International Workshop on Emerging Technologies for LTE-Advanced and Beyond-4G

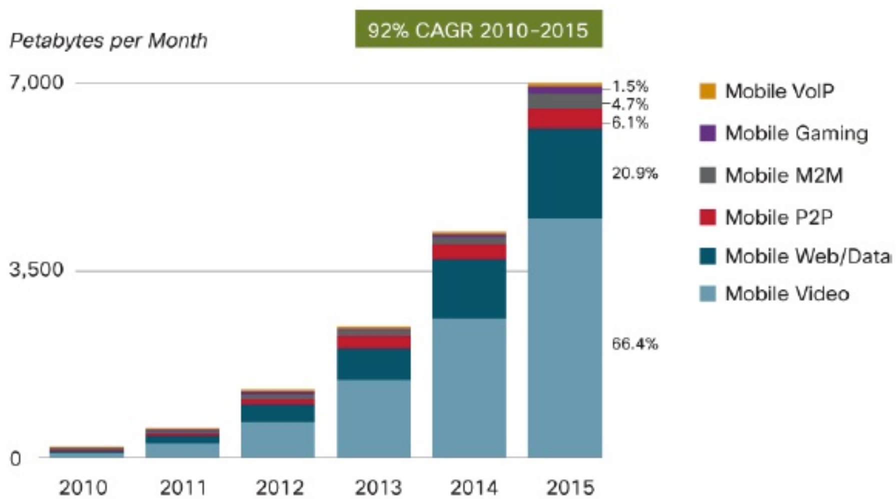
Emerging Technologies Beyond 4G: Massive MIMO, Dense Small Cells, Virtual MIMO, D2D and Distributed Caching

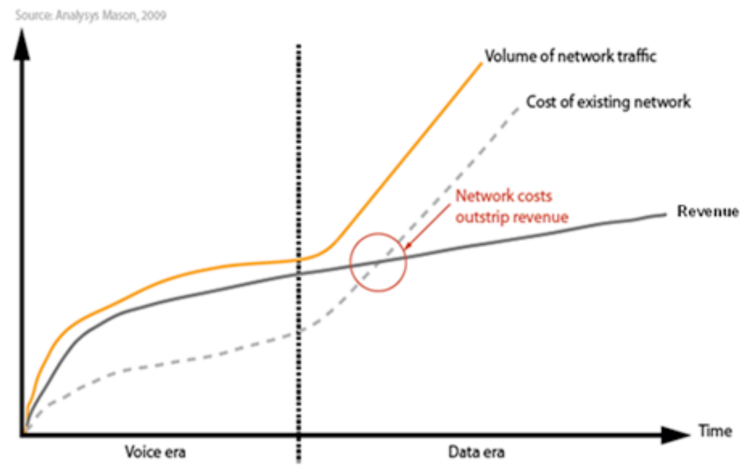
Giuseppe Caire

University of Southern California, Viterbi School of Engineering, Los Angeles, CA

Globecom 2012, Anaheim CA, Dec. 3, 2012

Wireless operators' nightmare





Mobile Broadband's Profitability Gap

- 100x Data traffic increase, due to the introduction of powerful multimedia capable user devices.
- Operating costs trends not matched by revenue trends.

Possible answers

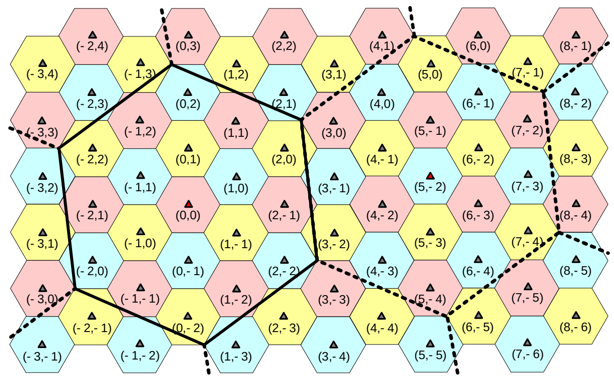
- Release of new wireless spectrum: today, cellular $+$ Wifi accounts for ≈ 500 MHz of bandwidth. Releasing new spectrum will at most double the available overall spectrum (at most x2 increase with same technology).

- Following current technology trend: LTE ... painstakingly slow, incremental gain due to backward compatibility.

- Disruptive technology approach: Massive MIMO, Dense Small Cells, Virtual MU-MIMO, D2D, Wireless Caching.

What Can We Expect?

MU-MIMO Cellular Networks



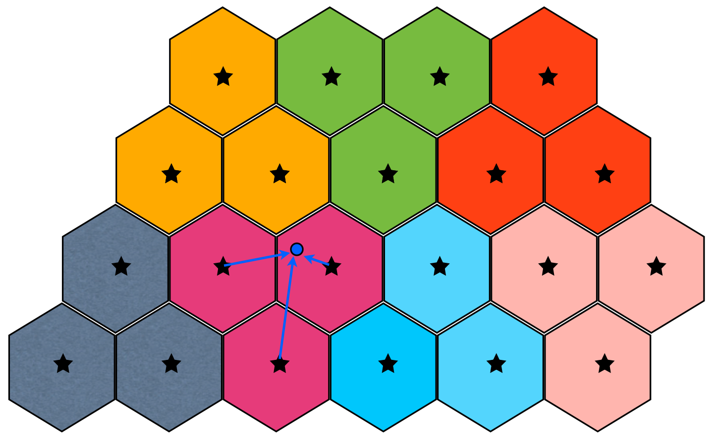
- Multiuser MIMO cooperative upper bound:

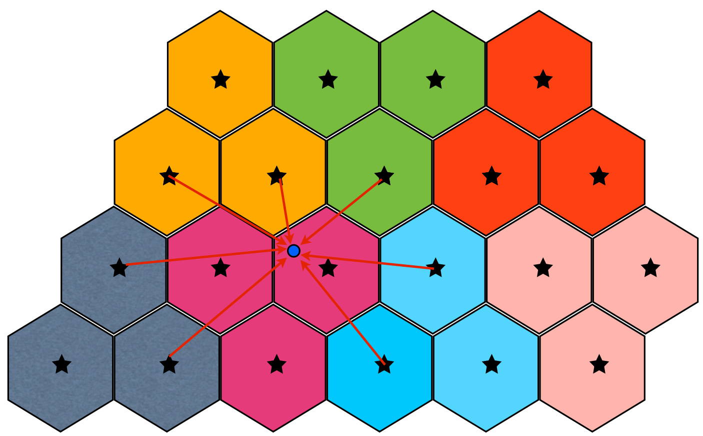
$$C = W \times M^* \left(1 - \frac{M^*}{T}\right) \times \log \text{SINR} + O(1),$$

  were $M^* = \min\{M, KN, T/2\}$.

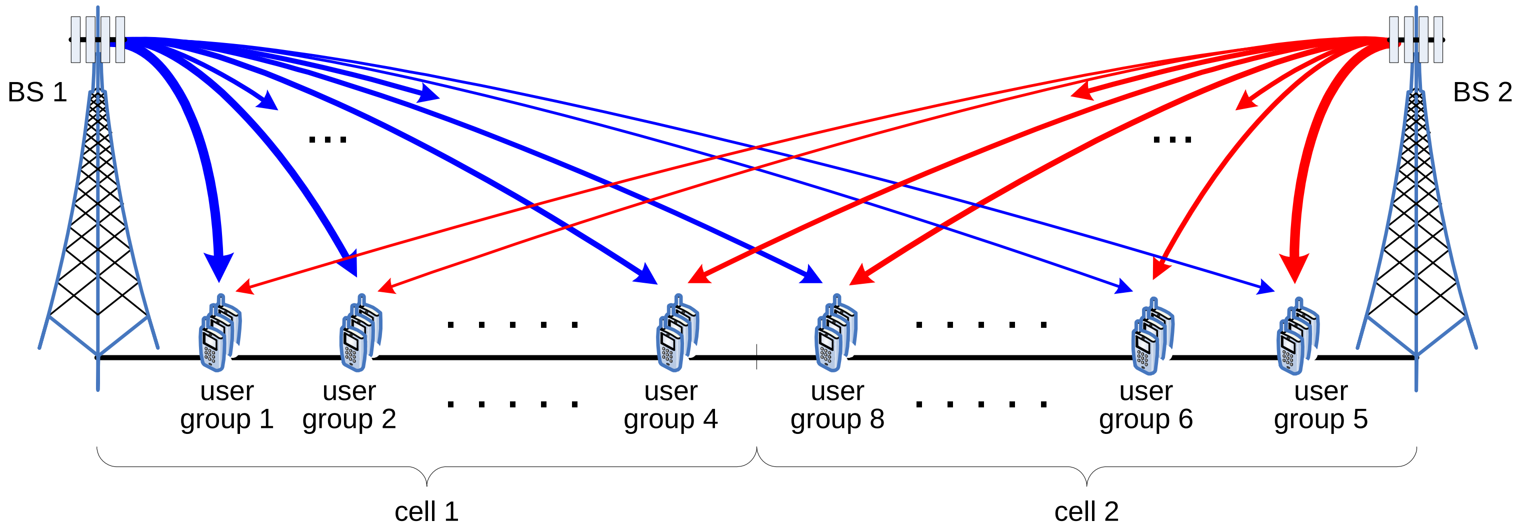
- Fundamental dimensionality bottleneck: channel state estimation overhead (information theoretic upper bound).
- See also high-SNR saturation effect in "Fundamental Limits of Cooperation," [Lozano, Heath, Andrews, arXiv:1204.0011].
- Per-user throughput vanishes as $O(\frac{1}{K})$.

Network MIMO: A Large-System Analysis





Discretization of the Users Distribution



- We assume that the users are partitioned in co-located groups with N single-antenna terminals each.

- We have A user groups per cluster, and clusters of B cells.

- We have $M = \gamma N$ base station antennas per cell.

Cluster of Cooperating Base Stations

- Modified path coefficients $\beta_{m,k} = \frac{\alpha_{m,k}}{\sigma_k}$ taking into account the ICI power.
- Cluster channel matrix

$$\mathbf{H} = \begin{bmatrix} \beta_{1,1}\mathbf{H}_{1,1} & \cdots & \beta_{1,A}\mathbf{H}_{1,A} \\ \vdots & \ddots & \vdots \\ \beta_{B,1}\mathbf{H}_{B,1} & \cdots & \beta_{B,A}\mathbf{H}_{B,A} \end{bmatrix}.$$

- Reference cluster channel model

$$\mathbf{y} = \mathbf{H}^{\mathsf{H}}\mathbf{x} + \mathbf{z}$$

where $\mathbf{y} = \mathbb{C}^{AN}$, $\mathbf{x} = \mathbb{C}^{\gamma BN}$, and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$.

Linear ZF Downlink Beamforming

- We consider the weighted rate sum maximization for given channel matrix:

$$\begin{aligned} \text{maximize} \quad & \sum_{k=1}^{A} \sum_{i=1}^{N} W_k^{(i)} R_k^{(i)} \\ \text{subject to} \quad & \mathbf{R} \in \mathcal{R}_{\mathrm{lzfb}}(\mathbf{H}) \end{aligned}$$

where $W_k^{(i)}$ denotes the rate weight for user i in group k, and $\mathcal{R}_{\mathrm{lzfb}}(\mathbf{H})$ is the achievable *instantaneous* rate region of LZFB for given channel matrix $\mathbf{H}$.

Some Simplifying Assumptions

- The scheduler picks a fraction μ_k of users in group k by random selection inside the group.

- LZFB precoder obtained by normalizing the columns of the Moore-Penrose pseudo-inverse of the channel matrix.

- Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A)$ denote the fractions of active users in groups $1, \ldots, A$, respectively. For given $\boldsymbol{\mu}$, the corresponding effective channel matrix is given by

$$\mathbf{H}_{\boldsymbol{\mu}} = \left[\begin{array}{ccc} \beta_{1,1}\mathbf{H}_{1,1}(\mu_1) & \cdots & \beta_{1,A}\mathbf{H}_{1,A}(\mu_A) \\ \vdots & & \vdots \\ \beta_{B,1}\mathbf{H}_{B,1}(\mu_1) & \cdots & \beta_{B,A}\mathbf{H}_{B,A}(\mu_A) \end{array} \right],$$

LZFB "parallel channels"

- Letting $\mathbf{V}_{\boldsymbol{\mu}} = \mathbf{H}_{\boldsymbol{\mu}}^{+} \boldsymbol{\Lambda}_{\boldsymbol{\mu}}^{1/2}$, we obtain the "parallel" channel model

$$\mathbf{y}_{\boldsymbol{\mu}} = \mathbf{H}_{\boldsymbol{\mu}}^{\mathsf{H}} \mathbf{V}_{\boldsymbol{\mu}} \mathbf{Q}^{1/2} \mathbf{u} + \mathbf{z}_{\boldsymbol{\mu}} = \boldsymbol{\Lambda}_{\boldsymbol{\mu}}^{1/2} \mathbf{Q}^{1/2} \mathbf{u} + \mathbf{z}_{\boldsymbol{\mu}}.$$

Theorem 1. *For all $i = 1, \ldots, \mu_k N$, the following limit holds almost surely:*

$$\lim_{N \to \infty} \Lambda_k^{(i)}(\boldsymbol{\mu}) = \Lambda_k(\boldsymbol{\mu}) = \gamma \sum_{m=1}^{B} \beta_{m,k}^2 \eta_m(\boldsymbol{\mu})$$

where $(\eta_1(\boldsymbol{\mu}), \ldots, \eta_B(\boldsymbol{\mu}))$ is the unique solution in $[0,1]^B$ of the fixed point equations

$$\eta_m = 1 - \sum_{q=1}^{A} \mu_q \frac{\eta_m \beta_{m,q}^2}{\gamma \sum_{\ell=1}^{B} \eta_\ell \beta_{\ell,q}^2}, \quad m = 1, \ldots, B$$

with respect to the variables $\boldsymbol{\eta} = \{\eta_m\}$. ■

Channel Estimation and Non-Perfect CSIT

- We assume that the channels are constant over time-frequency blocks of size WT complex dimensions.
- For each such block, $\gamma_p BN$ dimensions are dedicated to downlink training.
- Since the channel vectors are Gaussian, linear MMSE estimation is optimal with respect to the MSE criterion.
- The MMSE can be made arbitrarily small as $\sigma_k^2 \to 0$ (vanishing noise plus ICI) if and only if $\gamma_p \geq \gamma$.
- The ratio γ_p/γ denotes the "pilot dimensionality overhead".

- From the well-known MMSE decomposition, the channel matrix $\mathbf{H}$ can be written as $\mathbf{H} = \widehat{\mathbf{H}} + \mathbf{E}$, where

$$\widehat{\mathbf{H}} = \begin{bmatrix} \widehat{\beta}_{1,1}\mathbf{H}_{1,1} & \cdots & \widehat{\beta}_{1,A}\mathbf{H}_{1,A} \\ \vdots & & \vdots \\ \widehat{\beta}_{B,1}\mathbf{H}_{B,1} & \cdots & \widehat{\beta}_{B,A}\mathbf{H}_{B,A} \end{bmatrix},$$

with

$$\widehat{\beta}_{m,k} = \frac{\beta_{m,k}^2}{\sqrt{1/p + \beta_{m,k}^2}},$$

and where

$$\mathbf{E} = \begin{bmatrix} \bar{\beta}_{1,1}\mathbf{E}_{1,1} & \cdots & \bar{\beta}_{1,A}\mathbf{E}_{1,A} \\ \vdots & & \vdots \\ \bar{\beta}_{B,1}\mathbf{E}_{B,1} & \cdots & \bar{\beta}_{B,A}\mathbf{E}_{B,A} \end{bmatrix},$$

with

$$\bar{\beta}_{m,k} = \sqrt{\beta_{m,k}^2 - \widehat{\beta}_{m,k}^2} = \frac{\beta_{m,k}}{\sqrt{1 + p\beta_{m,k}^2}},$$

and the blocks $\mathbf{E}_{m,k}$ and independent with i.i.d. $\mathcal{CN}(0,1)$ elements.

Achievable rate lower bound

Theorem 2. *Under the downlink training scheme described above and assuming genie-aided CSIT feedback, the achievable rate of users in group* k *is lower bounded by*

$$R_k \geq \log \left(1 + \frac{\widehat{\Lambda}_k(\boldsymbol{\mu}) q_k}{1 + \sum_{m=1}^{B} \bar{\beta}_{m,k}^2 P_m} \right)$$

where

$$\widehat{\Lambda}_k(\boldsymbol{\mu}) = \gamma \sum_{m=1}^{B} \widehat{\beta}_{m,k}^2 \eta_m(\boldsymbol{\mu})$$

where $(\eta_1(\boldsymbol{\mu}), \ldots, \eta_B(\boldsymbol{\mu}))$ *is the unique solution with components in* $[0, 1]$ *of the fixed point equation*

$$\eta_m = 1 - \sum_{q=1}^{A} \mu_q \frac{\eta_m \widehat{\beta}_{m,q}^2}{\gamma \sum_{\ell=1}^{B} \eta_\ell \widehat{\beta}_{\ell,q}^2}, \quad m = 1, \ldots, B$$

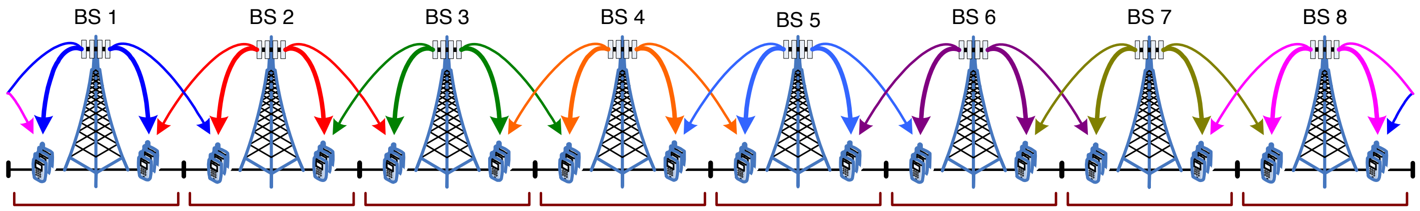
with respect to the variables $\boldsymbol{\eta} = \{\eta_m\}$*.* ■

Degrees of freedom and cost of training

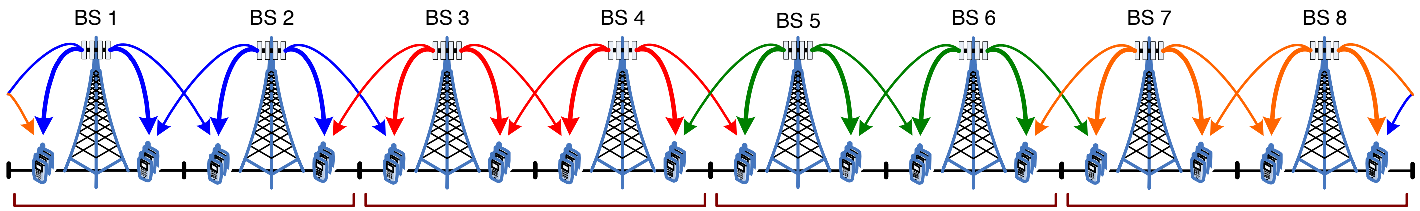
- The system spectral efficiency must be scaled by the factor $\left[1 - \frac{\gamma_p N B}{WT}\right]_+$, that takes into account the downlink training overhead, i.e., fraction of dimensions per block dedicated to (downlink) training.
- In particular, letting $\tau = \frac{N}{WT}$ denote the ratio between the number of users per group, N, and the dimensions in a time-frequency slot, we can investigate the system spectral efficiency for fixed τ, in the limit of $N \to \infty$.
- The ratio τ captures the "dimensional crowding" of the system.

Linear cellular layout

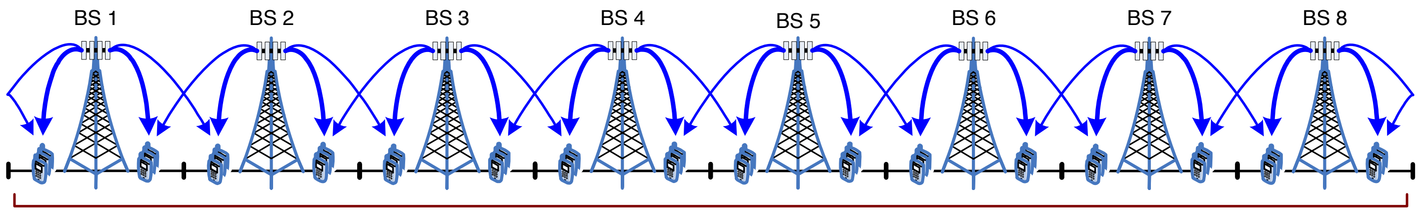
Example: linear cellular layout with $M = 8$ cells



$|\mathcal{M}_\ell| = B = 1$ cell cooperation

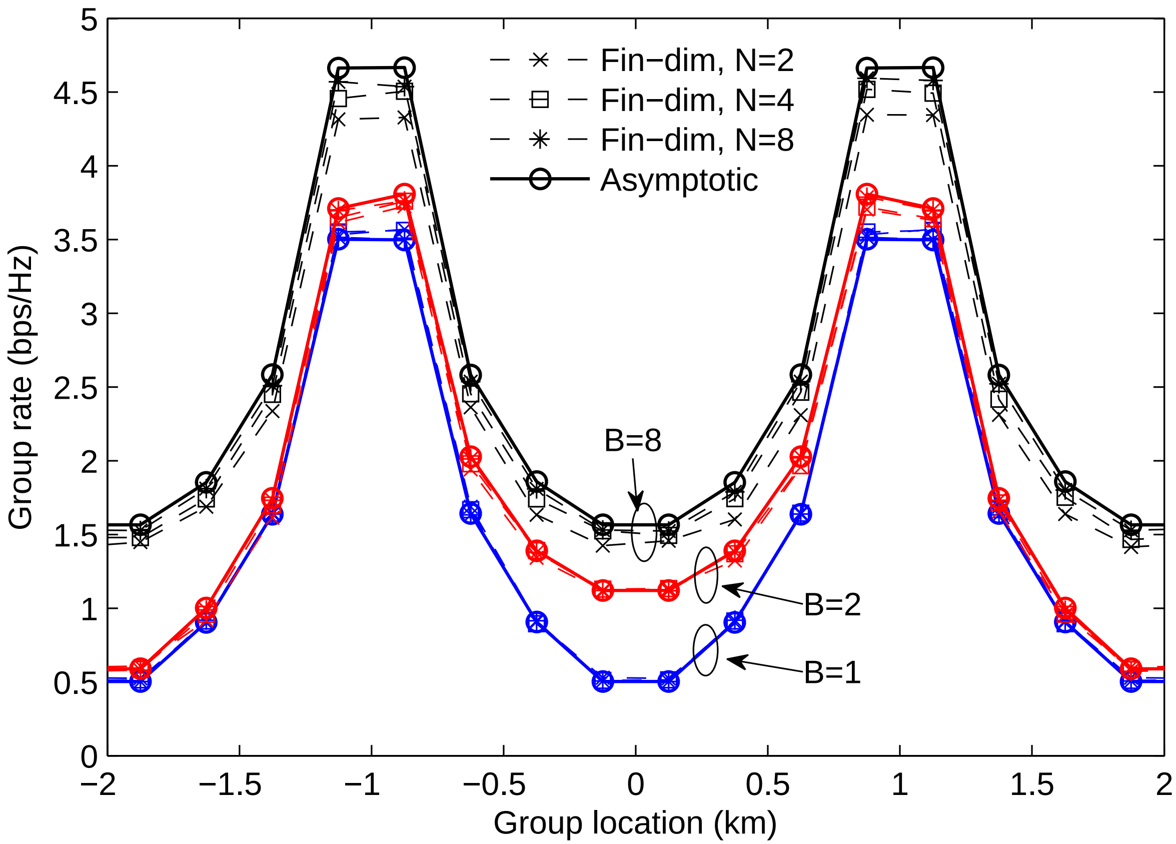


$|\mathcal{M}_\ell| = B = 2$ cell cooperation



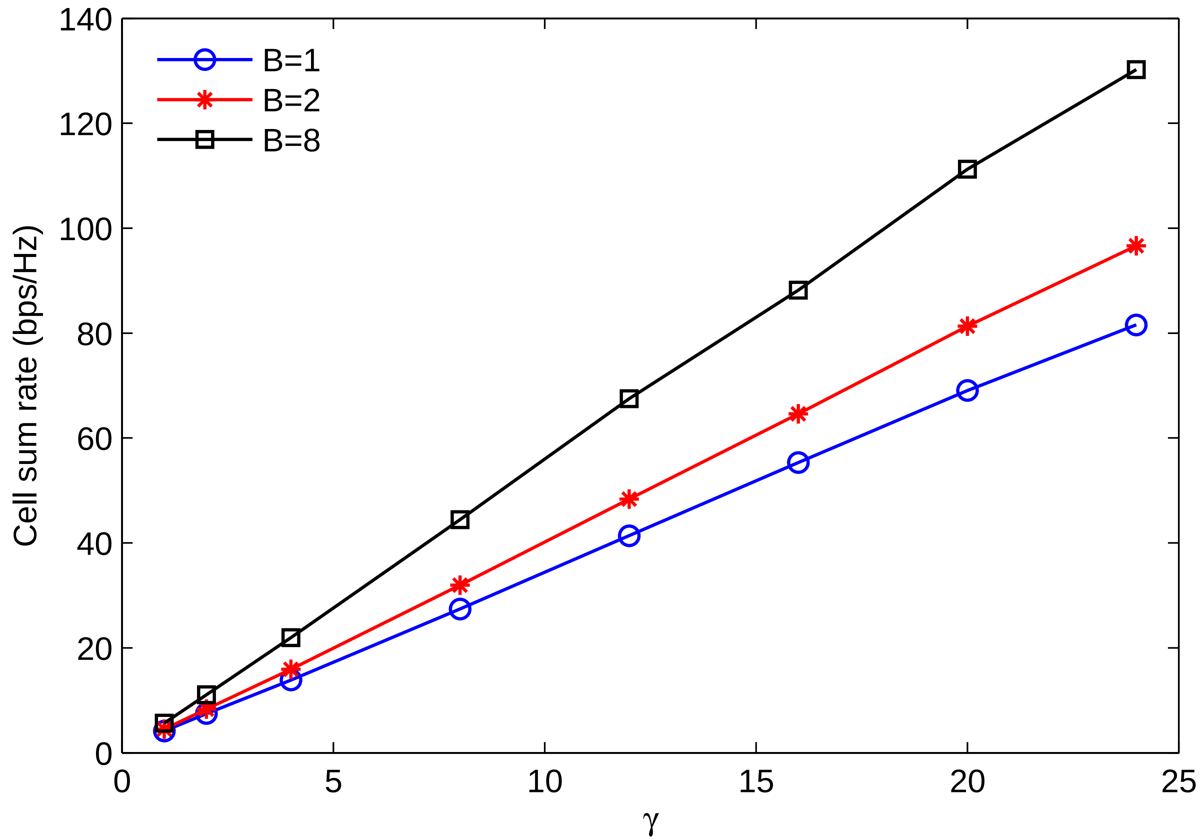
$|\mathcal{M}_\ell| = B = 8$ cell cooperation

Comparison with finite dimensional systems



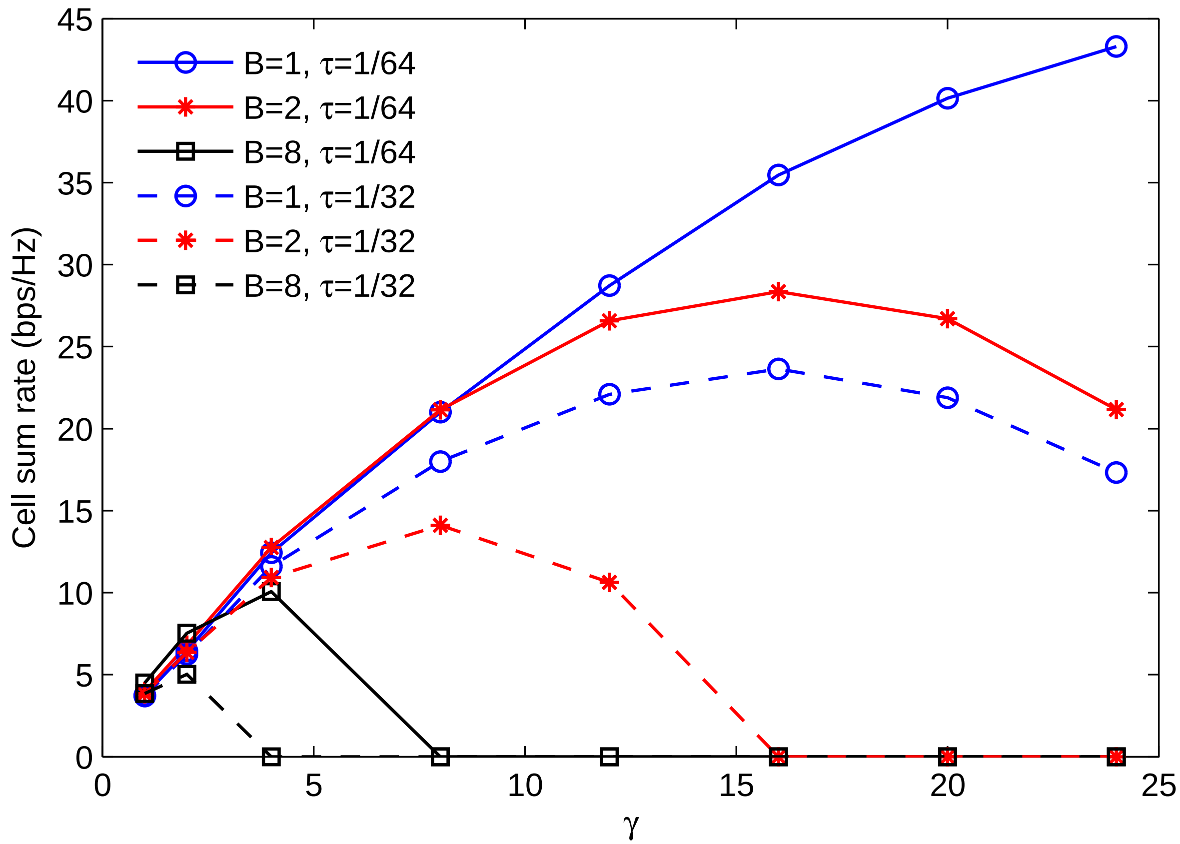
User group rate in finite dimension (N = 2, 4, and 8) for cooperation clusters of size B=1, 2, and 8, with perfect CSIT. $M = 8$ cells and $K = 64$ user groups.

Cost of CSIT and choice of network MIMO architecture (1)



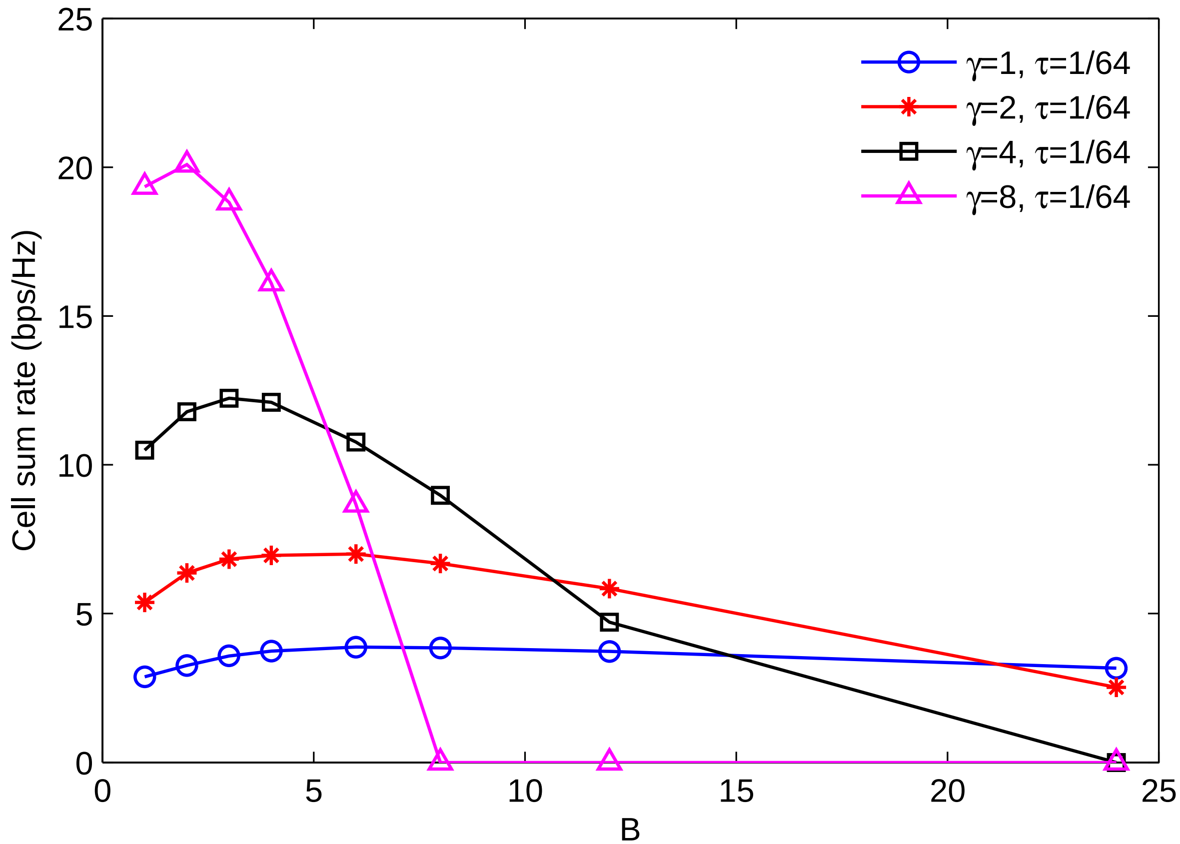
Cell sum rate versus the antenna ratio γ for cooperation clusters of size B=1, 2, and 8. $M = 8$ cells and $K = 192$ user groups.

Cost of CSIT and choice of network MIMO architecture (2)



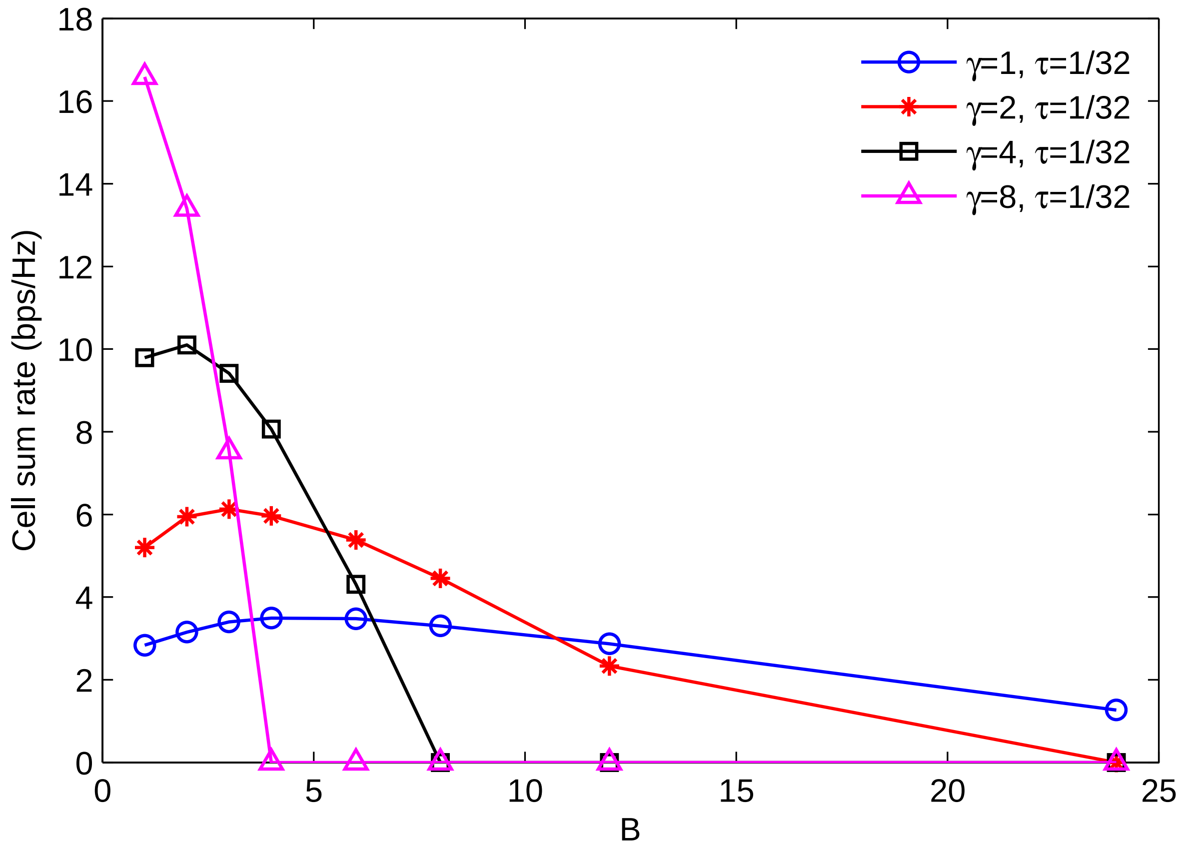
Cell sum rate versus the antenna ratio γ for cooperation clusters of size B=1, 2, and 8. $M = 8$ cells and $K = 192$ user groups.

Large cooperating clusters?



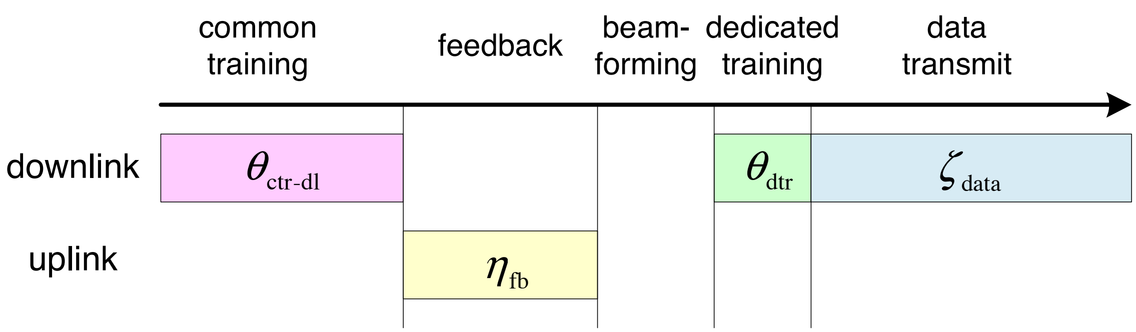
Cell sum rate versus the cluster size B for the antenna ratio γ=1, 2, 4, and 8 $\gamma_p = \gamma$, $M = 24$ cells and $K = 192$ user groups and $\tau = 1/64$.

Large cooperating clusters?



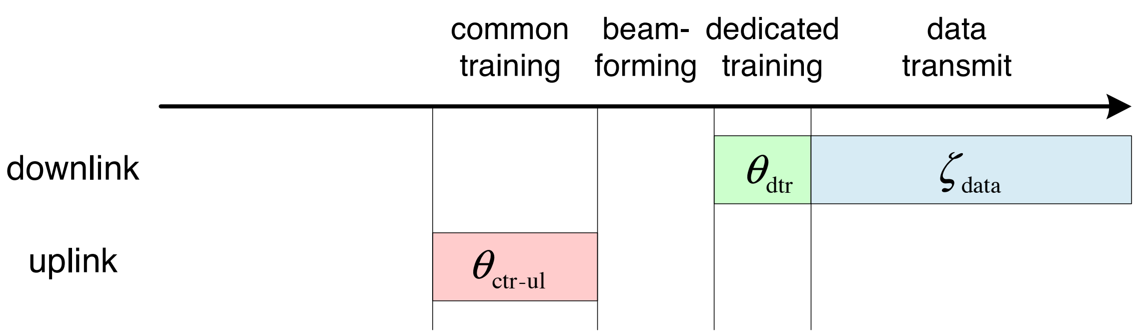
Cell sum rate versus the cluster size B for the antenna ratio γ=1, 2, 4, and 8 $\gamma_p = \gamma$, $M = 24$ cells and $K = 192$ user groups and $\tau = 1/32$.

FDD versus TDD



Frequency-division duplex (FDD)

- Estimation error
- Training overhead proportional to the number of transmit antennas



Time-division duplex (TDD)

- Estimation error
- Training overhead proportional to the number of served users

Marzetta's Massive MIMO Scheme

- No BS cooperation, each cell on its own.

- On each slot, a fraction T_{tr}/T is dedicated to uplink training, and $(1 - T_{\mathrm{tr}}/T)$ is dedicated to downlink data transmission.

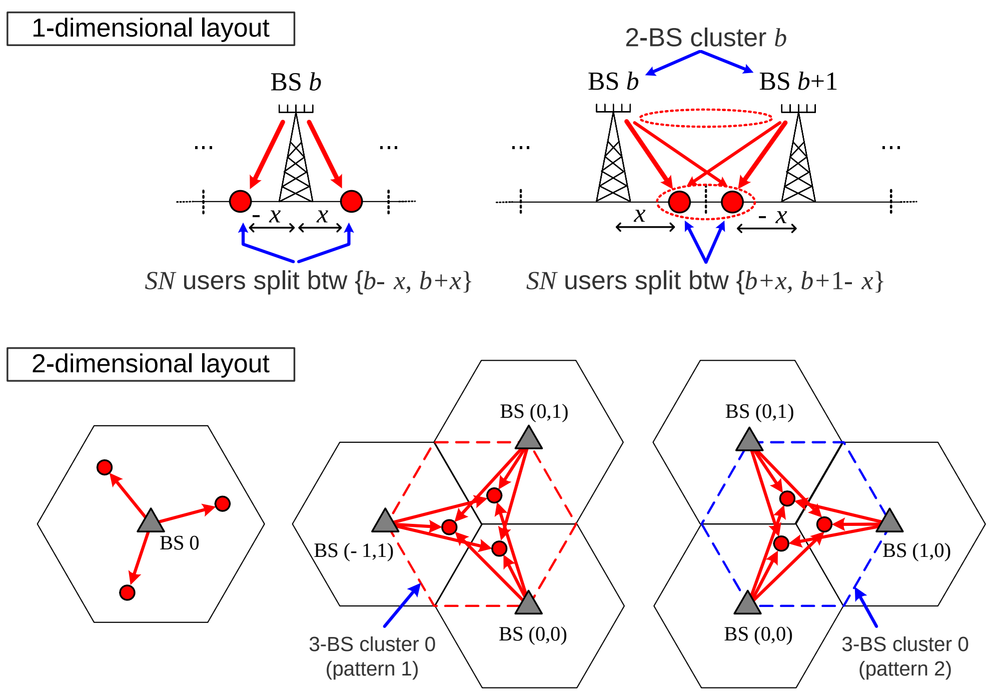
- Single-user downlink beamforming: transmit with the Hermitian transpose of the estimated channel matrix.

- Marzetta considers the limit for a finite number K of users per cell, and the number of BS antennas $M \to \infty$.

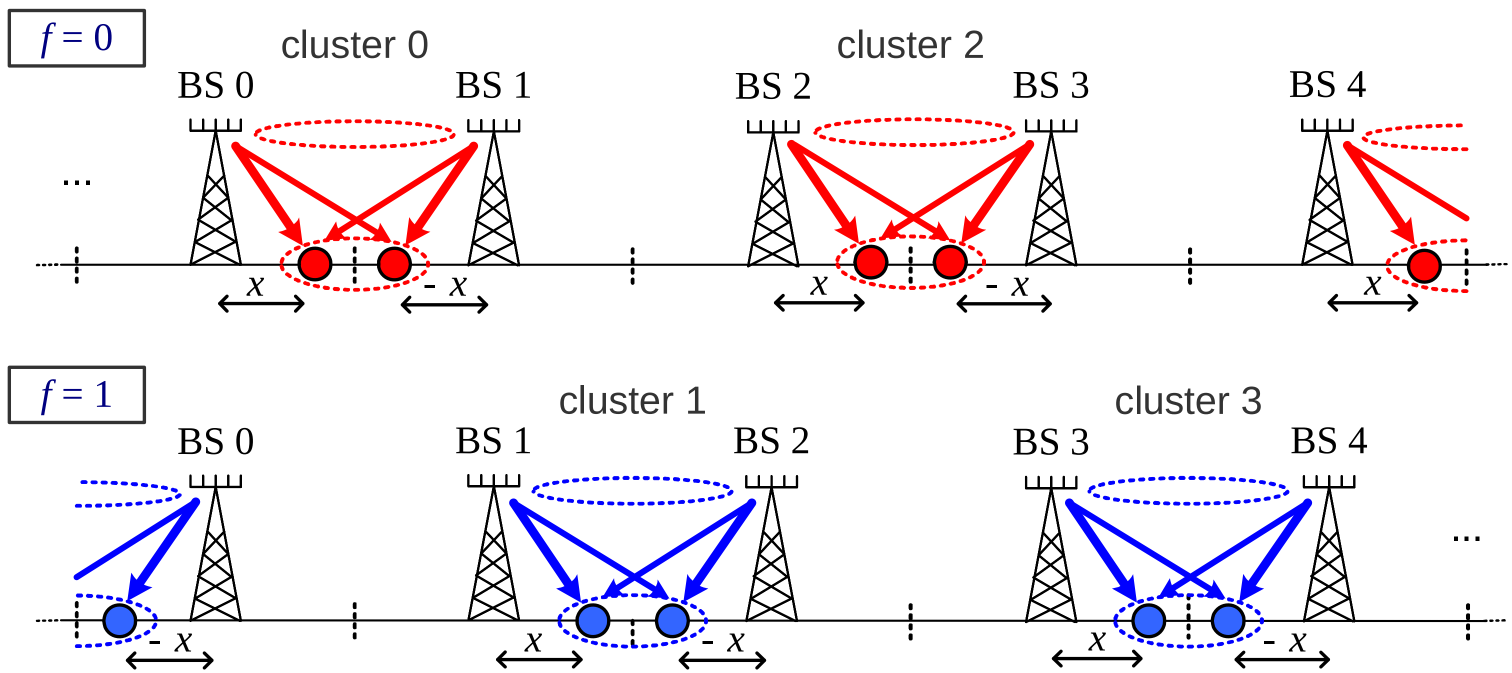
- In this regime, intra and inter cell interference and noise disappear, except for the inter-cell interference due to PILOT CONTAMINATION.

What happens for finite M/N?

- We partition the user population in "bins" of co-located users. Users in the same bin are (roughly) statistically equivalent.
- For each "bin" we consider an optimized MU-MIMO scheme.
- Scheduling over the user bins to maximize a desired Network Utility Function.

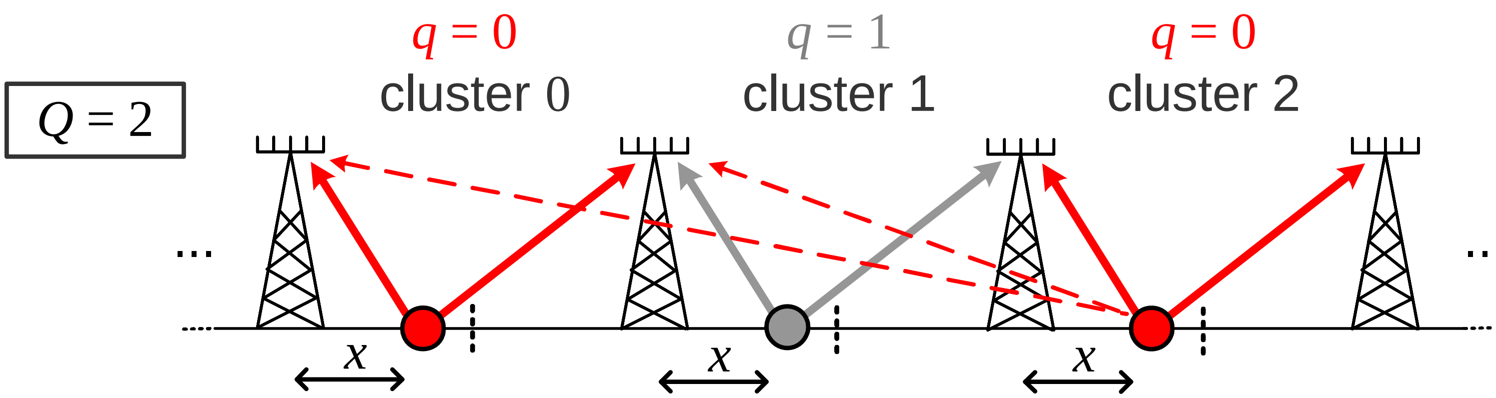


Frequecy reuse



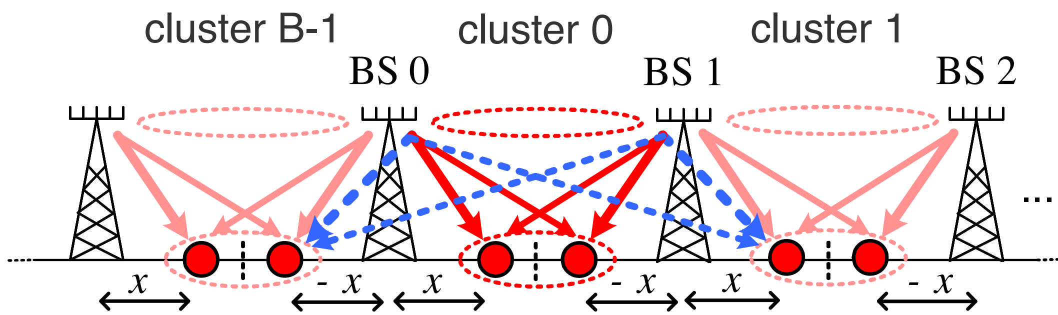
- 1-dimensional layout with $C = 2$ and $F = 2$.

Pilot reuse

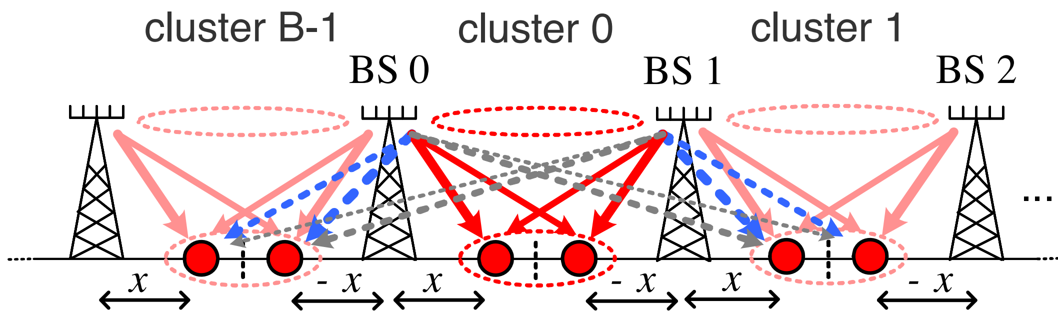


- Pilot reuse and contamination for $C = 2$, $F = 1$, and $Q = 2$. The dashed lines show the pilot contamination at cluster 0 from a user in cluster 2, sharing the same pilot signal.

Zero-Forcing and interference mitigation



(a) $J = Q$



(b) $J = C(Q-1)+1$

Multi-Mode MU-MIMO downlink scheduling

- Consider a system with K bins, $\{v(\mathcal{X}_0), \ldots, v(\mathcal{X}_{K-1})\}$, chosen to sample uniformly the coverage area $\mathcal{V}$.

- The net bin spectral efficiency (in bit/s/Hz)

$$\max\{1 - QS/T, 0\} \times R_{\mathcal{X}_k,\mathcal{C}}(F, C, J),$$

- Let $R^\star(\mathcal{X}_k)$ denote the maximum of the above for given $\mathcal{X}_k$, optimized over the the parameters S, C, J, Q, F.

- A scheduler gives fraction ρ_k of the total time-frequency transmission resource to bin $v(\mathcal{X}_k)$ in order to maximize a desired Network Utility Function.

- The scheduler determines the transmission resource allocation $\{\rho_k\}$ by solving the following convex problem:

$$\begin{array}{ll} \text{maximize} & \mathcal{G}(R_0, \ldots, R_{K-1}) \\ \text{subject to} & R_k \leq \rho_k R^\star(\mathcal{X}_k), \quad \sum_{k=0}^{K-1} \rho_k \leq 1, \quad \rho_k \geq 0. \end{array}$$

- We obtain a multi-modal network MIMO architecture.

- Optimization can be done easily based on large-system limit closed form results.

- Example: *Proportional Fairness* (PF) criterion corresponds to the choice

$$\mathcal{G}(R_0, \ldots, R_{K-1}) = \sum_{k=0}^{K-1} \log R_k,$$

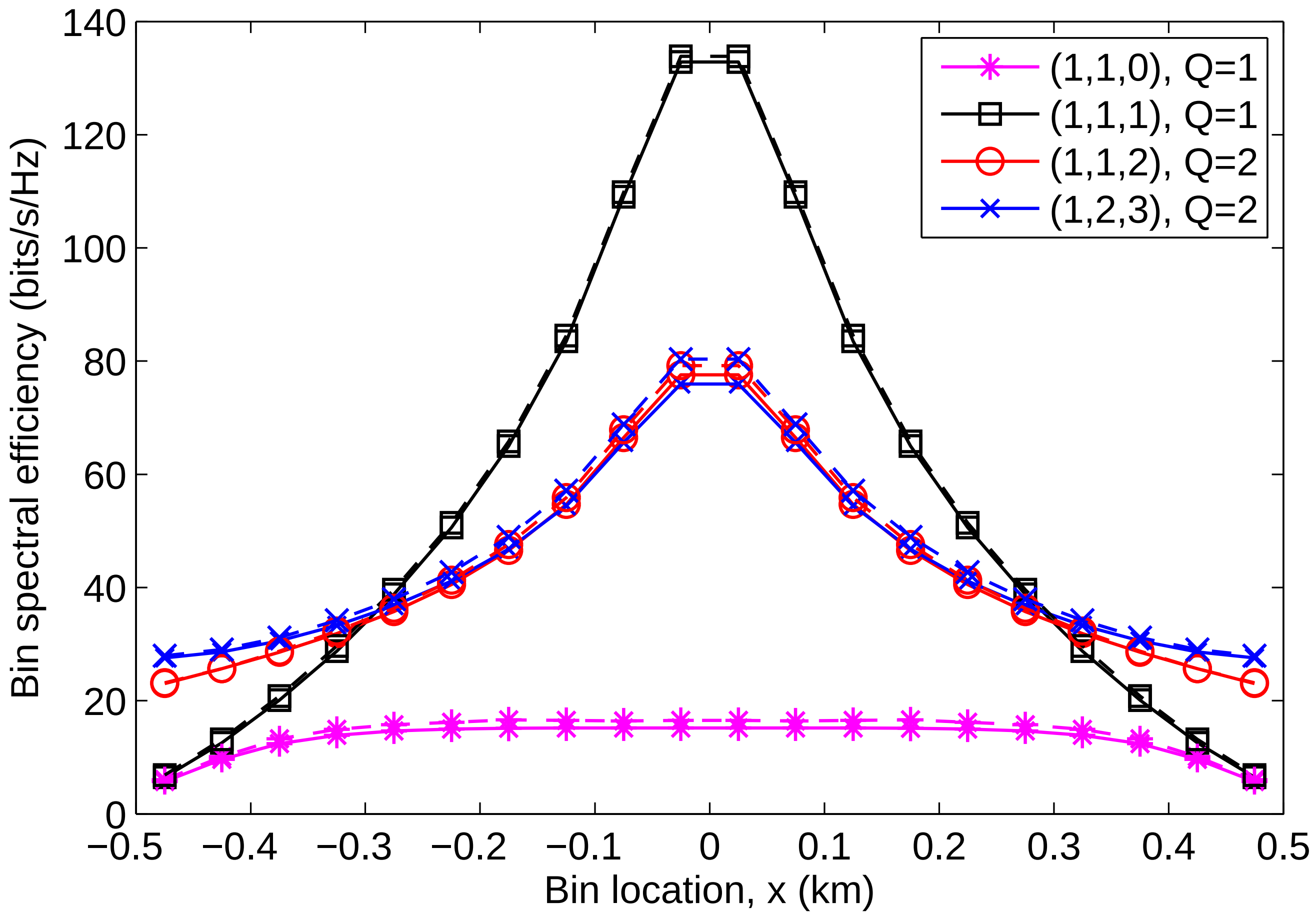
and yields $\rho_k = 1/K$ (each bin is given an equal amount of slots).

- Example: *Max-Min fairness* criterion corresponds to the choice

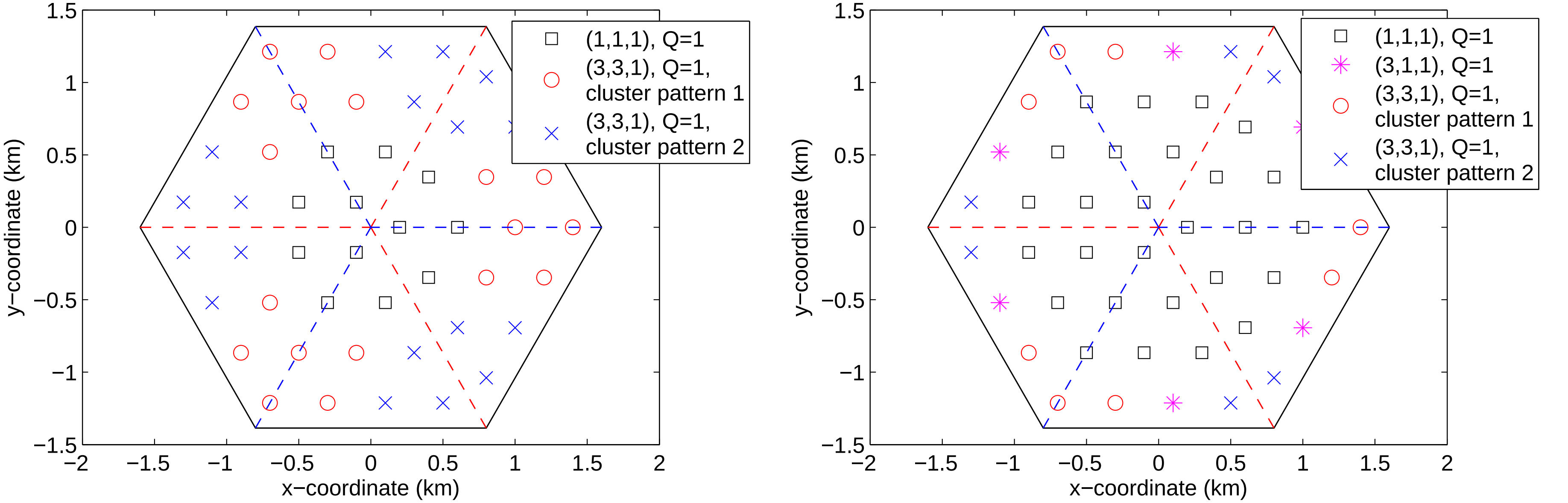
$$\mathcal{G}(R_0, \ldots, R_{K-1}) = \min_{k=0,\ldots,K-1} R_k,$$

and yields $\rho_k = \frac{\frac{1}{R^\star(\mathcal{X}_k)}}{\sum_{j=0}^{K-1} \frac{1}{R^\star(\mathcal{X}_j)}}$.

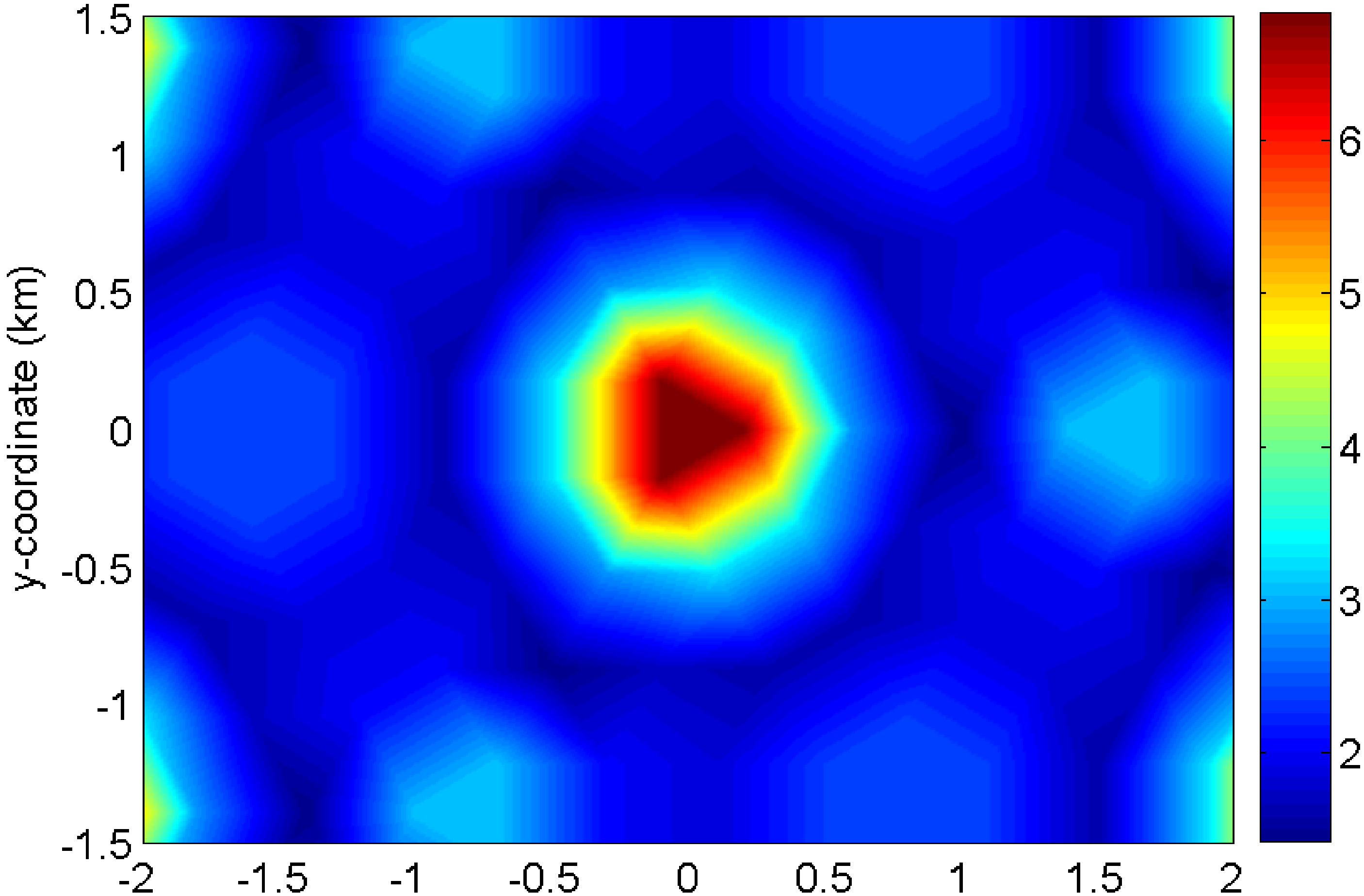
Results



Bin spectral efficiency vs. location within a cell obtained from the large system analysis (solid) and the finite dimension ($N = 1$) simulation (dotted) for various (F, C, J). $M = 30$ and $L = 40$.

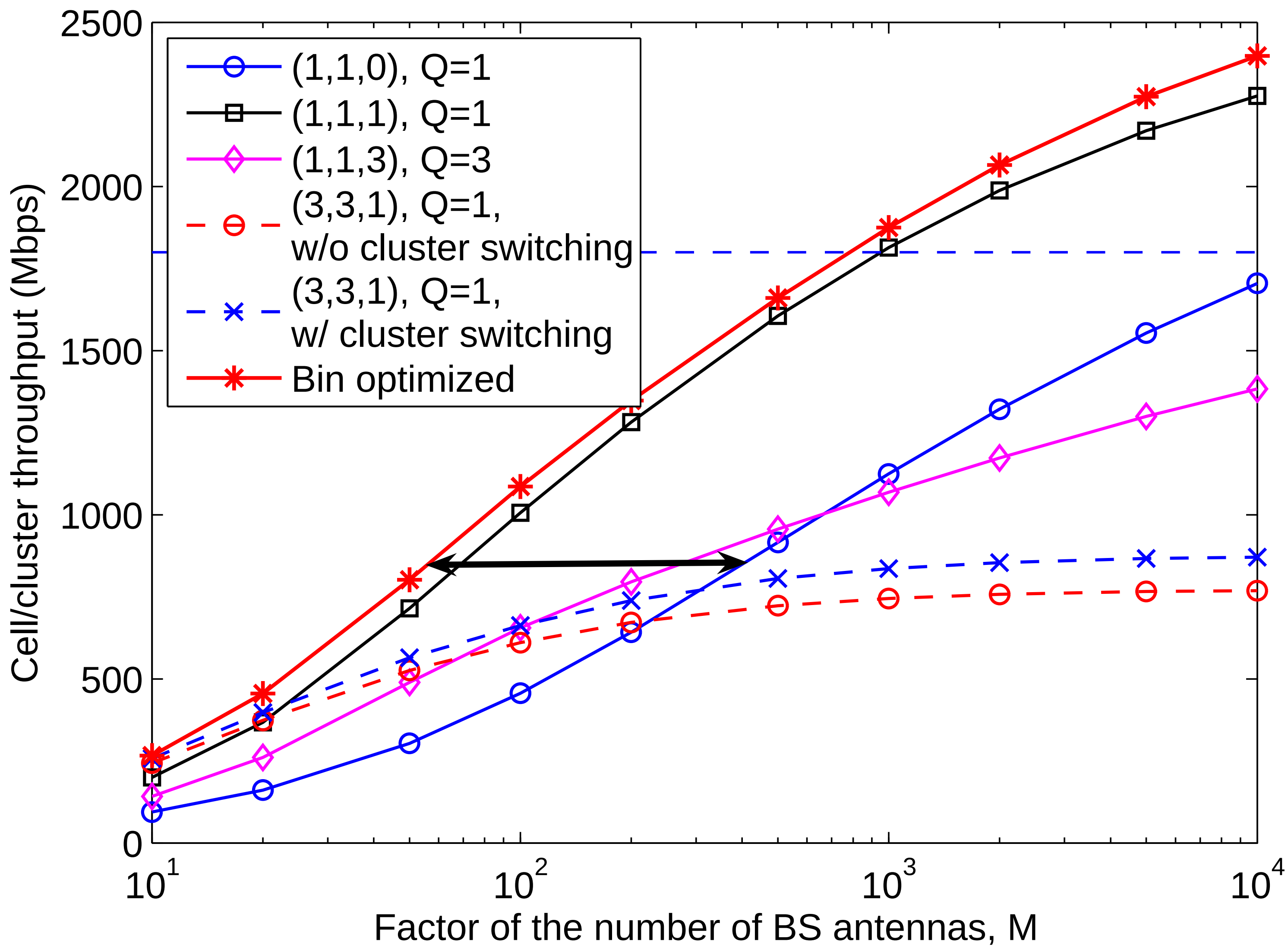


Optimal scheme at each user locations. $M = 20$ and 100, $K = 16$, and $L = 84$.

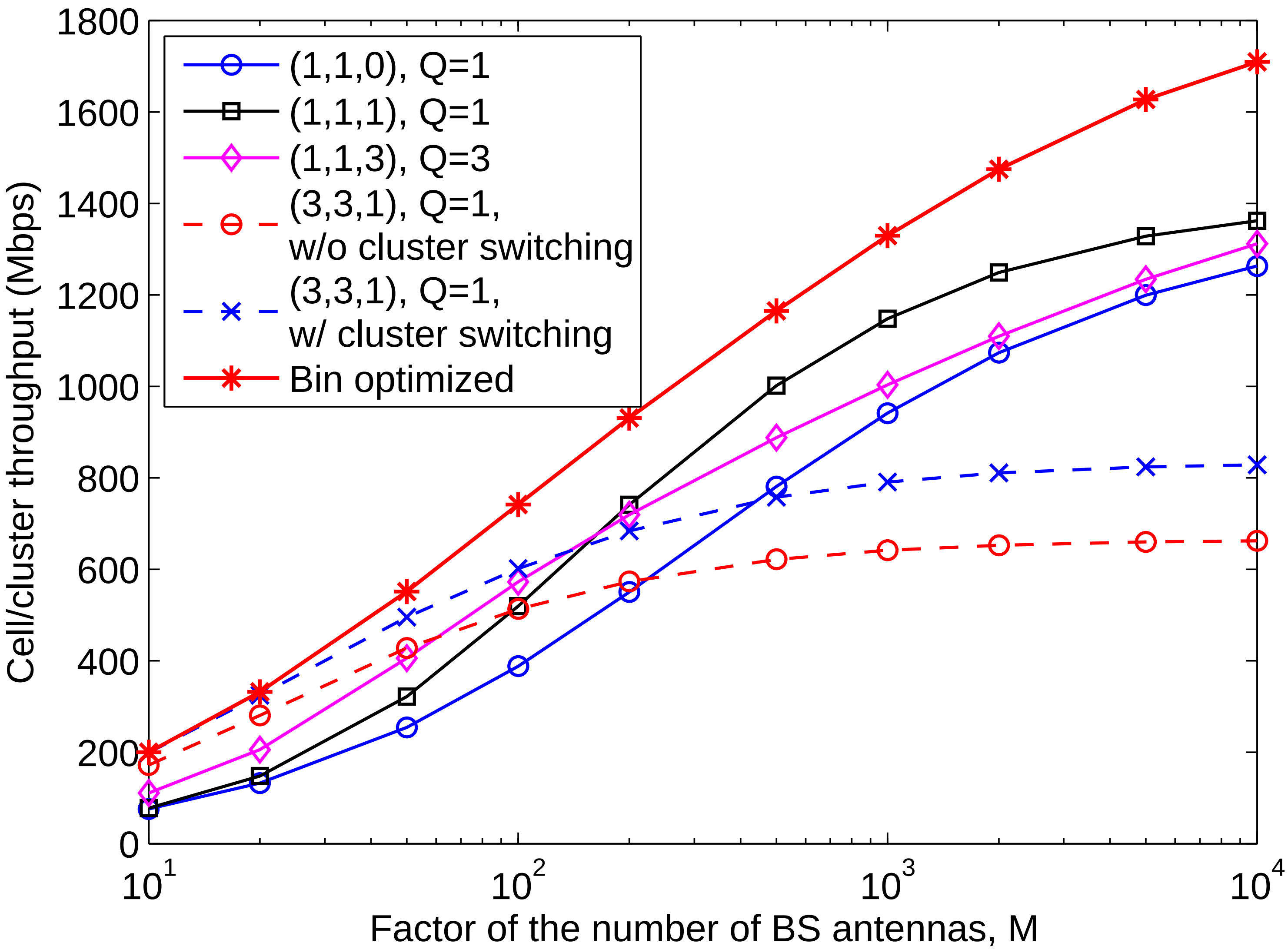


Bin-optimized spectral efficiencies normalized by the (1,1,0) (Marzetta) spectral efficiencies, for $M = 50$, $K = 48$,and $L = 84$.

Performance under PFS

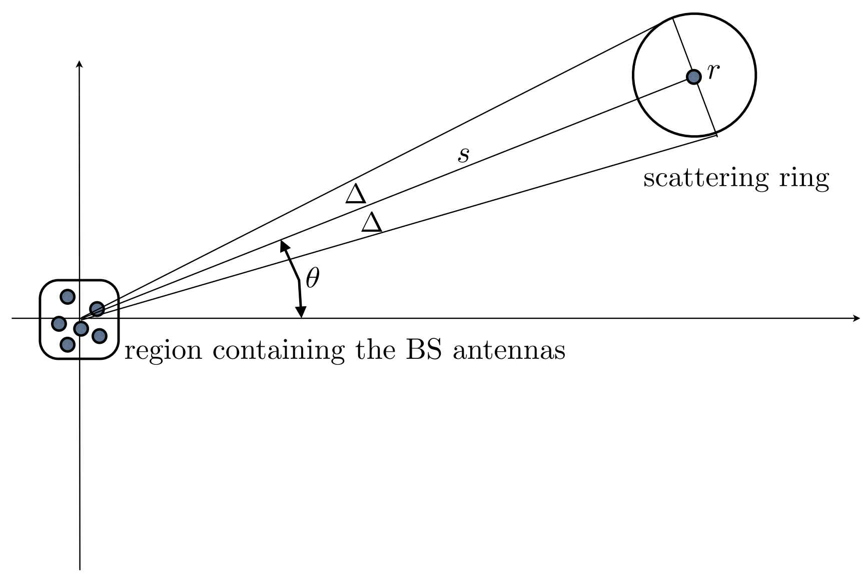


Performance under Max-Min Fairness



What about FDD systems? Exploiting Tx antenna correlation

- Users separated by a few meters (say 10 λ) are practically uncorrelated.
- In contrast, the base station sees user groups at different AoAs under narrow Angular Spread $\Delta \approx \arctan(\mathsf{r}/\mathsf{s})$.



- Tx antenna correlation:

$$\mathbf{h} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{w}, \quad \mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{H}}$$

with

$$[\mathbf{R}]_{m,p} = \frac{1}{2\Delta}\int_{-\Delta}^{\Delta} e^{j\mathbf{k}^{\mathsf{T}}(\alpha+\theta)(\mathbf{u}_m - \mathbf{u}_p)} d\alpha.$$

- The downlink channel model is given by

$$\mathbf{y} = \underline{\mathbf{H}}^{\mathsf{H}}\mathbf{x} + \mathbf{z} = \underline{\mathbf{H}}^{\mathsf{H}}\mathbf{V}\mathbf{d} + \mathbf{z}$$

where $\underline{\mathbf{H}}$ is the $M \times K$ system channel matrix (channel vectors by columns).

Joint Space Division and Multiplexing (JSDM)

- K users selected to form G groups, with $\approx$ same channel correlation.

$$\underline{\mathbf{H}} = [\mathbf{H}_1, \ldots, \mathbf{H}_G], \text{ with } \mathbf{H}_g = \mathbf{U}_g \mathbf{\Lambda}_g^{1/2} \mathbf{W}_g.$$

- Two-stage precoding: $\mathbf{V} = \mathbf{B}\mathbf{P}$.
- $\mathbf{B} \in \mathbb{C}^{M \times b}$ is a pre-beamforming matrix function of $\{\mathbf{U}_g, \mathbf{\Lambda}_g\}$ only.
- $\mathbf{P} \in \mathbb{C}^{b \times s}$ is a precoding matrix that depends on the effective channel.
- The effective channel matrix is given by

$$\underline{\mathbf{H}}^{\mathsf{H}} = \begin{bmatrix} \mathbf{H}_1^{\mathsf{H}}\mathbf{B}_1 & \mathbf{H}_1^{\mathsf{H}}\mathbf{B}_2 & \cdots & \mathbf{H}_1^{\mathsf{H}}\mathbf{B}_G \\ \mathbf{H}_2^{\mathsf{H}}\mathbf{B}_1 & \mathbf{H}_2^{\mathsf{H}}\mathbf{B}_2 & \cdots & \mathbf{H}_2^{\mathsf{H}}\mathbf{B}_G \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_G^{\mathsf{H}}\mathbf{B}_1 & \mathbf{H}_G^{\mathsf{H}}\mathbf{B}_2 & \cdots & \mathbf{H}_G^{\mathsf{H}}\mathbf{B}_G \end{bmatrix}.$$

- Joint Group Processing: If the estimation and feedback of the transformed channel $\underline{\mathbf{H}}$ can be afforded, the precoding matrix $\mathbf{P}$ is determined as a function of $\underline{\mathbf{H}}$.

- Per-Group Processing: If estimation and feedback of the whole $\underline{\mathbf{H}}$ is still too costly, then each group estimates its own diagonal block $\underline{\mathbf{H}}_g = \mathbf{B}_g^{\mathsf{H}}\mathbf{H}_g$, and $\mathbf{P} = \mathrm{diag}(\mathbf{P}_1, \cdots, \mathbf{P}_G)$.

- This results in

$$\mathbf{y}_g = \mathbf{H}_g^{\mathsf{H}}\mathbf{B}_g\mathbf{P}_g\mathbf{d}_g + \sum_{g' \neq g} \mathbf{H}_g^{\mathsf{H}}\mathbf{B}_{g'}\mathbf{P}_{g'}\mathbf{d}_{g'} + \mathbf{z}_g$$

Achieving capacity with reduced CSI

- Assume that the G groups are such that $\underline{\mathbf{U}} = [\mathbf{U}_1, \cdots, \mathbf{U}_G]$ is $M \times rG$ *tall unitary* (i.e., $rG \leq M$ and $\underline{\mathbf{U}}^{\mathsf{H}}\underline{\mathbf{U}} = \mathbf{I}$).

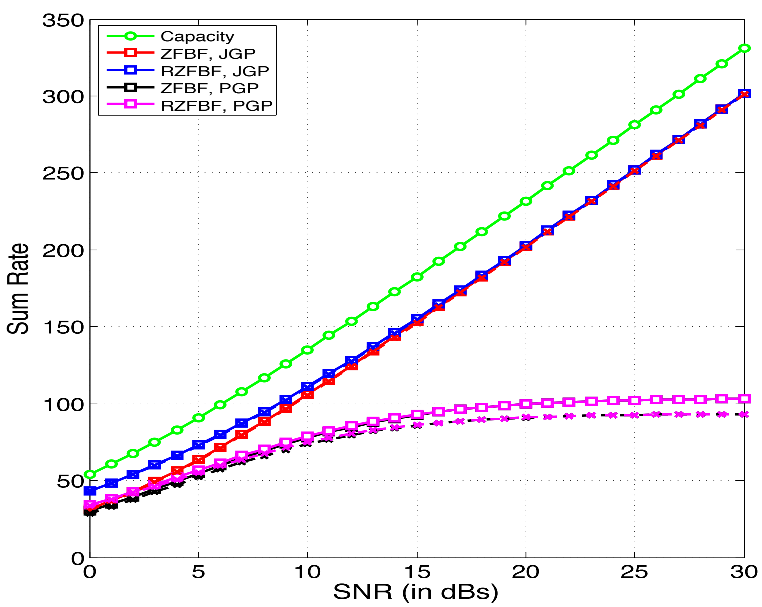
- We choose $b' = r$ and $\mathbf{B}_g = \mathbf{U}_g$ and obtain exact Block Diagonalization (BD):

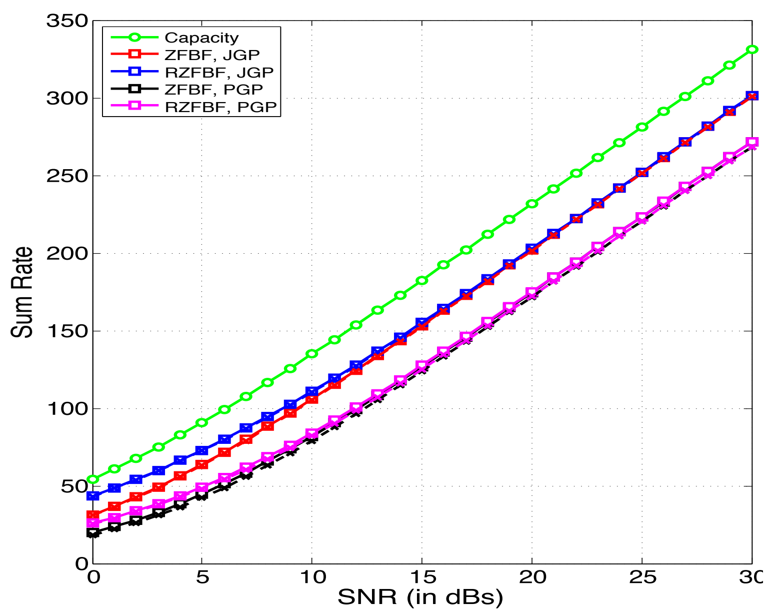
$$\mathbf{y}_g = \mathbf{H}_g^{\mathsf{H}}\mathbf{B}_g\mathbf{P}_g\mathbf{d}_g + \mathbf{z}_g = \mathbf{W}_g^{\mathsf{H}}\mathbf{\Lambda}_g^{1/2}\mathbf{P}_g\mathbf{d}_g + \mathbf{z}_g \tag{1}$$

Theorem 3. For $\underline{\mathbf{U}}$ tall unitary, the sum capacity of the original Gaussian vector broadcast channel with full CSI is equal to the sum capacity of the set of decoupled channels (1). ■

Analysis and results

- Analysis possible using the "deterministic equivalent method" (see [Couillet, Debbah, CUP 2011]).
- Example: $M = 100$, $G = 6$ user groups, Rank$(\mathbf{R}_g) = 21$, we serve 5 users per group with $b' = 10$.
- Sum throughput (bit/s/Hz) vs. SNR (dB) , approximated BD and regularized ZF, $r^\star = 6$ and $r^\star = 12$.





Remarks

- Full CSI: 100×30 channel matrix $\Rightarrow$ 3000 complex channel coefficients per coherence block (CSI feedback), with 100×100 unitary "common" pilot matrix for downlink channel estimation.

- JSDM with PGP: $6 \times 10 \times 5$ diagonal blocks $\Rightarrow$ 300 complex channel coefficients per coherence block (CSI feedback), with 10×10 unitary "dedicated" pilot matrices for downlink channel estimation, sent in parallel to each group through the pre-beamforming matrix.

- One order of magnitude saving in both downlink training and CSI feedback.

- 150 bit/s/Hz at SNR $= 18$ dB: 5 bit/s/Hz per user, for 30 users served simultaneously on the same time-frequency slot.

Is the tall unitary realistic?

- For a Uniform Linear Array (ULA), $\mathbf{R}$ is Toeplitz, with elements

$$[\mathbf{R}]_{m,p} = \frac{1}{2\Delta}\int_{-\Delta}^{\Delta} e^{-j2\pi D(m-p)\sin(\alpha+\theta)}d\alpha, \quad m,p \in \{0,1,\ldots,M-1\}$$

- We use Szego's asymptotic theory of Toeplitz matrices.

Theorem 4. *The empirical eigenvalue distribution of* $\mathbf{R}$ *can be approximated by*

$$\lim_{M\to\infty} F_{\mathbf{R}}(\lambda) = \mu\{S(\xi) \leq \lambda\}$$

where

$$S(\xi) = \sum_{m=-\infty}^{\infty} [\mathbf{R}]_{m,0} = \frac{1}{2\Delta} \sum_{m\in[D\sin(-\Delta+\theta)+\xi, D\sin(\Delta+\theta)+\xi]} \frac{1}{\sqrt{D^2-(m-\xi)^2}}.$$

■

Theorem 5. *The asymptotic normalized rank of the channel covariance matrix* $\mathbf{R}$ *with antenna separation* λD*, AoA* θ *and AS* Δ*, is given by*

$$\rho = \lim_{M\to\infty} \frac{1}{M}\mathrm{Rank}(\mathbf{R}) = \min\{1, B(D,\theta,\Delta)\},$$

where

$$B(D,\theta,\Delta) = \left|D\sin(-\Delta+\theta) - D\sin(\Delta+\theta)\right|.$$

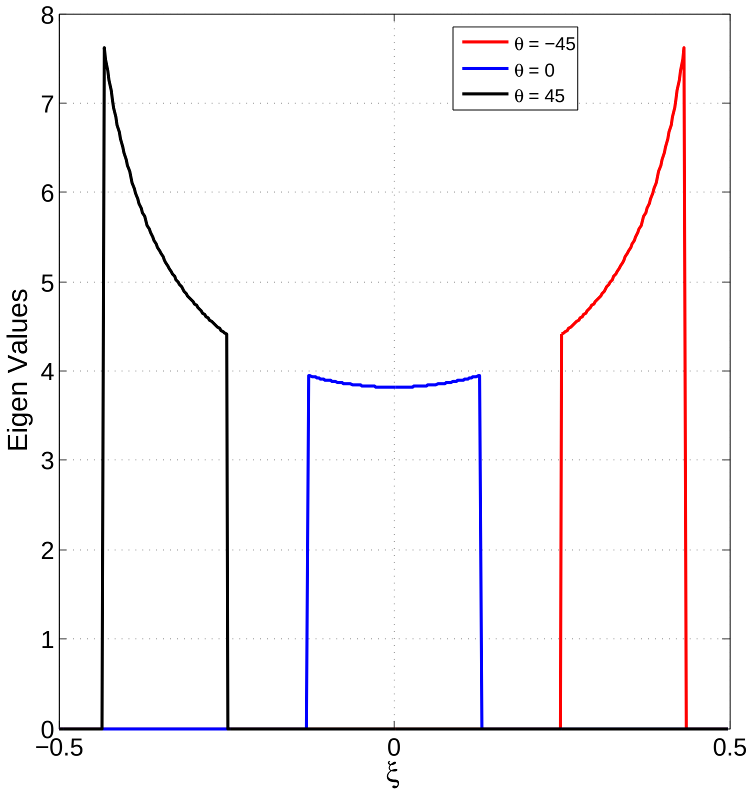
■

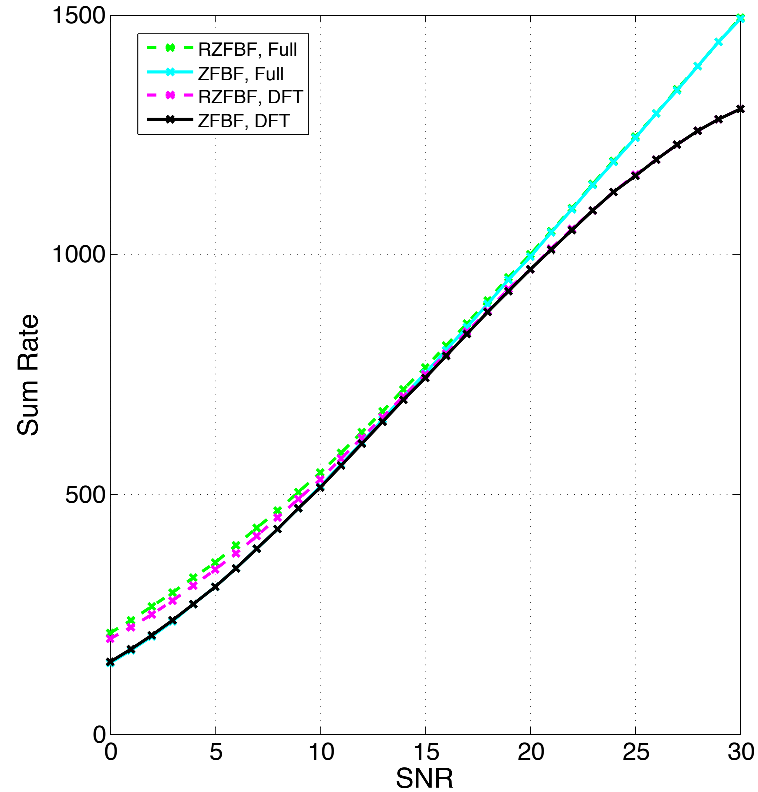
Theorem 6. *Let $\mathcal{S}$ denote the support of $S(\xi)$, let $\mathcal{J}_{\mathcal{S}} = \{m : [m/M] \in \mathcal{S}, m = 0, \ldots, M-1\}$ be the set of indices for which the corresponding "angular frequency" $\xi_m = [m/M]$ belongs to $\mathcal{S}$, let $\mathbf{f}_m$ denote the m-th column of the unitary DFT matrix $\mathbf{F}$, and let $\mathbf{F}_{\mathcal{S}} = (\mathbf{f}_m : m \in \mathcal{J}_{\mathcal{S}})$ be the DFT submatrix containing the columns with indices in $\mathcal{J}_{\mathcal{S}}$. Then,*

$$\lim_{M \to \infty} \frac{1}{M} \left\| \mathbf{U}\mathbf{U}^{\mathsf{H}} - \mathbf{F}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}^{\mathsf{H}} \right\|_F^2 = 0,$$

where $\mathbf{U}$ is the $M \times r$ "tall unitary" matrix of the non-zero eigenvectors of $\mathbf{R}$. ■

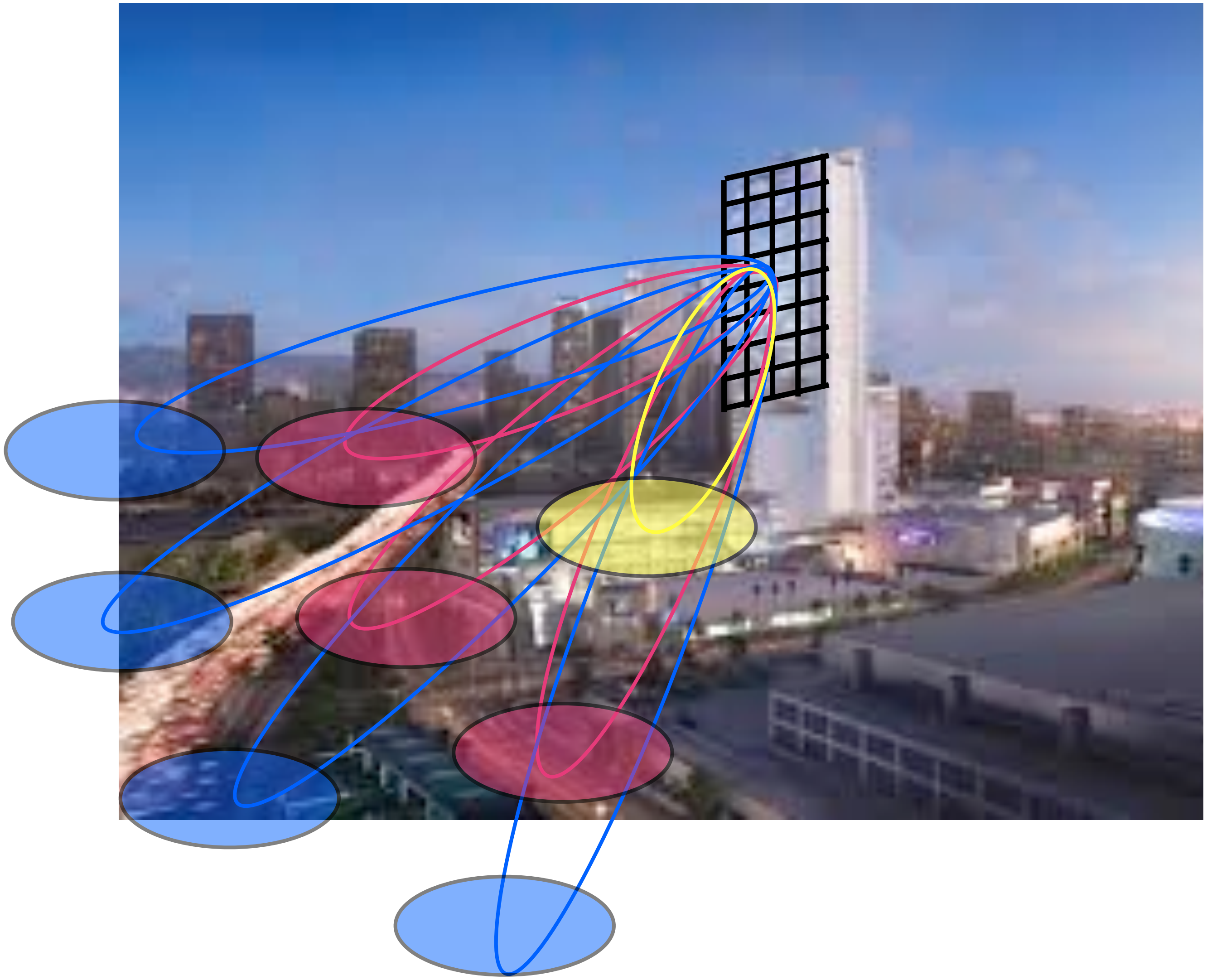
Corollary 1. *Groups g and g' with angle of arrival θ_g and $\theta_{g'}$ and common angular spread Δ have spectra with disjoint support if their AoA intervals $[\theta_g - \Delta, \theta_g + \Delta]$ and $[\theta_{g'} - \Delta, \theta_{g'} + \Delta]$ are disjoint.* ■





- ULA with $M = 400$, $G = 3$, $\theta_1 = \frac{-\pi}{4}, \theta_2 = 0, \theta_3 = \frac{\pi}{4}$, $D = 1/2$ and $\Delta = 15$ deg.

Super-Massive MIMO



- Idea: produce a 3D pre-beamforming by Kronecker product of a "vertical" beamforming, separating the sector into L concentric regions, and a "horizontal" beamforming, separating each ℓ-th region into G_ℓ groups.

- Horizontal beam forming is as before.

- For vertical beam forming we just need to find one dominating eigenmode per region, and use the BD approach.

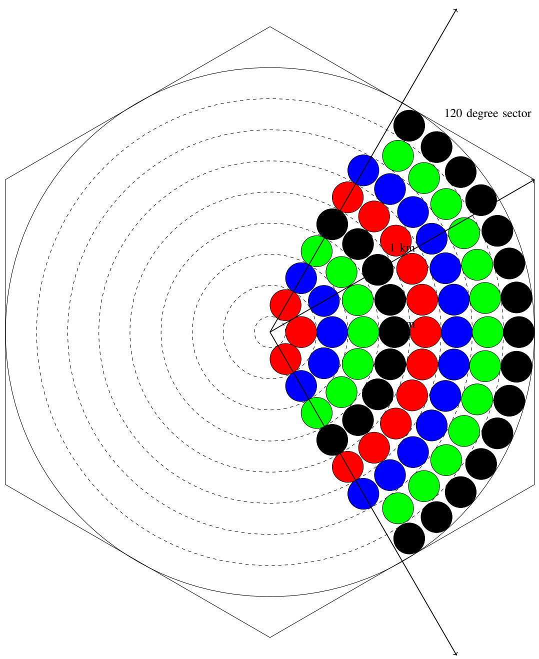
- A set of simultaneously served groups forms a "pattern".

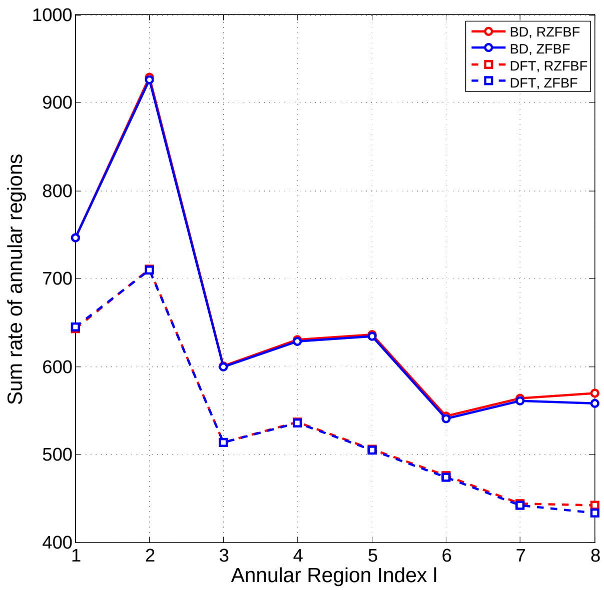
- Patterns need not cover the whole sector.

- Different intertwined patterns can be multiplexed in the time-frequency domain in order to guarantee a fair coverage.

An example

- Cell radius 600m, group ring radius 30m, array height 50m, $M = 200$ columns, $N = 300$ rows.
- Pathloss $g(x) = \frac{1}{1+(\frac{x}{d_0})^\delta}$ with $\delta = 3.8$ and $d_0 = 30$m.
- Same color regions are served simultaneously. Each ring is given equal power.



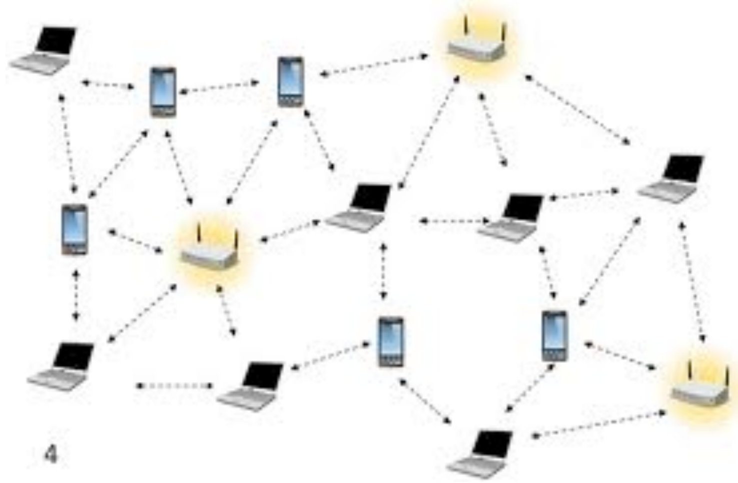


Sum throughput (bit/s/Hz) under PFS and Max-min Fairness

Scheme	Approximate BD	DFT based
PFS, RZFBF	1304.4611	1067.9604
PFS, ZFBF	1298.7944	1064.2678
MAXMIN, RZFBF	1273.7203	1042.1833
MAXMIN, ZFBF	1267.2368	1037.2915

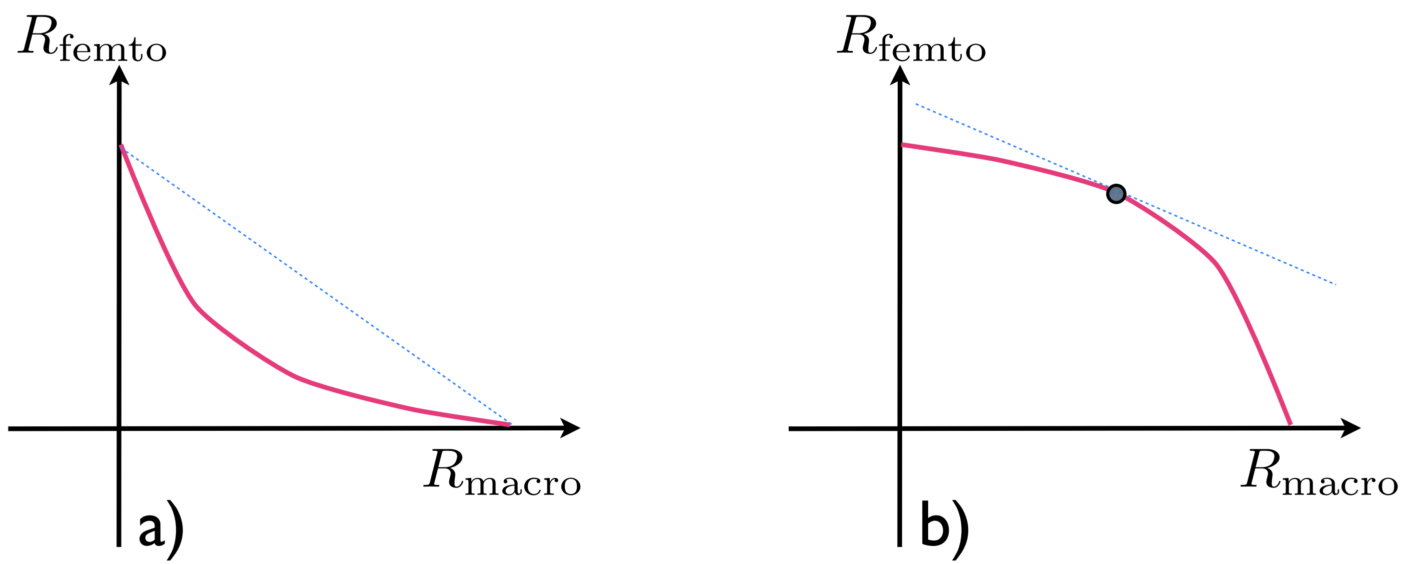
1000 bit/s/Hz $\times$ 40 MHz of bandwidth $= 40$ Gb/s per sector.

Heterogeneous/D2D Wireless Networks



- Scaling laws of D2D wireless networks: $C = O(\sqrt{K})$ bit$\times$meter/second.
- If source-destination are at distance $O(1)$, then the per-connection throughput vanishes as $O(\frac{1}{\sqrt{K}})$.
- If source-destination are at distance $O(1/\sqrt{K})$, then the per-connection throughput is constant $O(1)$.
- Source-destination pairs at 1 hop $\Longrightarrow$ Small Cells or D2D with Caching.

User-deployed small cells tier: In-Band or Out-of-Band?



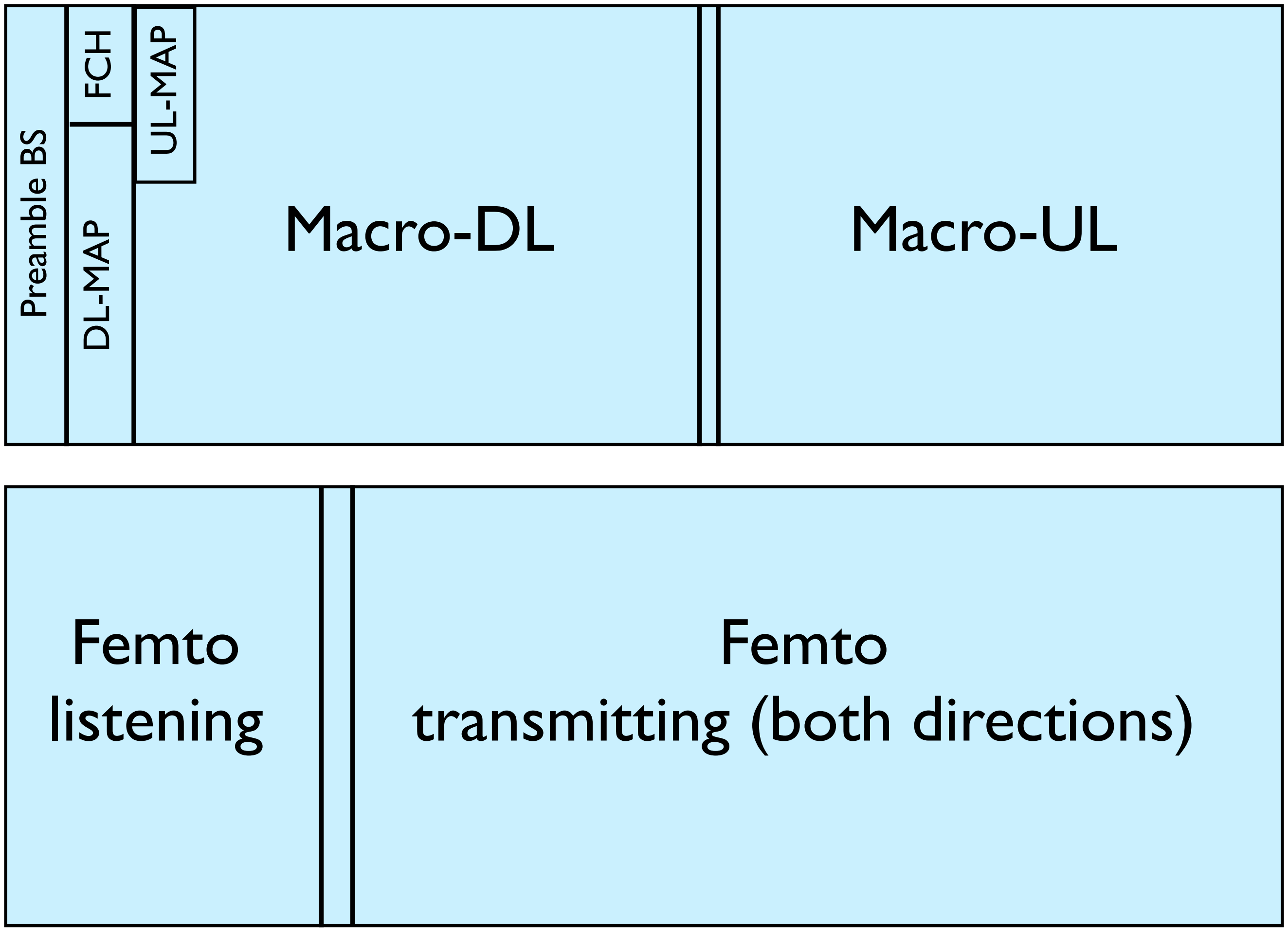
- In case a), time-sharing or bandwidth splitting is optimal (tier 1 and tier 2 on orthogonal dimensions).

- In case b), orthogonalization is not optimal, and tier 1 and tier2 should *interfere*.

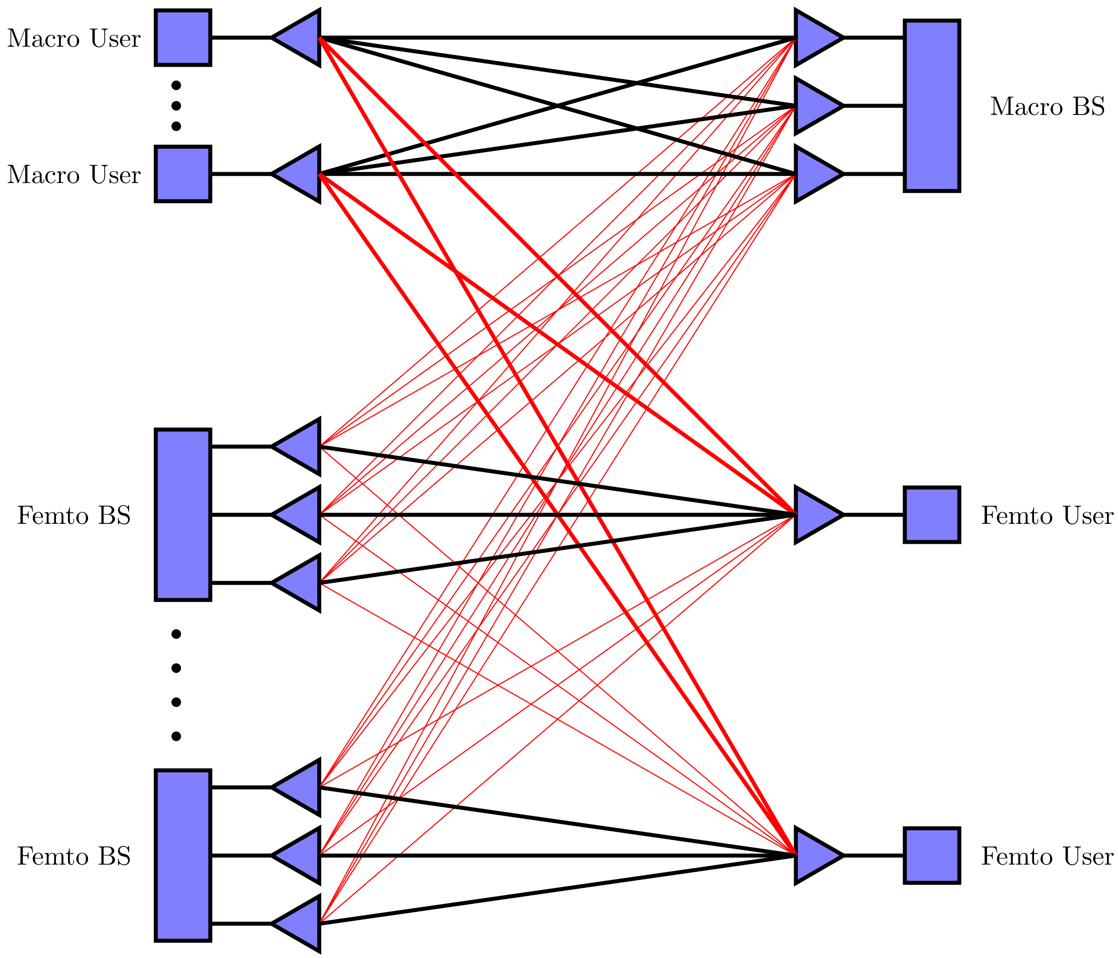
Smart Spectrum Reuse via “Cognition”

- Small Cells operate in TDD, macrocell operates either in TDD or in FDD (in this talk we focus on TDD macrocell).
- Small Cells overhear the macrocell control channel (similar to relays in WiMax 802.16j).
- Small Cells are aware of their location, and of the location of the macrocell users being scheduled.
- Open-access: any macrocell user inside the range of a small cell is absorbed.
- Closed-access: macrocell users can be anywhere, even inside a small cell.

Frame Structure for Cognition



Reverse TDD



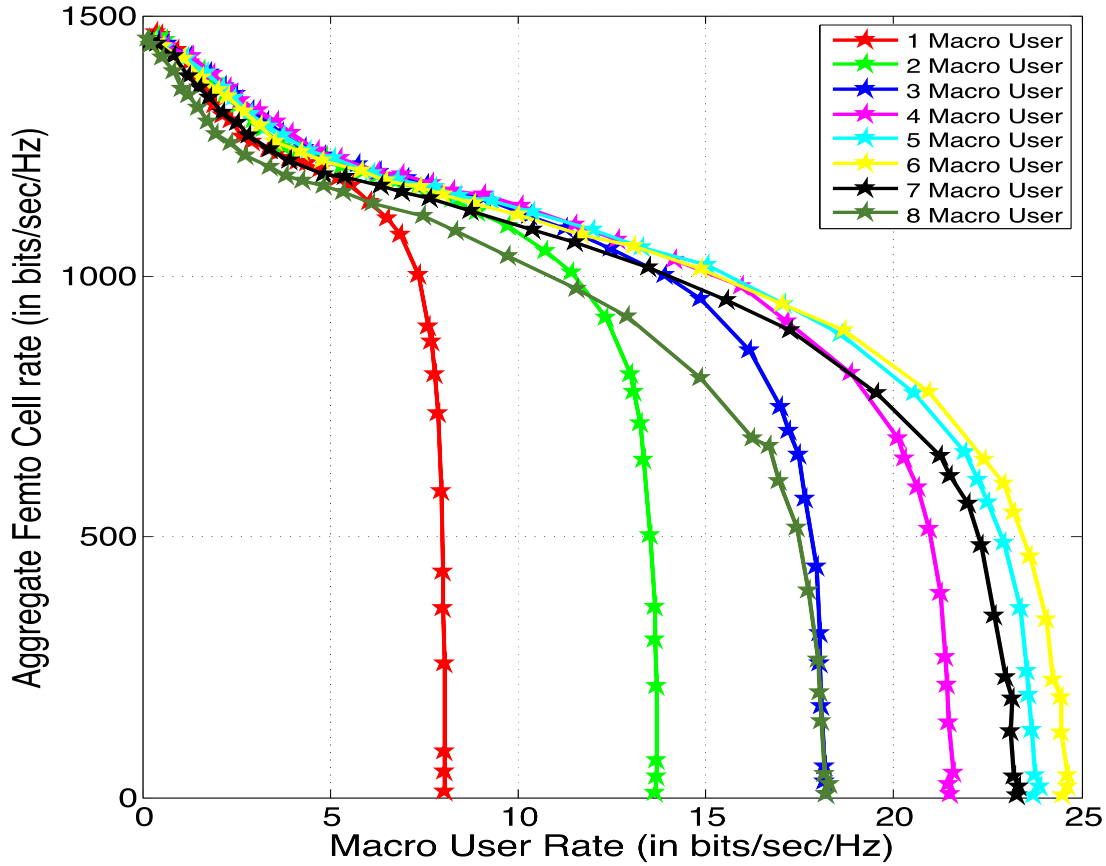
- We align the Macro DL with the Femto UL, and Vice-Versa.

MISO/SIMO Interference Channel and UL/DL Duality

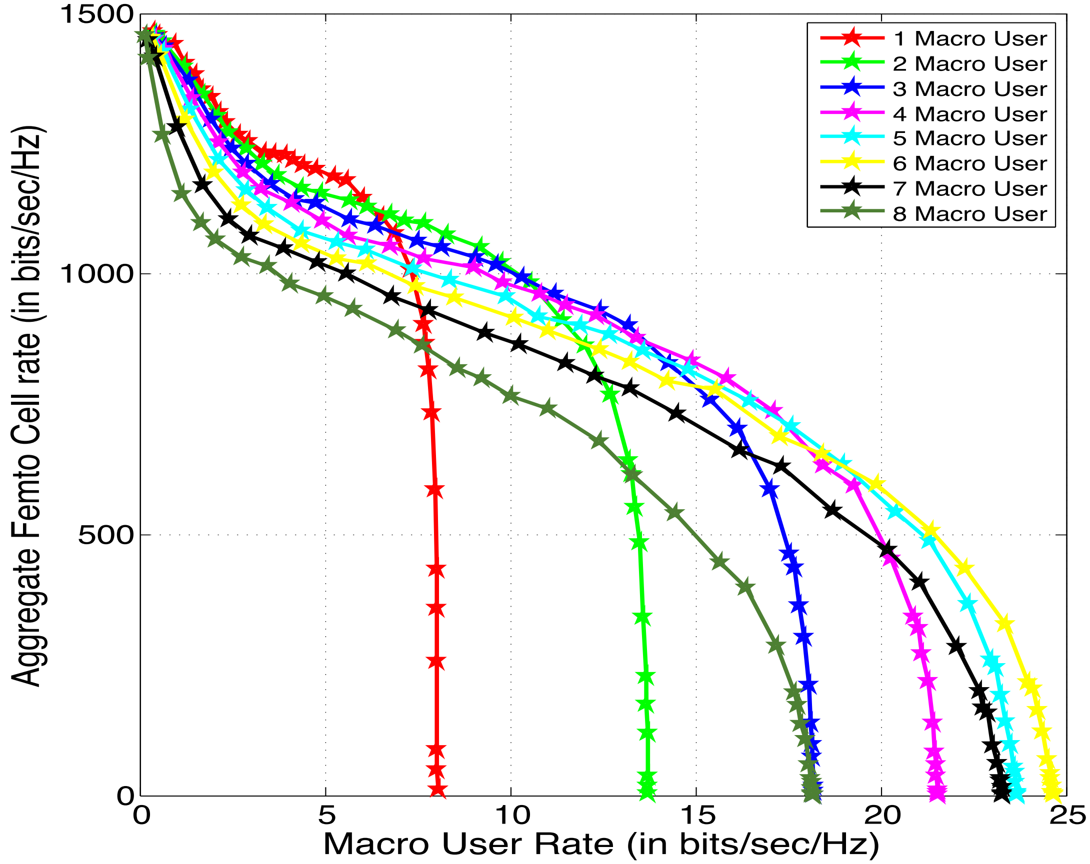
- In the macro-DL/femto-UL, we use interference temperature PC as in the baseline scheme.
- Femto APs use linear MMSE (optimal linear receivers).
- In the macro-UL/femto-DL, we use the MMSE receiving vectors as transmit beamforming vectors.
- By UL/DL duality, there exist a power assignment of the Femto APs and of the Macro user powers such that:

1. The sum-power is the same.
2. The SINR are the same.

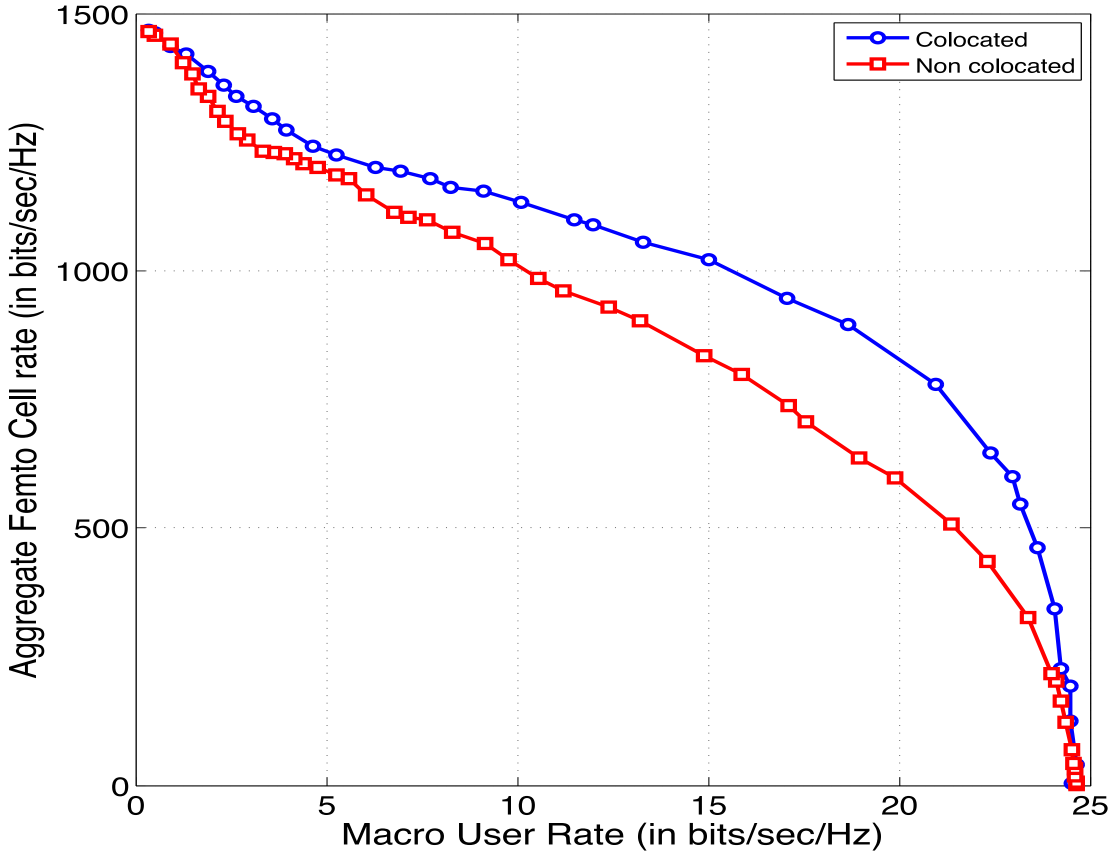
Femto-UL/Macro-DL with co-located macro UTs



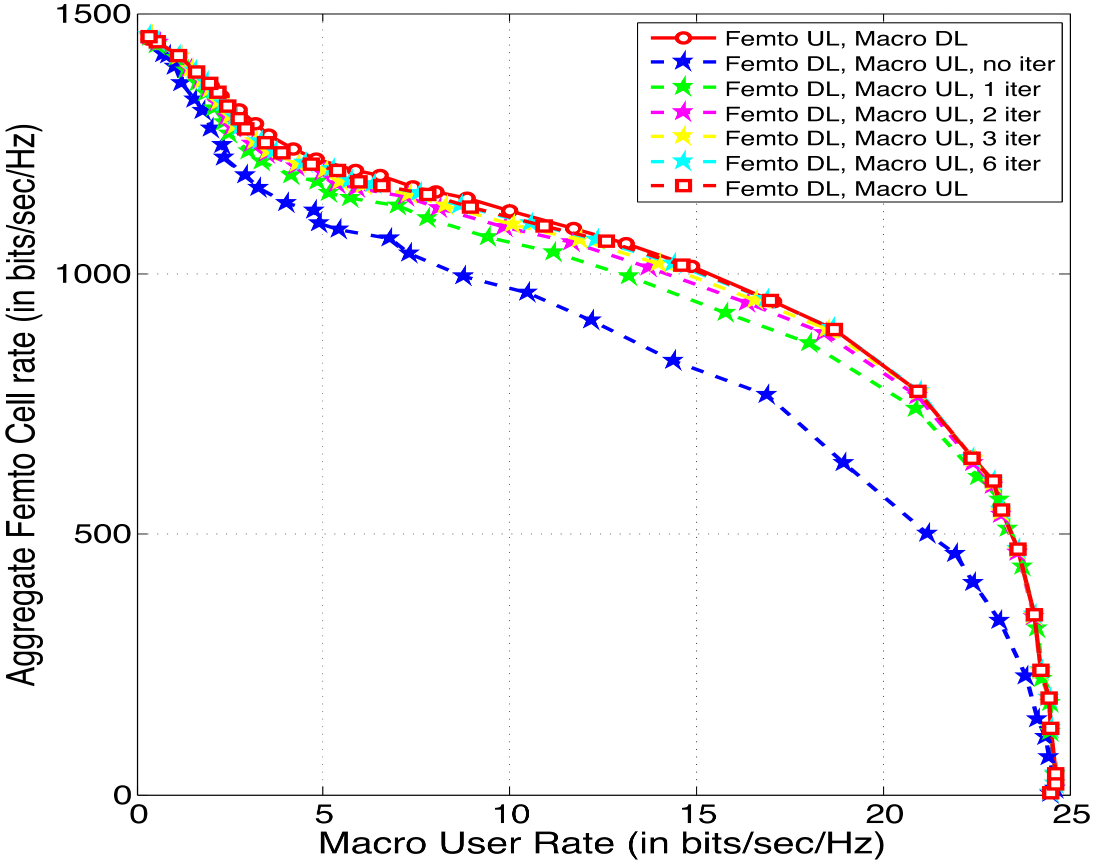
Femto-UL/Macro-DL with non co-located macro UTs



Co-located vs. non co-located: comparison



Femto-DL/Macro-UL: iterative power allocation

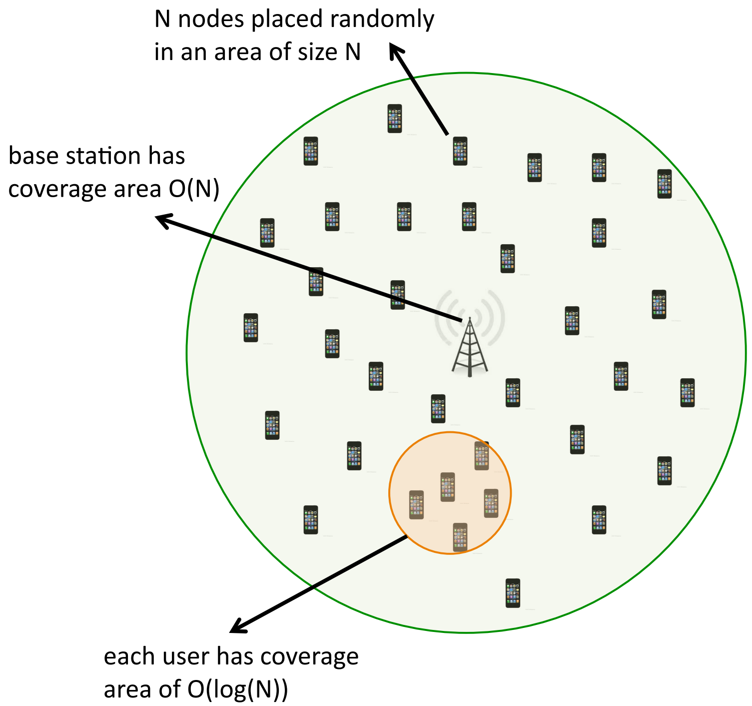


Performance

- Let's focus on the point $(15, 1000)$ and assume 40 MHz of system bandwidth.
- Macro users: 6 users per cell at 2.5 bit/s/Hz yields 100 Mb/s per user.
- Femto users: 625 femtocells per cell at 1.6 bit/s/Hz yields 64 Mb/s per femtocell.
- These rates are in line with today's target peak rates for LTE and WLANs (Wifi).
- The two systems can co-exist in the same system bandwidth.
- In terms of system special efficiency, we go well above the desired x100 increase, with relatively conventional technology.
- **Key point to take home: multiuser MIMO and inter-tier interference management must be at the core of the system design, not added later as "afterthoughts".**

Even Denser Spatial Reuse: Distributed Caching in Wireless Devices

- Cache predictable Internet content (web-pages, coded video) into the user devices and auxiliary wireless "helpers".



- N nodes in an area of size N.
- The base station has total downlink capacity $C_{\rm base}$ bit/s/Hz.
- Short-range D2D links between the user terminals can support capacity $C_{\rm d2d}$.
- Following the current LTE-Advanced eICIC the base station leaves a fraction β of the time-frequency slots free for D2D communication.
- Random placement of content in the caches. Probability that a given cache satisfies a random demand: $0 < p_{\rm cache} \leq 1$.
- Range of D2D communication such that the number of nodes reachable from any given node in one hop is $c \log N$, for some $c > 0$.
- Probability of not finding the requested file in the neighboring caches is $\bar{p} = (1 - p_{\rm cache})^{c \log N}$.
- Let the fraction of users originating demands (active users) be α, and let the individual rate per user be r bit/s/Hz.

- The demands not found in the local caches are handled directly by the base station. Hence, we have the constraint

$$r\alpha N\bar{p} \leq (1-\beta)C_{\text{base}}. \tag{2}$$

- The demands found in the neighboring caches are handled by D2D communication.

- Using simple *interference avoidance*, we can schedule $\frac{N}{c \log N}$ non-interfering links simultaneously on each time-frequency slot freed by the base station.

- The source rate constraint for the traffic handled by the caches is

$$r\alpha N(1-\bar{p}) \leq \beta \frac{N}{c \log N} C_{\text{d2d}}. \tag{3}$$

- By solving for the optimum β and replacing it in (3), we find

$$r \leq \frac{1}{1 - \bar{p} + \frac{\bar{p}N}{c \log N} \frac{C_{\rm d2d}}{C_{\rm base}}} \frac{C_{\rm d2d}}{\alpha c \log N}. \tag{4}$$

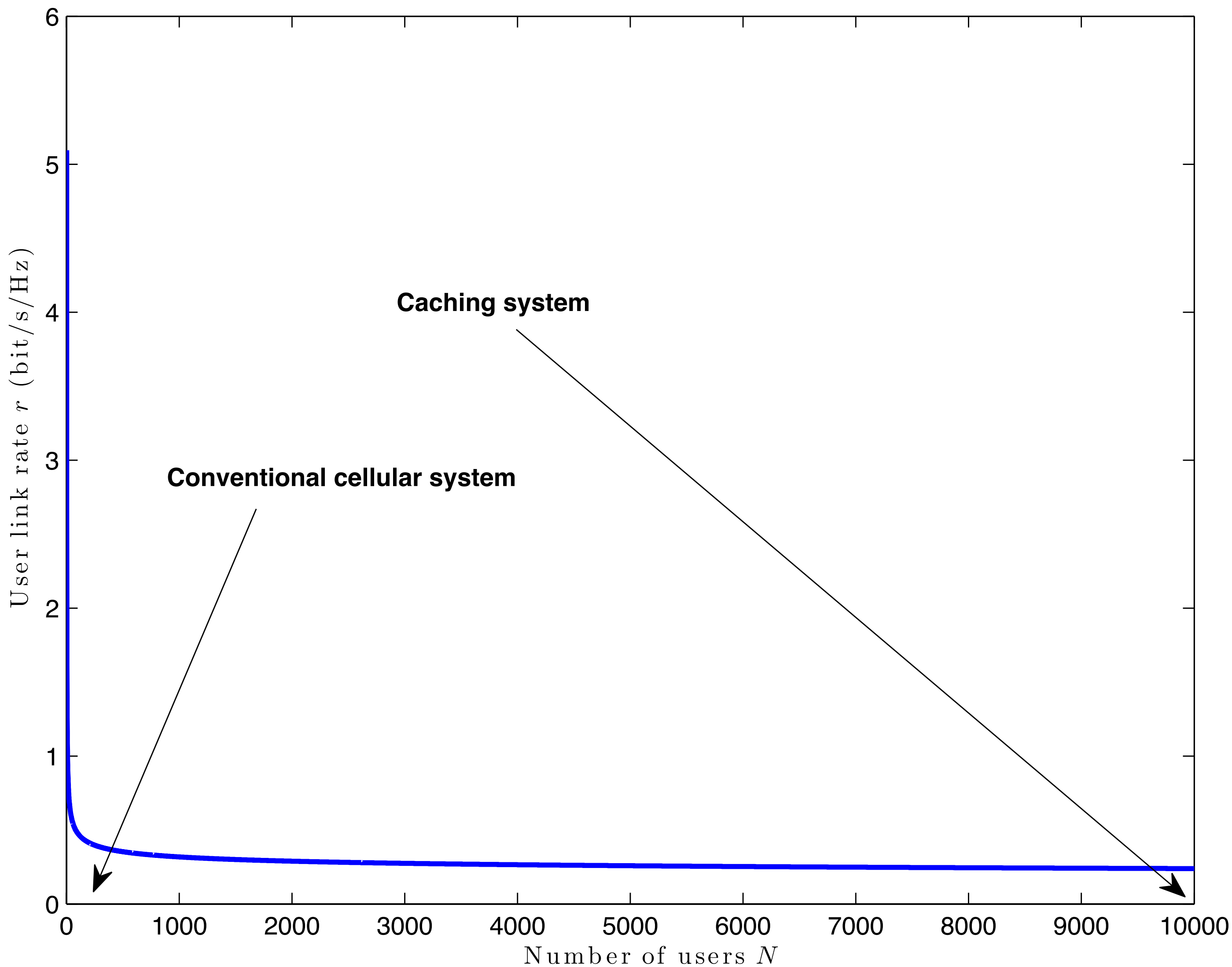
- Since $\bar{p} = N^{-c \log \frac{1}{1-p_{\rm cache}}}$, for $c \log \frac{1}{1-p_{\rm cache}} > 1$, we have that $N\bar{p} \to 0$ polynomially with N.

- As a consequence, $r \approx \frac{C_{\rm d2d}}{\alpha c \log N}$ vanishes only logarithmically with N.

- The gain over a conventional cellular system, achieving system: $r_{\rm conv} = \frac{C_{\rm base}}{\alpha N}$, is unbounded !!!

Example

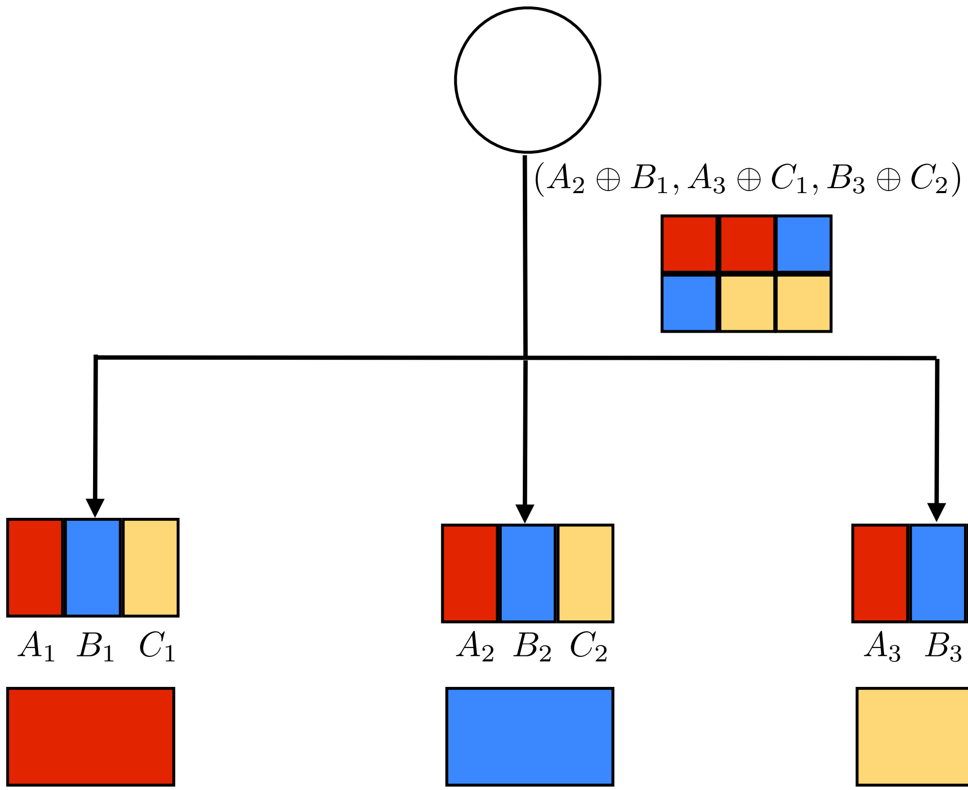
- Realistic LTE downlink capacity and a D2D link capacity inspired by Qualcomm FlashLinQ.
- $C_{\text{base}} = 5$ bit/s/Hz and $C_{\text{d2d}} = 2$ bit/s/Hz.
- Assume a cell bandwidth of 40 MHz and a target per-user rate of 10 Mb/s resulting in $r = 0.25$ bit/s/Hz.
- Assuming a user activity factor $\alpha = 0.2$, a conventional system would serve $N = 100$ users.
- With a modest cache hit probability $p_{\text{cache}} = 0.2$, requiring $c > \frac{1}{\log(1.25)} = 4.4814$, and letting $c = 4.5$, the proposed system serves $N = 10000$ users (we meet the target 100x capacity boost).



6
5
4
3
2
1
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
User link rate r (bit/s/Hz)
Number of users N
Caching system
Conventional cellular system

Global Caching Gain

- Recent result [Maddah-Ali, Neesen, arXiv:1209.5807] .... caching turns broadcast into multicast.



Conclusions

- Network MIMO (CoMP): appears to be fundamentally limited in conventional cellular systems.
- Large number of antennas at each BS: essentially no need for BS cooperation, beyond simple coordination of scheduling/frequency/pilots/beams.
- Large number of antennas naturally suited to TDD: but also possible with FDD, if Tx antenna correlation is properly exploited (JSDM).
- Further improvements: reduce the distance between source and destination.
- HetNets: cognitive multi-antenna small cells can share the same macro bandwidth.
- D2D: for throughput is meaningful if coupled with caching.
- Further caching gains from (index) coding.

Thank You