Fog Radio Access Networks: Architectures and Key Techniques

Mugen Peng
(pmg@bupt.edu.cn)
Beijing University of Posts & Telecommunications

Oct. 19, 2018

International Conference on Cyber-enabled distributed computing and knowledge discovery Zhengzhou, October 18 - 20, 2018

Outline

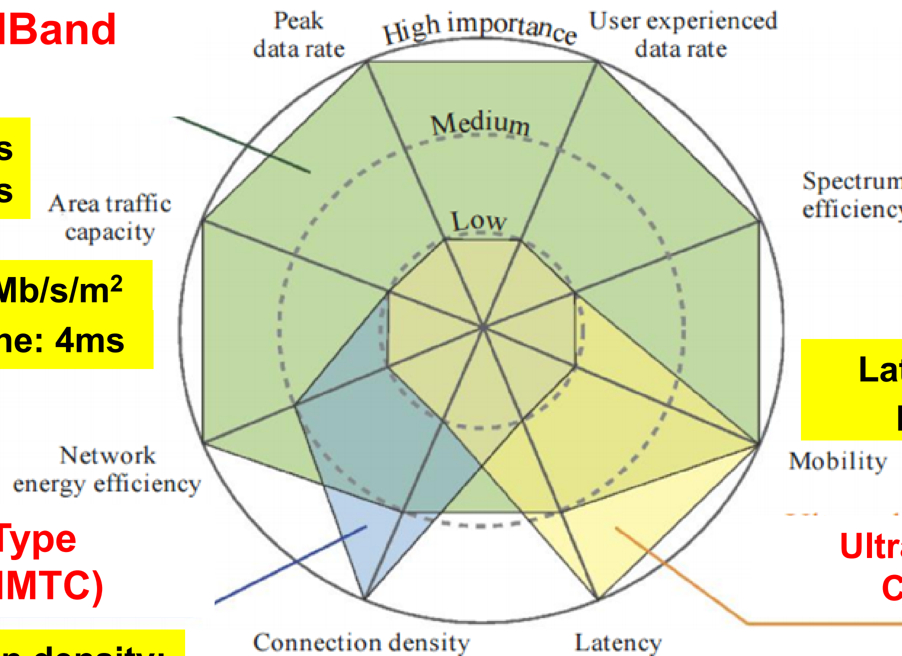
- 5G Evolution and F-RAN Architecture
- Performance Analysis of Access Model in F-RANs
- Performance Analysis of Edge Cache in F-RANs
- Resource Allocation Optimization in F-RANs
- Conclusions

Key Capabilities of 5G

Enhanced Mobile BroadBand (eMBB)

DL peak rate: 20 Gbit/s
UL peak rate: 10 Gbit/s

Area Capacity: 10 Mb/s/m^2
Latency of user plane: 4ms



Latency of user plane: 1ms

Massive Machine Type Communications (MMTC)

Connection density: 1 million per km^2

Ultra-Reliable and Low Latency Communications (URLLC)

Architecture evolution is urgent to meet performance requirements in three typical application scenarios

Architecture Evolution

Requirements driven

- 5G scenarios and KPI
- Operation enhancement
- Smooth evolution consideration

Technologies driven

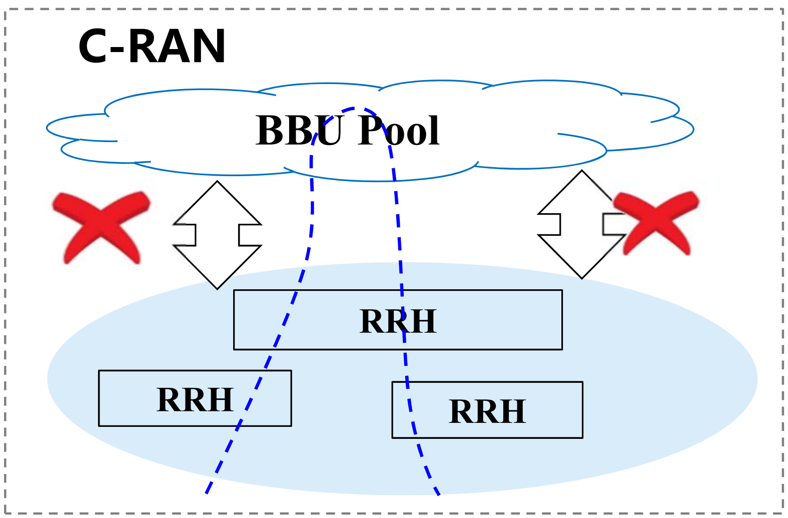
- **NFV**
  separation of software and hardware, provide flexible infrastructure platform
- **SDN**
  separation of control function and forwarding function, impact on architecture design

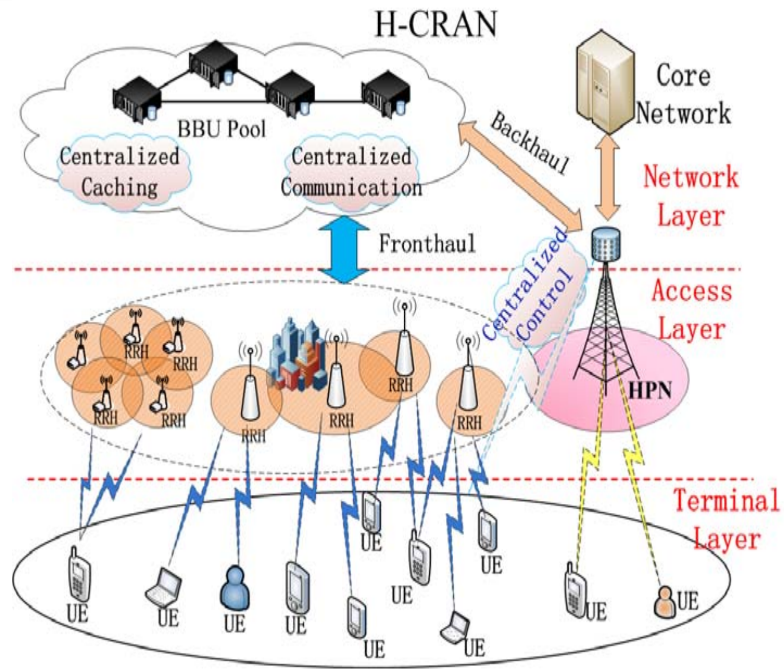
Driven by requirements and new IT technologies, 5G network can be re-constructed into diversified networking.

- Support diverse networking mode: C-RAN, D-RAN, mesh,D2D, BS plug-in
- AI is expected to increase the network's intelligence

5G Solution: C-RAN & H-CRAN

High SE/EE





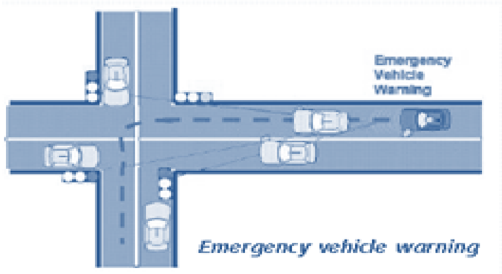
- **Decouple control plane from C-RANs into HPN**
- **HPN is used to alleviate the burdens of fronthaul links and support the seamless coverage**

Support eMBB, not M-MTC and URLLC

5G Challenges: V2X Example

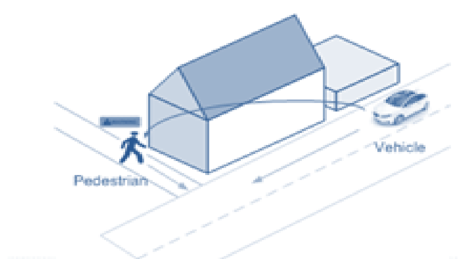
Vehicle to Everything (V2X) Generals

- **Safety, Traffic Efficiency, and Emergency Control**

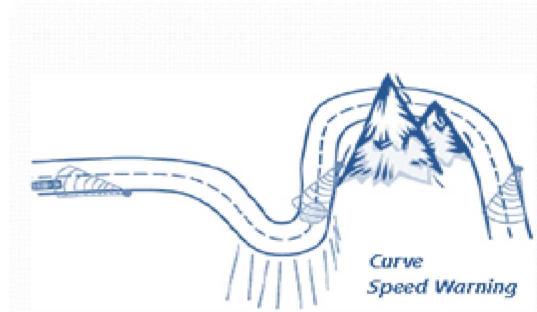


V2V Safety

V2V/V2I Traffic Efficiency



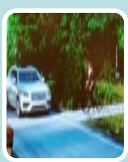
V2P Road Safety



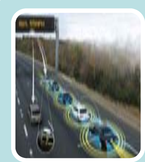
V2I Road Safety

- **Technical feature**
  - **Low latency: 20ms**
  - **High reliability: with the high-speed of 200km/h**

- **V2X Merits**



Beyond Visual Range
- See invisible
- Hear inaudible



Global Coordinating
- Clear intention exchange
- Global traffic optimization

Solution: Cloud to Fog Computing

2000 – 2015

2015 – 2030 ?

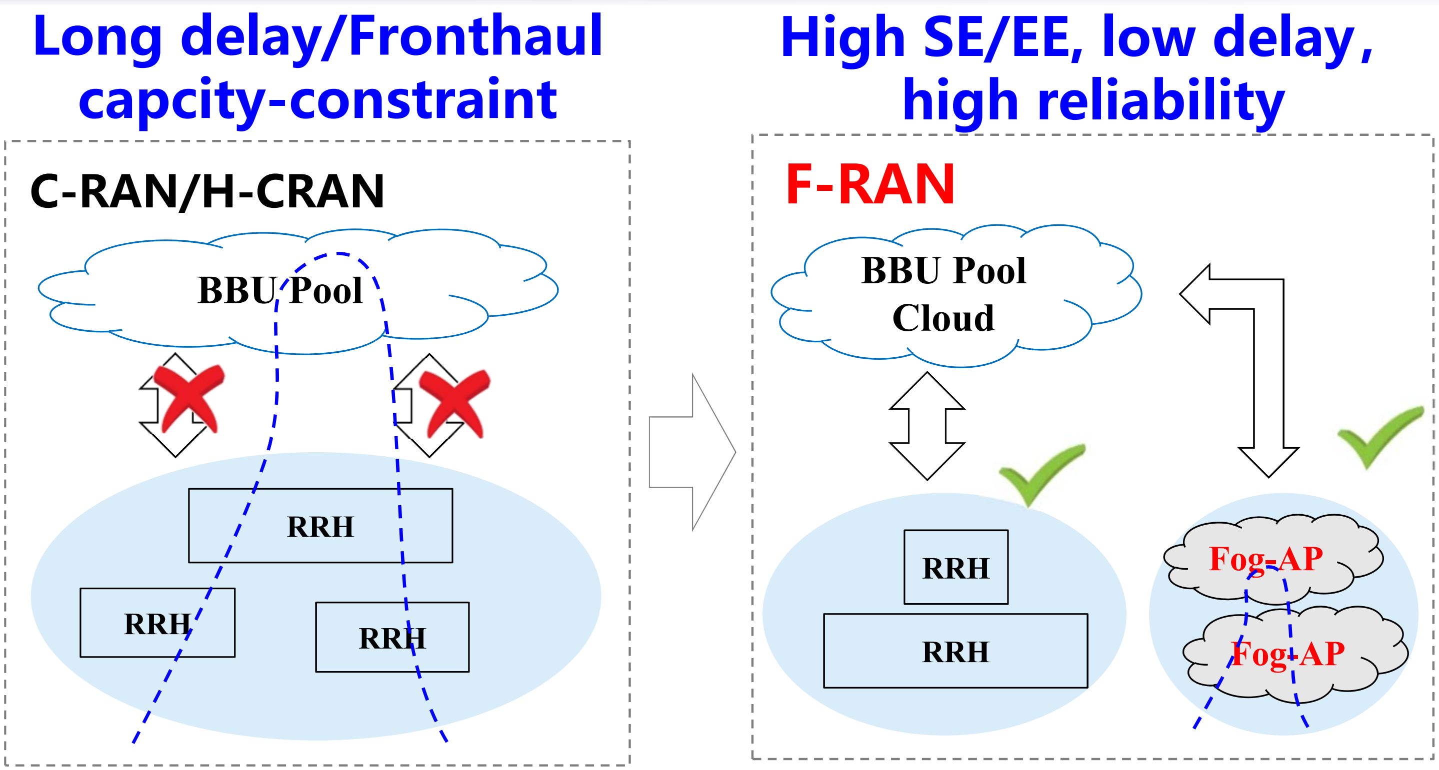
Prof. Mung Chiang (Princeton University) : Fog network architecture that uses one or a collaborative multitude of end-user clients or near-user edge devices to carry out a substantial amount of storage (rather than stored primarily in cloud data centers), communication (rather than routed over backbone networks), control, configuration, measurement and management (rather than controlled primarily by network gateways such as those in LTE core).

Source: fogresearch.org

Difference: Fog Computing & Network

- Fog Computing
  - Cloud computing is extended to the edge of the network.
  - To create a highly virtualized platform that provides compute, storage, and networking services between end-devices and traditional cloud data centers
- Fog Network
  - Use one or a collaborative multitude of end-user clients or near user edge devices to carry out storage, communications, control, configuration, measurement and management
- F-RAN (Fog Radio Access Network)
  - To improve SE/EE and decrease latency of C-RANs/H-CRANs
  - First proposed in 2015

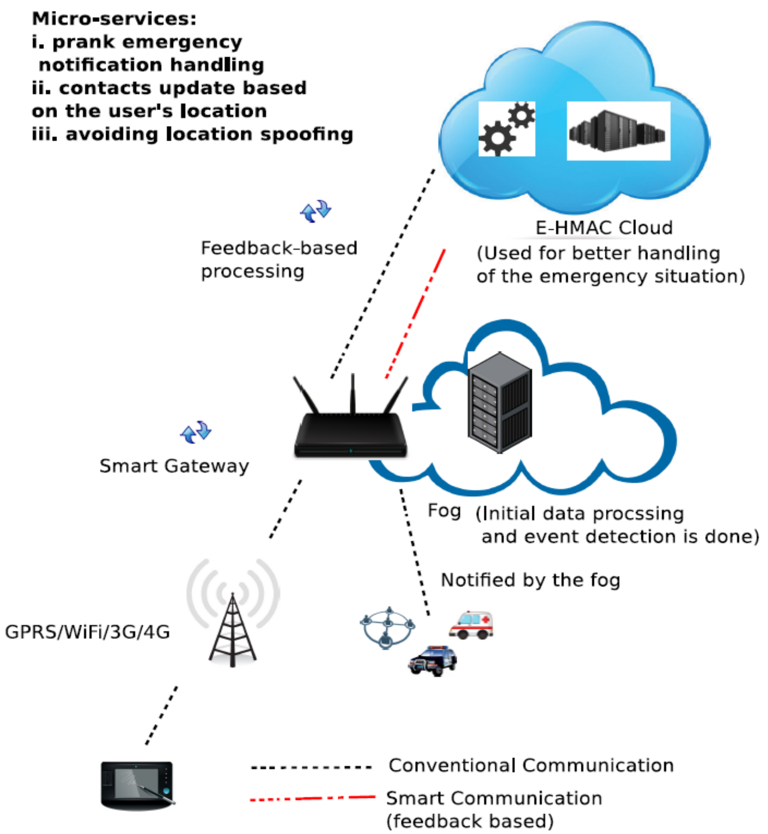
F-RAN: Architecture and Features



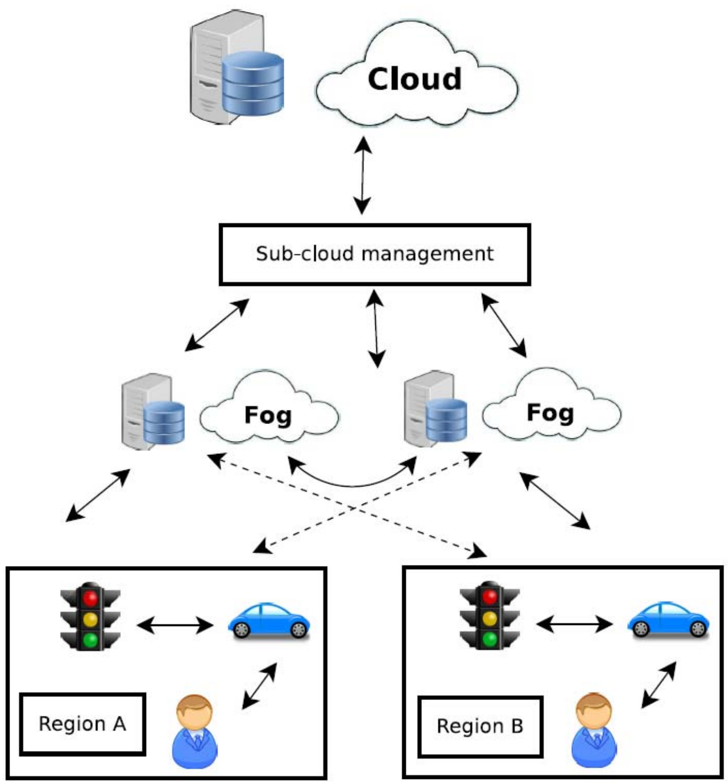
- F-RAN Characteristics
  - ✓C-RAN
  - ✓Small cells with caches
  - ✓Device-to-Device
  - ✓3G/4G Cellular
  - ✓Content Local Delivery

M. Peng, S. Yan, C. Wang, "Fog Computing based Radio Access Networks: Issues and Challenges", *IEEE Network Mag.*, submitted in Mar 2015, published in Jul. 2016.

F-RAN: Applications for IoT

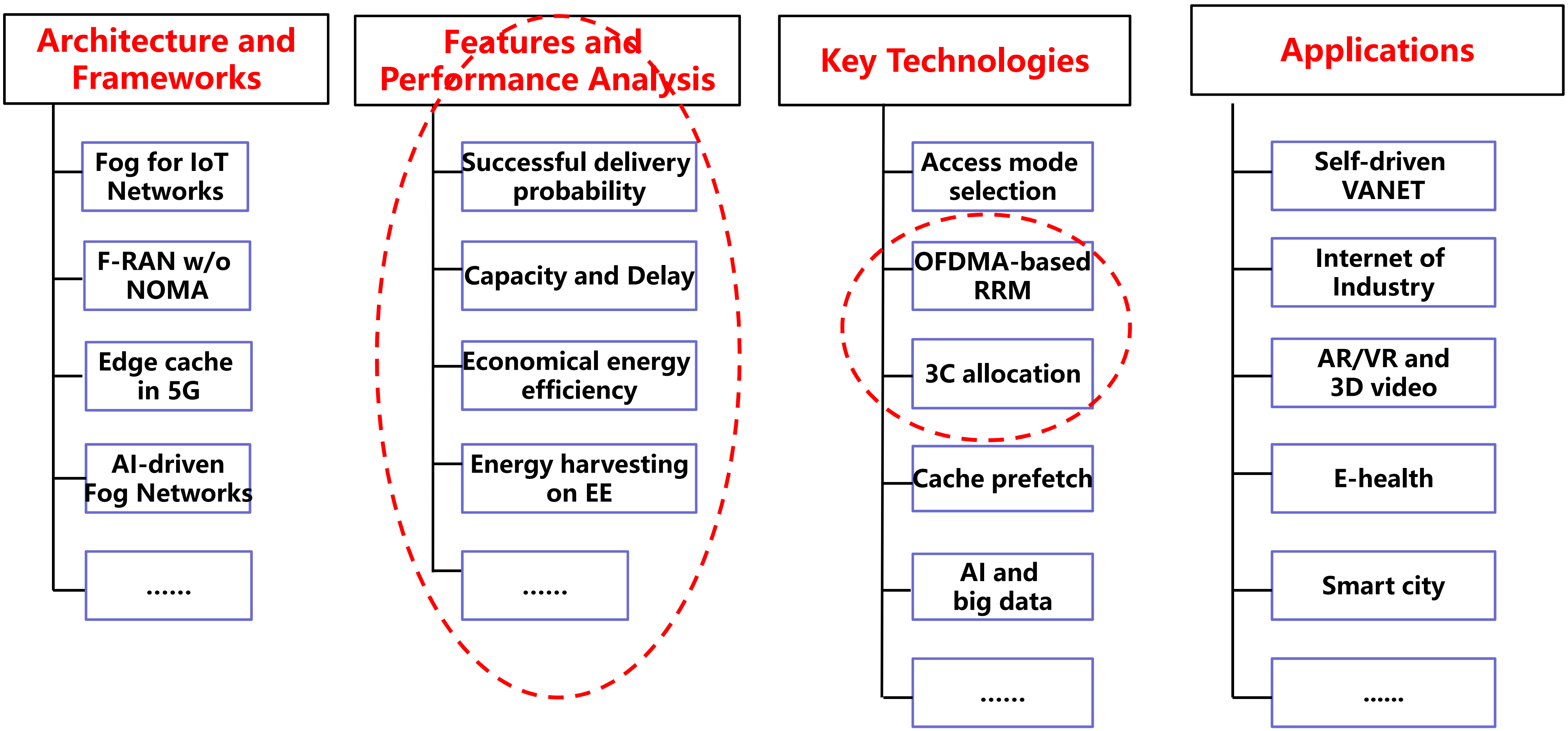


Smart Gateway based F-RAN for Emergency Control



Internet of Vehicular based on F-RANs

F-RAN: Research Framework



Outline

- 5G Evolution and F-RAN Architecture
- Performance Analysis of Access Model in F-RANs
- Performance Analysis of Edge Cache in F-RANs
- Resource Allocation Optimization in F-RANs
- Conclusions

Questions to be Addressed

- ✓ For a typical user, which kind of access modes is preferred when the edge cache is considered: D2D, best small cell BS, C-RAN, MBS?

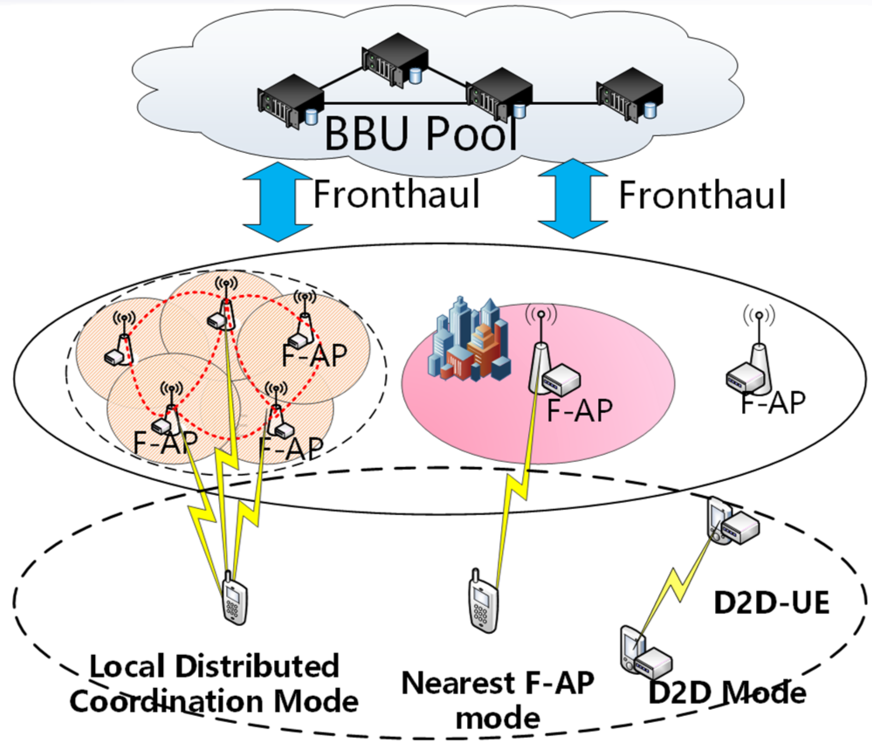
- ✓ Can we develop a mathematical performance analysis framework for access mode selection in F-RANs?

- ✓ What are the exact performance gains for each access mode selection?

Radio Access Model

System Architecture

- A downlink F-RAN with a group of F-APs working in parallel with several D2D users
- F-APs and D2D users are spatially distributed following two independent two-dimensional Poisson point process （PPPs）



SIR Calculation

Transmit Power
Channel Gain
Path loss

$$SIR(x_i \rightarrow u) = \frac{P_i h_i \|x_i\|^{-\alpha_i}}{\sum_{y\in\Phi_i \setminus x} P_y h_y \|x_y\|^{-\alpha_i} + \sum_{z\in\Phi_j} P_z h_z \|x_z\|^{-\alpha_j}}$$

Intra-tier Interference
Inter-tier Interference

Edge Cache Model

Assumptions

- Each D2D user and F-AP has a limited caching storage
- Only a small portion of *N* contents are frequently accessed by the majority of users

Zipf Distribution

$$f_i(\sigma, N) = \frac{1/i^{\sigma}}{\sum_{k=1}^{N} 1/k^{\sigma}}$$

Number of video content (N)

Zipf exponent (σ)

Content Caching Probability

$$p_c^D = \Pr(V \in C_d) = \sum_{i=1}^{C_d} f_i(\sigma_d, N)$$

$$p_c^F = \Pr(V \in C_f) = \sum_{i=1}^{C_f} f_i(\sigma_f, N)$$

Cache of D2D user (C_d)

Cache of F-AP (C_f)

Access Modes

D2D Mode:

- D2D mode is enabled when U can successfully obtain the requested content from another D2D user in a known location within a distance threshold, meanwhile, the SIR is larger than a pre-set SIR threshold T_d

$$\Psi_D = \{X_d : X_d \in \Phi_{du}, \|X\| \le L_d, V \in C_d, \gamma_d \ge T_d\}.$$

Nearest F-AP mode (i.e., Best small cell BS):

- When the desired user U does not support D2D mode, U tries to access the nearest F-AP which can respond to the desired user's content request, and the SIR should be larger than the SIR threshold T_f

$$\Psi_F = \{X_f : \arg\min_{X \in \Phi_f} (\|X\|), V \in C_f, \gamma_f \ge T_f,\ U \notin \Phi_{du} \cup V \notin C_d \cup \gamma_d < T_d,\}$$

Local distributed coordination mode (i.e., C-RAN):

- This mode means that U associates with multiple RRHs in a user-centric cluster with a radius *L*c.

$$\Psi_C = \{X_c : X \in \Phi_f, \forall X \in B(U, R) \cap \Phi_f\}$$

Performance Analysis of D2D Mode

D2D Mode:

- The coverage probability is defined as the probability that the desired user achieves an SIR higher than the target SIR threshold

$$P_D(T_d, \alpha_f, \alpha_d, \|X_d\|) = \Pr\left(\frac{P_d h_d \|X_d\|^{-\alpha_d}}{I_{d,du} + I_{f,du}} \geq T_d\right) = \exp\left(-\pi \|X_d\|^{\frac{2\alpha_d}{\alpha_f}}\left(\lambda_{du} + \left(\frac{P_f}{P_d}\right)^{\frac{2}{\alpha_f}} \lambda_f\right) C(\alpha_f) T_d^{\frac{2}{\alpha_f}}\right)$$

- The ergodic rate is defined as $R_x = p_x \mathrm{E}\left[\ln\left(1+\mathrm{SIR}(\mathrm{U} \to \mathrm{X}_x)\right) | \mathrm{SIR}(\mathrm{U} \to \mathrm{X}_x) > \mathrm{T}_x\right]$
  where p_x denotes the probability of U select the x mode

$$R_d = p_D \mathrm{E}\left[\ln\left(1+\gamma_d\right) | \gamma_d \geq T_d\right]$$

$$\approx p_D \ln(T_d) P_D(T_d, \alpha_f, \alpha_d \|X_d\|) - \frac{p_D \alpha_f}{2} \mathrm{Ei}\left[-T_d^{\frac{2}{\alpha_f}} \pi \|X_d\|^{\frac{2\alpha_d}{\alpha_f}}\left(\lambda_{du} + \left(\frac{P_f}{P_d}\right)^{\frac{2}{\alpha_f}} \lambda_f\right) C(\alpha_f)\right]$$

$$p_D = 1 - \exp(-\pi \lambda_{du} p_c^D L_d^2) \qquad C(\alpha) = \frac{2\pi \csc(2\pi/\alpha)}{\alpha}$$

Performance Analysis of F-AP Mode

Nearest F-AP Mode:

- The nearest F-AP mode will be triggered if U cannot meet the conditions of D2D mode
- The coverage probability of the nearest F-AP mode can be calculated as

$$P_F\left(T_f,\alpha_f,p_c^F\right)=\Pr\left(\frac{P_f h_f \left\|X_f\right\|^{-\alpha_f}}{I_{f,fu}+I_{d,fu}}\geq T_f\right)=\frac{1}{1+\rho\left(T_f,\alpha_f\right)+\frac{\lambda_{du}}{p_c^F\lambda_f}C\left(\alpha_f\right)\left(\frac{P_d T_f}{P_f}\right)^{2/\alpha_f}}$$

- The ergodic rate is $R_f=\int_{\ln(T_f)}^{\infty}p_F P_F(e^{\theta},\alpha_f,p_c^F)\mathrm{d}\theta+p_F\ln(T_f)P_F(T_f,\alpha_f,p_c^F)$
- Special Case: $T_f>1$, $\alpha_f=4$

$$R_f^{\alpha_f=4}=p_F\mathrm{E}\left[\ln\left(1+\gamma_f\right)\mid\gamma_f\geq T_f\right]=\frac{2p_F(2+\ln(T_f))}{\pi\sqrt{T_f}\left(1+\frac{\lambda_{du}}{p_c^F\lambda_f}\sqrt{\frac{P_d}{P_f}}\right)}$$

$$p_F=1-p_D P_D(T_d,\alpha_f,\alpha_d,\left\|X_d\right\|) \qquad C(\alpha)=\frac{2\pi\csc(2\pi/\alpha)}{\alpha}$$

Performance Analysis of C-RAN Mode

Local distributed coordination mode:

- The local distributed coordination mode is triggered if U don't meet the conditions of both the first two modes
- U associates to multiple F-APs near to it in a user-centric cluster with a distance threshold L_c
- The ergodic rate is defined as $\mathrm{E}[\ln(1+A)] = \int_0^\infty \frac{1}{z}\left(1-e^{-Az}\right)e^{-z}dz$

$$R_c = p_C \mathrm{E}\left[\ln\left(1+\frac{\sum_{c\in\Psi_C} P_f h_c \|X_c\|^{-\alpha_f}}{I_{f,cu}+I_{d,fu}}\right)\right] = p_C \mathrm{E}\left[\int_0^\infty \frac{e^{-z}}{z}\left(1-\exp\left(-\frac{z\sum_{c\in\Psi_C} P_f h_c \|X_c\|^{-\alpha_f}}{I_{f,cu}+I_{d,fu}}\right)\right)\mathrm{d}z\right]$$

$$= p_C \int_0^\infty \frac{1}{s}\exp\left(-\pi\lambda_{du} C(\alpha_f)(P_d s)^{\frac{2}{\alpha_f}}\right)\left[\exp\left(-2\pi p_c^F \lambda_f \int_R^\infty \frac{P_f s v}{v^{\alpha_f}+P_f s}\mathrm{d}v\right) - \exp\left(-\pi p_c^F \lambda_f C(\alpha_f)(P_f s)^{\frac{2}{\alpha_f}}\right)\right]\mathrm{d}s$$

$$p_C = p_f\left(1-P_F(T_f, \alpha_f, p_c^F)\right) \qquad C(\alpha) = \frac{2\pi\csc(2\pi/\alpha)}{\alpha}$$

Adaptation Mode Selection

User Access Mode Selection Scheme

- Based on communication distance, nodes location, QoS requirements, and caching capabilities, an adaptive mode selection scheme is to take full advantages of these three modes.

D2D mode

The nearest F-AP mode

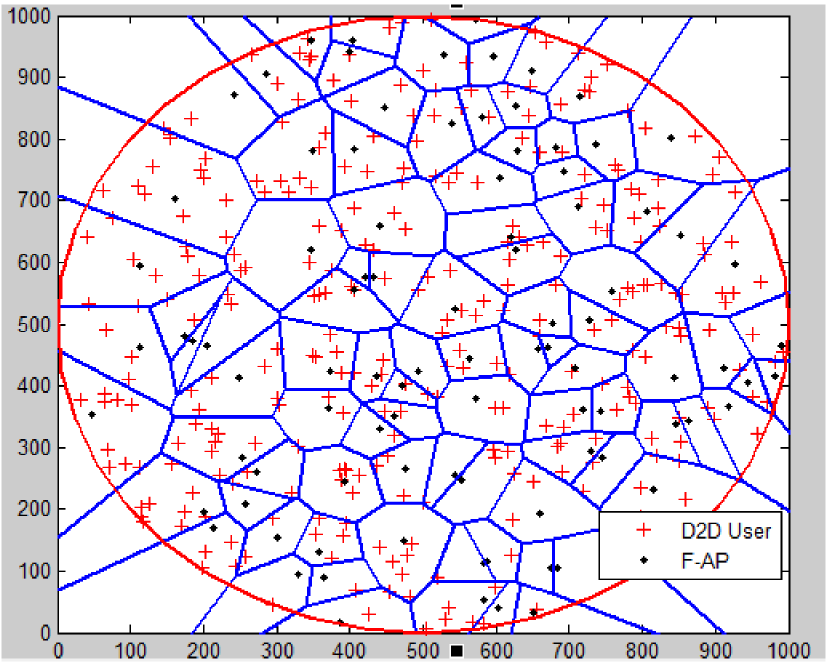
Local distributed coordination mode

Algorithm 1 User Access Mode Selection Mechanism

Initialize $\Psi_D = \emptyset, \Psi_F = \emptyset, \Psi_C = \emptyset$.
Step 1 Check the cache content of another D2D user nearby the desired user U with a radius threshold L_d, $B(U, L_d) \cap \Phi_{du} = \{X_1, X_2, ..., X_D\}$.
for $i = 1, 2, ..., D$ do.
  if $V \in C_d^i$
    Calculate the SIR γ_d^i from (7).
    if $\gamma_d^i \geq T_d$.
      $X_i \in \Psi_D$ User select D2D mode with X_i.
    end if
  else
    go to **Step 2**
  end if
end for
Step 2 Check the cache content of the F-AP nearby the desired user $U, \Phi_f = \{X_1, X_2, ..., X_F\}$.
if $V \in C_f^j$
Set $\|X_f\| = \|X_1\|$.
  for $j = 1, 2, ..., F$ do.
    if $\|X_j\| < \|X_f\|$.
      Nearest node $\|X_f\| = \|X_j\|$.
    end if
  end for
end if
Calculate the SIR γ_f with X_f from (8).
if $\gamma_f \geq T_f$.
  $X_f \in \Psi_F$ User select 1 FAP mode with X_f.
else
  User select local distributed coordination mode with $\Psi_F = B(U, L_f) \cap \Phi_f = \{X_1, X_2, ..., X_D\}$.
end if

Simulation Parameters

Monte Carlo simulation method

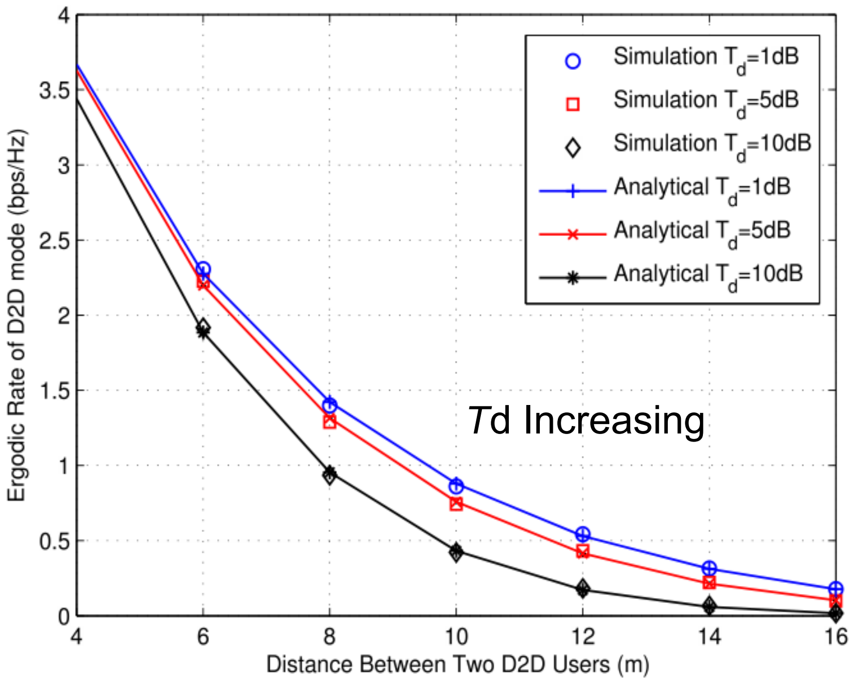


Poisson distributed F-APs and D2D users. The F-AP coverage boundaries form a Voronoi tessellation

SIMULATION PARAMETERS

Parameters	Value
Number of video content N	1000
Caching size of D2D user C_d	50
Caching size of F-AP C_f	$200 \sim 800$
Intensity of D2D users $p\lambda_u$	1×10^{-3}
Intensity of F-AP nodes λ_f	$1 \times 10^{-4} \sim 1 \times 10^{-3}$
Path loss exponent α	4 [15]
D2D user Zipf exponent σ_d	0.8
F-AP Zipf exponent σ_f	1
Transmit power of D2D user P_d	3dBm [16]
Transmit power of F-AP P_f	23dBm
Cluster distance threshold L_c	$20 \sim 50$m

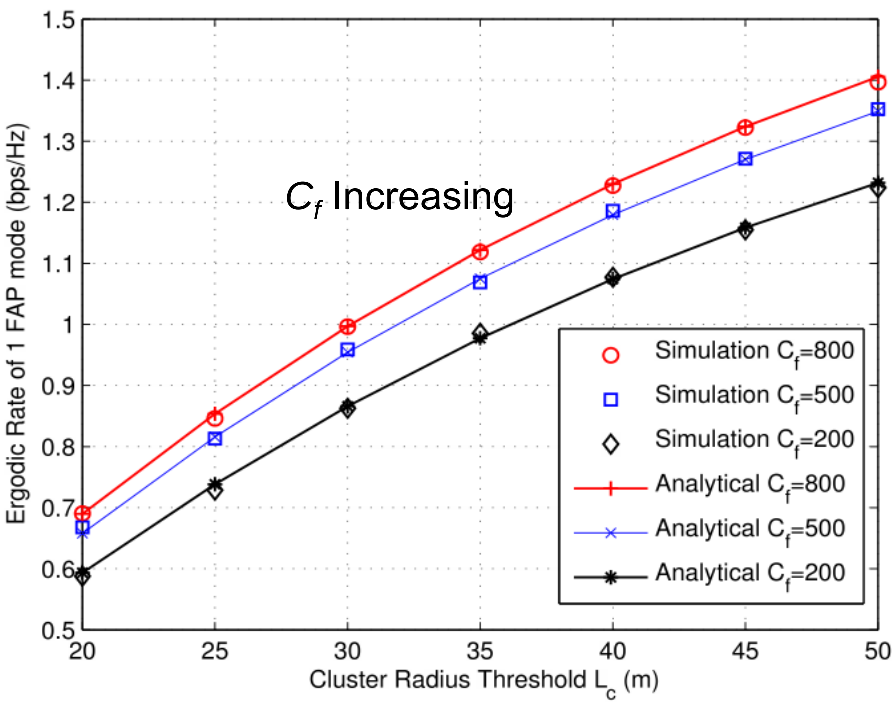
Simulation Results (1)



The ergodic rate of D2D mode with different SIR thresholds versus distance between two D2D users

- The analytical results closely match with the corresponding simulation results
- The ergodic rate of D2D mode decreases as the distance between D2D pair increases
- The larger *Td* means less user selects D2D mode, i.e., D2D mode selection strictly depends on SIR threshold, and 5-10 dB is preferred.

Simulation Results (2)



The ergodic rate of local distributed coordination mode with different F-AP cache size versus cluster radius threshold

- The number of F-APs in the cluster increases with cluster radius threshold, which leads to the increase of cluster size and results in the improvement of ergodic rate
- The larger cache size of F-AP C_f suggests that there are more opportunities for the desired user to get the video content it needs, which leads to a higher ergodic rate.

Conclusions

- ✓ The **outage probability, ergodic rates** of D2D mode, nearest F-AP mode, local distributed coordination mode are characterized, in which intra-tier and inter-tier interference, and distributed cache are considered.
- ✓ Based on the proposed performance metrics, the impacts of the **cache size, user node density, and QoS constraints** are characterized.
- ✓ An **adaptation access mode selection mechanism** is proposed to improve performance. The Monte Carlo simulation results evaluate impacts of the F-AP nodes density, SIR threshold, cache size and association schemes on the ergodic rate.

Outline

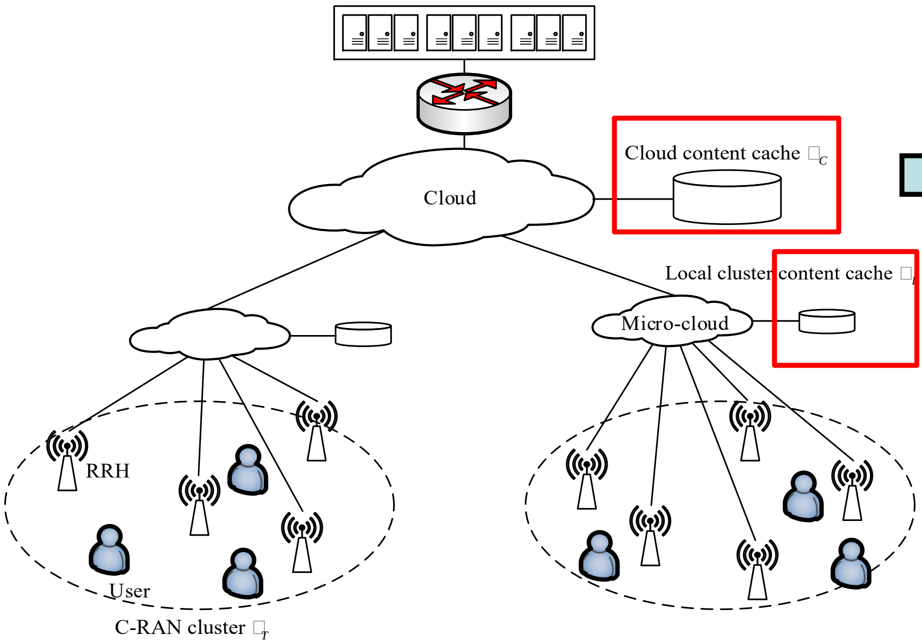
- 5G Evolution and F-RAN Architecture
- Performance Analysis of Access Model in F-RANs
- Performance Analysis of Edge Cache in F-RANs
- Resource Allocation Optimization in F-RANs
- Conclusions

Questions to be Addressed

- ✓ Propose an advanced edge cache architecture to tackle the disadvantage of homogeneous centralized cache ?
- ✓ Can we develop a mathematical performance analysis model for hierarchal cache in F-RANs?
- ✓ How much are the exact performance gains for the proposed hierarchal cache?

Local Cluster Caching Model

System Model



Stores all the content objects
$\Omega_C = \{S_1, \ldots, S_L\}$

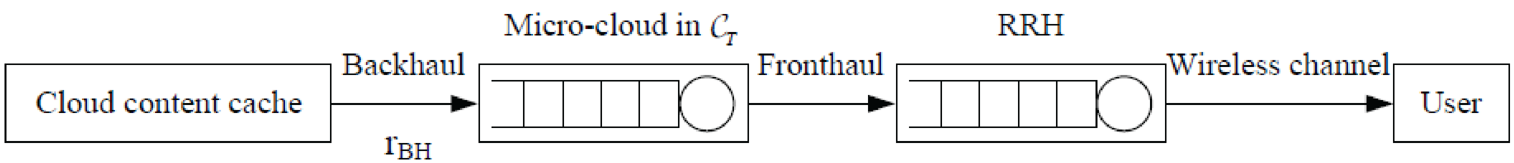
Stores some content objects
$\Omega_T = \{S_{T_1}, \ldots, S_{T_K}\} \subseteq \Omega_C$

Receive SINR: $\gamma = \dfrac{\rho d_m^{-\beta} |h_m|^2}{\sum_{j \neq m} \rho d_j^{-\beta} |h_j|^2 + \sigma^2}$

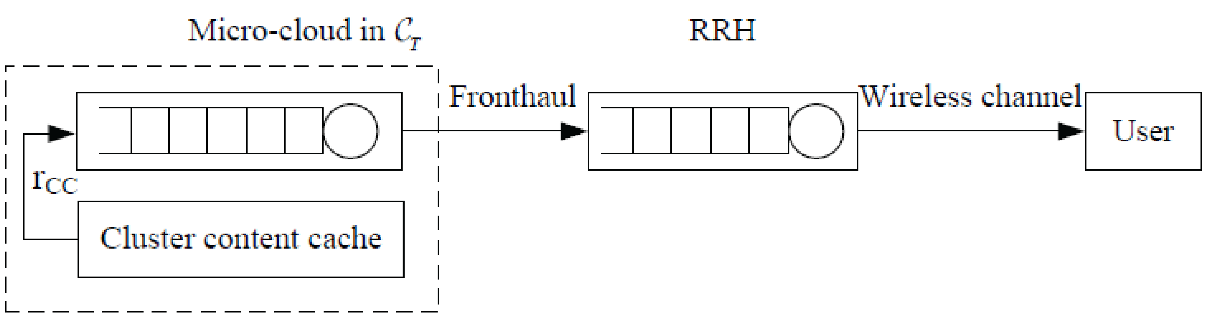
- ✓ Nodes are modeled as a homogenous PPP Φ_R with density λ_R
- ✓ Users are modeled as a homogenous marked PPP $\Phi_U(M_n)$ with density λ_U
- ✓ M_n denotes the type of content U_n requires

System Model for Edge Cache

- **The content requests from UEs are aggregated at the edge cache in C_T**
  - C_T checks its edge cache U_T , and can be served immediately if the desired content is available at U_T. Otherwise, the requests will be forwarded to the centralized cloud through the backhaul link



Cloud Cache



Edge Cache

Effective Capacity with Delay-Rate

- **The received signal for a typical user u_T in C_T can be expressed as**

$$y_T = \sqrt{\rho} h_m d_m^{-\beta/2} s_m + \sum_{j \neq m} \sqrt{\rho} h_j d_j^{-\beta/2} s_j + n_T,$$

- **The Shannon capacity with unit bandwidth:**

$$C = \mu \log(1+\gamma), \text{ where } \gamma = \frac{\rho d_m^{-\beta} |h_m|^2}{\sum_{j \neq m} \rho d_j^{-\beta} |h_j|^2 + \sigma^2},$$

μ denotes the spectral efficiency that is inversely proportional to the number of occupied orthogonal radio resource units for content transmissions in $\mathcal{C}_T$

- **The effective capacity is defined as a log-moment generation function**

$$E(\theta) = -\lim_{t \to \infty} \frac{1}{\theta t} \log \mathbb{E}\{e^{-\theta S(t)}\}$$

where $S(t) = \sum_{0=t_0<t_1<\cdots<t_n=t} \int_{t_{i-1}}^{t_i} r(\tau) \mathrm{d}\tau$ is the transmitted service

Effective Capacity with Delay-Rate

- **In each time unit, the effective capacity can be further derived as**

$$E(\theta) = -\frac{1}{\theta T} \log \mathbb{E}\{e^{-\mu\theta T C}\} \stackrel{(a)}{=} -\frac{1}{\theta \bar{T}} \ln \mathbb{E}\{(1+\gamma)^{-\mu\theta\bar{T}}\}, \quad \text{where } \bar{T} = T/\ln 2$$

Proposition 1: (Proposition 5, [25]) Assume that a network carries packetized traffic, and consists of N_{h} hops. Given an external arrival process with constant data rate r and constant packet size B, the end-to-end delay D experienced by the traffic traversing the network can be expressed as

$$\lim_{D_{\max}\to\infty} \frac{\log \Pr\{D > D_{\max}\}}{D_{\max} - (N_{\mathrm{h}}B)/r} = -\theta,$$

[25] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," *J. ACM Mobile Networks and Applications*, vol. 11, no. 1, pp. 91-99, Feb. 2006.

Performance Evaluation

- **Effective capacity of a typical user：**

$$E_{i,m}(\theta_j, d_m) = -\frac{1}{\theta_j \bar{T}} \ln(\mathcal{G}(\theta_j, d_m)),$$

where

$$\mathcal{G}(\theta_j, d_m) = \sum_{n=1}^{N} \left(e^{-2\pi A(\beta)\gamma_n^{\frac{2}{\beta}} \lambda_{\mathrm{R}} d_m^2 - \frac{\gamma_n d_m^{\beta} \sigma^2}{\rho}} - e^{-2\pi A(\beta)\gamma_{n+1}^{\frac{2}{\beta}} \lambda_{\mathrm{R}} d_m^2 - \frac{\gamma_{n+1} d_m^{\beta} \sigma^2}{\rho}} \right) (1 + \bar{\gamma}_n)^{-\mu\theta_j \bar{T}}$$

- **Average effective capacity of a typical cluster：**

$$\bar{E}_T = P_{\text{hit}} \sum_{l=1}^{L} \bar{E}(\theta_l^{\mathrm{T}}) + (1 - P_{\text{hit}}) \sum_{l=1}^{L} \bar{E}(\theta_l^{\mathrm{C}})$$

Hit ratio

where

$$\bar{E}(\theta_l) = P_l \sum_{n=1}^{N} \left[\mathcal{L}_l(\gamma_n) - \mathcal{L}_l(\gamma_{n+1})\right] (1 + \bar{\gamma}_n)^{-\mu\theta_l \bar{T}}, \ \theta_l = \theta_l^{\mathrm{T}}, \theta_l^{\mathrm{C}}$$

popularity

and

$$\mathcal{L}_l(\gamma_n) = 1 - 2\pi\lambda_l \int_0^{\infty} d_m e^{-(2\pi A(\beta)\gamma_n^{\frac{2}{\beta}}(\lambda_{\mathrm{R}} - \lambda_l) + \pi\lambda_l u(\gamma_n, \beta) + \pi\lambda_l) d_m^2} e^{-\frac{\gamma_n d_m^{\beta} \sigma^2}{\rho}} \mathrm{d}d_m$$

Optimization Problems

⌖ Two important factors:

- ✓ The conditions of radio access links
- ✓ Where to get content (Edge/centralized)

⇨

- ✓ Resource Block (RB) allocation
- ✓ RRH/F-AP association

Main problems:

- ➢ RB allocations and RRH/F-AP association are coupled tightly
- ➢ Centralized strategy is not applicable:
  - Just has local information
  - Global optimization is NP-hard

Optimization Solution

- **RRH/F-AP association:**

$$\phi_k(\mathcal{R}_{j_m}) = \begin{cases} \bar{E}(\mathcal{R}_{j_m} \cup \mathrm{R}_k) - \bar{E}(\mathcal{R}_{j_m}) - \underbrace{c_{\mathrm{RH}}\left(P_{\mathrm{R}}^{\mathrm{act}} + \frac{1}{\mathcal{O}(\mathcal{R}_{j_m})} P_{\mathrm{T}}\right)}_{\text{The cost part } \tau_k}, & S_{j_m} \text{ is in } \mathcal{U}_T \\ \bar{E}(\mathcal{R}_{j_m} \cup \mathrm{R}_k) - \bar{E}(\mathcal{R}_{j_m}) - \underbrace{c_{\mathrm{RH}}\left(P_{\mathrm{R}}^{\mathrm{act}} + \frac{1}{\mathcal{O}(\mathcal{R}_{j_m})} P_{\mathrm{C}}\right)}_{\text{The cost part } \tau_k}, & S_{j_m} \text{ is in } \mathcal{U}_C \end{cases}$$

Increment of effective capacity when R_k serves S_{jm}

Power consumption

Hedonic coalition formation

- **RB allocation:**

$$\psi(\mathcal{T}_i) = \left[\sum_{S_{i_n} \in \mathcal{T}_i} \bar{E}(\mathcal{R}_{i_n}) - \underbrace{c_{\mathrm{RB}}\left(\sum_{S_{i_n} \in \mathcal{T}_i} \mathcal{O}(\mathcal{R}_{i_n}) P_{\mathrm{R}} + \mathcal{O}(S_{i_n} \in \mathcal{U}_T) P_{\mathrm{T}} + \mathcal{O}(S_{i_n} \in \mathcal{U}_C) P_{\mathrm{C}}\right)}_{\text{The cost part } \varrho_i}\right]^+$$

Effective capacity of contents using RB_i

Power consumption

Merge and split algorithm

- **Maximized utility function**

$$\psi(\mathcal{T}_i) = \left[\sum_{S_{i_n} \in \mathcal{T}_i} v(\mathcal{R}_{i_n})\right]^+ = \left[\sum_{S_{i_n} \in \mathcal{T}_i} \sum_{\mathrm{R}_k \in \mathcal{R}_{i_n}} \phi_k(\mathcal{R}_{i_n})\right]^+$$

Nested coalition formation game

Resource Allocation

A nested coalition formation game-based algorithm:

Algorithm 2 (A nested coalition formation game-based algorithm)

Step 1. Joint allocation of RRB and RRH

Initialization: Formulate K disjoint coalitions of content objects $\mathcal{T}_1, \ldots, \mathcal{T}_K$ randomly, $1 \leqslant K \leqslant L$;

Repeat: For each content coalition $\mathcal{T}_i$ ($\mathcal{T}_i \neq \varnothing$)

1) Merge operation: Negotiate with other content coalitions, i.e., $\mathcal{T}_{j_1}, \ldots, \mathcal{T}_{j_l}$, $j_1, \ldots, j_l \neq i$;
   - Obtain the utility values of $\psi(\mathcal{T}_i)$, $\psi(\mathcal{T}_{j_1}), \ldots, \psi(\mathcal{T}_{j_l})$ and $\psi(\mathcal{T}_i \cup \mathcal{T}_{j_1} \cdots \cup \mathcal{T}_{j_l})$ based on (43) and Algorithm 1, in which the initialization of RRH partitions is given in (45), and both the RRH coalitions and their members follow the negotiation order defined in (44);
   - If $\psi(\mathcal{T}_i) + \sum_{p=1}^{l} \psi(\mathcal{T}_{j_p}) < \psi(\mathcal{T}_i \cup \mathcal{T}_{j_1} \cdots \cup \mathcal{T}_{j_l})$, $\mathcal{T}_i = \{\mathcal{T}_i \cup \mathcal{T}_{j_1} \cdots \cup \mathcal{T}_{j_l}\}$, $\mathcal{T}_{j_1} = \cdots = \mathcal{T}_{j_l} = \varnothing$;
2) Split operation: For each subset $\mathcal{T}_{i_n}^{\text{sub}}$ in $\mathcal{T}_i$
   - Obtain the utility values of $\psi(\mathcal{T}_i)$, $\psi(\mathcal{T}_{i_n}^{\text{sub}})$ and $\psi(\mathcal{T}_i / \mathcal{T}_{i_n}^{\text{sub}})$ based on (43) and Algorithm 1, in which the initialization of RRH partitions is given in (45), and both the RRH coalitions and their members follow the negotiation order defined in (44);
   - If $\psi(\mathcal{T}_{i_n}^{\text{sub}}) + \psi(\mathcal{T}_i / \mathcal{T}_{i_n}^{\text{sub}}) > \psi(\mathcal{T}_i)$, $\mathcal{T}_i = \mathcal{T}_i / \mathcal{T}_{i_n}^{\text{sub}}$, and formulate a new content coalition $\mathcal{T}_k = \mathcal{T}_{i_n}^{\text{sub}}$;

Termination: When the members of each coalition do not change.

Step 2. Identify RRHs that are not required by any user, and put them into sleep mode.

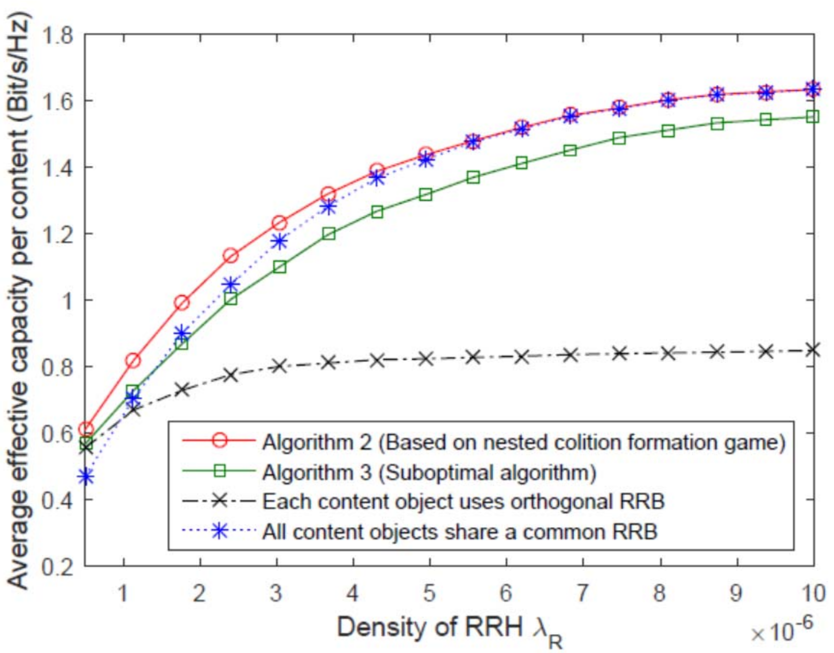
Merge and split operation for RB allocation

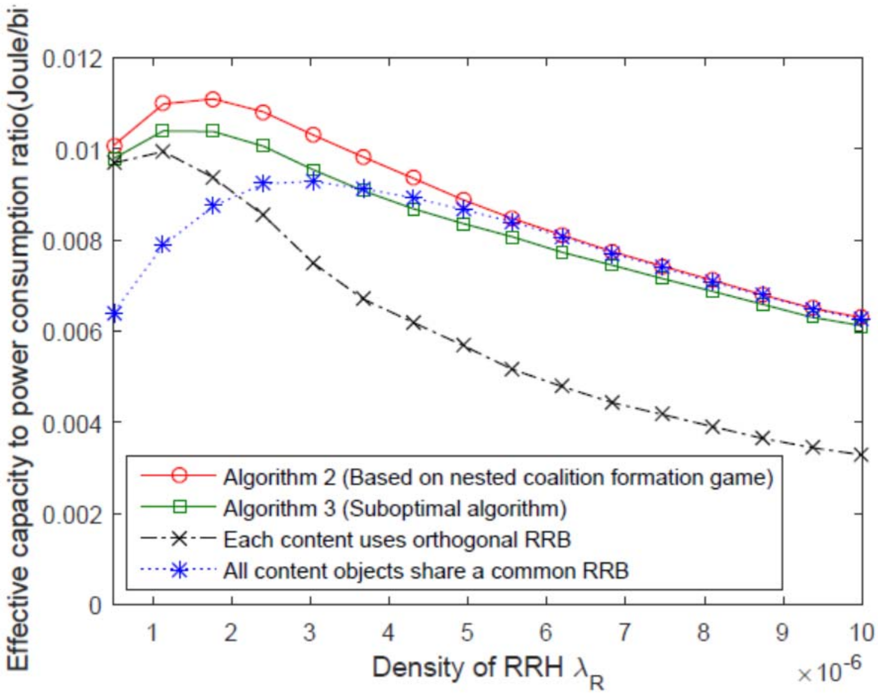
Hedonic coalition formation for RRH/F-AP association

Algorithm converges and D-hp stable

Simulation Results

- ✓ Proposed nested coalition formation Alg. vs. suboptimal Alg. vs orthogonal RB allocation vs. full RB reuse
- ✓ Effective capacity and energy efficiency





Outline

- 5G Evolution and F-RAN Architecture
- Performance Analysis of Access Model in F-RANs
- Performance Analysis of Edge Cache in F-RANs
- Resource Allocation Optimization in F-RANs
- Conclusions

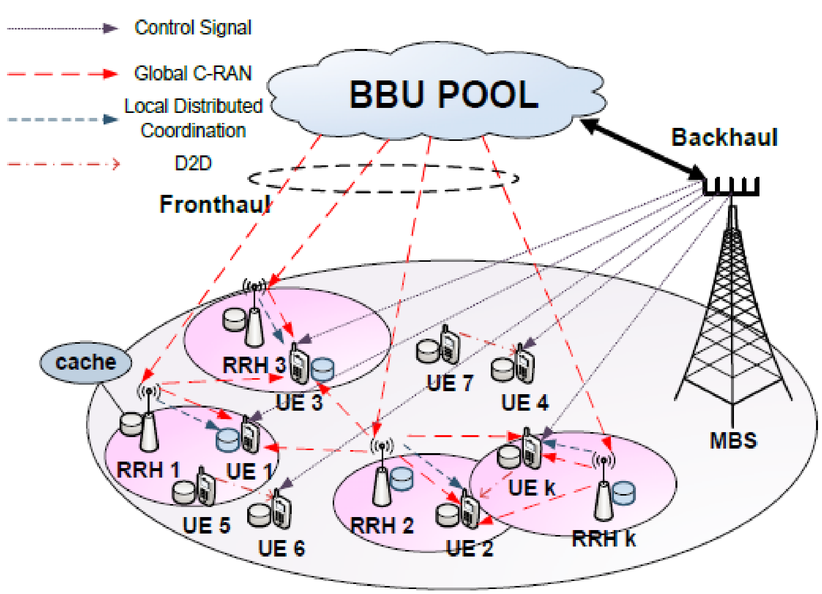
Questions to be Addressed

- ✓ How to design the joint mode selection and radio resource allocation to maximize SE/EE in OFDMA based F-RANs?
- ✓ NP-hard problem due to integer variables, and how to solve the non-convex optimization problem?
- ✓ Can the sub-optimal solution be proposed, and what is the exact performance gain?

System Model

F-RAN architecture

- MBS for control plane while F-APs for user plane
- Cache at F-APs and F-UEs enable the data transmission in the local area



- ✓ The downlink of an OFDM F-RAN supported D2D is considered, with N subchannels
- ✓ There are I video files in the cloud, part of each files are stored in the cache of M F-APs and K F-UEs (m,k)
- ✓ F-UEs (D2D pairs $\boldsymbol{D}$ and D2D-unable F-UEs $\boldsymbol{C}$) access the files via F-UEs and RRHs, respectively

Optimization Problem

- **Joint mode selection and resource allocation problem (MSRAP) to maximize EE in F-RANs :**

$$\max_{\{\mathbf{q}, \mathcal{X}(t), \mathcal{V}(t)\}} \quad \mathbb{E}\{\eta(t)\} \qquad (1)$$

$$
\begin{aligned}
C_1 &\quad \sum_{k=1}^{K} x_{k,n}^{\{0\}}(t) \le 1 && \forall n, \\
C_2 &\quad \sum_{k=1}^{K} x_{k,n}^{\{1\}}(t) \le 1 && \forall n, \\
C_3 &\quad \sum_{ss=0}^{1} x_{k,n}^{\{ss\}}(t) \le 1 && \forall k, n. \\
C_4 &\quad x_{k,n}^{\{0\}}(t) = 0 && \forall k \in \mathbf{C}, n, \\
C_5 &\quad x_{k,n}^{\{ss\}}(t) \in \{0, 1\} && \forall k, n, ss, \\
C_6 &\quad \mathbf{H}(t) \text{ are mean rate stable,} && \\
C_7 &\quad \mathbb{E}\{P_k^{\{0\}}(t)\} \le \tilde{P}_{\text{thres}}^{\{0\}} && \forall k \in \mathbf{D}, \\
C_8 &\quad \mathbb{E}\{P_m^{\{1\}}(t)\} \le \tilde{P}_{\text{thres}}^{\{1\}} && \forall m \in \{1, \ldots, M\}.
\end{aligned}
$$

Subcarrier allocation constraints (C_1–C_5)

Queue's length constraint (C_6)

Power allocation constraints (C_7, C_8)

Optimization Analysis

- **Based on the Lyapunov optimization and the concept of opportunistically minimizing an expectation, the supremum minimization of drift plus penalty becomes a new MSRAP:**

$$\max_{\{\mathcal{X}(t),\mathcal{V}(t)\}} \sum_{n=1}^{N}\Big\{\sum_{k=1}^{K}\alpha_k R_{k,n}^{\{0\}}(t)+\sum_{k=1}^{K}\alpha_k R_{k,n}^{\{1\}}(t) - \sum_{k=1}^{K}\beta_k P_{k,n}^{\{0\}}(t)-\sum_{m=1}^{M}\gamma_m P_{m,n}^{\{1\}}(t)\Big\} \quad (2)$$

$$s.t. \quad C_1,\ C_2\ ,C_3,\ C_4, C_5$$

where $\alpha_k = H_k(t)+\frac{V\alpha}{K}, \beta_k = F_k(t)+\frac{V\beta}{K}, \gamma_m = G_m(t)+\frac{V\beta}{M}$

Transferred Problems

Problem (1) can be decomposed into two subproblems

- ✓ Finding the optimal $\mathcal{X}(t)$ under the fixed $\mathcal{V}(t)$

$$\max_{\{\mathcal{X}(t)\}} \sum_{n=1}^{N} \Big\{ \sum_{k=1}^{K} \alpha_k R_{k,n}^{\{0\}}(t) + \sum_{k=1}^{K} \alpha_k R_{k,n}^{\{1\}}(t) - \sum_{k=1}^{K} \beta_k P_{k,n}^{\{0\}}(t) - \sum_{m=1}^{M} \gamma_m P_{m,n}^{\{1\}}(t) \Big\} \quad (3)$$

- ✓ Finding the optimal $\mathcal{V}(t)$ under the fixed $\mathcal{X}(t)$

$$\max_{\{\mathcal{V}(t)\}} \alpha_{k_0} R_{k_0,n_0}^{\{0\}}(t) + \alpha_{k_1} R_{k_1,n_0}^{\{1\}}(t) - \beta_{k_0} \| v_{k_0,n_0}(t) \|_2^2 - \sum_{m=1}^{M} \gamma_m \| \mathbf{D}_m \mathbf{v}_{k_1,n_0}(t) \|_2^2. \quad (4)$$

Solution to Problem 3

- **Problem (3) can be solved via particle swarm optimization**
  - ✓ Assume there are B particles, and the velocity of particles are

$$\mathbf{v}_b = (v_b^1, v_b^2, \ldots, v_b^{2N})$$

  - ✓ The mapping between the

$$x_{k,n}^{\{ss\}}(t) = \begin{cases} 1 & \text{if } k = \lfloor (K+1)x_b^n \rfloor, ss = 1, x_b^n \in (0,1), \\ 1 & \text{if } k = S_D(t) \lceil (|\mathrm{D}|+1)x_b^{N+n} + |\mathrm{C}| \rceil, \\ & \quad ss = 0, x_b^{N+n} \in (0,1), \\ 0 & \text{others}. \end{cases}$$

  - ✓ When F-UE accesses both the pairing F-UE and F-APs in one subchannel, which violates C_5, we force F-UE access the pairing F-UE via D2D.

Challenge 1 is solved by particle swarm optimization

Solution to Problem 4

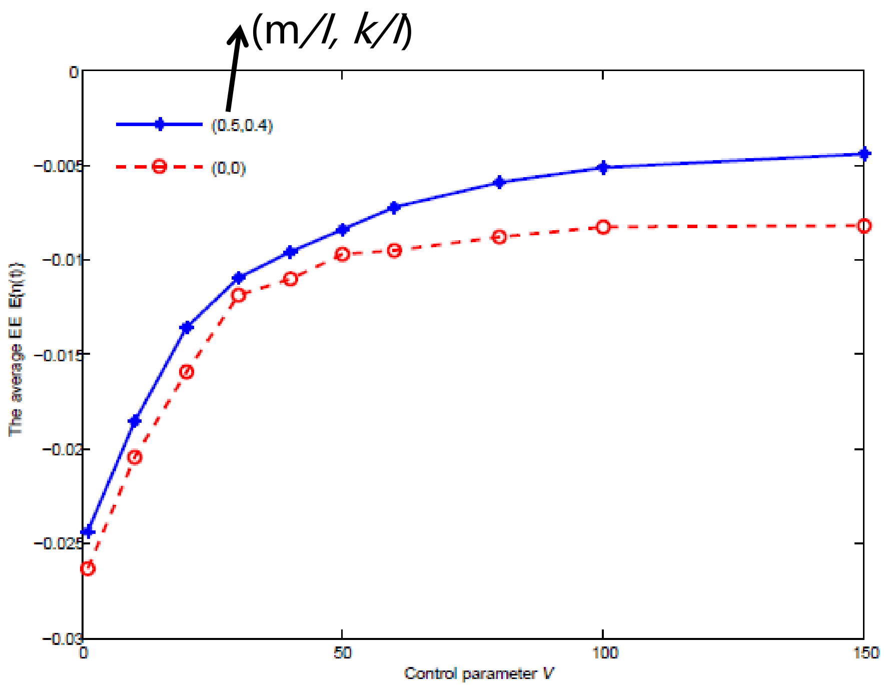
- **Problem (4) can be solved via WMMSE method**
  - ✓ Problem (4) has the same optimal solution as the following WMMSE minimization problem

$$
\begin{aligned}
\min_{\{\omega_{k,n_0}^{\{ss\}}, u_{k,n_0}^{\{ss\}}, \mathcal{V}(t)\}} \quad & \alpha_{k_0}\{\omega_{k_0,n_0}^{\{0\}} e_{k_0,n_0}^{\{0\}} - \log \omega_{k_0,n_0}^{\{0\}}\} \\
& + \alpha_{k_1}\{\omega_{k_1,n_0}^{\{1\}} e_{k_1,n_0}^{\{1\}} - \log \omega_{k_1,n_0}^{\{1\}}\} \\
& + \beta_{k_0}\|v_{k_0,n_0}(t)\|_2^2 \\
& + \sum_{m=1}^{M} \gamma_m \|\mathbf{D}_m \mathbf{v}_{k_1,n_0}(t)\|_2^2,
\end{aligned}
$$

  - ✓ Convex in each of the optimization variables and can be solved via the block coordinate descent method

Challenge 2 is solved by WMMSE method

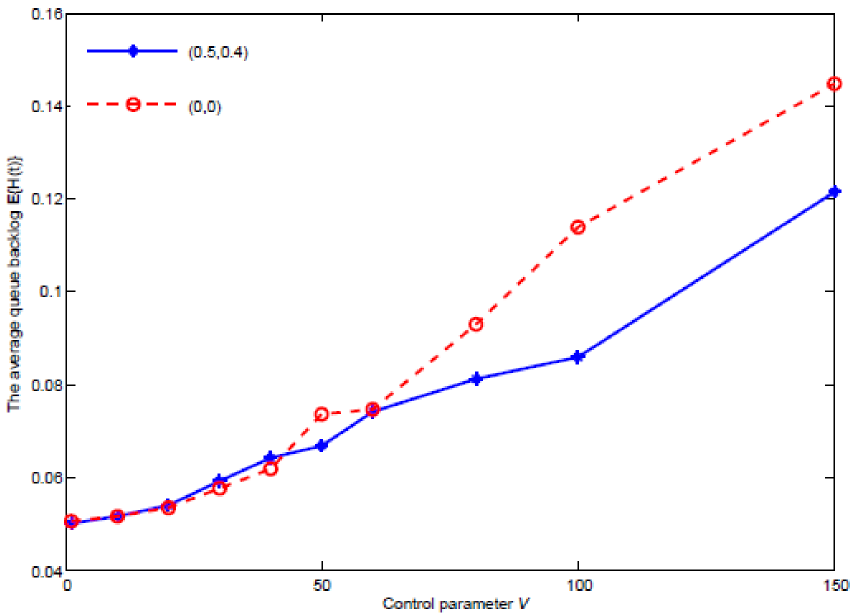
Simulation Results (1)



The average EE versus the control parameter V

- ✓ The average EE performance increases with V till saturation.
- ✓ It is shown that the F-RAN with cache owns a significant performance gain over traditional C-RANs as the incorporation of cache, especially when the fronthaul consumptions is taken into account.

Simulation Results (2)



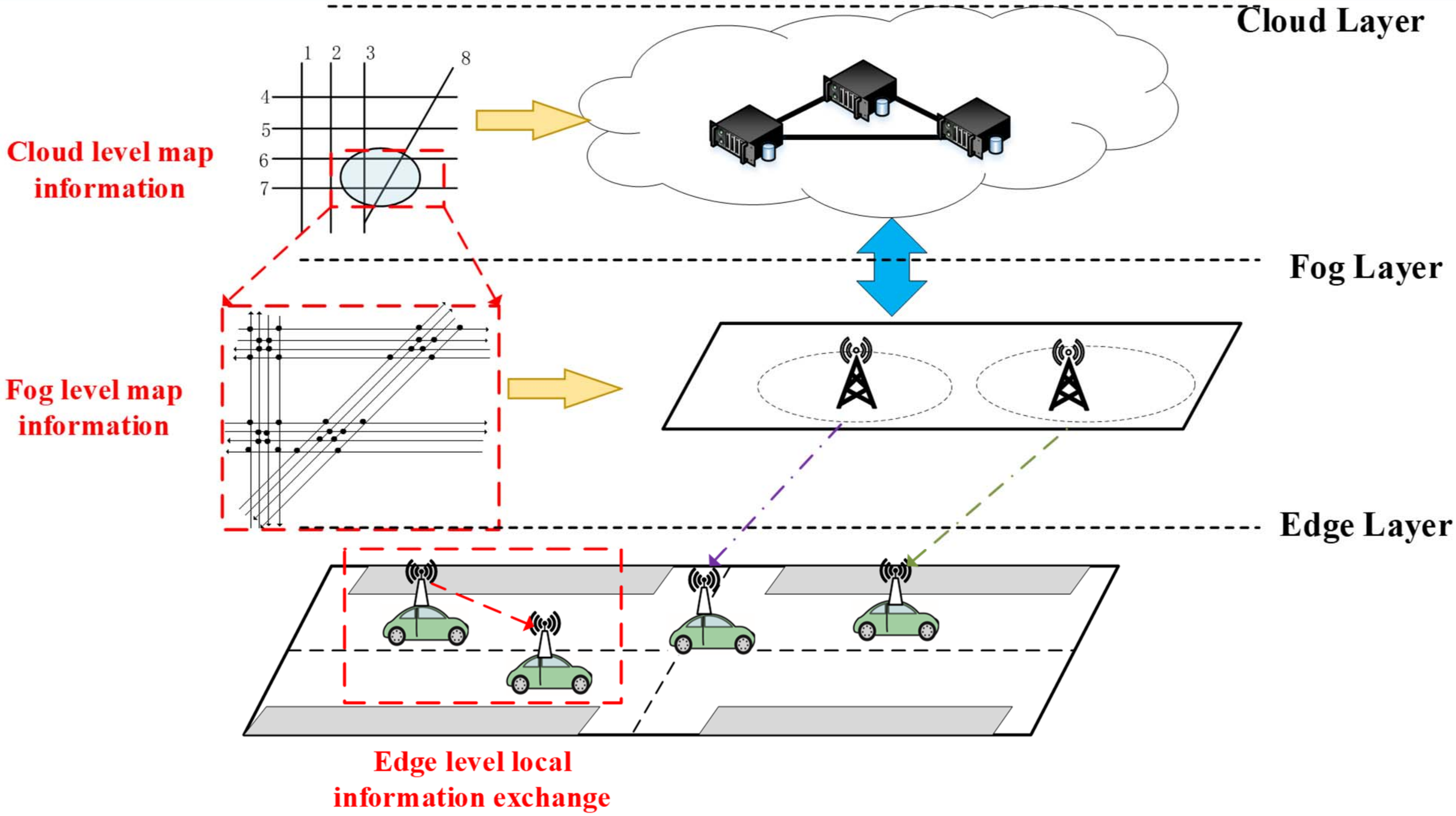
The average queue backlog versus the control parameter V

- ✓ The average queue backlog grows linearly in $\mathcal{O}(V)$. As a larger V leads more emphasizes on EE at the cost of incurring worse queueing delays.
- ✓ It is shown that the F-RAN outperforms the C-RAN in the aspect of delay due to the various access modes.

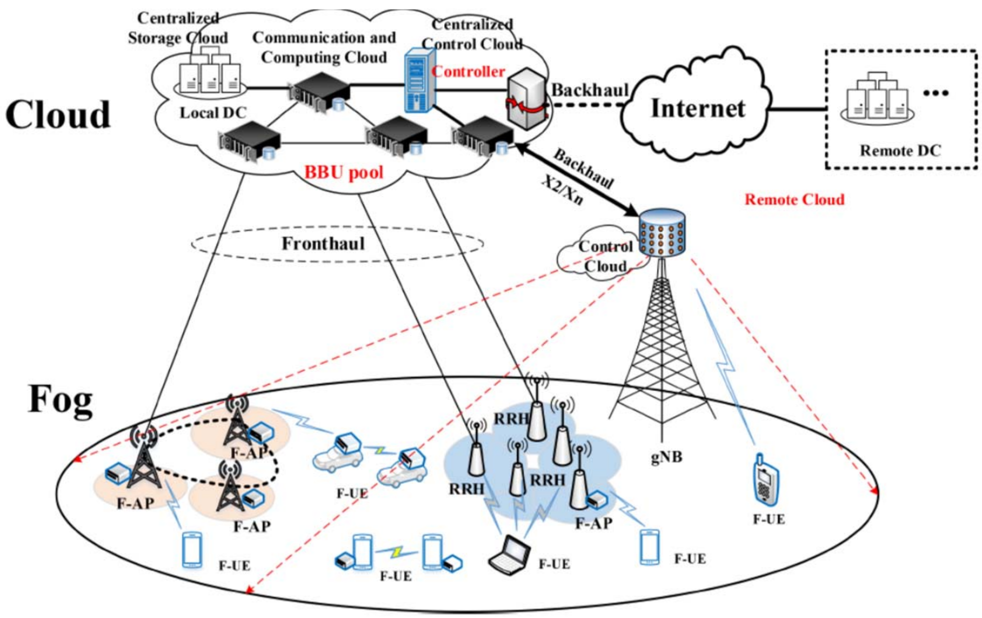
Outline

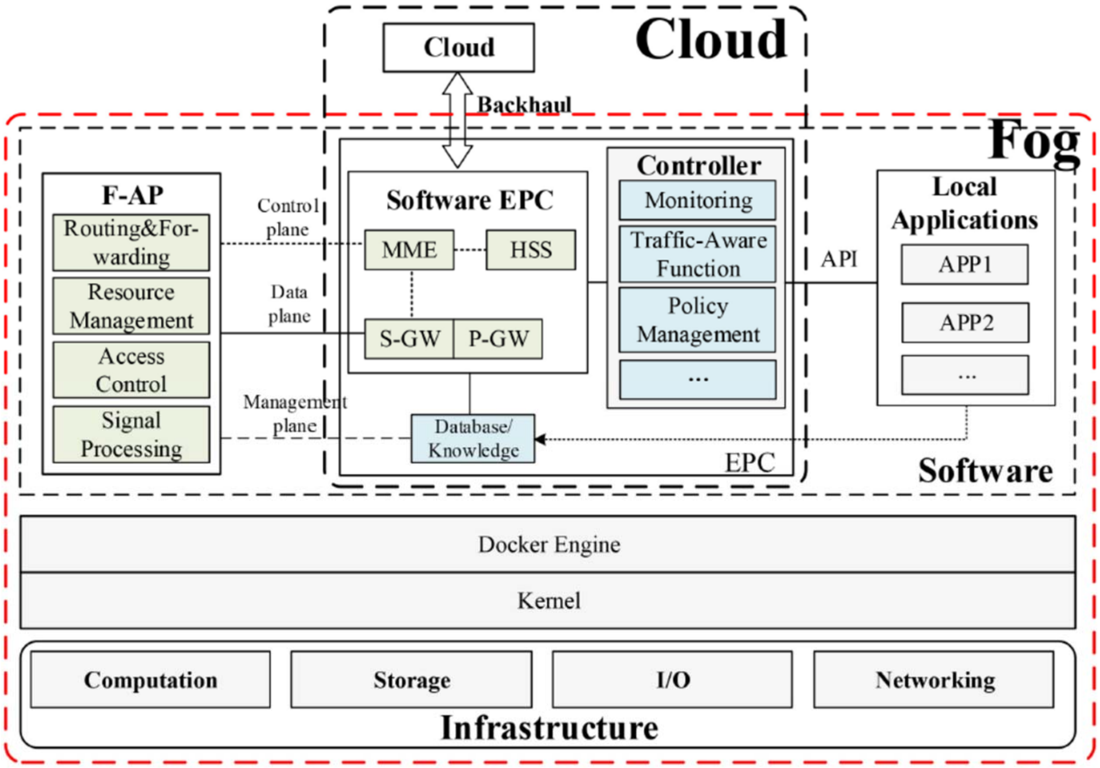
- 5G Evolution and F-RAN Architecture
- Performance Analysis of Access Model in F-RANs
- Performance Analysis of Edge Cache in F-RANs
- Resource Allocation Optimization in F-RANs
- Conclusions

F-RAN for Self-driven Vehicular Networks



OpenAirInterface (OAI)-based Hardware Testbed





Selected Related Publications

System Architecture

- "Fog Computing based Radio Access Networks: Issues and Challenges", *IEEE Network Mag.*
- "Network Slicing in Fog Radio Access Networks: Issues and Challenges", *IEEE Commun. Mag.*
- "Fully Exploiting Cloud Computing to Achieve a Green and Flexible C-RAN", *IEEE Wireless Commun.*
- "Recent Advances in Fog Radio Access Networks: Performance Analysis and Radio Resource Allocation", *IEEE Access*

Performance Analysis

- "Economical Energy Efficiency: An Advanced Performance Metric for 5G Systems", *IEEE Wireless Commun.*
- "Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks", *IEEE J. Sel. Areas Commun.*
- "A Non-Orthogonal Multiple Access-Based Multicast Scheme in Wireless Content Caching Networks", *IEEE J. Sel. Areas Commun.*
- "Outage Probability Analysis of Non-Orthogonal Multiple Access in Cloud Radio Access Networks", *IEEE Commun. Let.*
- "Channel Matrix Sparsity With Imperfect Channel State Information in Cloud Radio Access Networks", *IEEE Trans. Veh. Tech.*

Selected Related Publications

Energy Harvesting

- “Wireless-Powered Cooperative Communications: Power-Splitting Relaying with Energy Accumulation", *IEEE J. Sel. Areas Commun.*
- "Joint Power Splitting and Antenna Selection in Energy Harvesting Relay Channels," *IEEE Signal Processing Let.*

Radio Resource Allocation

- "Energy-Efficient Joint Congestion Control and Resource Optimization in Heterogeneous Cloud Radio Access Networks," *IEEE Trans. Veh. Tech.*
- "Energy-Efficient Resource Allocation Optimization for Multimedia Heterogeneous Cloud Radio Access Networks," *IEEE Trans.Multimedia*
- "Queue-Aware Energy-Efficient Joint Remote Radio Head Activation and Beamforming in Cloud Radio Access Networks," *IEEE Trans. Wireless Commun.*
- “Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks”, *IEEE Trans. Veh. Tech.*
- “An Evolutionary Game for User Access Mode Selection in Fog Radio Access Networks”, IEEE Access
- “Delay-Aware Uplink Fronthaul Allocation in Cloud Radio Access Networks”, *IEEE Trans. Wirele. Com.*
- “Cost-Efficient Resource Allocation in Cloud Radio Access Networks with Heterogeneous Fronthaul Expenditures”, *IEEE Trans. Wireless Commun.*

Questions & Comments

Thanks
for your attention