NJIT®
New Jersey Institute of Technology

THE ELISHA YEGAL BAR-NESS CENTER FOR WIRELESS COMMUNICATIONS AND SIGNAL PROCESSING RESEARCH

Fog-Aided Wireless Networks

O. Simeone
New Jersey Institute of Technology (NJIT)

Joint work with
Ravi Tandon

Towards 5G Systems

Cloud processing: centralization

Edge processing: localization

Towards 5G Systems

Cloud processing: centralization

→ cloud radio access network (C-RAN)

Edge processing: localization

→ edge caching

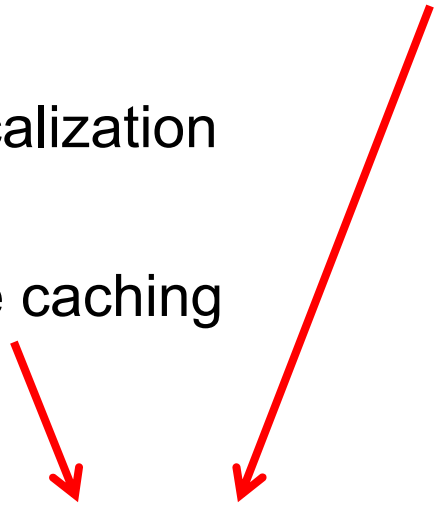
Towards 5G Systems

Cloud processing: centralization

→ cloud radio access network (C-RAN)

Edge processing: localization

→ edge caching



Fog-RAN (F-RAN): cache-aided C-RAN

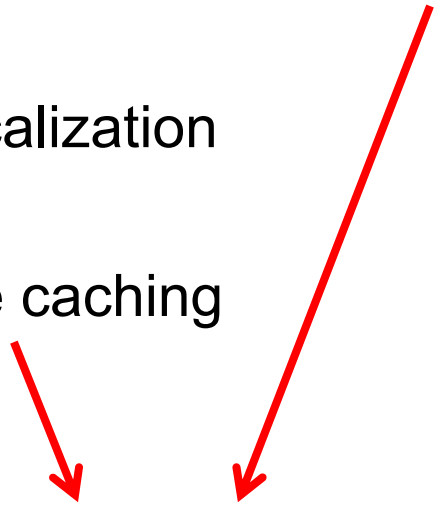
Towards 5G Systems

Cloud processing: centralization

→ cloud radio access network (C-RAN)

Edge processing: localization

→ edge caching

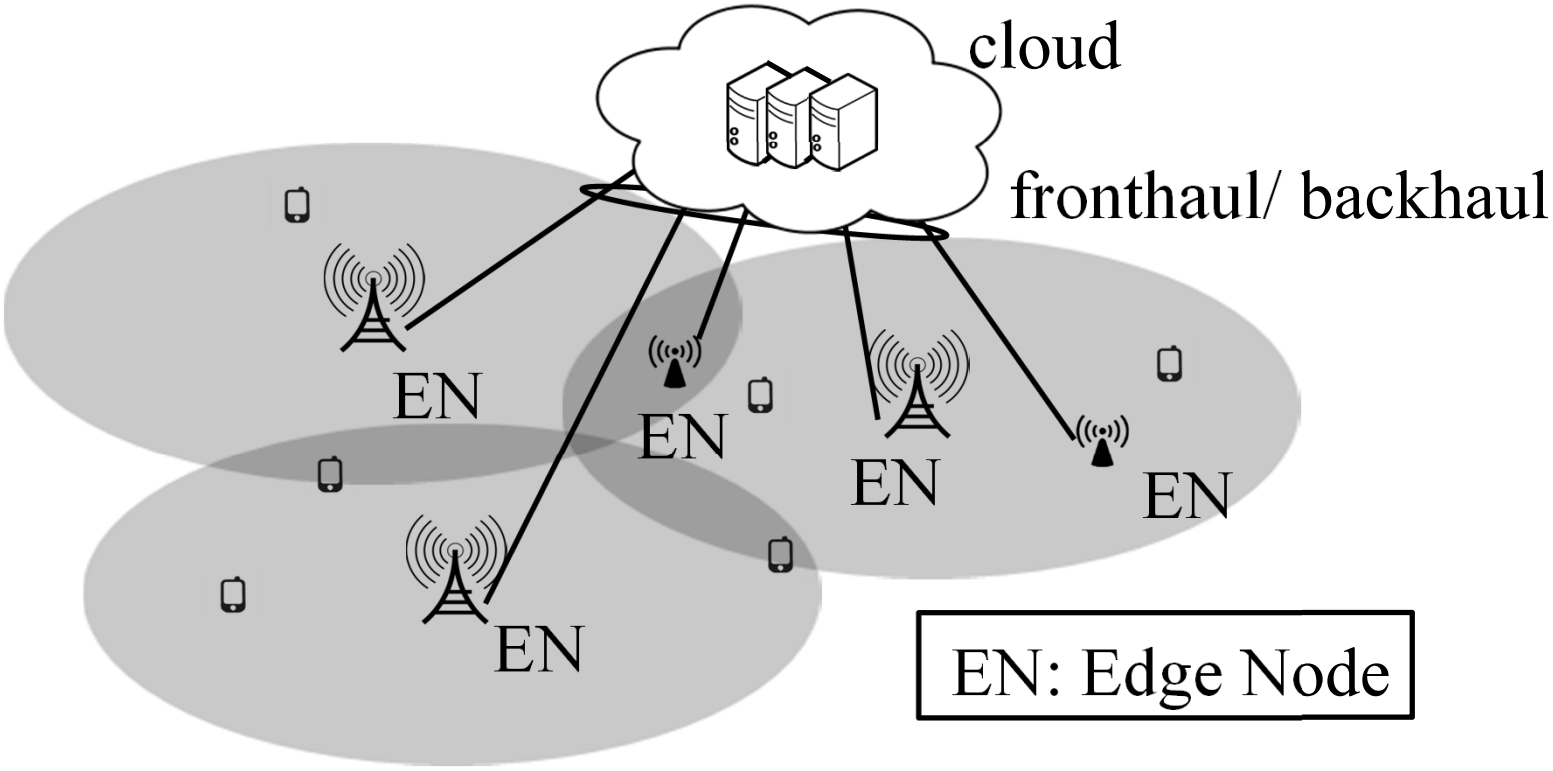


information-theoretic analysis of

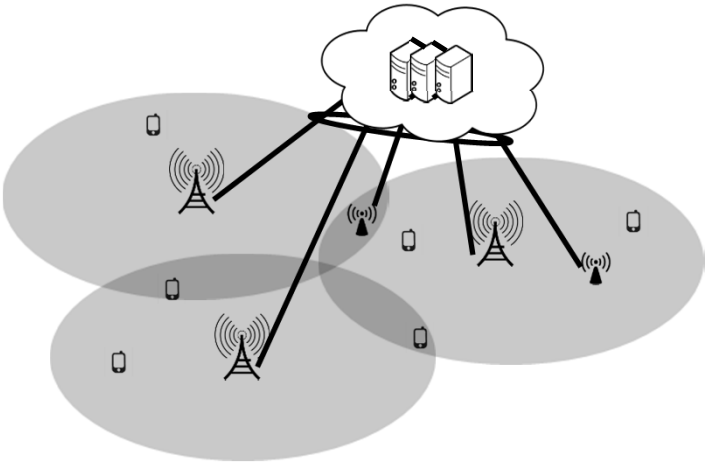
Fog-RAN (F-RAN): cache-aided C-RAN

Cloud Radio Access Network (C-RAN)

Centralization of network functionalities via virtualization



Cloud Radio Access Network (C-RAN)

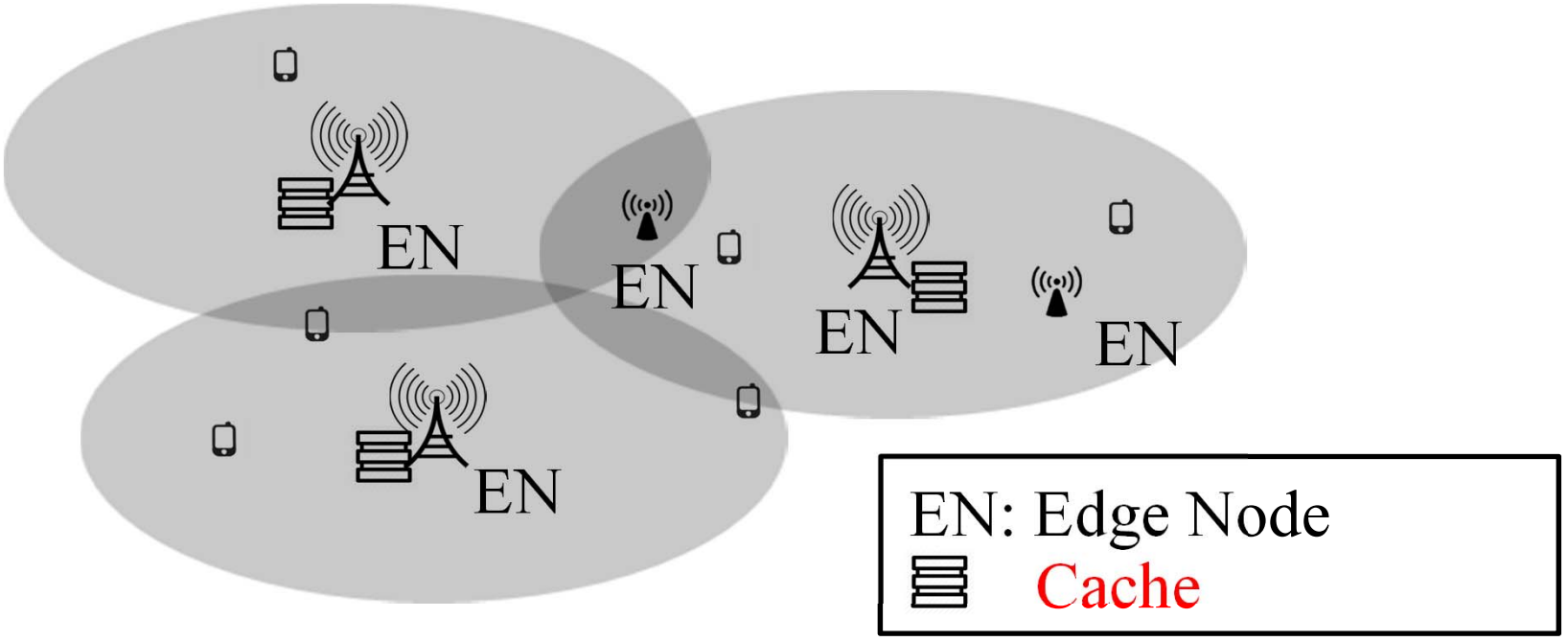


- Reduced CAPEX and OPEX
- Greening and statistical multiplexing
- Interference management

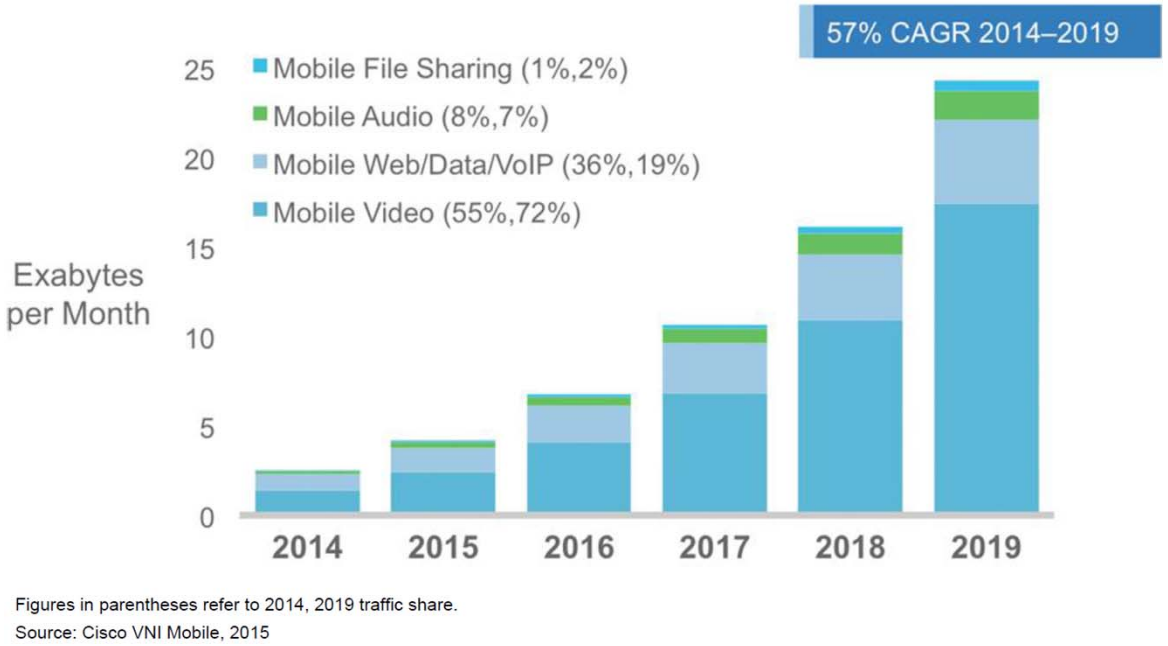
- Fronthaul/ backhaul overhead
- Latency

Edge Caching

- Storage of popular content at the edge



Edge Caching

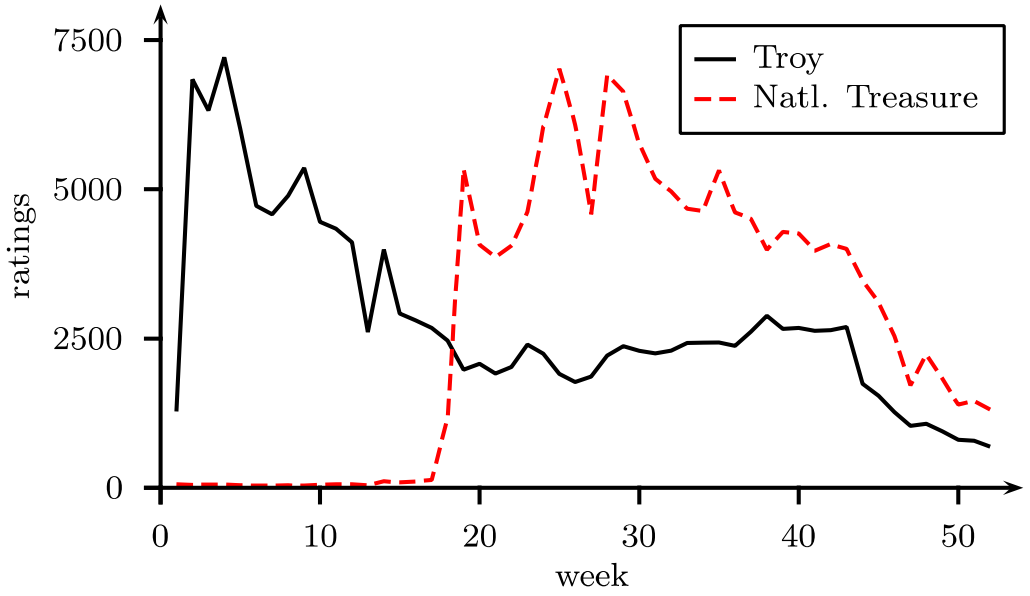


Figures in parentheses refer to 2014, 2019 traffic share.
Source: Cisco VNI Mobile, 2015

- Video-on-demand is driving wireless traffic growth
- Predicted to account for almost ¾ traffic by 2019

Image source: Cisco (2015)

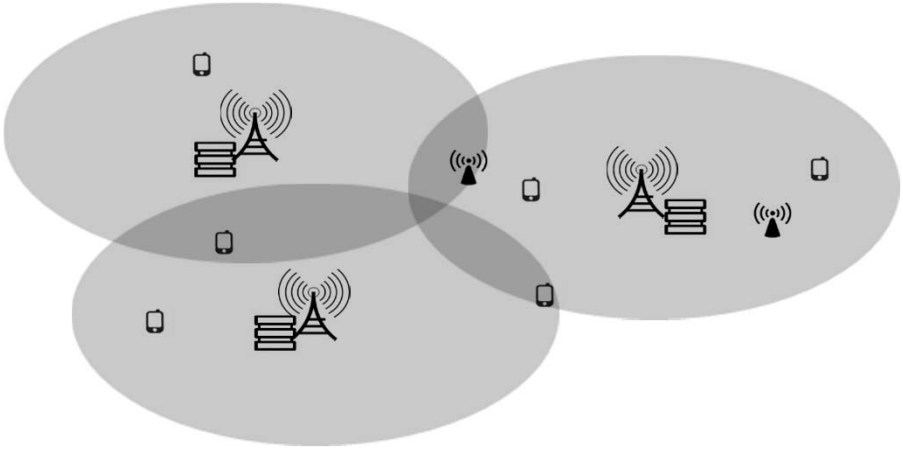
Edge Caching



- Content popularity is slowly time-varying and hence predictable

Image source: Pedarsani et al. (2013)
Data source: netflixprize.com

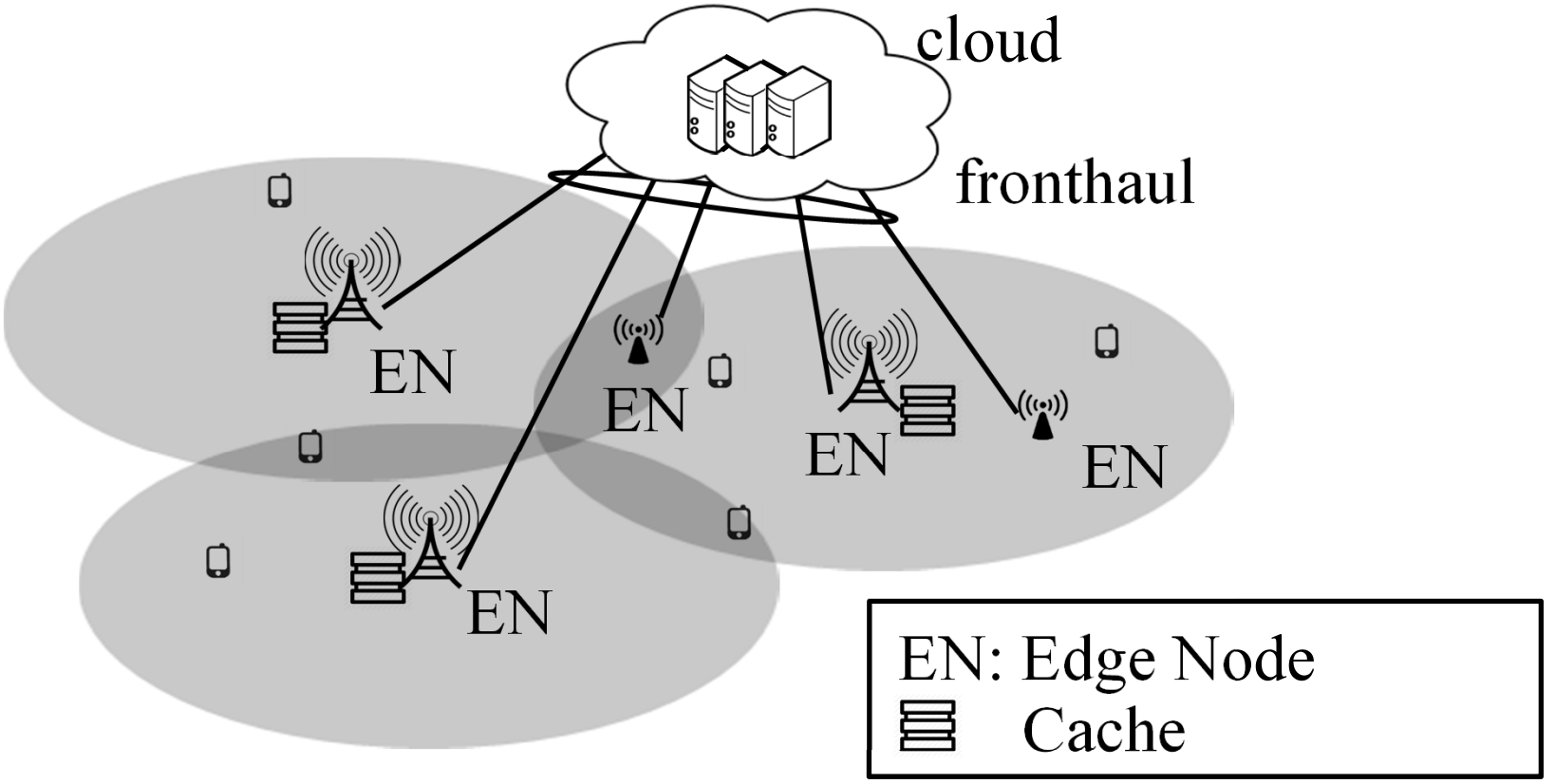
Edge Caching



- Reduced backhaul overhead
- Reduced latency

- Interference limited
- Unable to support generic traffic requirements

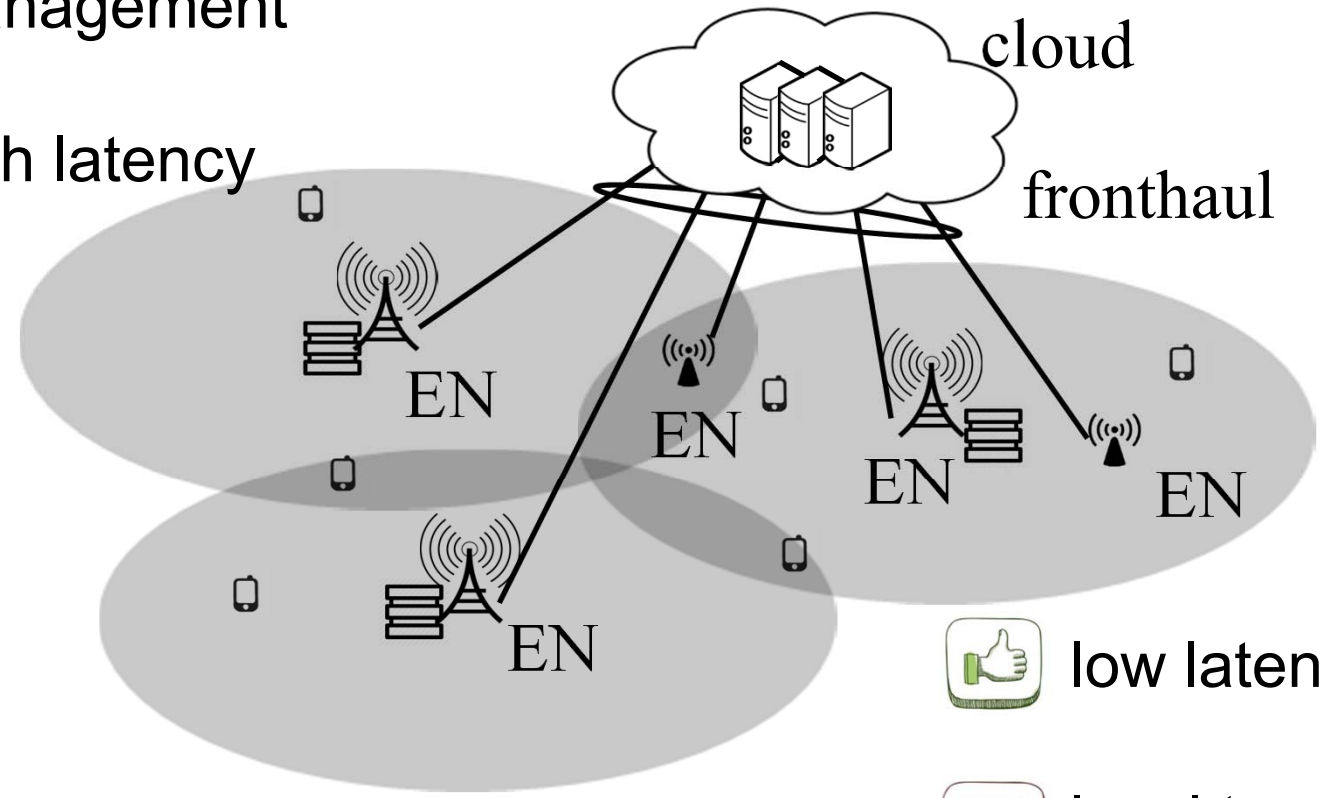
Fog-RAN (F-RAN)



Fog-RAN (F-RAN)

centralized interference management

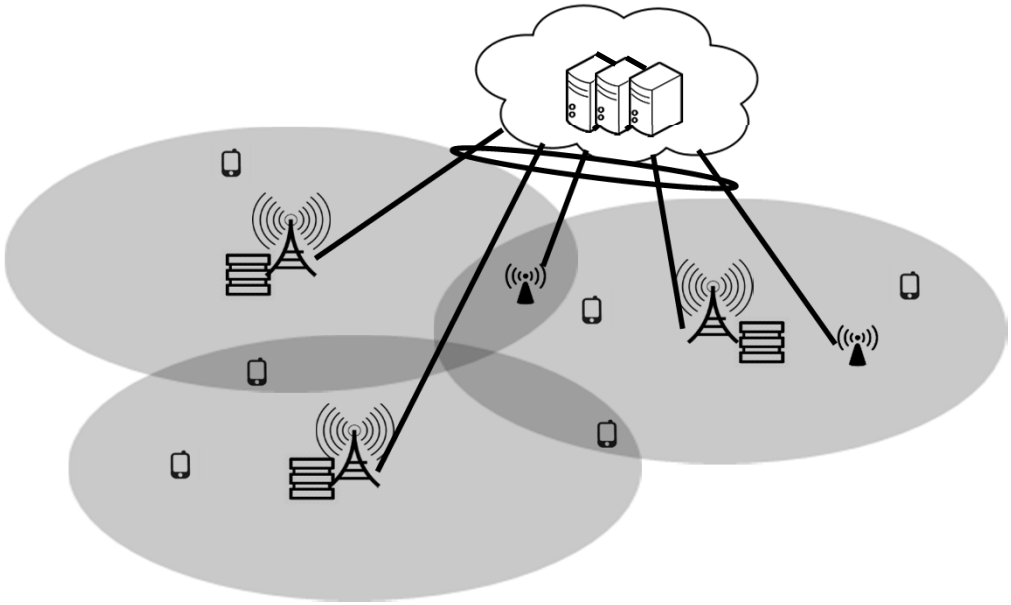
high latency



low latency

local transmission

Fog-RAN (F-RAN)



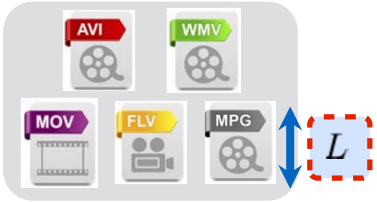
- Optimal operation of an F-RAN: complex design problem across cloud, fronthaul and edge segments
- Focus: *Information-theoretic model and performance metric* to obtain fundamental insights

System Model

Key F-RAN Parameters

N : # of popular files

L : size of a File



Set of popular files

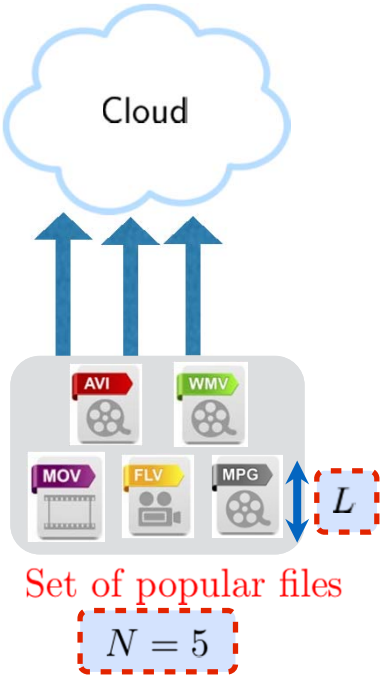
$N = 5$

System Model

Key F-RAN Parameters

N : # of popular files

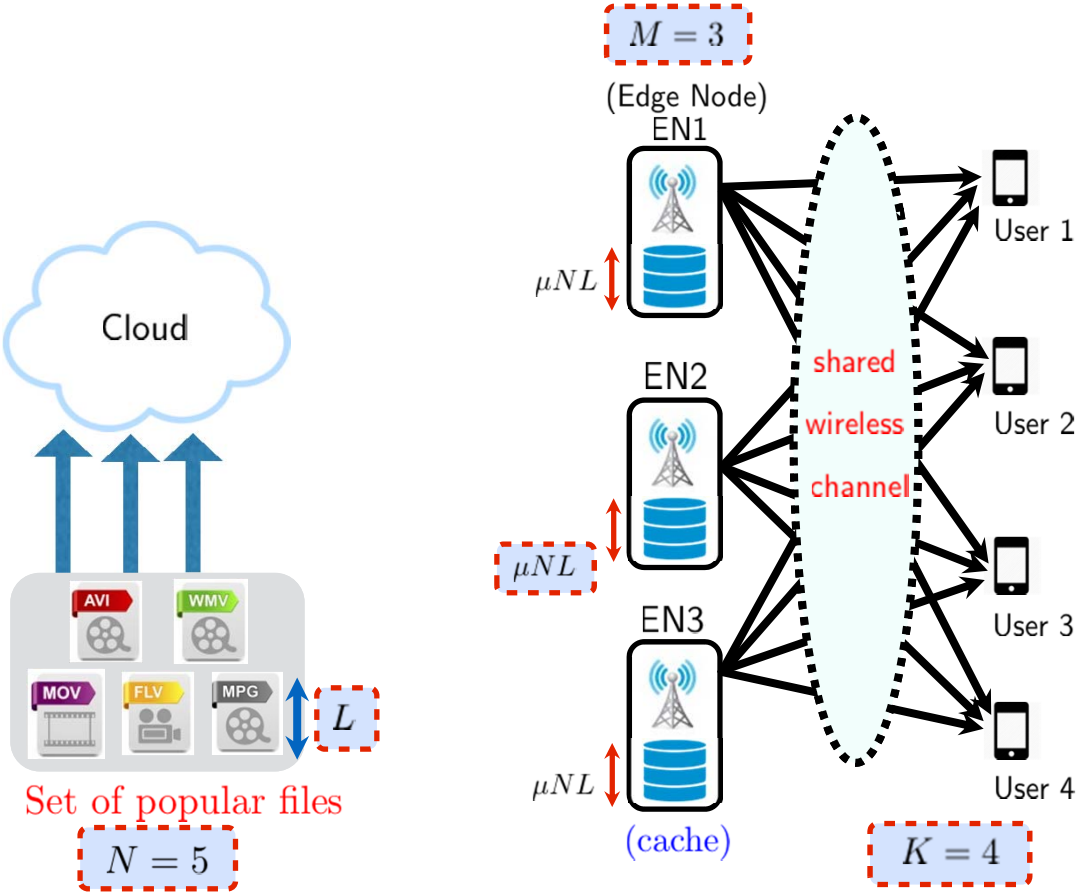
L : size of a File



System Model

Key F-RAN Parameters

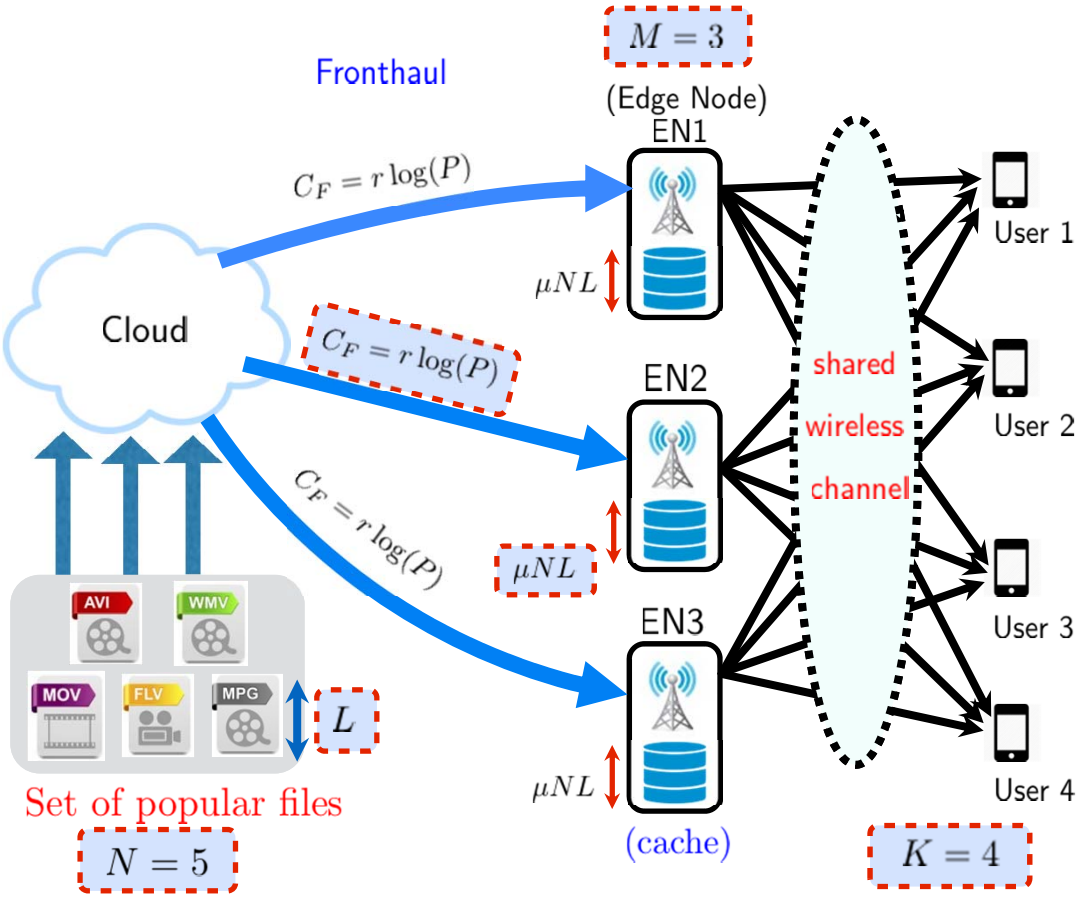
N : # of popular files	M : # of edge nodes (ENs)	K : # of users
L : size of a File	μNL : cache storage capacity	



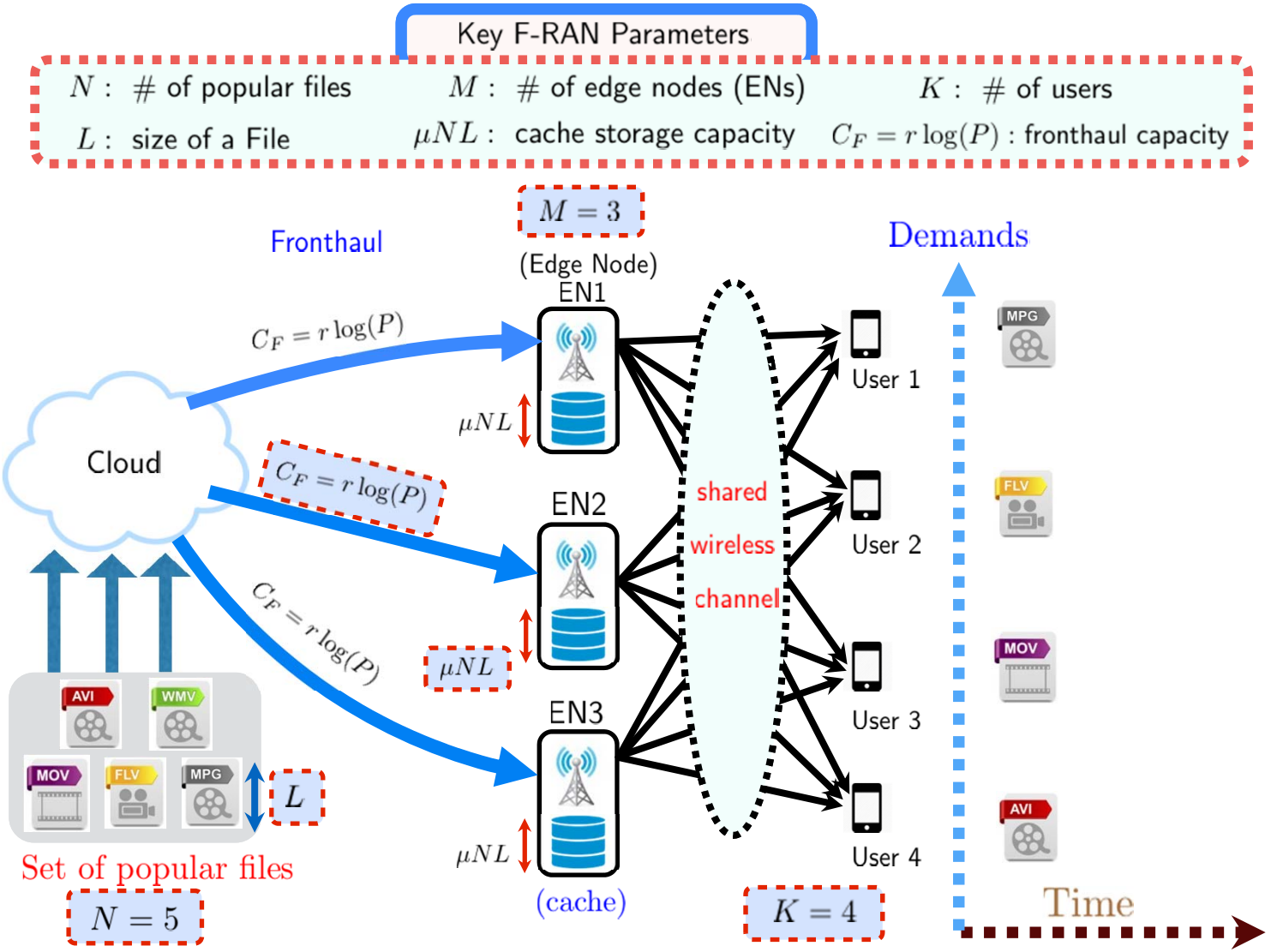
System Model

Key F-RAN Parameters

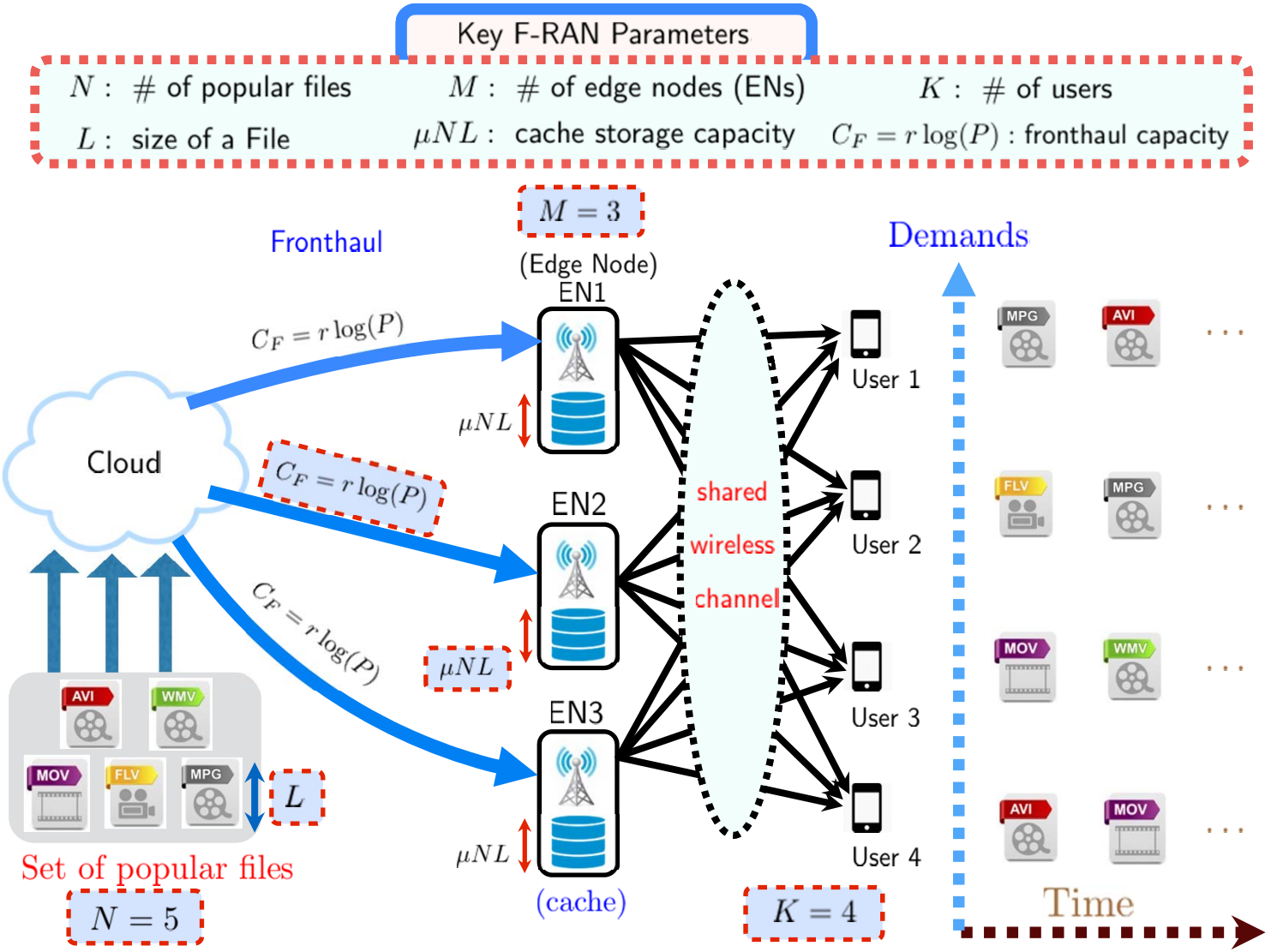
N : # of popular files	M : # of edge nodes (ENs)	K : # of users
L : size of a File	μNL : cache storage capacity	$C_F = r\log(P)$: fronthaul capacity



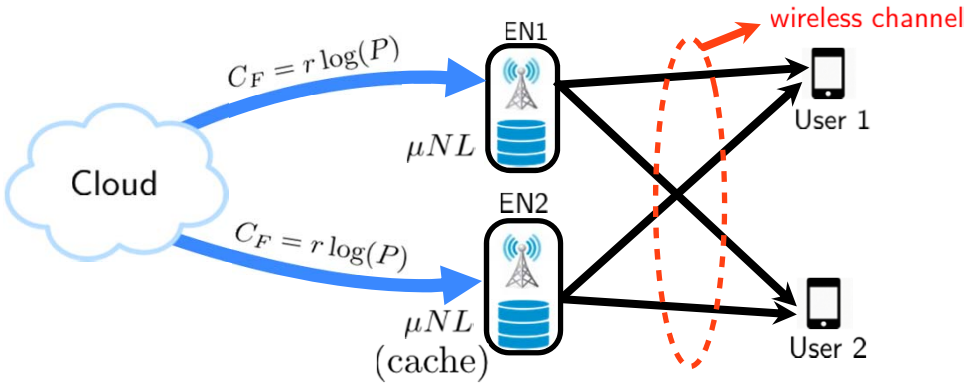
System Model

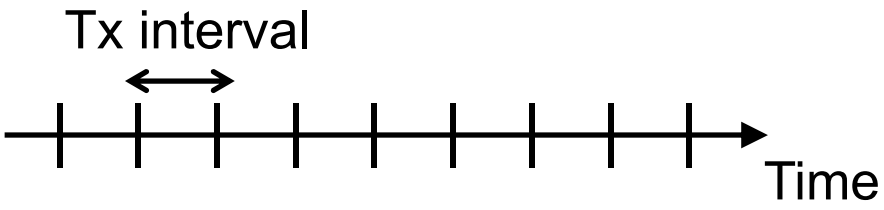


System Model

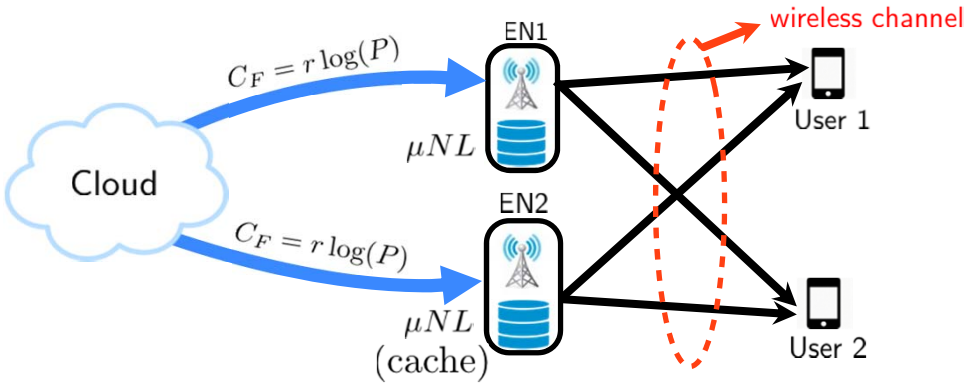


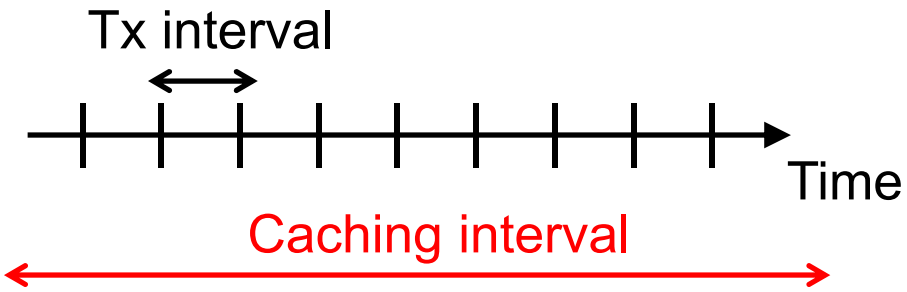
Design Space





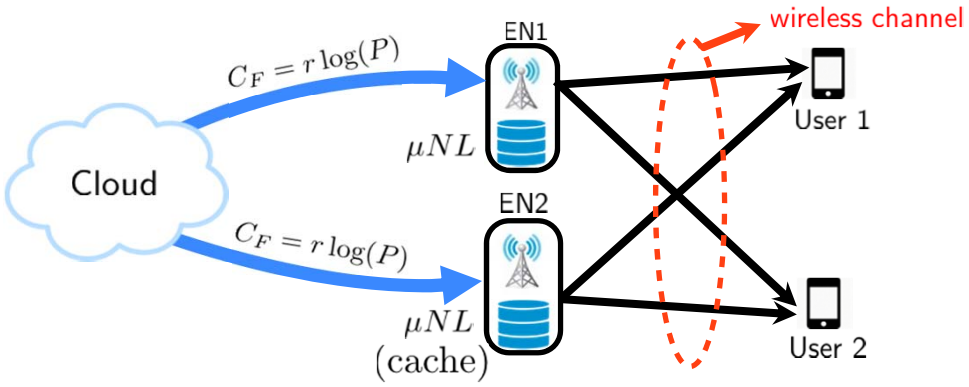
Design Space

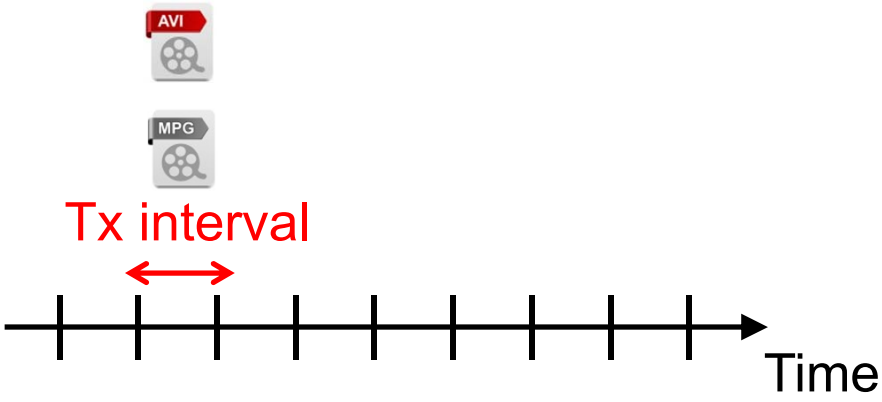




- Cache storage policy: What to cache

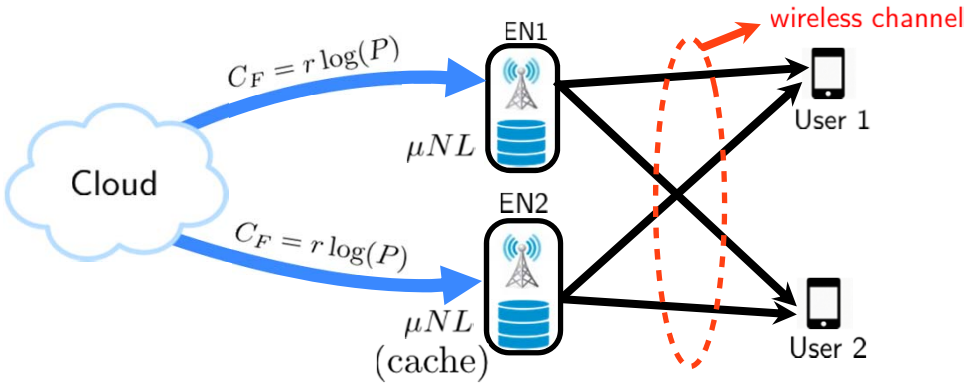
Design Space

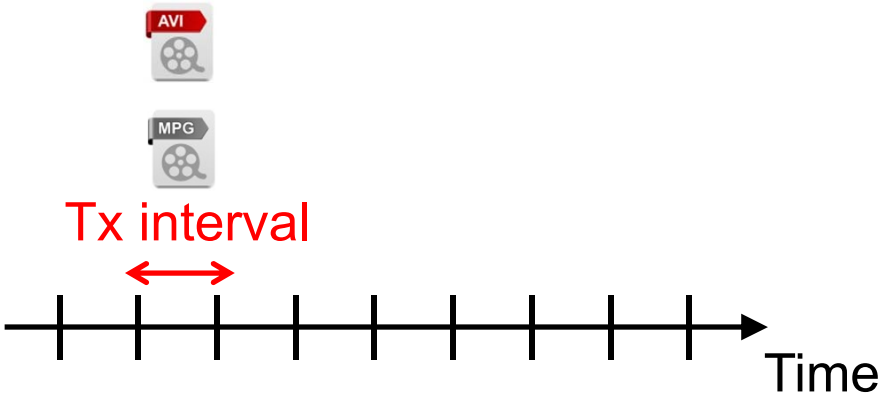




- Cache storage policy: What to cache
- Fronthaul policy: What to transmit on the fronthaul links
    - Hard/ soft-transfer mode

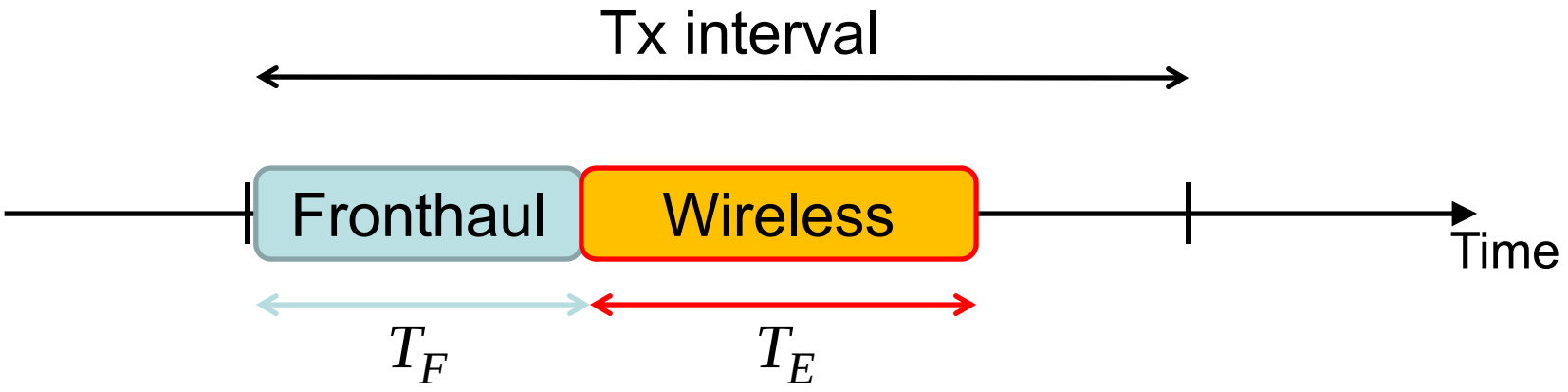
Design Space





- Cache storage policy: What to cache
- Fronthaul policy: What to transmit on the fronthaul links
    - Hard/ soft-transfer mode
- Edge transmission policy: What to transmit on the wireless channel

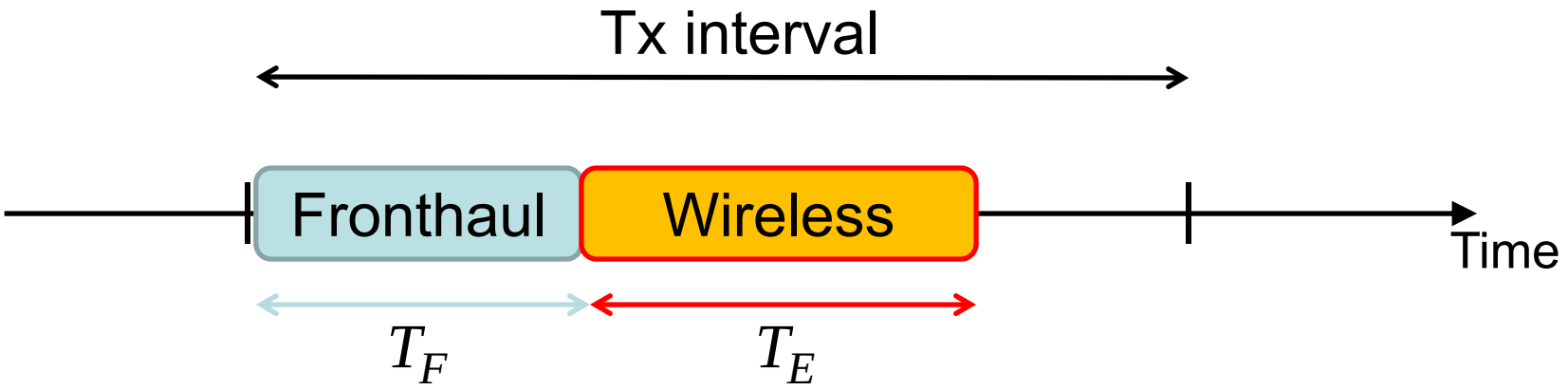
Performance Metric



- Delivery time per bit

$$\Delta(\mu, C_F, P) = \min_{\text{user's requests}} \frac{T_F + T_E}{L}$$

Performance Metric



- Delivery time per bit

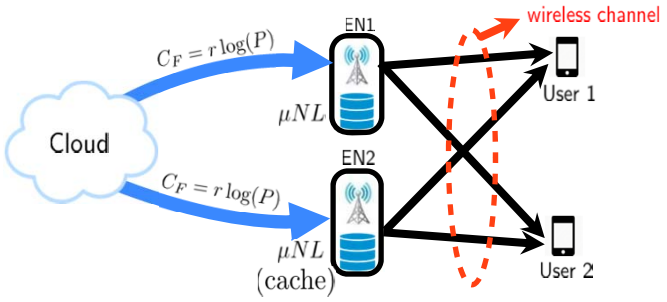
$$\Delta(\mu, C_F, P) = \min_{\text{user's requests}} \frac{T_F + T_E}{L}$$

- Normalized Delivery Time (NDT):

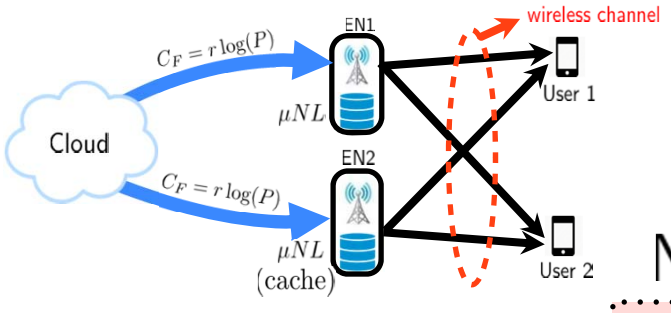
$$\delta^*(\mu, r) = \lim_{P\to\infty} \frac{\Delta(\mu, r\log P, P)}{1/\log P}$$

Ideal system: interference-free and unlimited caching

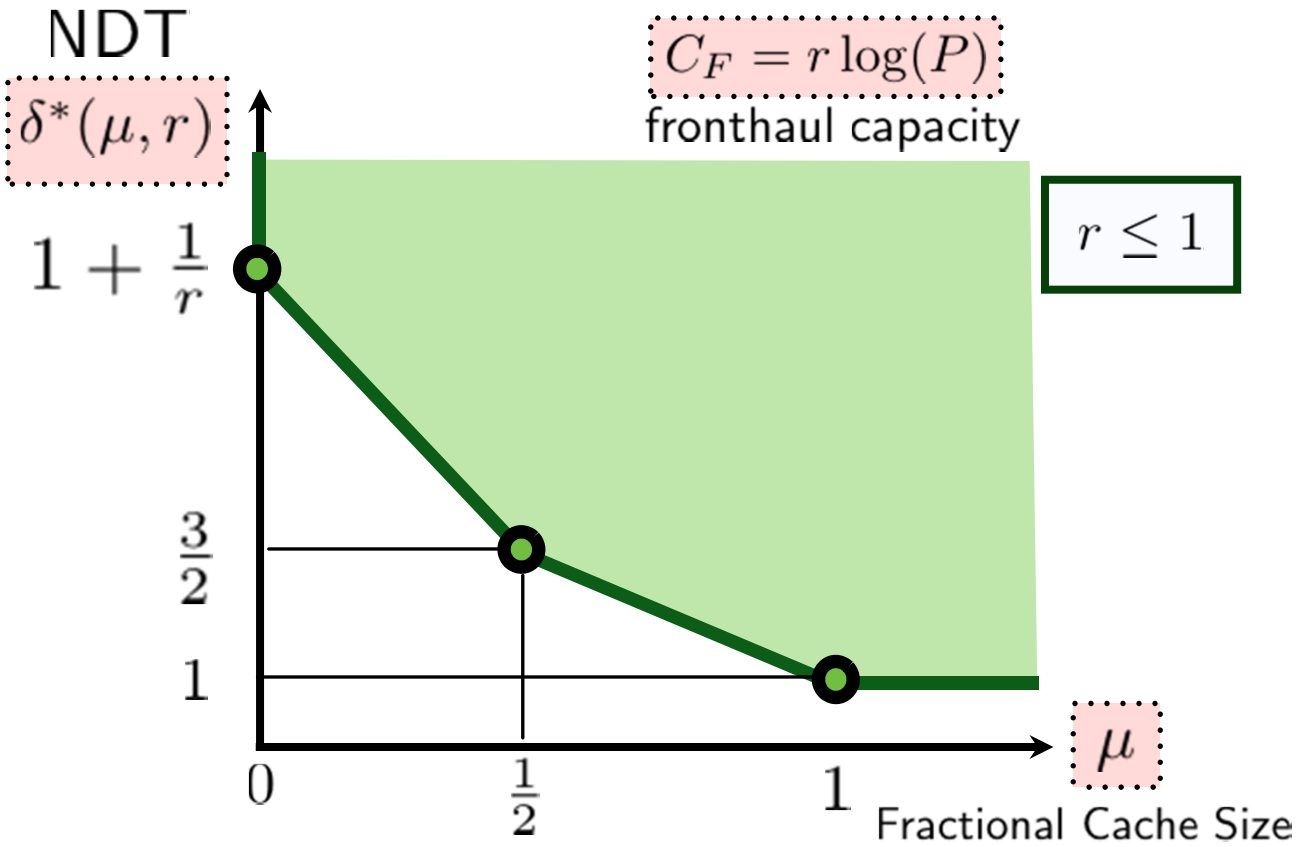
Normalized Delivery Time



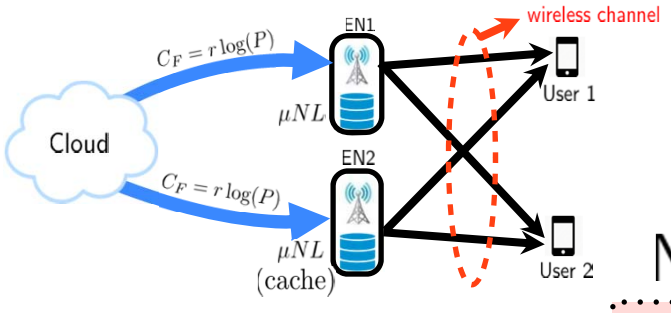
Normalized Delivery Time



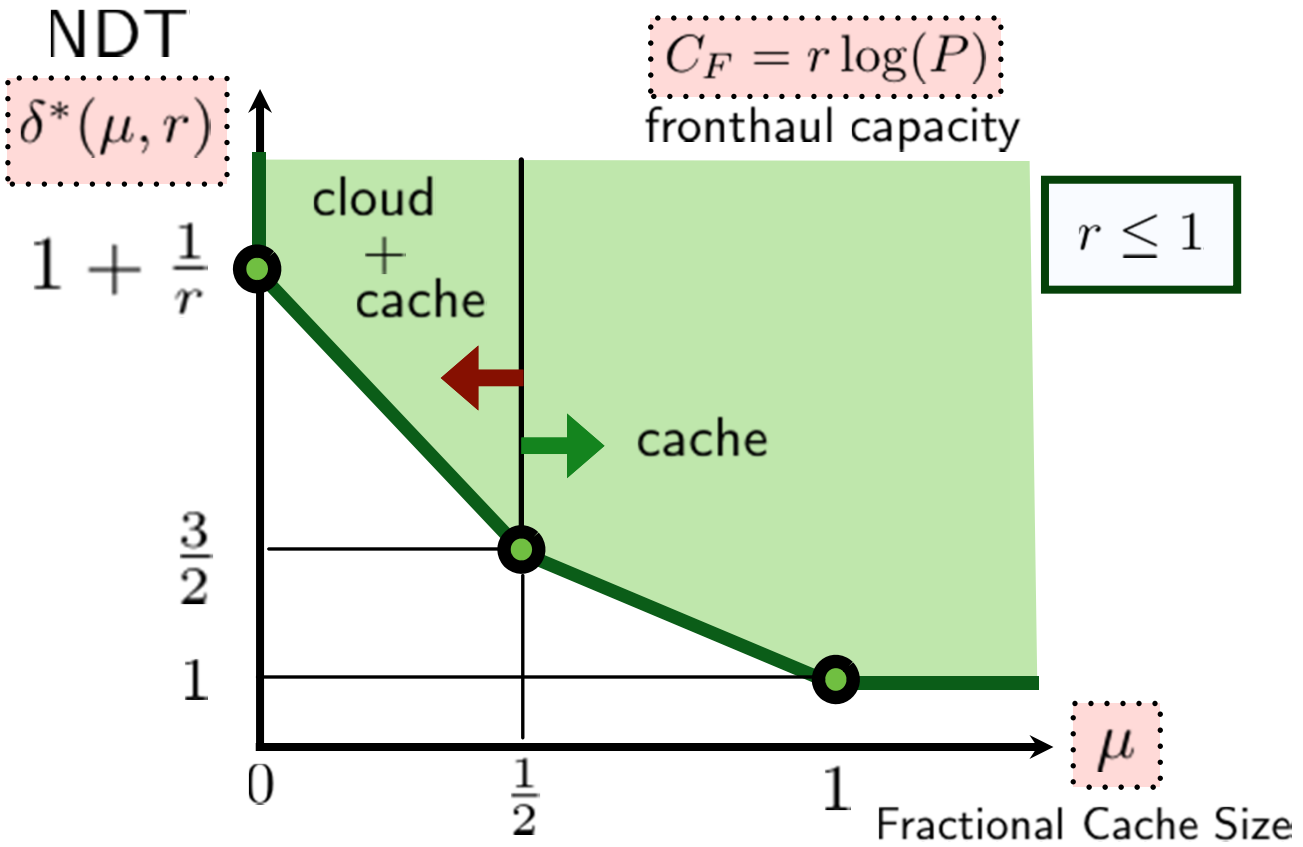
Low Fronthaul Capacity Regime



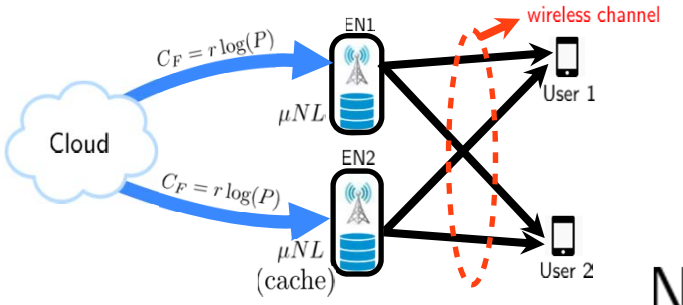
Normalized Delivery Time



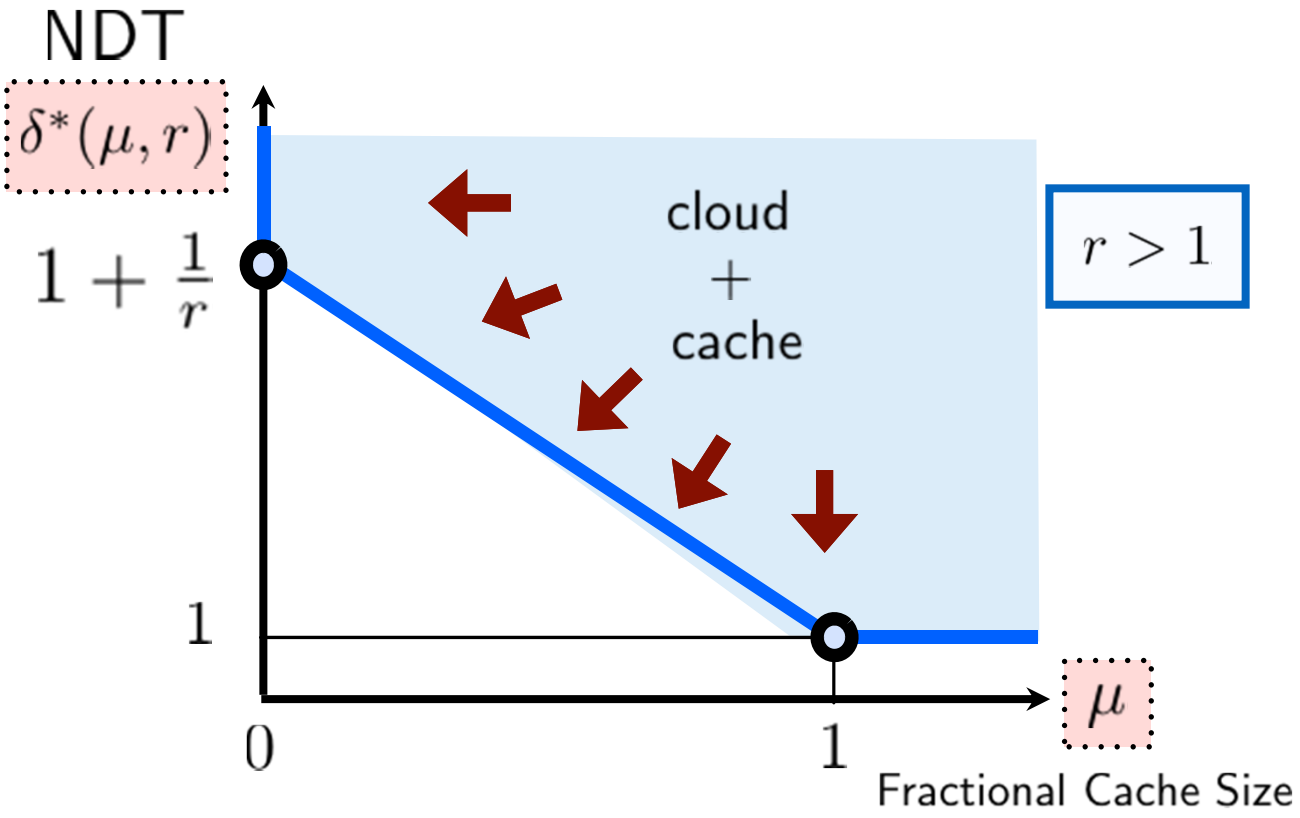
Low Fronthaul Capacity Regime



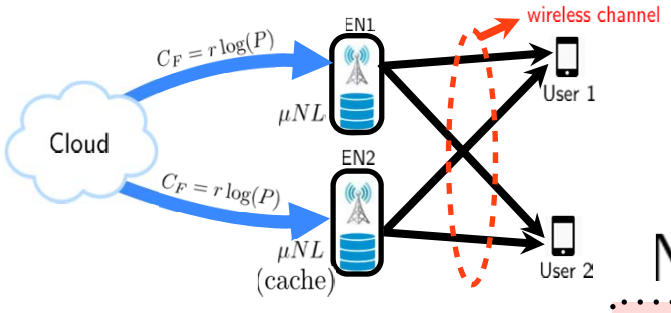
Normalized Delivery Time



High Fronthaul Capacity Regime



Normalized Delivery Time



Low Fronthaul Capacity Regime

NDT

$\delta^*(\mu, r)$

$C_F = r\log(P)$

fronthaul capacity

$r \leq 1$

$1 + \frac{1}{r}$

cloud + cache

cache

$\frac{3}{2}$

1

0

$\frac{1}{2}$

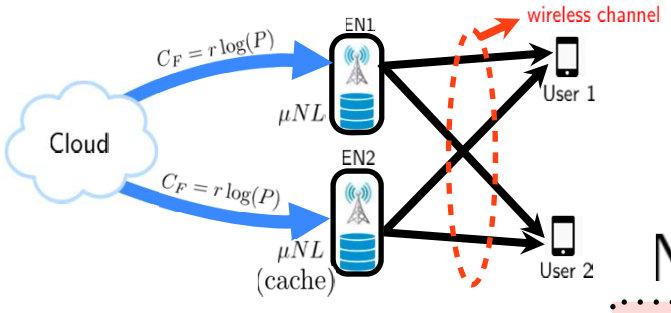
1

μ

Fractional Cache Size

Full caching: Cooperative zero-forcing beamforming at the ENs

Normalized Delivery Time



Low Fronthaul Capacity Regime

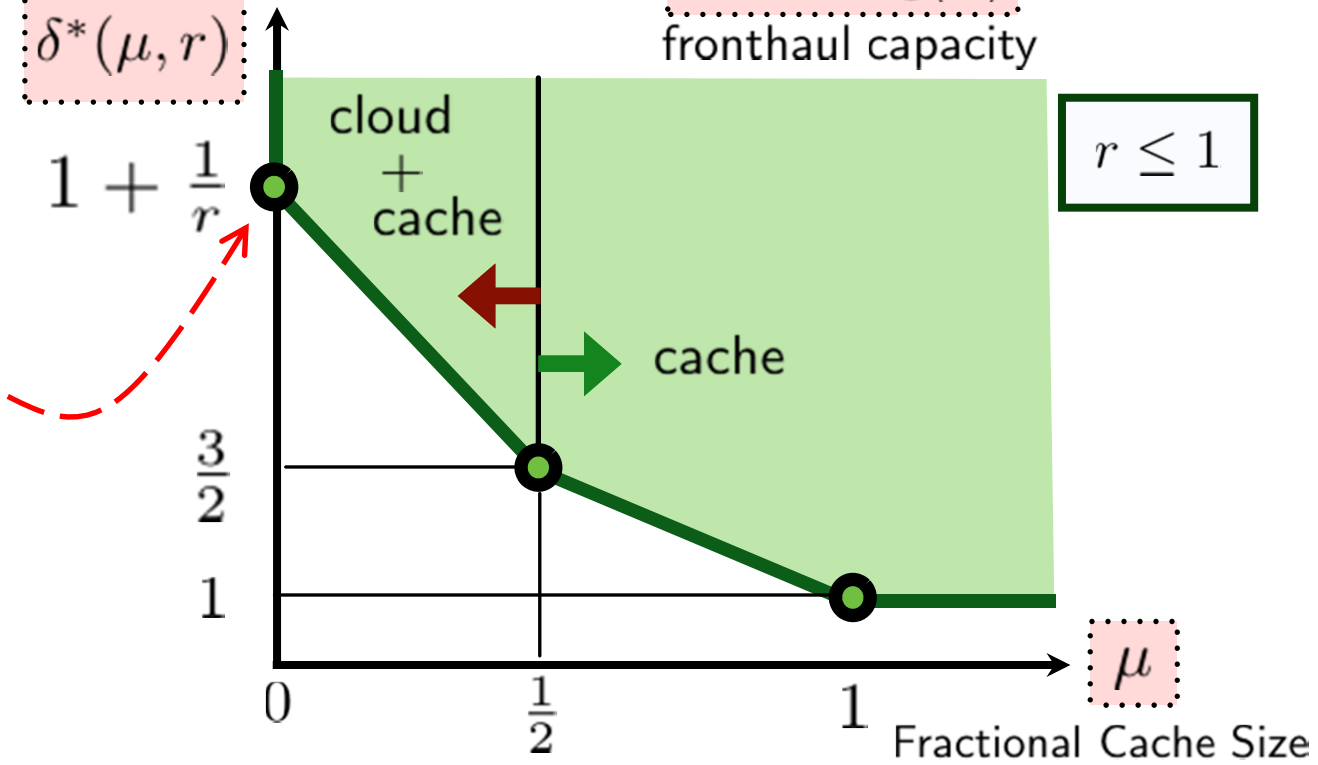
NDT

$\delta^*(\mu, r)$

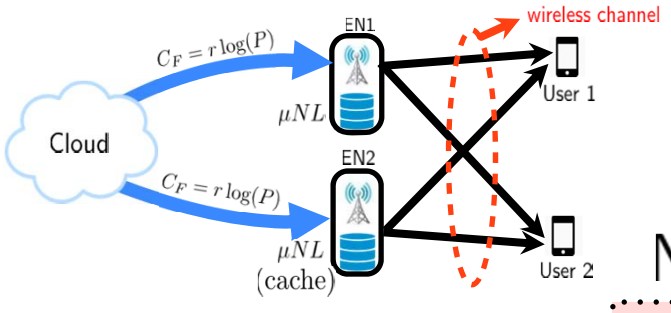
$C_F = r\log(P)$

fronthaul capacity

No caching: Zero-forcing beamforming at the cloud + soft-transfer fronthauling



Normalized Delivery Time



Low Fronthaul Capacity Regime

NDT

$\delta^*(\mu, r)$

$C_F = r\log(P)$

fronthaul capacity

$r \leq 1$

$1 + \frac{1}{r}$

cloud + cache

cache

$\frac{3}{2}$

1

0

$\frac{1}{2}$

1

μ

Fractional Cache Size

Caching of half file: Interference alignment on an "X-channel"

Conclusions

- F-RAN leverages the synergy and complementarity of cloud processing and edge caching

- Information-theoretic framework to obtain fundamental insights into the optimal trade-offs between latency and system resources (fronthaul and caching)

The Science and Technology University of New Jersey