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

Wireless operators’ nightmare

- 100x Data traffic increase, due to the introduction of powerful multimedia capable user devices.
- Operating costs trends not matched by revenue trends.
• 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?
• 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 \( O\left(\frac{1}{K}\right) \).
Network MIMO: A Large-System Analysis
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

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

- Reference cluster channel model

$$y = H^H x + z$$

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

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

$$\text{maximize} \quad \sum_{k=1}^{A} \sum_{i=1}^{N} W_k^{(i)} R_k^{(i)}$$

subject to $\mathbf{R} \in \mathcal{R}_{\text{lzfb}}(\mathbf{H})$

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

- The scheduler picks a fraction $\mu_k$ 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 $\mu = (\mu_1, \ldots, \mu_A)$ denote the fractions of active users in groups $1, \ldots, A$, respectively. For given $\mu$, the corresponding effective channel matrix is given by

$$H\mu = \begin{bmatrix}
\beta_{1,1}H_{1,1}(\mu_1) & \cdots & \beta_{1,A}H_{1,A}(\mu_A) \\
\vdots & \ddots & \vdots \\
\beta_{B,1}H_{B,1}(\mu_1) & \cdots & \beta_{B,A}H_{B,A}(\mu_A)
\end{bmatrix},$$
LZFB “parallel channels”

- Letting $V_\mu = H_\mu^+ \Lambda_\mu^{1/2}$, we obtain the “parallel” channel model

$$y_\mu = H_\mu^H V_\mu Q^{1/2} u + z_\mu = \Lambda_\mu^{1/2} Q^{1/2} u + z_\mu.$$

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

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

where $(\eta_1(\mu), \ldots, \eta_B(\mu))$ 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 $\eta = \{\eta_m\}$.  ■
We assume that the channels are constant over time-frequency blocks of size $WT$ complex dimensions.

For each such block, $\gamma_pBN$ 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 $\gamma_p/\gamma$ denotes the “pilot dimensionality overhead”.
From the well-known MMSE decomposition, the channel matrix $H$ can be written as $H = \hat{H} + E$, where

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

with

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

and where

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

with

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

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

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

where

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

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

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

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


The system spectral efficiency must be scaled by the factor \(1 - \frac{\gamma_p^N B}{W T}\), that takes into account the downlink training overhead, i.e., fraction of dimensions per block dedicated to (downlink) training.

In particular, letting \(\tau = \frac{N}{W T}\) 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 \(\tau\), in the limit of \(N \to \infty\).

The ratio \(\tau\) captures the “dimensional crowding” of the system.
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 ($N = 2, 4, \text{ and } 8$) for cooperation clusters of size $B=1, 2, \text{ and } 8$, with perfect CSIT. $M = 8$ cells and $K = 64$ user groups.
Cell sum rate versus the antenna ratio $\gamma$ for cooperation clusters of size $B=1$, 2, and 8. $M = 8$ cells and $K = 192$ user groups.
Cell sum rate versus the antenna ratio $\gamma$ for cooperation clusters of size $B=1, 2, \text{ and } 8$. $M=8$ cells and $K=192$ user groups.
Large cooperating clusters?

Cell sum rate versus the cluster size $B$ for the antenna ratio $\gamma=1, 2, 4, \text{ and } 8$

$\gamma_p = \gamma$, $M = 24$ 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 $\gamma=1$, 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 $(1 - T_{tr}/T)$ 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.
• 1-dimensional layout with $C = 2$ and $F = 2$. 

- $f = 0$

- $f = 1$
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

Fig. 3. 1-dimensional layout with $C = 2$ and $F = 2$.

Fig. 4. 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.

Fig. 5. Two cases of a precoding scheme for $C = 2$, $F = 1$, and $Q = 2$, with $J = Q$ (a) and $J = C(Q - 1) + 1$ (b). The dashed blue lines indicate the channel vectors of out-of-cluster user locations imposing a ZF constraint. In Figure (b), the light-shaded grey dashed lines indicate the channel vectors assumed zero in the beamforming calculation.
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)

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

- Let $R^*(\mathcal{X}_k)$ 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 $\rho_k$ 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:

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

• 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

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

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

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

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

and yields \( \rho_k = \frac{1}{R^*(X_k)} \cdot \frac{1}{\sum_{j=0}^{K-1} \frac{1}{R^*(X_j)}}. \)
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

Cell/cluster throughput (Mbps) vs. Factor of the number of BS antennas, M

- (1,1,0), Q=1
- (1,1,1), Q=1
- (1,1,3), Q=3
- (3,3,1), Q=1, w/o cluster switching
- (3,3,1), Q=1, w/ cluster switching
- Bin optimized
Performance under Max-Min Fairness

![Graph showing cell/cluster throughput (Mbps) vs. factor of the number of BS antennas, M.](image)

- (1,1,0), Q=1
- (1,1,1), Q=1
- (1,1,3), Q=3
- (3,3,1), Q=1, w/o cluster switching
- (3,3,1), Q=1, w/ cluster switching
- Bin optimized
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 $\Delta \approx \arctan(r/s)$. 

![Diagram showing scattering ring and region containing the BS antennas]
• Tx antenna correlation:

\[ h = U \Lambda^{1/2} w, \quad R = U \Lambda U^H \]

with

\[ [R]_{m,p} = \frac{1}{2\Delta} \int_{-\Delta}^{\Delta} e^{jk^T(\alpha+\theta)(u_m-u_p)} d\alpha. \]

• The downlink channel model is given by

\[ y = H^H x + z = H^H V d + z \]

where \( H \) 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.

$$H = [H_1, \ldots, H_G], \text{ with } H_g = U_g \Lambda_g^{1/2} W_g.$$  

- Two-stage precoding: $V = BP$.

- $B \in \mathbb{C}^{M \times b}$ is a pre-beamforming matrix function of $\{U_g, \Lambda_g\}$ only.

- $P \in \mathbb{C}^{b \times s}$ is a precoding matrix that depends on the effective channel.

- The effective channel matrix is given by

$$H^H = \begin{bmatrix}
H_1^H B_1 & H_1^H B_2 & \cdots & H_1^H B_G \\
H_2^H B_1 & H_2^H B_2 & \cdots & H_2^H B_G \\
\vdots & \vdots & \ddots & \vdots \\
H_G^H B_1 & H_G^H B_2 & \cdots & H_G^H B_G 
\end{bmatrix}.$$
• **Joint Group Processing:** If the estimation and feedback of the transformed channel $\mathbf{H}$ can be afforded, the precoding matrix $\mathbf{P}$ is determined as a function of $\mathbf{H}$.

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

• This results in

$$y_g = \mathbf{H}_g^H \mathbf{B}_g \mathbf{P}_g \mathbf{d}_g + \sum_{g' \neq g} \mathbf{H}_g^H \mathbf{B}_{g'} \mathbf{P}_{g'} \mathbf{d}_{g'} + \mathbf{z}_g$$
• Assume that the $G$ groups are such that $\mathbf{U} = [\mathbf{U}_1, \cdots, \mathbf{U}_G]$ is $M \times rG$ tall unitary (i.e., $rG \leq M$ and $\mathbf{U}^H \mathbf{U} = \mathbf{I}$).

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

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

**Theorem 3.** For $\mathbf{U}$ 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, \( \text{Rank}(\mathbf{R}_g) = 21 \), we serve 5 users per group with \( b' = 10 \).

• Sum throughput (bit/s/Hz) vs. SNR (dB), approximated BD and regularized ZF, \( r^* = 6 \) and \( r^* = 12 \).
Remarks

• Full CSI: $100 \times 30$ channel matrix $\Rightarrow$ 3000 complex channel coefficients per coherence block (CSI feedback), with $100 \times 100$ 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 $10 \times 10$ 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 $\text{SNR} = 18$ dB: 5 bit/s/Hz per user, for 30 users served simultaneously on the same time-frequency slot.
• 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 $\lambda D$, AoA $\theta$ and AS $\Delta$, is given by

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

where

$$
B(D, \theta, \Delta) = |D \sin(-\Delta + \theta) - D \sin(\Delta + \theta)|.
$$
Theorem 6. Let $S$ denote the support of $S(\xi)$, let $\mathcal{J}_S = \{m : [m/M] \in S, m = 0, \ldots, M - 1\}$ be the set of indices for which the corresponding “angular frequency” $\xi_m = [m/M]$ belongs to $S$, let $f_m$ denote the $m$-th column of the unitary DFT matrix $F$, and let $F_S = (f_m : m \in \mathcal{J}_S)$ be the DFT submatrix containing the columns with indices in $\mathcal{J}_S$. Then,

$$
\lim_{M \to \infty} \frac{1}{M} \|UU^H - F_SF_S^H\|_F^2 = 0,
$$

where $U$ is the $M \times r$ “tall unitary” matrix of the non-zero eigenvectors of $R$. ■
**Corollary 1.** Groups $g$ and $g'$ with angle of arrival $\theta_g$ and $\theta_{g'}$ and common angular spread $\Delta$ 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 $\ell$-th region into $G_\ell$ 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 $d_0 = 30m$.
- Same color regions are served simultaneously. Each ring is given equal power.
Sum throughput (bit/s/Hz) under PFS and Max-min Fairness

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Approximate BD</th>
<th>DFT based</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFS, RZFBF</td>
<td>1304.4611</td>
<td>1067.9604</td>
</tr>
<tr>
<td>PFS, ZFBF</td>
<td>1298.7944</td>
<td>1064.2678</td>
</tr>
<tr>
<td>MAXMIN, RZFBF</td>
<td>1273.7203</td>
<td>1042.1833</td>
</tr>
<tr>
<td>MAXMIN, ZFBF</td>
<td>1267.2368</td>
<td>1037.2915</td>
</tr>
</tbody>
</table>

1000 bit/s/Hz × 40 MHz of bandwidth = 40 Gb/s per sector.
• Scaling laws of D2D wireless networks: $C = O(\sqrt{K}) \text{ bit} \times \text{meter/second}$.

• 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 $\Rightarrow$ Small Cells or D2D with Caching.
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

Preamble BS
DL-MAP
UL-MAP
FCH
Macro-DL
Macro-UL
Femto listening
Femto transmitting (both directions)
We align the Macro DL with the Femto UL, and Vice-Versa.
• 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

![Graph showing comparison between co-located and non-co-located Macro User Rate and Aggregate Femto Cell rate.](image-url)
Femto-DL/Macro-UL: iterative power allocation
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”.
• 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 \( C_{\text{base}} \) bit/s/Hz.

• Short-range D2D links between the user terminals can support capacity \( C_{\text{d2d}} \).

• Following the current LTE-Advanced eICIC the base station leaves a fraction \( \beta \) 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: \( 0 < p_{\text{cache}} \leq 1 \).

• 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 \( \bar{p} = (1 - p_{\text{cache}})^{c \log N} \).

• Let the fraction of users originating demands (active users) be \( \alpha \), 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

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

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

• Using simple \textit{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

\[ r\alpha N(1 - \bar{p}) \leq \beta \frac{N}{c\log N}C_{d2d}. \] (3)
• By solving for the optimum $\beta$ and replacing it in (3), we find

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

(4)

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

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

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

- Realistic LTE downlink capacity and a D2D link capacity inspired by Qualcomm FlashLinQ.

- $C_{\text{base}} = 5 \text{ bit/s/Hz}$ and $C_{\text{d2d}} = 2 \text{ bit/s/Hz}$.

- Assume a cell bandwidth of 40 MHz and a target per-user rate of 10 Mb/s resulting in $r = 0.25 \text{ bit/s/Hz}$.

- 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).
Figure 3: a) A system setup demonstrating unbounded gains achievable with caching in the user devices and D2D communication. b) per-user rate (bit/s/Hz) versus $N$ for the caching system.

A large number of users, even simple randomized caching of the most popular content and simple "independent set scheduling" interference avoidance for the D2D links suffices to achieve and surpass the target capacity increase. However, a number of technical challenges must be addressed for an optimized system design that does not only "scale well" in the large $N$ regime, but also provide near-optimal performance for fixed $N$ over a wide range of Signal-to-Noise ratios, fading channel statistics and realistic pathloss models. In addition, the problem of feeding the distributed caches is far from trivial, since this requires the transfer of a massive amount of data to a massive number of wireless nodes. If this is not implemented efficiently, the caching phase will become the system bottleneck. Finally and perhaps most importantly, incentive mechanisms must be designed for distributed caching and user cooperation (i.e., serving demands of other users), since otherwise the users are not motivated to commit their memory and battery power to the system. In the next sections, we tackle research problems targeting these problems, the solution of which are fundamental enablers for the proposed system concept.

2 Proposed Research (9 pages)

We start by presenting a general formulation of the wireless caching problem in an abstract "network level" form, such that it can be clearly defined. This allows us to discuss some fundamental features of the problem and support through a simple example that our approach (i.e., trading memory for spectral efficiency) can indeed meet the desired target spectral efficiency increase. In order to make this possible, though, several problems at the level of communication theory, information theory, scheduling for interference avoidance and design of incentive mechanisms for cooperative distributed caching must be solved. Then, Sections 2.4, ?? and 2.2 are devoted to the description of concrete research sub-problems.

2.1 The centralized wireless caching problem

Consider $N$ wireless users, $M$ helper nodes, and $K$ content files. The wireless users can demand one of the $K$ content files, which could be web pages or chunks of a video file. This request is sent to a server in the core network, which without loss of generality we identify with the serving cellular base station. The base station could then route the request to one of the helper nodes where the demanded file is cached. As noticed in Section 1 [REMEMBER TO ADD THIS POINT IN THE INTRO], a node may be both a helper and a wireless user, depending on whether it generates a demand or it serves other users' demands to files in its cache. We will index the base station as helper 0, while other helpers are indexed $j = 1, \ldots, M$. For simplicity of exposition, we assume that all content files have the same size of 1 data unit. We will denote the cache size of helper $j$ by $B_j$ data units, and the maximum number of simultaneous requests that helper $j$ can serve by $R_j$ data units. We also introduce a state variable $\xi_j$ that accounts for the helper node $j$'s channel...
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