One6G – Open Lectures

Fairness scheduling and fronthaul optimization in cell-free user-centric scalable mMIMO networks

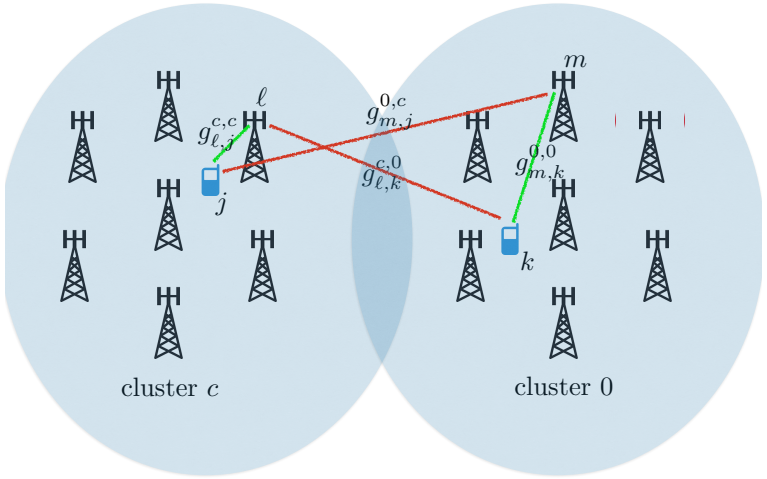
Giuseppe Caire

Communications and Information Theory Chair, EECS, TU Berlin

20 June, 2024

Joint Processing of Distributed Antennas

- [Wyner, TIT 1994]: centralized processing of all antennas in the uplink, Vector Gaussian MAC, capacity region was already known.
- [GC, Shamai, TIT 2003 – Weingarten, Steinberg, Shamai, TIT 2006]: Vector Gaussian BC, sum capacity and capacity region, in the downlink.
- Some past attempts: Coordinated MultiPoint (CoMP) .... not so successful, per-site (non-cooperative) massive MIMO has taken over the scene...
- Some successes: C-RAN, distributed antenna systems with joint processing, virtualization of the PHY/MAC in the CP.



Cell-Free User-Centric Wireless Networks (1)

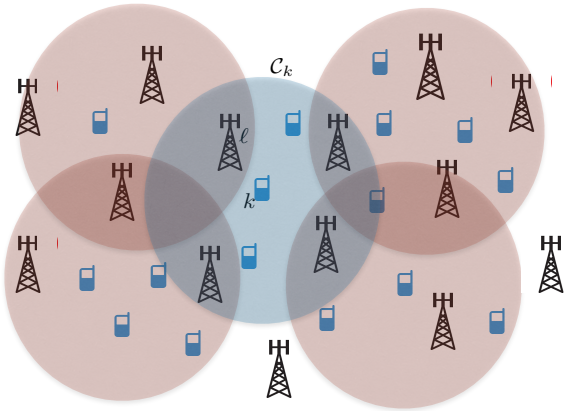
- Typically operating in FR1 (sub-6GHz), TDD reciprocity, UL/DL duality, pilot contamination/decontamination, linear precoding/detection.

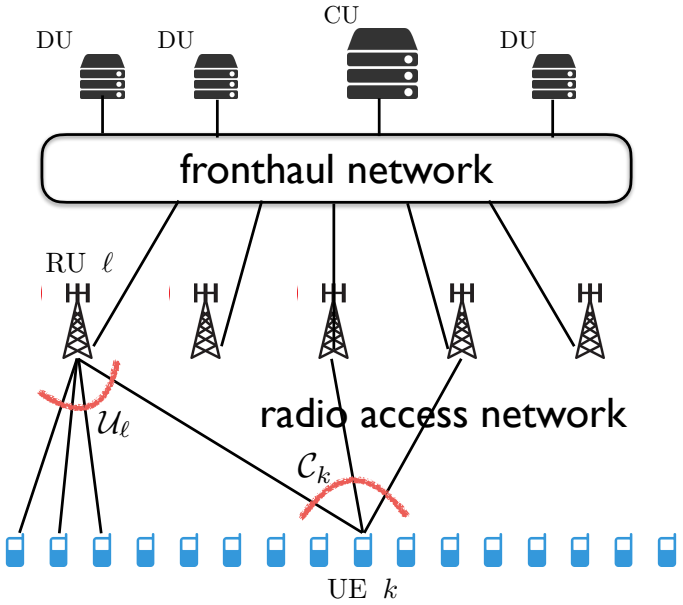
- Expected to become central in 6G systems operating in FR3 (7-24 GHz).

- Ultra-dense scenarios: campus networks, super-high spectral efficiency ... e.g. a sport arena with 10,000 users, on a 20-60 MHz bandwidth, served by 20 RUs with 10 antennas each, achieving ~ 50 bit/s/Hz per 10×10 m^2.

Cell-Free User-Centric Wireless Networks (2)

- Each UE is served by a user-centric cluster of RUs; each RU participates in multiple user-centric clusters.
- The UE-RU association is described by a bipartite graph.
- RUs are connected with DUs via a flexible fronthaul network, and implement the user-centric cluster processors (PHY layer) as SDVNF.
- A CU implements higher level centralized functions.

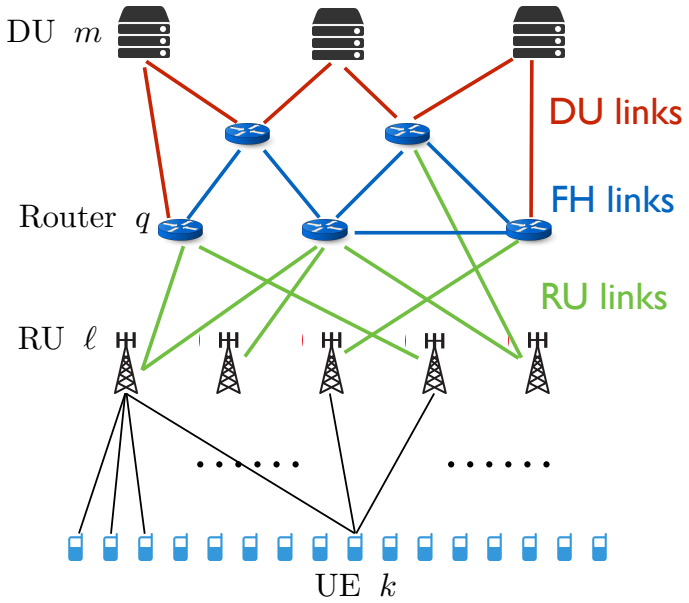




Scalability

- Scalable CF user-centric networks: decentralized processing units (DUs) handle the user-centric cluster processors. DUs and RUs are connected via a routing fronthaul network.

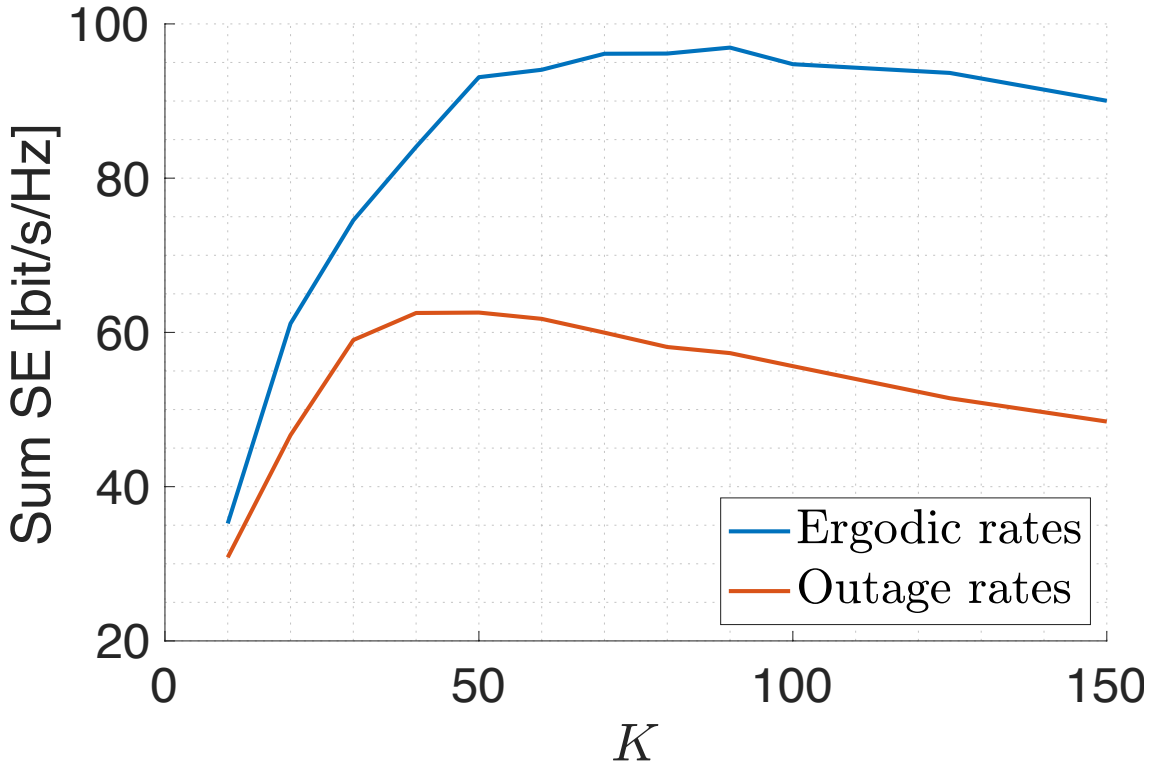
- As the coverage area $A \to \infty$, with given RU density λ_a, DU density λ_d, and UE density λ_u, the load of the fronthaul at any node and the computational load at any processor remain finite.



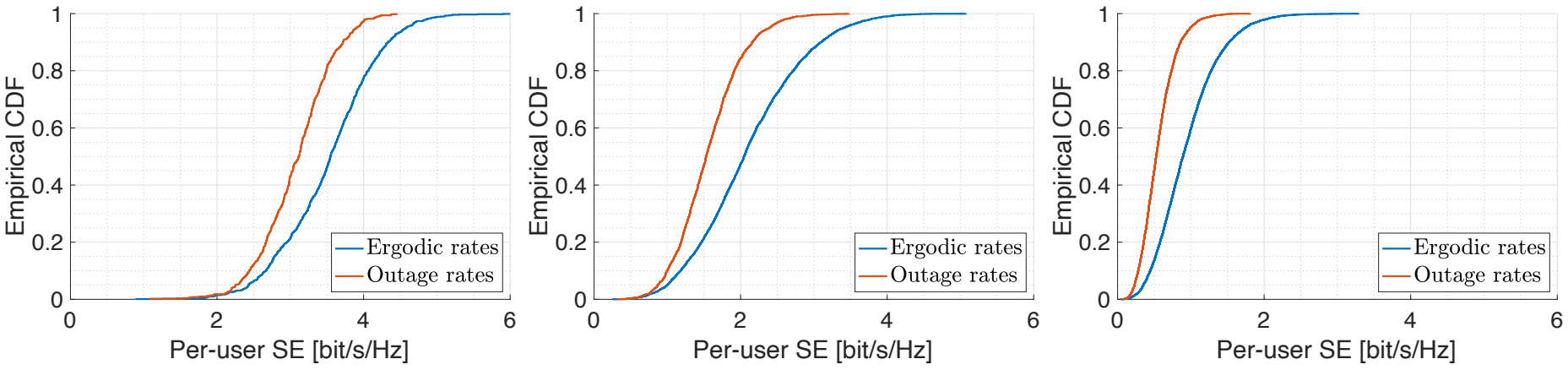
Fairness Scheduling

Total spectral efficiency vs number of active users

- Running example with $L = 12$ RUs, $M = 8$ antennas per RU, and $\tau_p = 20$ pilot dimension for a coherence block (RB) $T = 200$ symbols.
- The current literature investigated the "massive MIMO" regime of $LM \gg K$, i.e., a lightly loaded system with small sum SE.



Lightly loaded versus highly loaded regimes



- As the number of active users $K = 10, 40, 100$ increases near the maximum SE, the per-user rate collapses.

- In real-world systems, $K \gg LM$ and we need to schedule subsets of K_{act} users on different time-frequency slots, such that the per-user throughput is fair among all users.

- With scheduling, ergodic rates are not so meaningful: outage rates are more meaningful!.

Instantaneous outage rate (1)

- For a given precoding/detection scheme, assuming Gaussian signaling and treating interference as noise, the "instantaneous" mutual information $I(\{\hat{s}_k(t,f) : f \in [1:F]\}; \{s_k(t,f) : f \in [1:F]\})$ in slot t is a function of the precoding/combining vectors and channel matrix $\{\mathbb{v}(t), \mathbb{H}(t)\}$.

- E.g., Uplink:

$$\mathcal{I}_k\left(\mathbb{v}_k(t), \mathbb{H}(t)\right) = \frac{1}{F}\sum_{f=1}^{F} \log\left(1 + \mathsf{SINR}_k(t,f)\right),$$

where

$$\mathsf{SINR}_k(t,f) = \frac{|\mathbb{v}_k(t,f)^{\mathsf{H}}\mathbb{h}_k(t,f)|^2}{\mathsf{SNR}^{-1} + \sum_{j\neq k}|\mathbb{v}_k(t,f)^{\mathsf{H}}\mathbb{h}_j(t,f)|^2}.$$

Instantaneous outage rate (2)

- The instantaneous service rate of UE k in time slot t (expressed in bit per time-frequency channel use) is thus given by

$$\mu_k(t) = \begin{cases} (1 - \frac{\tau_p}{T}) R_k(t), & \text{if } k \in \mathcal{A}(t), \\ 0, & \text{if } k \notin \mathcal{A}(t), \end{cases}$$

  where $R_k(t)$ is the random variable

$$R_k(t) = r_k(t) \times \mathbb{1}\left\{ r_k(t) < \mathcal{I}_k\left(\mathbb{v}_k(t), \mathbb{H}(t) \right) \right\},$$

  and where $r_k(t)$ is the scheduled coding rate.

- Here we assume that rate adaptation (i.e., the choice of $r_k(t)$ for $k \in \mathcal{A}(t)$) is made based on the statistics of the mutual information, and not on its instantaneous value.

- The outage rate captures the fact that, when coding over a finite number F of channel states, the block error rate is generally not vanishing.

Fairness scheduling (1)

- A scheduling policy consists of choosing the active user set $\mathcal{A}(t)$ and the coding rates $\{r_k(t)\}$ at each slot time t as a function of the channel states $\{\mathbb{H}(\tau) : \tau = 0, \ldots, t-1\}$.

- Define the per-user throughput rate as the long-term average service rate

$$\bar{\mu}_k = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mu_k(\tau) = \mathbb{E}\left[\mu_k\left(\mathbb{H}\right)\right],$$

  where, with some abuse of notation, we denote by $\mu_k(\mathbb{H})$ the random variable induced by the scheduling policy.

- The goal is to design a scheduling policy that solves the NUM problem:

$$\text{maximize } \; g(\bar{\boldsymbol{\mu}}), \quad \text{subject to } \; \bar{\boldsymbol{\mu}} \in \mathcal{R}$$

  where $\mathcal{R}$ is the achievable throughput region of the system.

Fairness scheduling (2)

- Statistical decoupling assumption: we assume that the marginal CDF of $\mathcal{I}_k\left(\mathbb{v}_k, \mathbb{H}\right)$ depends only on k and not on the set of active users $\mathcal{A}$.
- Defining the complementary CDF

$$P_k(z) = \mathbb{P}(\mathcal{I}_k\left(\mathbb{v}_k, \mathbb{H}\right) > z),$$

  we have that $\mathbb{E}[R_k(t)] = r_k(t) P_k(r_k(t))$.
- Hence, the optimization of the coding rate $r_k(t)$ is immediate and yields

$$r_k^*(t) = r_k^* = \arg\max_{z \geq 0} \; z \times P_k(z).$$

- In practice, we use the empirical "local" distribution of the mutual information at the k-th user receiver collected over a window of channel samples.
- This is fully consistent with present rate adaptation algorithms based on the CQI on a window of past slots.

Fairness scheduling (3)

- From the standard Lyapunov drift plus penalty (DPP) approach, we have that the following dynamic algorithm approximates the NUM solution:

  at each scheduling slot t repeat:

  1. Solve (with respect to the active user set $\mathcal{A}(t)$) the max weighted sum outage rate: $\sum_{k\in\mathcal{A}(t)} Q_k(t)\overline{R}_k$ where $\overline{R}_k := r_k^* P_k(r_k^*)$
  2. Solve the "virtual arrivals" auxiliary problem: $\mathbf{a}(t)$ is the solution of

$$\begin{array}{ll} \text{maximize} & Vg(\mathbf{a}) - \sum_{k\in[K]]} Q_k(t)\, a_k \\ \text{subject to} & \mathbf{a} \in [0, A_{\max}]^K \end{array}$$

  3. Update the virtual queues: $Q_k(t+1) = \max\{Q_k(t) - \mu_k(t), 0\} + a_k(t)$

Weighted sum rate maximization (1)

- The (computationally) critical step is the weighted sum rate maximization.
- In order to make the statistical decoupling (approximately) hold, we need to ensure that the selected active user set avoid "conflicts", i.e., users with strong pilot contamination.
- We first define a conflict graph $\mathcal{G} = ([K], \mathcal{E})$ where a UE-pair $(k, k') \in \mathcal{E}$ (i.e., it has a scheduling conflict) if:
  1. their RU clusters have at least one common RU, i.e., $\mathcal{C}_k \cap \mathcal{C}_{k'} \neq \emptyset$;
  2. they are associated to the same UL pilot sequence;
  3. their channel subspaces to at least one common RU have overlap.

Weighted sum rate maximization (2)

- The resulting weighted sum rate maximization is a linear integer program:

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{k\in[K]} Q_k(t)\bar{R}_k x_k \\ \text{subject to} \quad & \textstyle\sum_{k\in[K]} x_k \leq K_{\text{act}}, \\ & x_k \in \{0,1\}, \\ & x_k + x_{k'} \leq 1, \ \forall (k,k') \in \mathcal{E}. \end{aligned}$$

- This can be solved directly by standard tools (e.g., Gurobi) or relaxed to a LP followed by quantization.

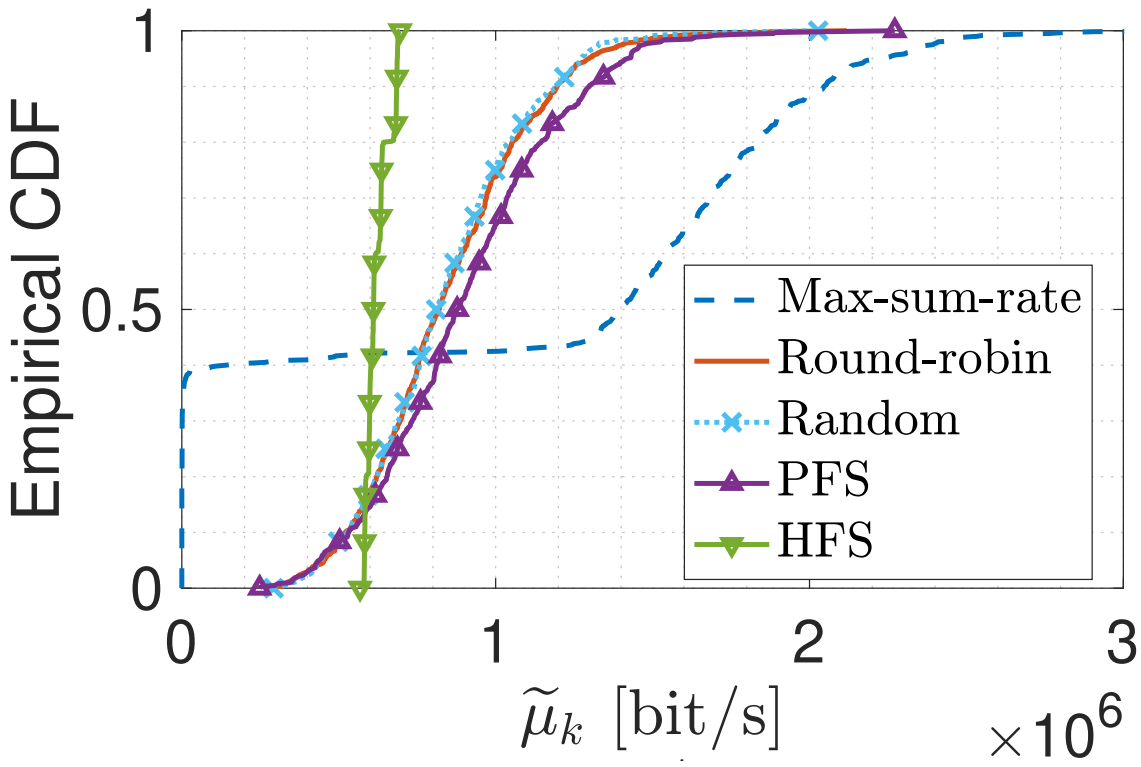
- The max size of the active set K_{act} is chosen in order to operate (slightly to the left) of the peak of the SE.

- A rule of thumb for typical system parameters is $K_{\text{act}} \approx LM/2$.

Example

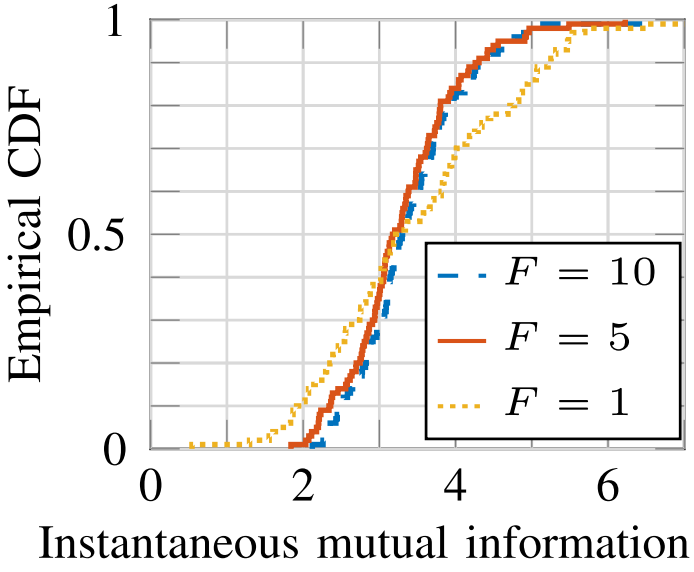
- $A = 200 \times 200$ m^2, $L = 20$ RUs with $M = 10$ antennas each, $W = 60$ MHz.
- RBs of $T = 200$ time-frequency symbols, of which $\tau_p = 20$ are UL pilots (at most 20 orthogonal pilots per RB for the whole networks).
- Each RB spans 12 subcarriers with spacing 60 kHz, i.e., ~ 80 RBs in frequency at each time slot.
- $K_{\text{tot}} = 10,000$ UEs allocated on subchannels of F RBs in frequency such that $K = K_{\text{tot}} F W_{\text{RB}}/W = K_{\text{tot}} F/80$ users per subchannel.
- Each subchannel is independently scheduled.
- F yields a tradeoff between frequency diversity and number of users per subchannel: for $F = 1$ we have $K = 125$ users per subchannel. For $F = 8$ we have $K = 1000$ users per subchannel.
- We consider UEs transmitting at 20 dBm (no power control), and a balanced system (total Tx power in UL = total Tx power in DL).
- The scheduler targets $K_{\text{act}} = ML/2 = 100$ active user per time slot per subchannel.

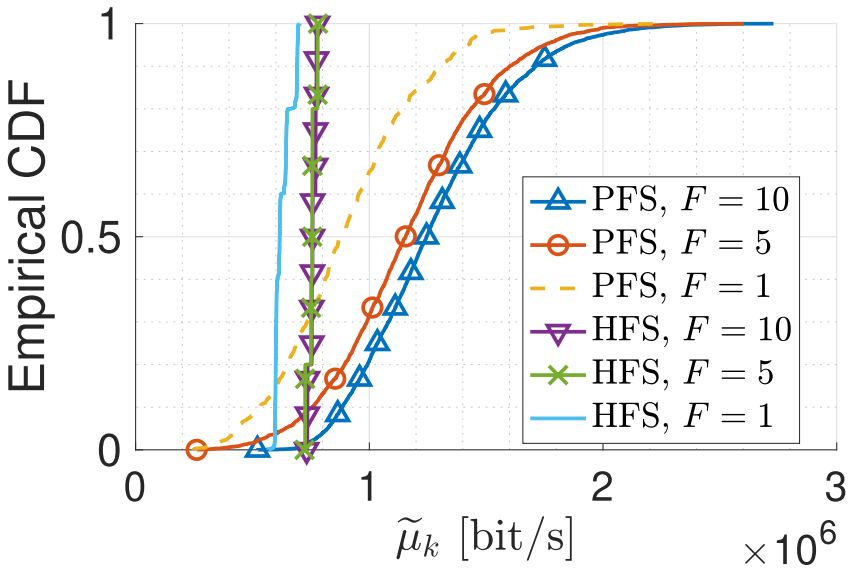
Example



- Proportional fairness scheduling: $g(\bar{\boldsymbol{\mu}}) = \sum_k \log \bar{\mu}_k$.
- Max-min fairness scheduling: $g(\bar{\boldsymbol{\mu}}) = \min_k \bar{\mu}_k$.

Examples

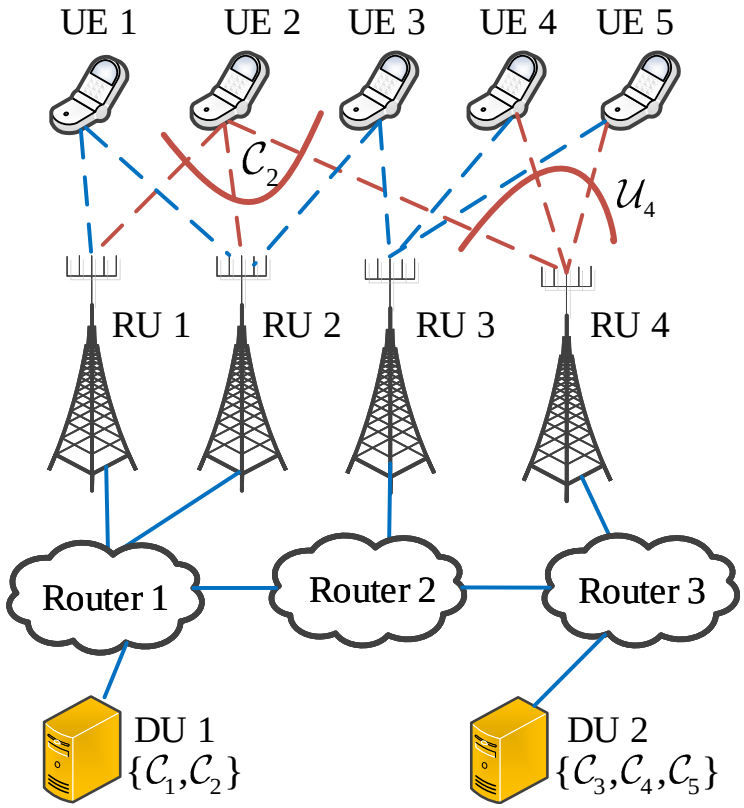




- Effect of frequency diversity $F = 1, 5, 10$.

Fronthaul Optimization

Load Balancing and Computation Resource Allocation



- Cluster processors are SDVNF allocated to the DUs.
- We wish to allocate the cluster processors to the DUs and the routing through the fronthaul such that the max link load is minimized.

Network Topology

- K UEs, L RUs, Q routers, N DUs.
- The Radio Access Network (RAN) is defined by a bipartite graph $\mathcal{G}_{\text{ran}}(\mathcal{K}, \mathcal{L}, \mathcal{E}_{\text{ran}})$ that we consider fixed (resulting from the user-centric cluster formation/UE-RU association).
- The fronthaul is described by a graph $\mathcal{G}_{\text{front}}(\mathcal{L}, \mathcal{Q}, \mathcal{N}, \mathcal{E}_{\text{front}})$.
- Asymmetric UL and DL traffic is defined by $\gamma_{\text{DL}} \in (0, 1)$.
- Fronthaul traffic in the UL direction (from RUs to DUs): Multiple unicast.
- Fronthaul traffic in the DL direction (from DUs to RUs): Multiple multicast.

PHY Rates and Fronthaul Rates

- In the UL direction, each RU ℓ produces a locally combined signal for each UE $k \in \mathcal{U}_\ell$.
- These locally combined signals are "sources" represented at (quantization) rate $B_{\ell,k}$ bit/channel use.
- All sources (ℓ, k) for $\ell \in \mathcal{C}_k$ must be routed and delivered to the DU hosting cluster processor $\mathcal{C}_k$, for all $k \in \mathcal{K}$ (7.3 Split of 3GPP).
- Let $\sigma^2_{\ell,k}$ denote the variance of observation (ℓ, k), and fix a target quantization distortion level D. Then,

$$B_{\ell,k} = \left[\log_2 \frac{\sigma^2_{\ell,k}}{D}\right]_+$$

- In addition, the quantized signal can be represented as

$$\widehat{r}_{\ell,k} = \alpha r_{\ell,k} + e_{\ell,k}, \quad \text{with} \quad \alpha = \frac{[\sigma^2_{\ell,k} - D]_+}{\sigma^2_{\ell,k}} \quad \text{and} \quad \mathbb{E}[|e_{\ell,k}|^2] = [1 - D/\sigma^2]_+ D$$

PHY Rates and Fronthaul Rates

- In the DL direction, the best fronthaul compression consists of multicasting the information bits from the DU hosting the cluster processor for $\mathcal{C}_k$ to all RUs $\ell \in \mathcal{C}_k$, for all $k \in \mathcal{K}$.

- This implies that the multicast rate for UE k is equal to the PHY rate R_k in the downlink (bit/channel use).

- This implies also that encoding and combining must be performed in the RUs, which is consistent with the original idea of Marzetta for CF networks (he studied only the DL, with local combining at each antenna unit).

- Sending the information bits (payload) corresponds to the so-called 7.2 Split of 3GPP.

Computation resource allocation

- It is clear that the load balancing multicommodity flow optimization problem (min max link load) depends on the allocation of the cluster processors to the DUs (computation allocation problem).
- The computation allocation is defined by binary allocation variables

$$b_{k,n} = \begin{cases} 1, & \text{if } \mathcal{C}_k \text{ is hosted by DU } n \\ 0, & \text{if } \mathcal{C}_k \text{ is not hosted by DU } n \end{cases}$$

- With the constraints

$$\sum_{n=1}^{N} b_{k,n} = 1, \forall\ k, \quad \text{and} \quad \sum_{n=1}^{K} b_{k,n} \leq Z_n$$

where Z_n denotes the computation capacity of DU n.

- The resulting joint optimization is a MILP and can be efficiently solved even for several hundreds of RUs, tens of routers and DUs.

Example

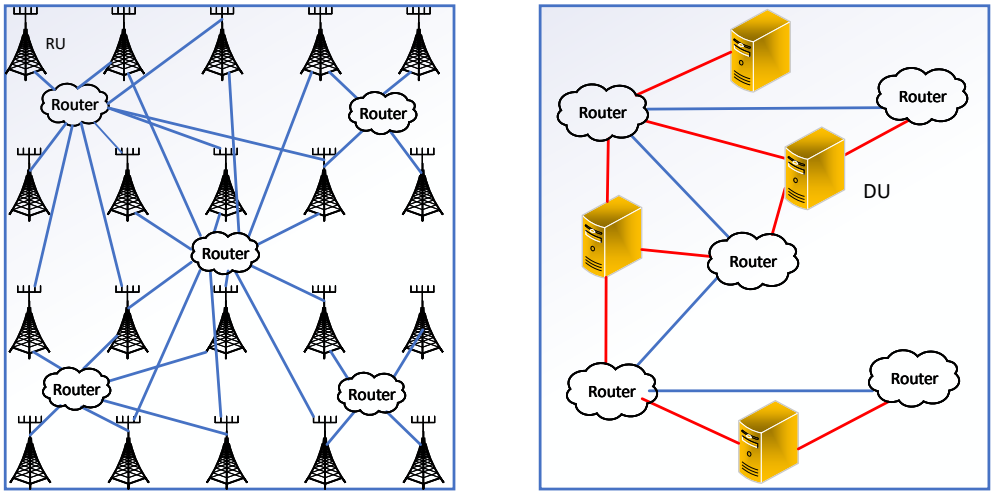
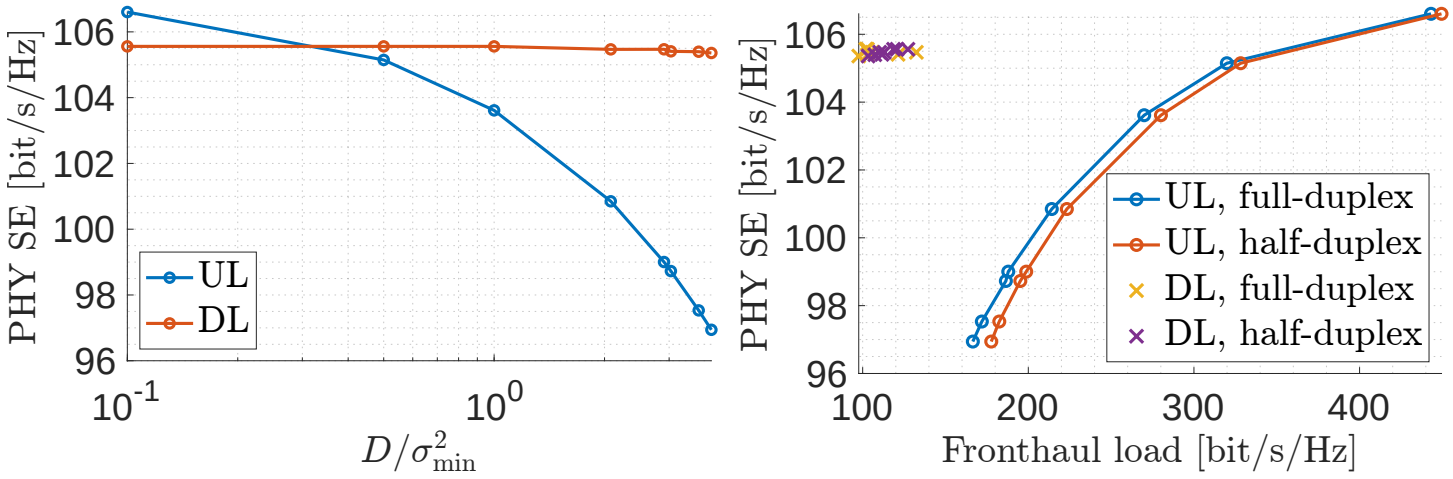


Fig. 3: Left: The fronthaul links between RUs and routers. Right: The fronthaul links between routers and DUs.

- $L = 20$, $M = 10$, $K = 100$ users per subchannel, very similar to before.

Example



(a) (b)

- (a) UL/DL PHY Spectral Efficiency vs. fronthaul quantization distortion level D; (b) UL/DL PHY Spectral Efficiency vs. fronthaul load.

- To be noted: 1) Aggressive quantization in the UL direction is good; 2) the frontload bottleneck is the UL (with proper compression in the DL!!!); 3) this is very good news because the DL/UL duty cycle γ_{DL} is generally large; 4) It is beneficial (more flexible) to use the fronhaul links in half-duplex mode.

Thank You