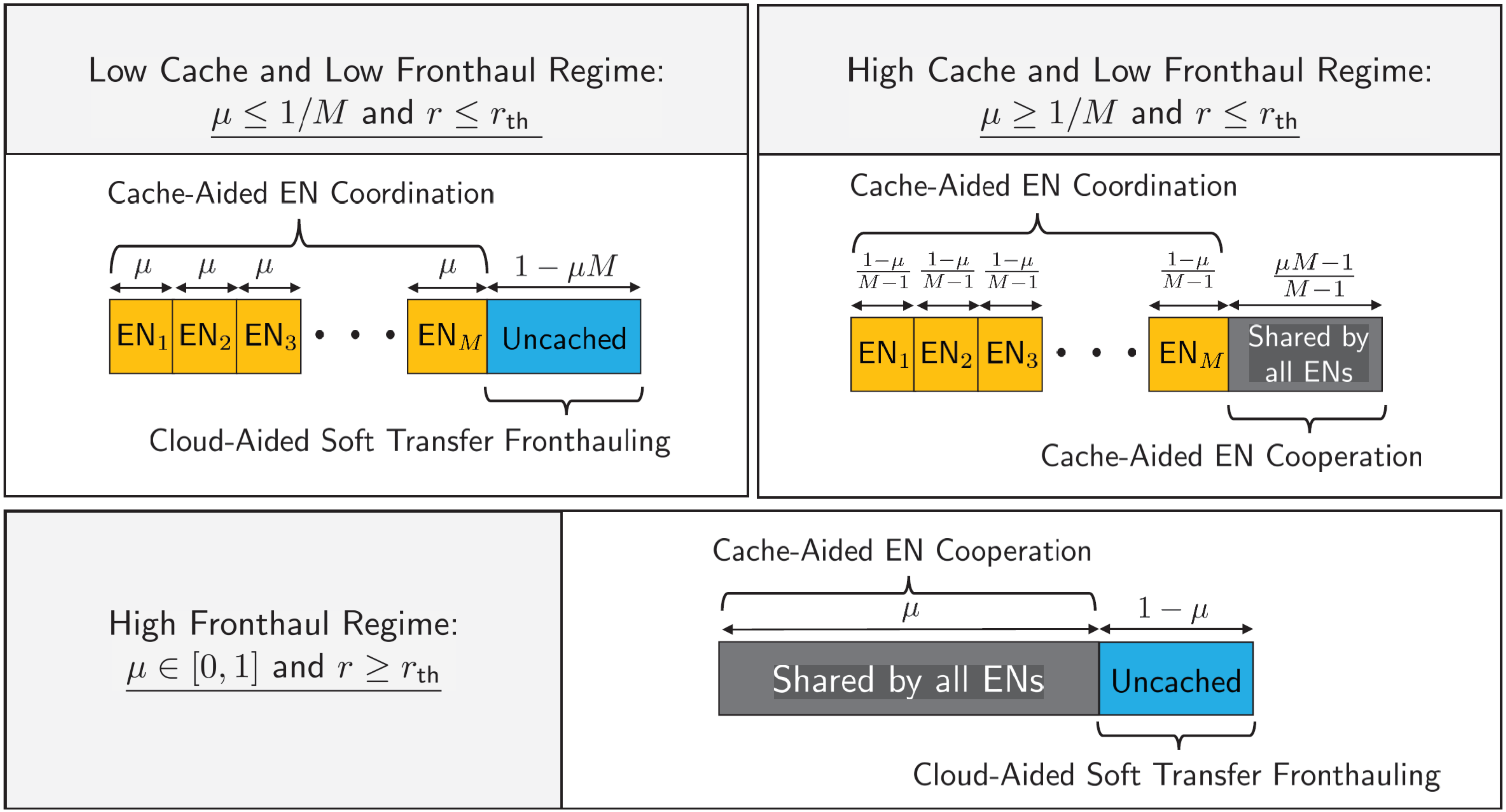
Fronthaul compr.
Uncoded fract.
$C_F = r\log_{10} P$
$C_F = 0.2$ bit/symbol
Full-caching lower bound
Average latency [symbol]
per-EN transmit power P [dB]
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
5
10
15
20
25
30
35
40

Finite Blocklength

- Gaussian approximation [Polyanskiy et al ‘10] treating interference as in [Scarlett et al ’17]

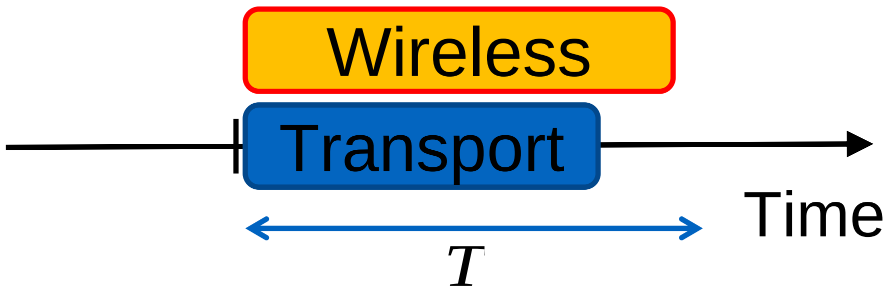


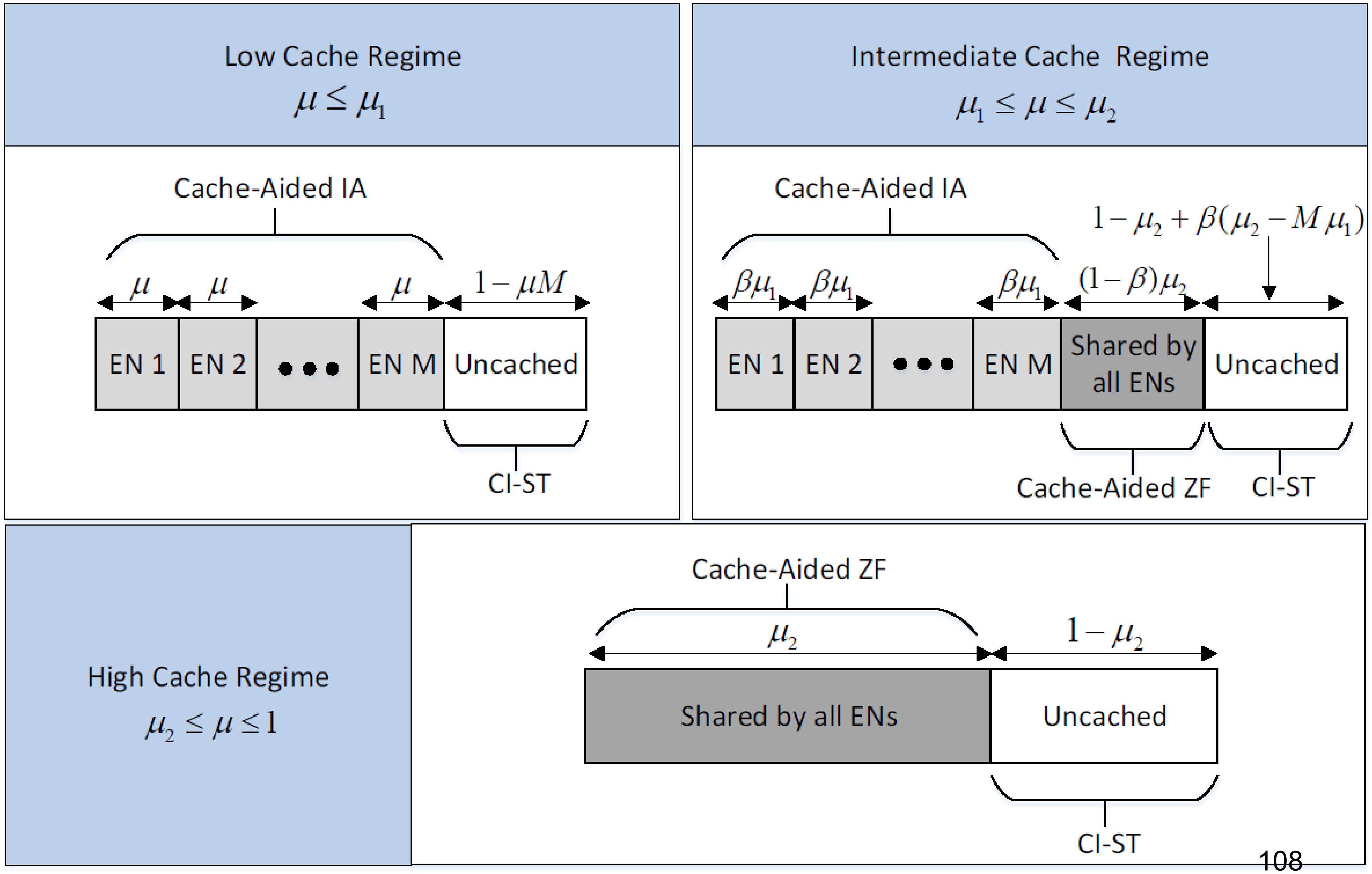
General Caching and Delivery Policy



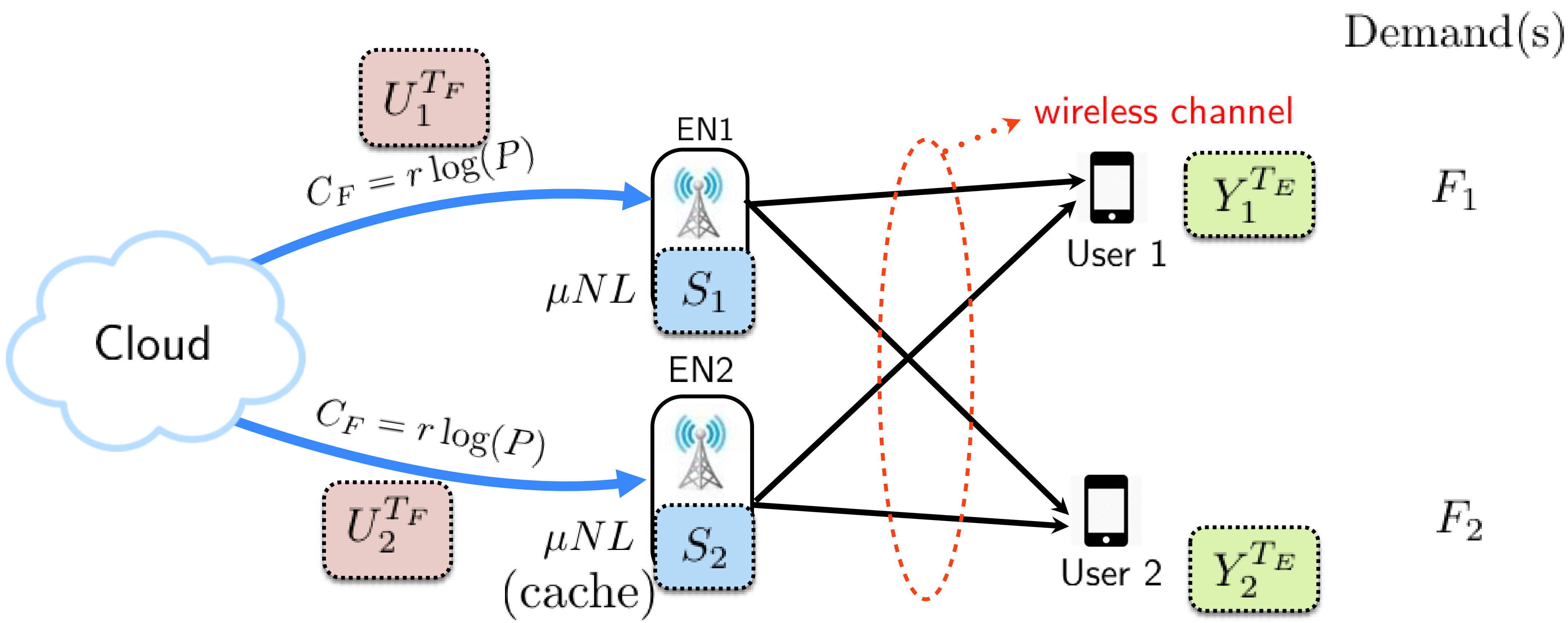
General Caching and Delivery Policy

- Pipelined delivery model [Sengupta et al ’17]

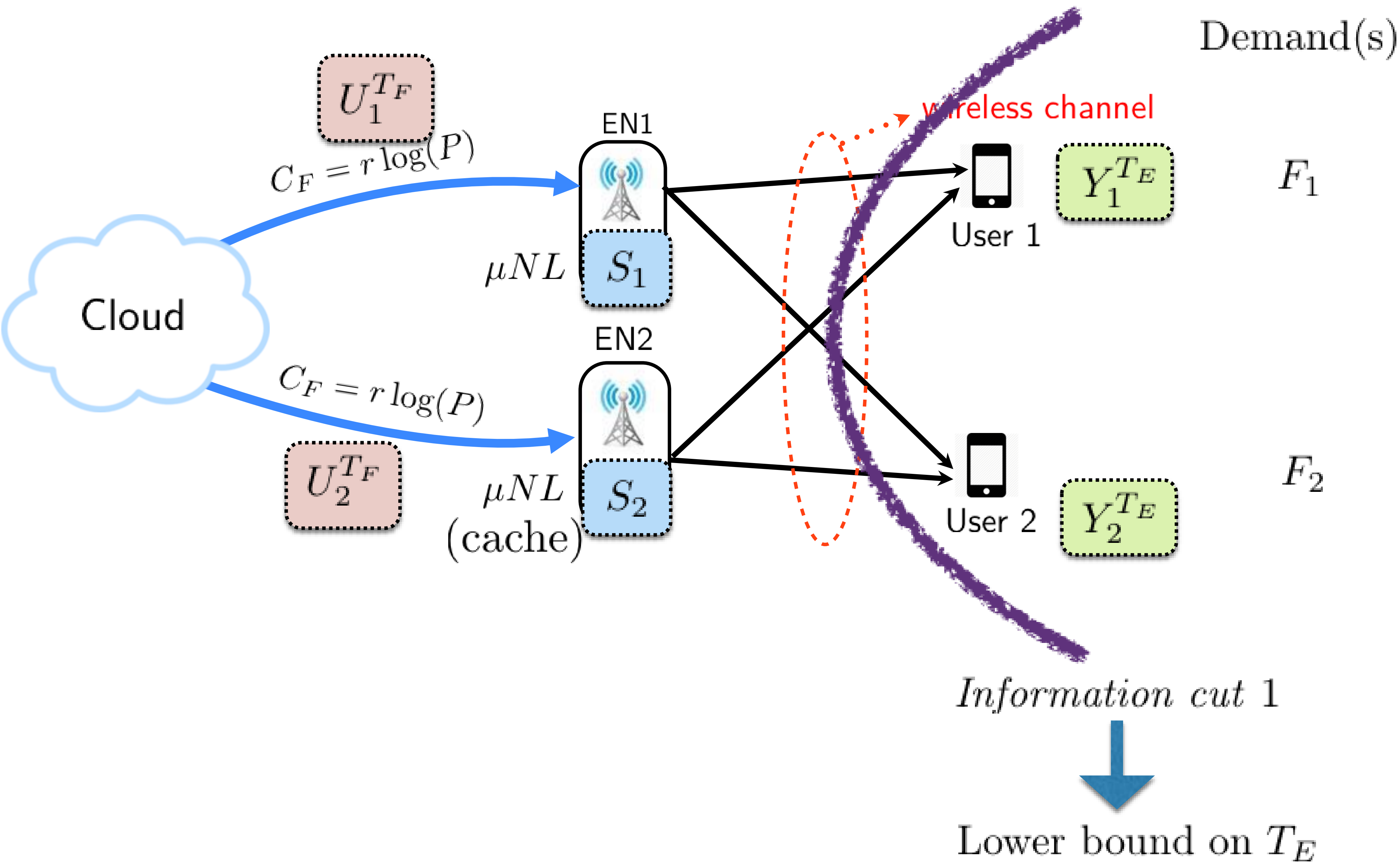




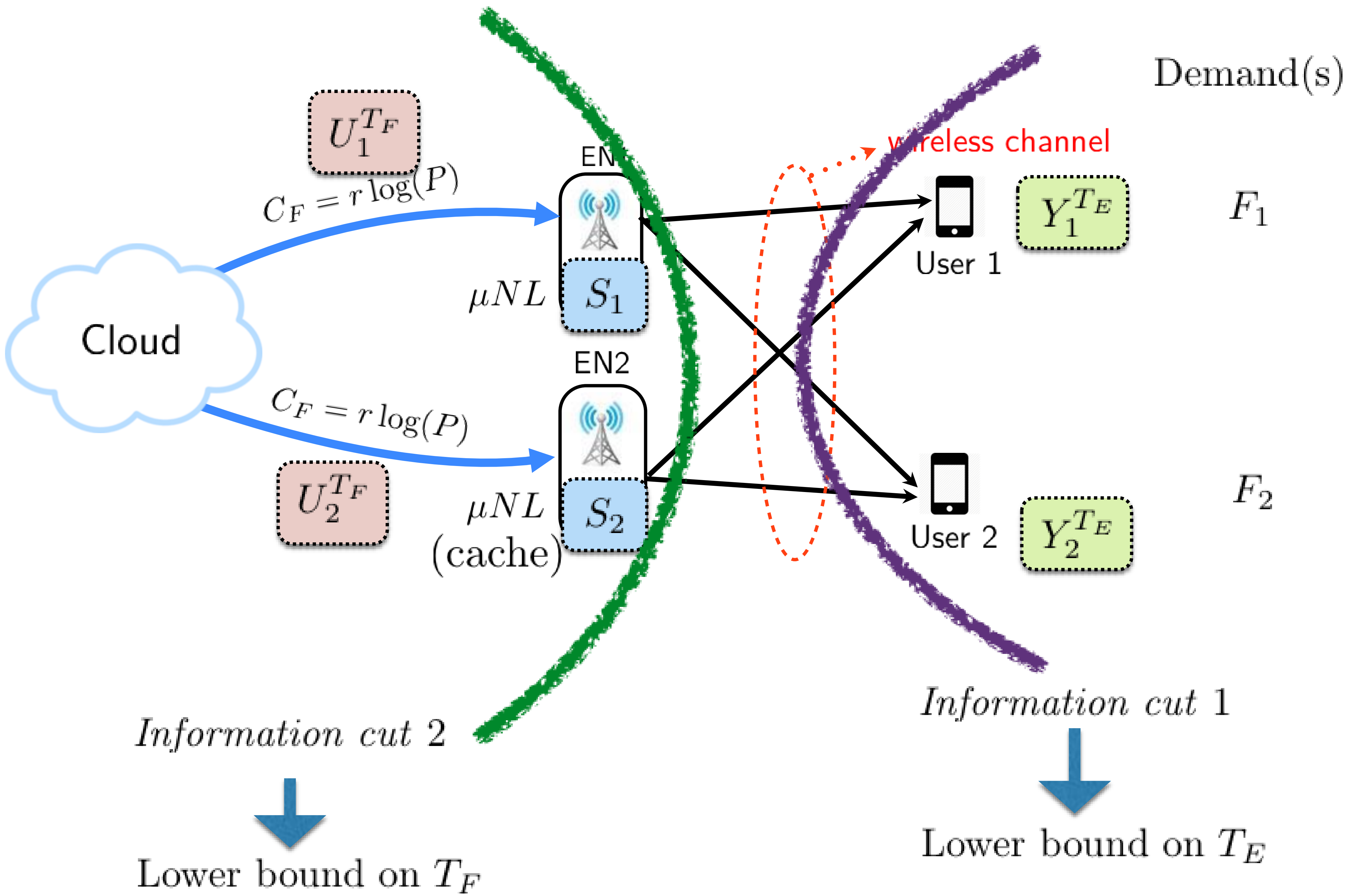
Minimum NDT: Serial Transmission



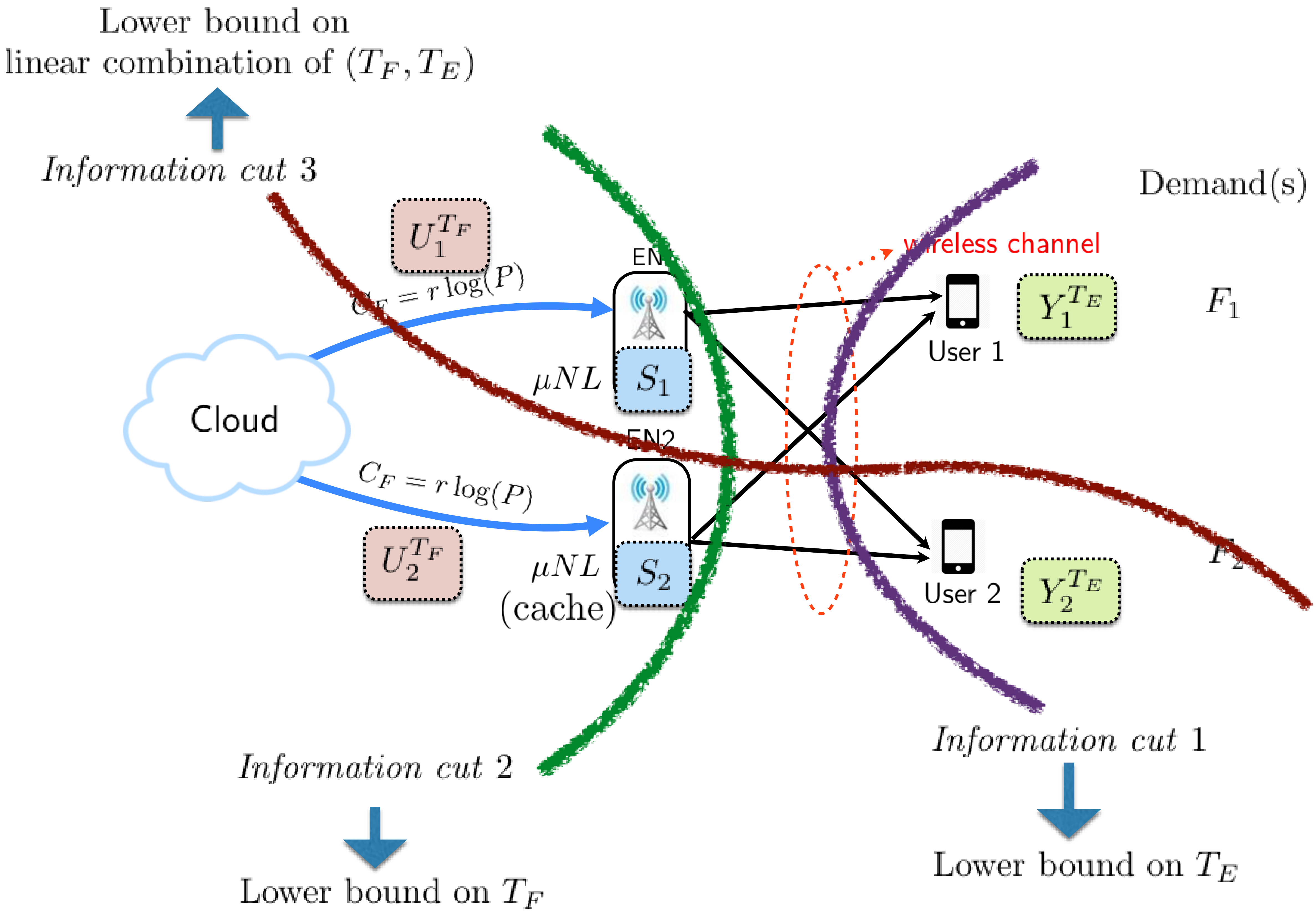
Minimum NDT: Serial Transmission



Minimum NDT: Serial Transmission



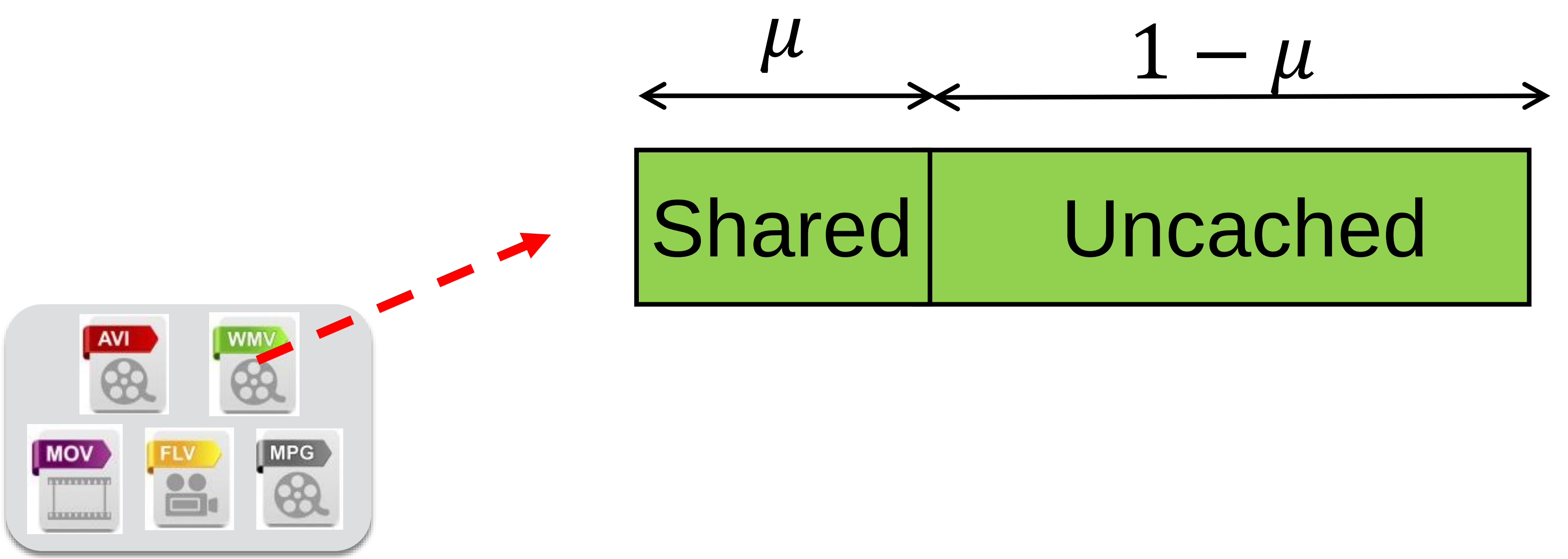
Minimum NDT: Serial Transmission



Minimum NDT: Serial Transmission

High transport capacity

$r \geq r_{th}$

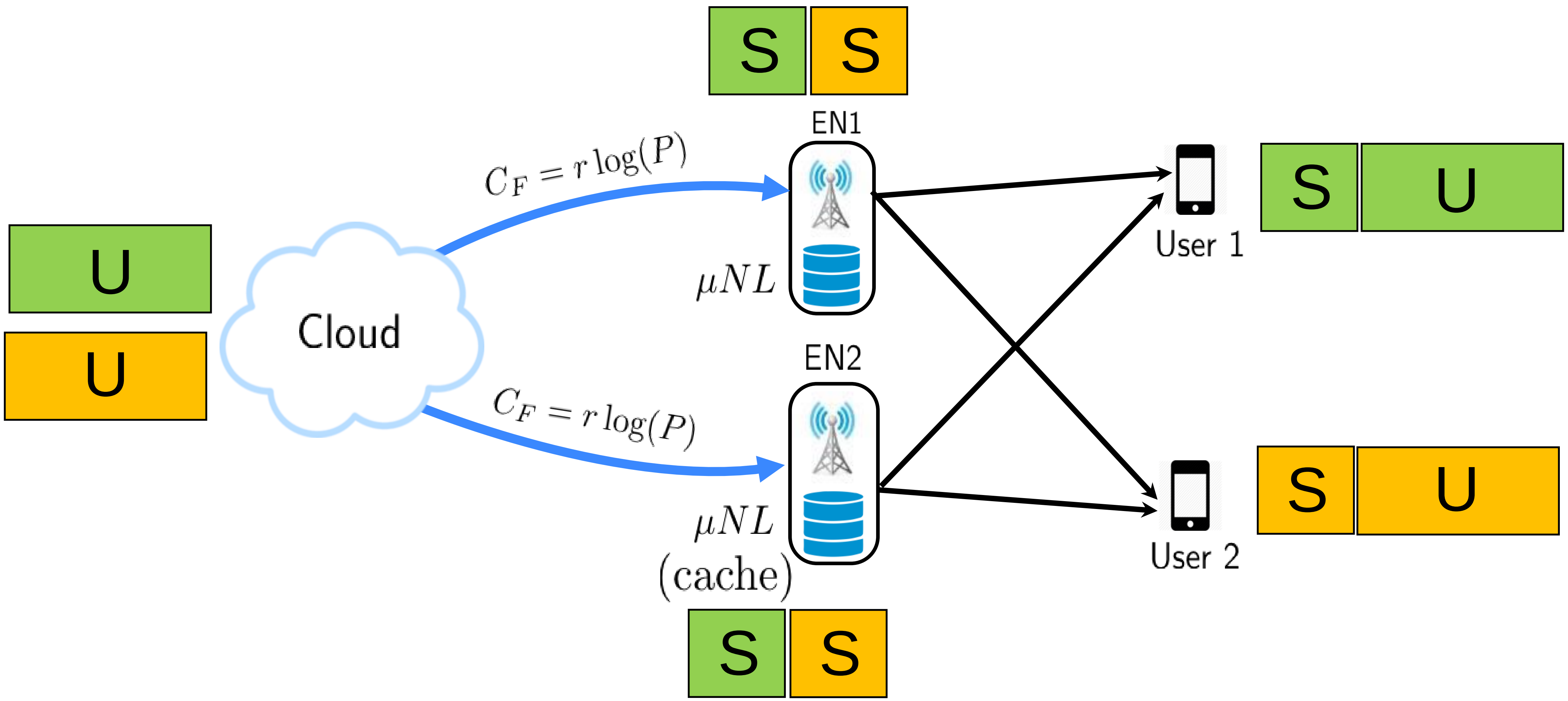


Placement phase

Minimum NDT: Serial Transmission

High transport capacity

$$r \geq r_{th}$$

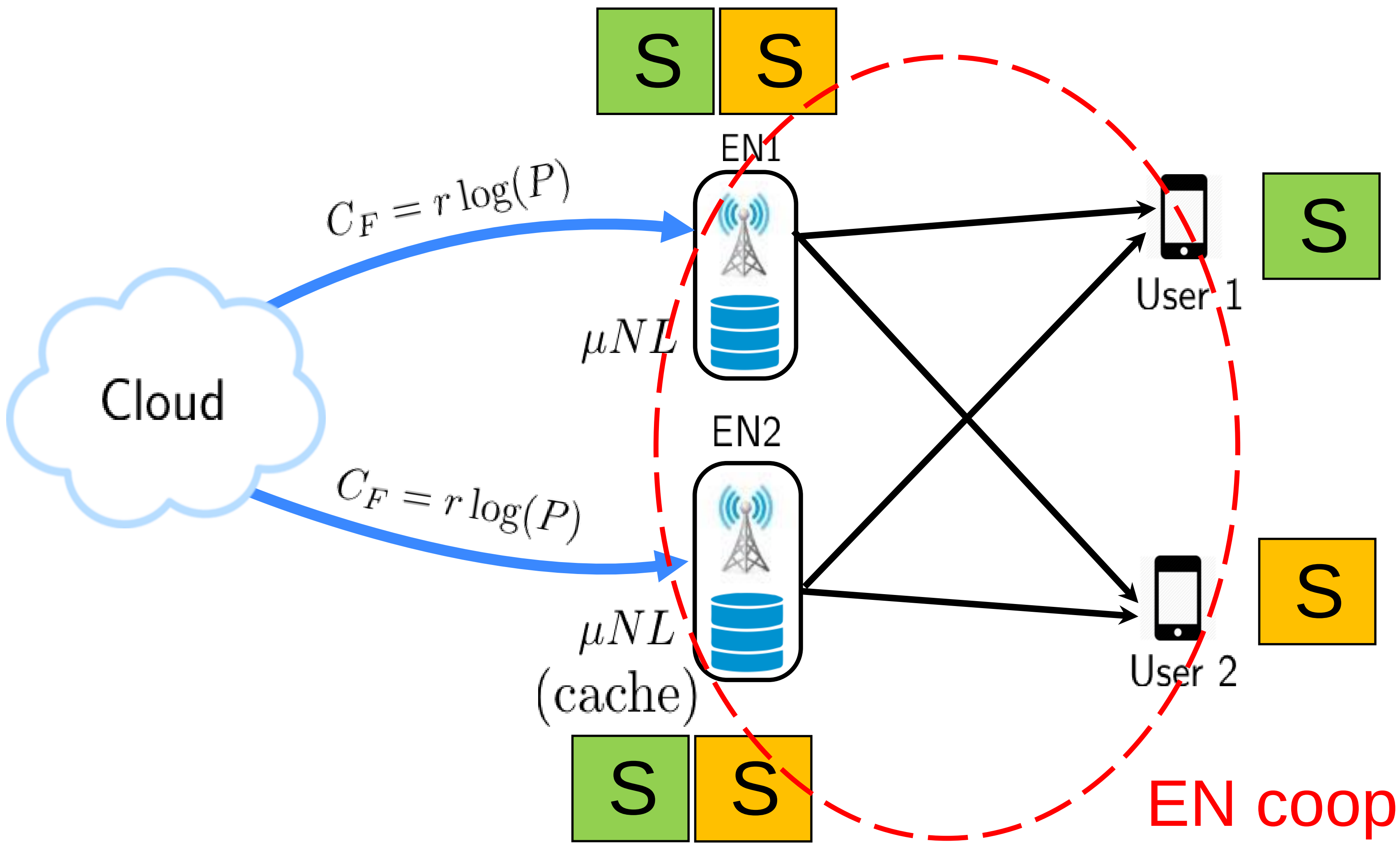


Delivery phase

Minimum NDT: Serial Transmission

High transport capacity

$r \geq r_{th}$



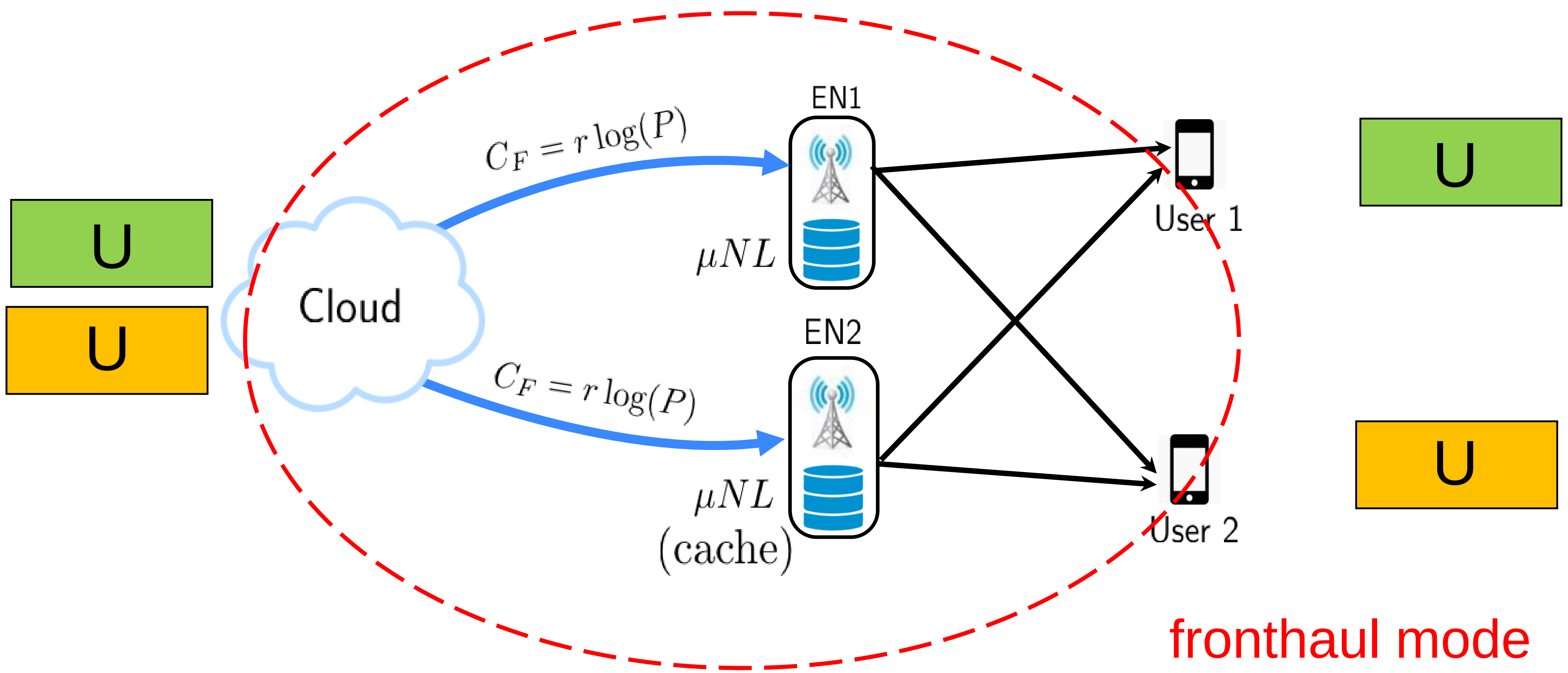
EN cooperation
(ZF precoding)

Delivery phase

Minimum NDT: Serial Transmission

High transport capacity

$r \geq r_{th}$



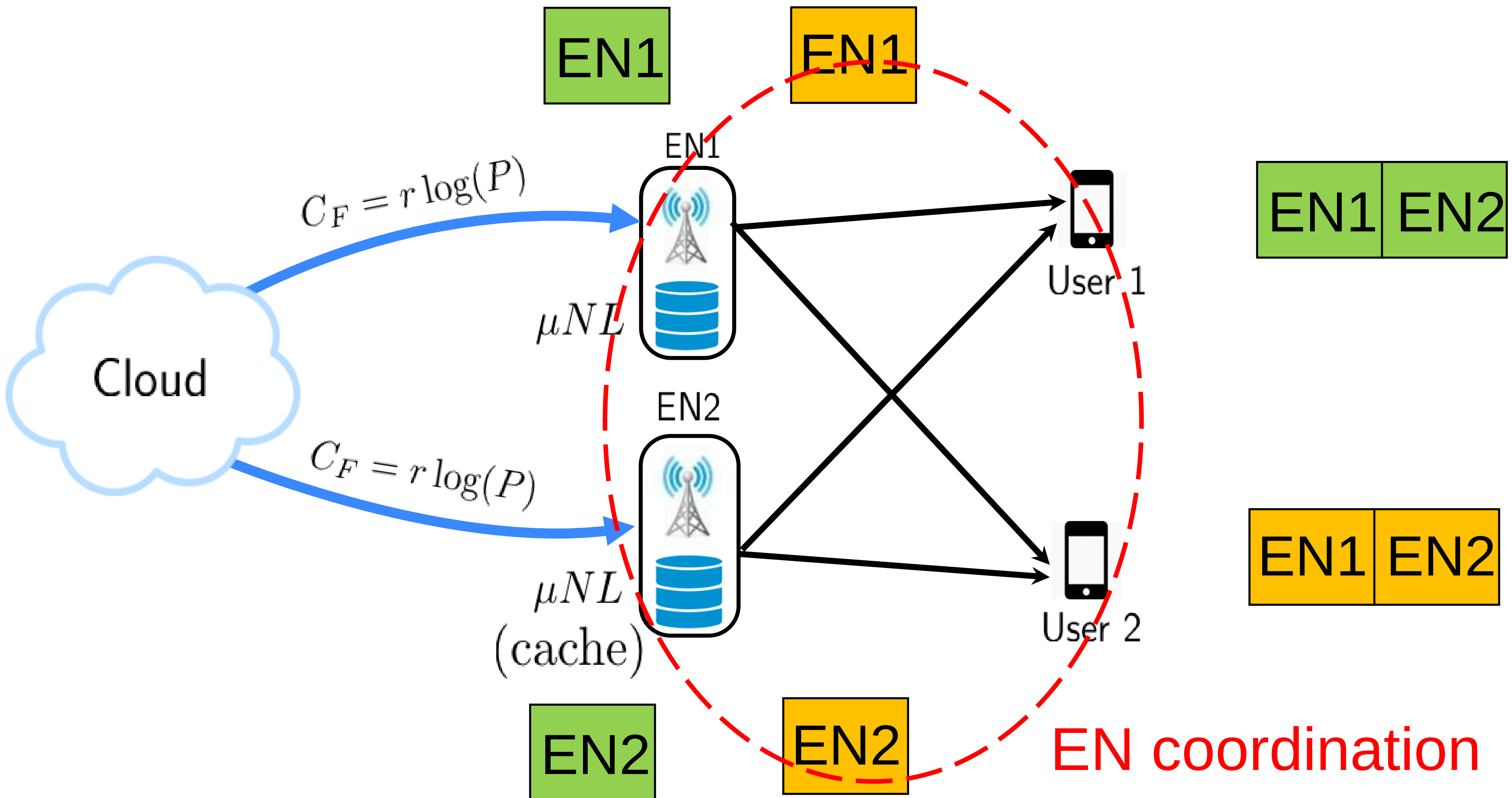
fronthaul mode
C-RAN transmission

Delivery phase

Minimum NDT: Serial Transmission

Low transport capacity

$r < r_{th}$



Delivery phase

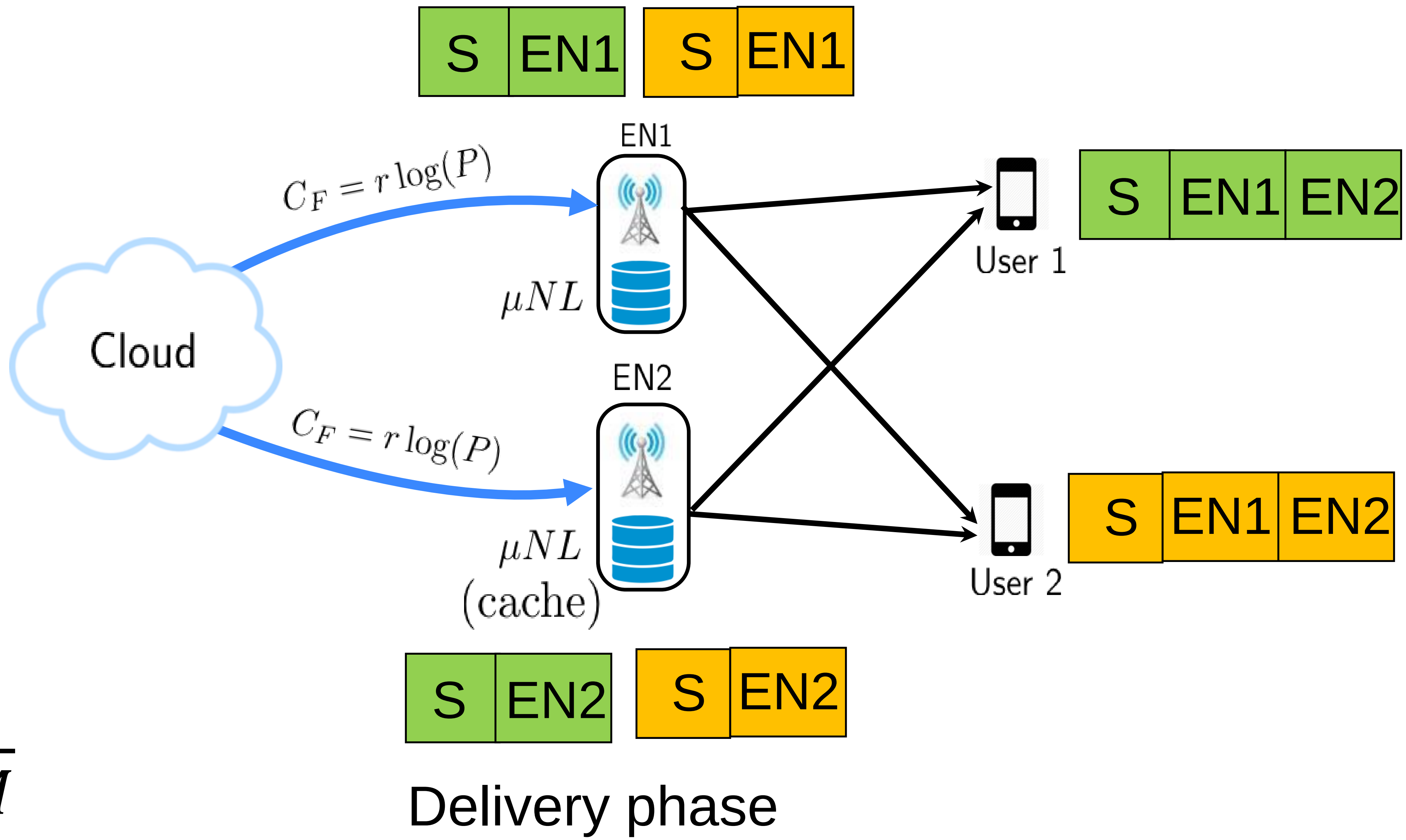
EN coordination
(interference alignment)

$$\mu \geq \frac{1}{M}$$

Minimum NDT: Serial Transmission

Low transport capacity

$r < r_{th}$

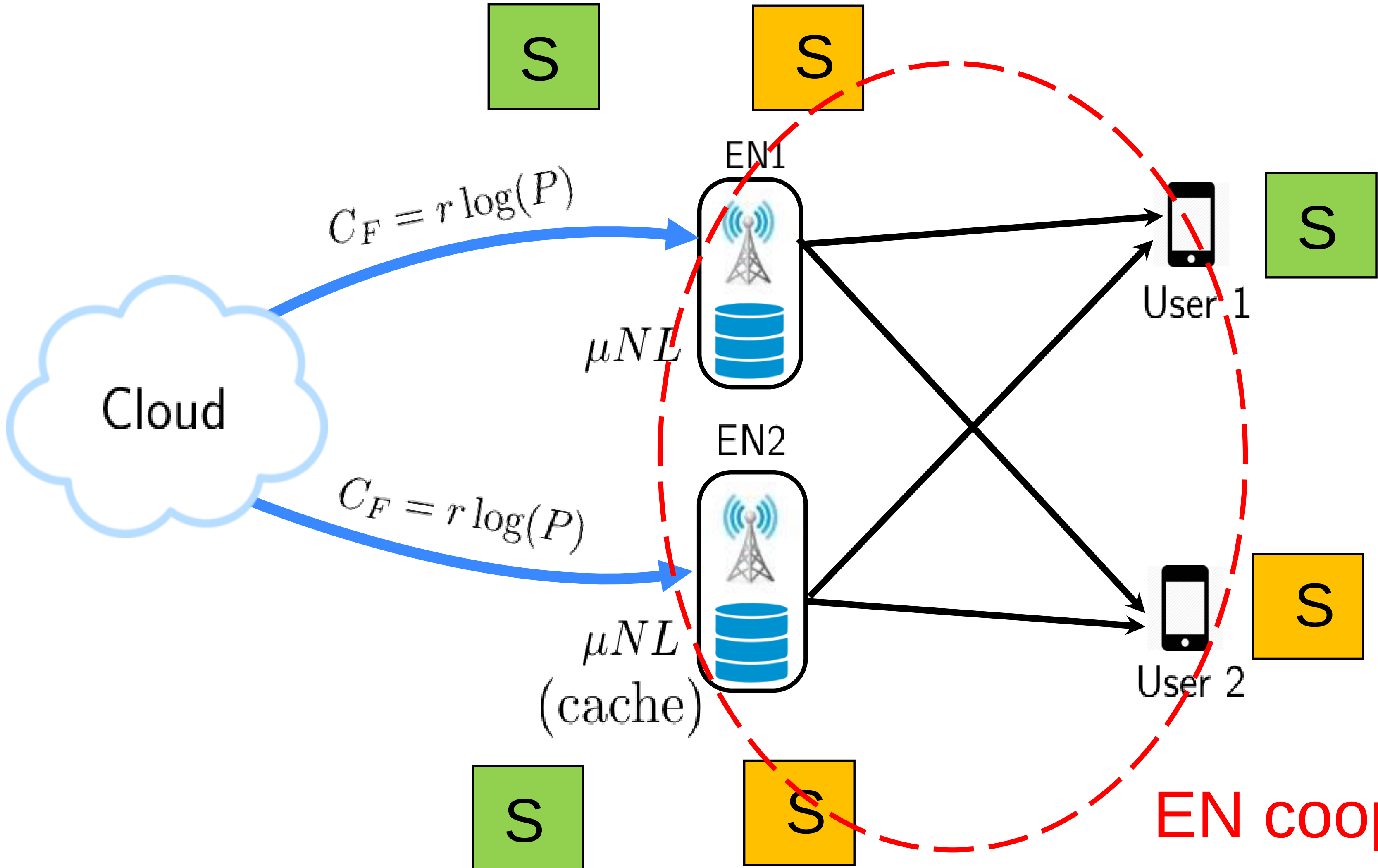


$$\mu \geq \frac{1}{M}$$

Minimum NDT: Serial Transmission

Low transport capacity

$r < r_{th}$

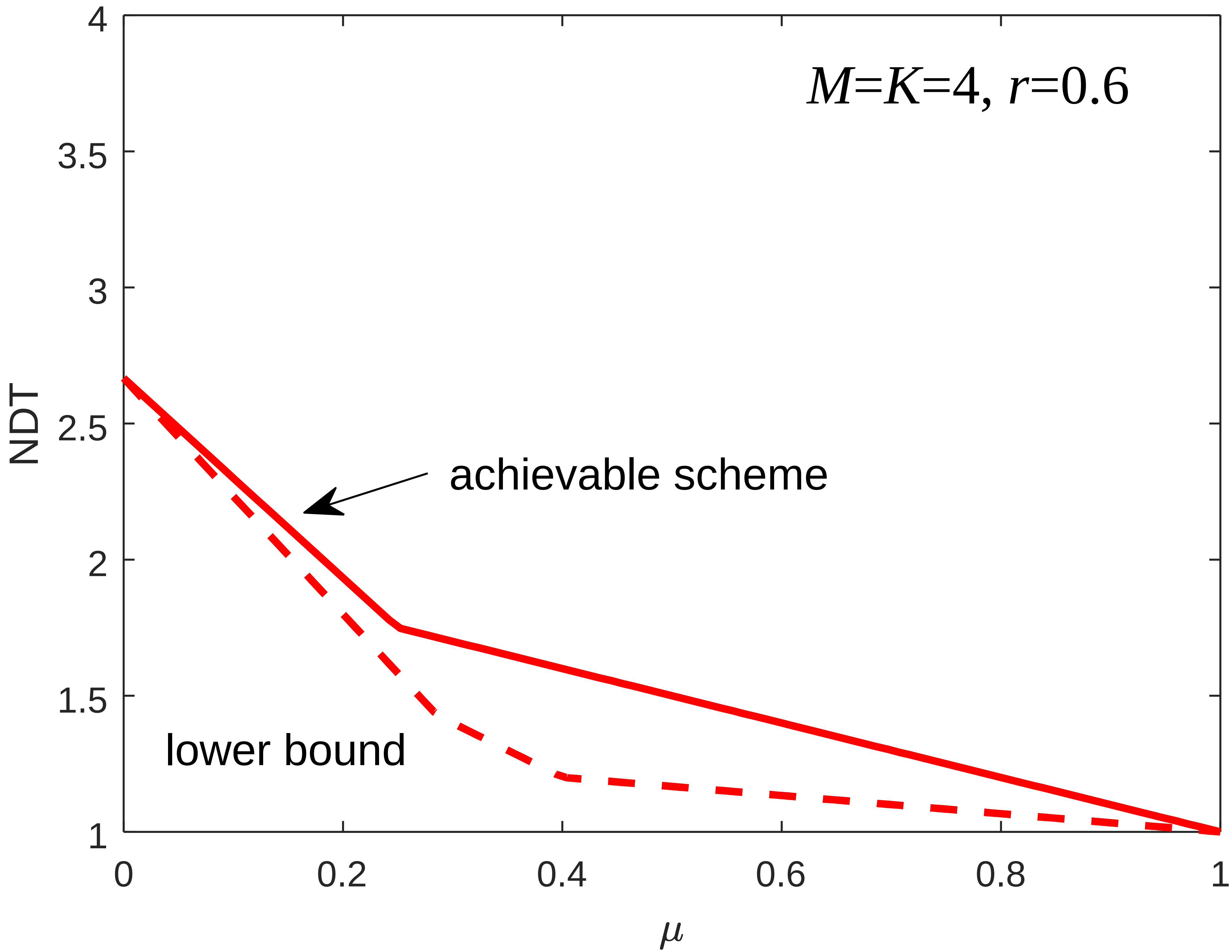


$\mu \geq \frac{1}{M}$

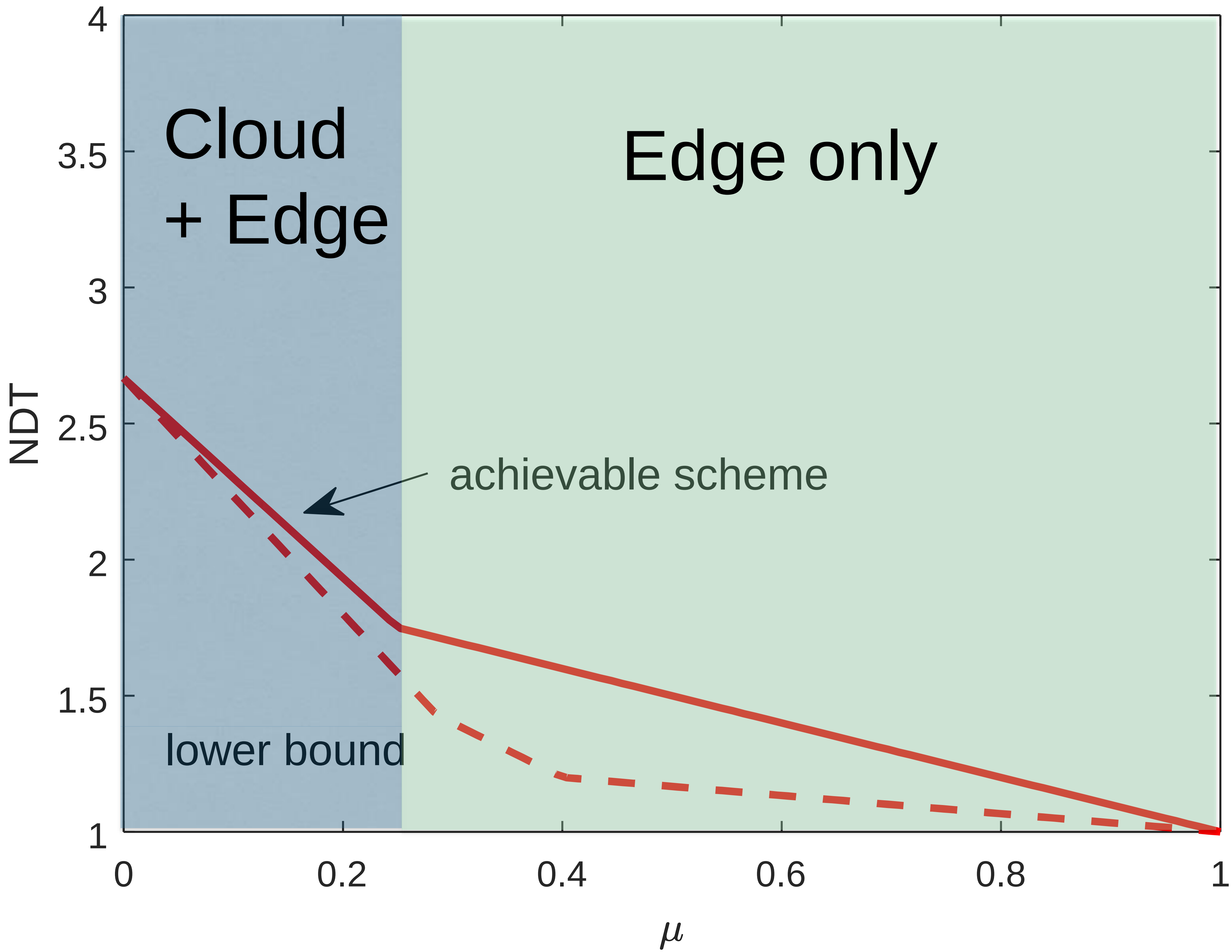
Delivery phase

EN cooperation
(ZF precoding)

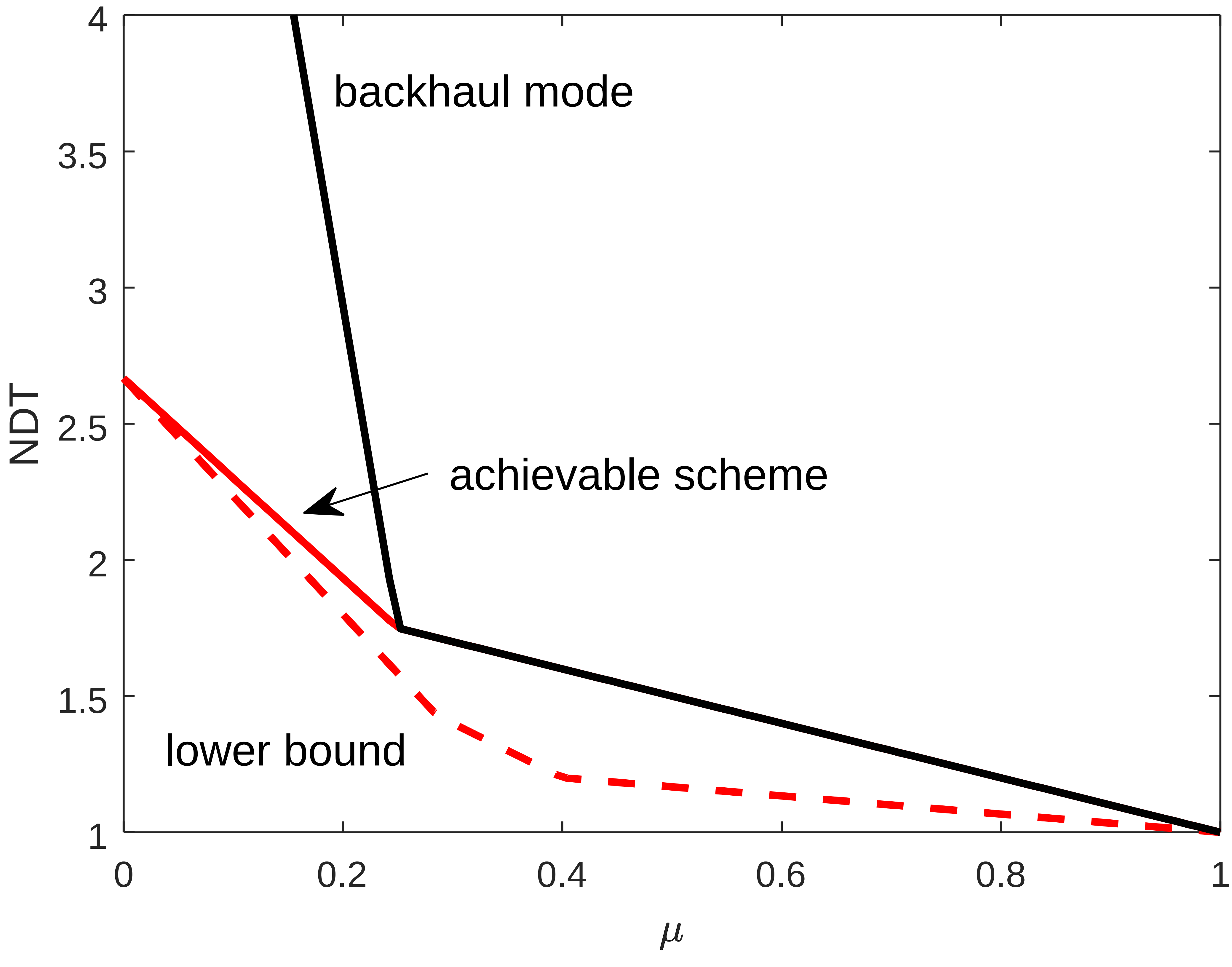
Minimum NDT: Serial Transmission



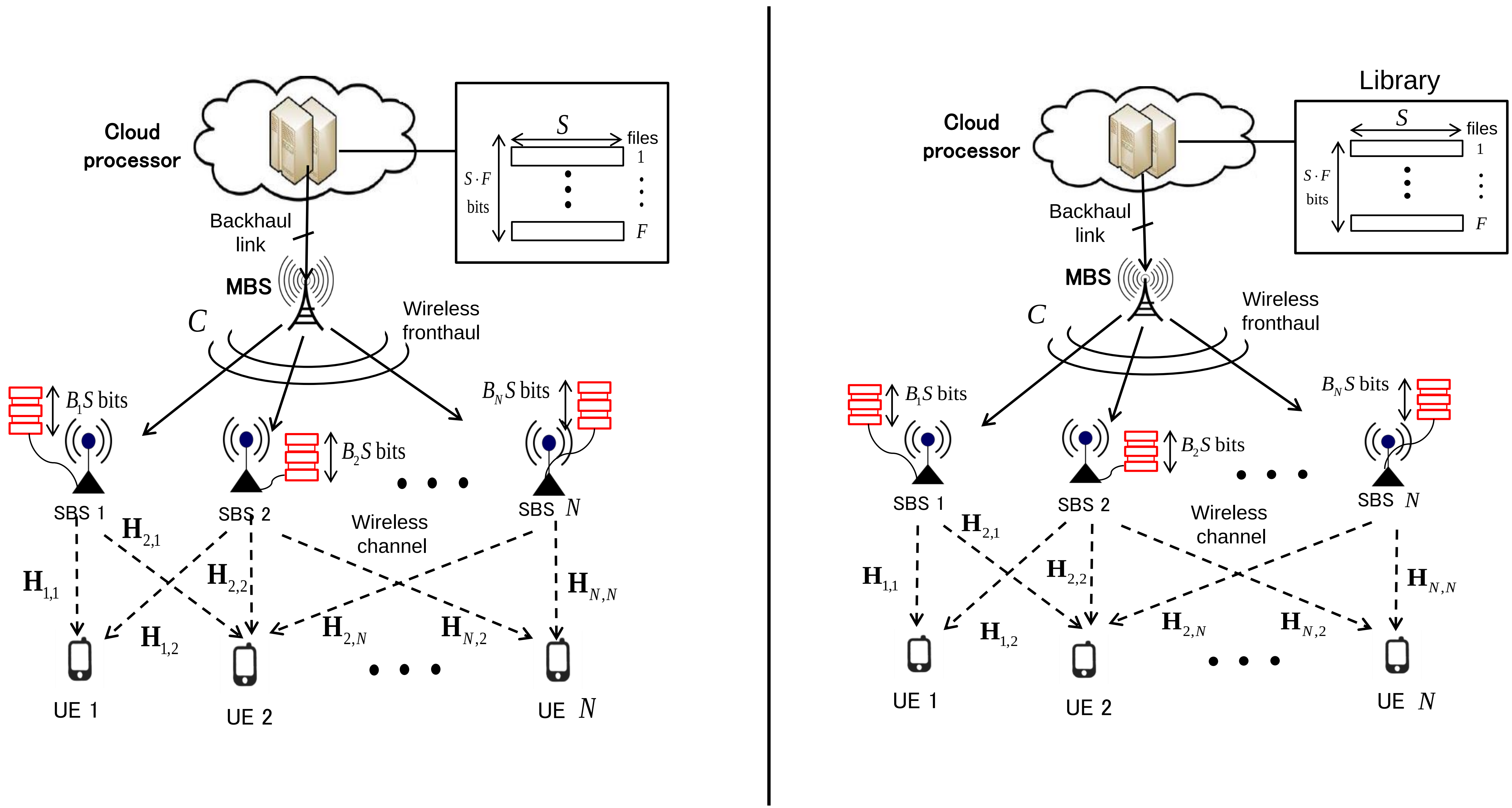
Minimum NDT: Serial Transmission



Minimum NDT: Serial Transmission



Multicast Fronthauling



S.-H. Park, W. Lee, O. Simeone and S. Shamai (Shitz), "Coded Multicast Fronthauling and Edge Caching for Multi-Connectivity Transmission in Fog Radio Access Networks," in Proc. IEEE SPAWC 2017.

Multicast Fronthauling

$F = 60$ files, $S = 100\,\text{MB}$, $L = 50$, $N = 4$ UEs/ENs, $n_R = n_U = 1$, $\alpha = 0.7$, $\gamma = 0.2$, $C = 2$, SNR $= 10\,\text{dB}$, $M = 2$