Cloud Processing for 5G Systems

O. Simeone
New Jersey Institute of Technology (NJIT)

Joint work with
S.-H. Park¹, O. Sahin², S. Shamai³, J. Kang⁴ and J. Kang⁴
Cloud Processing for 5G Systems

- What will 5G be? Highly integrative with a flexible and intelligent core network [Andrews et al ’14]
- Enabling technologies: mmwave, massive MIMO, extreme densification, cloud processing, software defined networking,...
Cloud Processing for 5G Systems

- Examples:
  
  **Cloud Radio Access Network (C-RAN):** Baseband processing offloading from the base stations to the cloud
    - Interference management (CoMP), simplified base station deployment

  **Cloud Mobile Computing:** Code offloading from mobile users to the cloud
    - Increase in battery lifetime, enabling computation-heavy mobile applications (e.g., video processing, object recognition, gaming, ...)

  
  ![Diagram of C-RAN](image1.png)

  ![Diagram of Cloud Mobile Computing](image2.png)
Cloud Processing for 5G Systems

• Examples:

**Cloud Radio Access Network (C-RAN):** Baseband processing offloading from the base stations to the cloud
  → Interference management (CoMP), simplified base station deployment

**Cloud Mobile Computing:** Code offloading from mobile users to the cloud
  → Increase in battery lifetime, enabling computation-heavy mobile applications (e.g., video processing, object recognition, gaming, …)
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Overview

- Introduction and Motivation
  - Uplink with Single-Hop Fronthaul Topology
    - Point-to-Point Fronthaul Compression
    - Distributed Fronthaul Compression
    - Joint Decompression and Decoding
    - Compute-and-Forward
    - Time-Varying Channels
  - Downlink with Single-Hop Fronthaul Topology
    - Point-to-Point Fronthaul Compression
    - Multivariate Fronthaul Compression
    - Compute-and-Forward
    - Time-Varying Channels
    - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Cloud Radio Access Networks

- Heterogeneous dense network
- Macro, femto, pico-BSs, relays
- C-RAN: Baseband processing takes place in the “cloud”
Cloud Radio Access Networks

- Base stations act as radio units (RUs), or remote radio heads (RRHs): up/down-converters with no need for codebook information.
- Fronthaul links carry complex (IQ) baseband signals.
Cloud Radio Access Networks

Advantages:
• Low-cost “green” BSs (RUs)
• Dense deployment with enhanced indoor coverage
• Flexible radio and computing resource allocation
• Effective interference mitigation via joint baseband processing (e.g., eICIC and CoMP in LTE-A)

Key challenge: Effective transfer of the IQ signals on the fronthaul links
Cloud Radio Access Networks

- Fronthaul/backhaul links

- Mmwave backhauling for 5G systems [Ghosh '13]
- Convergence of wireless access and backhaul [Dahlman et al 14]
- Copper (LAN cable) for indoor coverage [Lu et al ‘14]
Cloud Radio Access Networks

- CPRI standard based on ADC/DAC

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Settings</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sectors</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>LTE Carriers</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Bandwidth</td>
<td>100</td>
<td>MHz</td>
</tr>
<tr>
<td>MIMO</td>
<td>2x2</td>
<td>Tx-Rx</td>
</tr>
<tr>
<td>Bits-per-I/Q</td>
<td>15</td>
<td>Bits</td>
</tr>
<tr>
<td>Protocol</td>
<td>LTE-A</td>
<td></td>
</tr>
<tr>
<td>Throughput</td>
<td>13.8</td>
<td>Gbps</td>
</tr>
</tbody>
</table>

... Rate higher than supported by standard optical fiber channels (10GE)...

[IDT, Inc]
Scope of the talk

• How to design C-RANs? An information-theoretic view
• Implications on system design
• Topics:
  o Advanced fronthaul compression inspired by network information theory
  o Structured coding (compute-and-forward)
  o Joint design of fronthaul and wireless communication
  o Optimal functional split RU-CU: channel estimation, precoding
  o Single-hop vs. multi-hop fronthaul network
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
System Model

Single-cluster single-hop fronthaul topology
Point-to-Point Fronthaul Compression

• Point-to-point fronthaul compression:
  – Algorithms [Segel and Weldon] [Samardzija et al ‘12] [Nieman and Evans ‘13]
  – Testbed results [Irmer et al ‘11] [Vosoughi et al ‘12]
Point-to-Point Fronthaul Compression

RU 1

Compressor

$\mathbf{y}_1^{ul}$

Fronthaul

$C_1$

Decompressor

$\hat{\mathbf{y}}_1^{ul}$

Control Unit

RU 2

Compressor

$\mathbf{y}_2^{ul}$

Fronthaul

$C_2$

Decompressor

$\hat{\mathbf{y}}_2^{ul}$

...$

RU N_R$

Compressor

$\mathbf{y}_{N_R}^{ul}$

Fronthaul

$C_{N_R}$

Decompressor

$\hat{\mathbf{y}}_{N_R}^{ul}$

Decoder
Point-to-Point Fronthaul Compression

- Ex.: Scalar quantizer
Point-to-Point Fronthaul Compression

- Ex.: Vector quantizer
Point-to-Point Fronthaul Compression

- Information-theoretic analysis (rate-distortion theory)
  - Compression at RU $i$ described by a test channel $p(\hat{y}_i | y_i)$

\[ y_i \xrightarrow{p(\hat{y}_i | y_i)} \hat{y}_i \]

Quantization noise
Point-to-Point Fronthaul Compression

- The test channel \( p(\hat{y}_i | y_i) \) determines the shape of the quantization regions

- Fronthaul rate = \( I(y_i; \hat{y}_i) \leq C_i \)
Point-to-Point Fronthaul Compression

- MSE bound closely approximated by sufficiently large vector quantizers [Gray and Neuhoff ‘98] or scalar quantizers with entropy encoding [Ziv ‘85]

- Gaussian noise model asymptotically correct with dithered lattice quantizers such as TCQ, o at high resolution [Zamir and Feder ‘96]

- Graphical codes [Nagpal et al 09]
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Distributed Fronthaul Compression

[Sanderovich et al '09] [del Coso and Simoens '09] [Zhou and Yu '11]
Point-to-Point Fronthaul Compression

- Ex.: Scalar quantizer
Distributed Fronthaul Compression

- Ex.: Scalar WZ compression
Distributed Fronthaul Compression

- Ex.: Vector WZ compression
Distributed Fronthaul Compression

- Information-theoretic analysis (rate-distortion theory):

\[ I(\mathbf{y}_i; \hat{\mathbf{y}}_i | \hat{\mathbf{y}}_s) \leq C_i \]
Distributed Fronthaul Compression

- Practical implementations:
  - Quantization code that can be partitioned into cosets that are good channel codes: trellis codes [Pradhan and Ramchandran '03], polar codes [Korada and Urbanke '10], compound LDGM-LDPC code [Martinian and Wainwright '06]

  - Separate quantization and channel coding: scalar quantization and vector binning [Liu et al '06], scalar quantization and turbo codes [Aaron et al '02]
Distributed Fronthaul Compression

- **Max-Rate compression** problem at BS $i$

\[
\begin{align*}
\text{maximize} & \quad I(x; \hat{y}_i | \hat{y}_S) \\
\text{s.t.} & \quad I(y_i; \hat{y}_i | \hat{y}_S) \leq C_i
\end{align*}
\]

- $\mathcal{S}$ is the set of BSs previously selected by BS ordering
Proposition [del Coso and Simoens ’09].
The optimal test channel \( p(\hat{y}_i | y_i) \) for (P1) is Gaussian, i.e.,

\[
\hat{y}_i = A_i y_i + q_i \quad \text{with } q_i \sim \mathcal{CN}(0, I).
\]

– Linear transformation \( A_i \) given as

\[
A_i = \text{diag}\left( \alpha_1, \ldots, \alpha_{n_{B,i}} \right)^{1/2} U^\dagger
\]

where \( \alpha_i = \left[ \frac{1}{\mu} \left( 1 - \frac{1}{\lambda_i} \right) - 1 \right]^+ \),

\[
\sum_{y_i, \hat{y}_s} = U \text{diag}\left( \lambda_1, \ldots, \lambda_{n_{B,i}} \right) U^\dagger
\]
Distributed Fronthaul Compression

• Block diagram of Max-Rate compression

\[ \sum_{x|\hat{y}_s} = \sum_x \mathbf{H}_s^\dagger \left( \mathbf{H}_s \sum_x \mathbf{H}_s^\dagger + \sum_{t_s} \right)^{-1} \mathbf{H}_s \sum_x \]
Example

[Park et al '13]

• Simulation setting:
  – Three-cell heterogeneous network
  – One MBS, $N_B/3 - 1$ HBSs and $N_M$ MSs per cell
    • MBS located at the cell center
    • HBSs and MSs are uniformly distributed within the cells to which they belong.
  – Single-antenna MSs, i.e., $n_{M,i} = 1$

• Schemes:
  – Max-Rate, (direct/indirect) MMSE, DF
  – With side information and no side information
Example

[Park et al ’13]

$C = 10 \text{ bps/Hz} \quad N_B = 9, \quad N_M = 9, \quad \omega = 0.5$
Example

[Park et al ’13]

\[ N_B = 9, \; N_M = 9, \; n_{B,i} = 8 \; \quad C = 15 \; \text{bps/Hz} \; \quad P_{tx} = -5 \; \text{dB} \]
Distributed Fronthaul Compression

[Park et al '13]

- CSI requirements of max-rate compression

\[
\sum_{x \mid \hat{y}_s} = \sum_x - \sum_x \bar{H}_s^\dagger \left( \bar{H}_s \sum_x \bar{H}_s^\dagger + \sum_{x} \right)^{-1} \bar{H}_s \sum_x
\]
Distributed Fronthaul Compression

Original Max-Rate maximization

Robust Max-Rate maximization [Park et al ’13]
Distributed Fronthaul Compression

- Imperfect covariance feedback, i.e., $\hat{\Sigma}_{x|\hat{y}_s} \neq \Sigma_{x|\hat{y}_s}$.
- Additive model

$$\hat{\Sigma}_{x|\hat{y}_s} = \Sigma_{x|\hat{y}_s} + \Delta_{x|\hat{y}_s}$$

with uncertainty set $\mathcal{U}_\Delta$, i.e., $\Delta_{x|\hat{y}_s} \in \mathcal{U}_\Delta$ where

$$\lambda_{\text{min}} (\tilde{\Delta}_{x|\hat{y}_s}) \geq \lambda_{\text{LB}}, \quad \lambda_{\text{max}} (\tilde{\Delta}_{x|\hat{y}_s}) \geq \lambda_{\text{UB}}$$

$$\tilde{\Delta}_{x|\hat{y}_s} = H_i \Delta_{x|\hat{y}_s} H_i^\dagger$$
Distributed Fronthaul Compression

- Max-min robust formulation

\[
\begin{align*}
\text{maximize} & \quad \min_{p(\hat{y}_i | y_i)} \min_{\Delta_{x|\hat{y}_s} \in \mathcal{H}^{nm}} I(x; \hat{y}_i | \hat{y}_s) \\
\text{s.t.} & \quad I(y_i; \hat{y}_i | \hat{y}_s) \leq C_i, \\
& \quad \lambda_{\max} (\tilde{\Delta}_{x|\hat{y}_s}) \leq \lambda_{UB} \quad \text{and} \quad \lambda_{\min} (\tilde{\Delta}_{x|\hat{y}_s}) \geq \lambda_{LB}.
\end{align*}
\]
Distributed Fronthaul Compression

• Theorem [Park et al ’13].

A local optimum for problem (P2) is given as

\[ \hat{y}_i = A_i y_i + q_i \quad \text{with} \quad q_i \sim \mathcal{CN}(0, I). \]

where

\[ A_i = \text{diag} \left( \alpha_1, \ldots, \alpha_{n_{B,i}} \right)^{1/2} U^\dagger \]

\[ \hat{\Sigma}_{y_i|\hat{y}_s} = U \text{diag} \left( \lambda_1, \ldots, \lambda_{n_{B,i}} \right) U^\dagger \quad \begin{cases} 
UU^\dagger = U^\dagger U = I \\
\lambda_1 \geq \cdots \geq \lambda_{n_{B,i}} \geq 1
\end{cases} \]

\[ \alpha_1, \ldots, \alpha_{n_{B,i}} : \text{solution to the following mixed integer problem} \]

\[
\begin{aligned}
\max_{\mu, \alpha_1, \ldots, \alpha_{n_{B,i}}} & \sum_{l=1}^{n_{B,i}} (\log(1 + \alpha_l(\lambda_l + \lambda_{UB})) - \log(1 + \alpha_l)) \\
\text{s.t.} & \quad 0 < \mu < 1, \\
& \quad \alpha_l \in \mathcal{P}(\mu), \ l = 1, \ldots, n_{B,i}, \\
& \quad \sum_{l=1}^{n_{B,i}} \log(1 + \alpha_l(\lambda_l + \lambda_{UB})) = C_i.
\end{aligned}
\]

where \( \mathcal{P}(\mu) \) is a discrete set dependent on \( \mu \).
Distributed Fronthaul Compression

- Block diagram:

\[ y_i \xrightarrow{\text{Conditional KLT } U^\dagger} y'_1, \ldots, y'_{n_{B,j}} \xrightarrow{\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_{n_{B,j}}}} q_{i,1}, \ldots, q_{i,n_{B,j}} \xrightarrow{\hat{y}_{i,1}, \ldots, \hat{y}_{i,n_{B,j}}} \]

\[ \sum_{x|\hat{y}_s} \]

Solution of optimization problem
Example

• Simulation setting:
  – Single-cell heterogeneous network
  – One MBS and $N_B - 1$ HBSs
    • MBS located at the cell center
    • HBSs and MSs uniformly distributed within whole cell
  – Single-antenna MSs, i.e., $n_{M,i} = 1$

• Schemes:
  – Max-Rate, NSI
  – Max-Rate, imperfect SI
  – Max-Rate, imperfect SI, robust
  – Max-Rate perfect SI
Example

\begin{equation}
N_B = 4, \quad N_M = 8, \quad n_{B,i} = 2 \quad \omega = 0.5 \quad P_{tx} = 10 \text{ dB}
\end{equation}
Simulation Set-up

• In each macro-cell, $N$ pico-BSs and $K$ MSs are uniformly distributed.
Simulation Set-up

- Frequency reuse pattern with reuse factor $F = \frac{1}{3}$ for 1-cell cluster [Wang and Yeh ’11]
## Simulation Set-up

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>System bandwidth</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Path-loss (macro-BS - MS)</td>
<td>$PL(\text{dB}) = 128.1 + 37.6 \log_{10} R \ (R \text{ in km})$</td>
</tr>
<tr>
<td>Path-loss (pico-BS - MS)</td>
<td>$PL(\text{dB}) = 38 + 30 \log_{10} R \ (R \text{ in m})$</td>
</tr>
<tr>
<td>Antenna pattern for sectorized macro-BS antennas</td>
<td>$A(\theta) = -\min[12(\theta / \theta_{3\text{dB}})^2, A_m]$</td>
</tr>
<tr>
<td></td>
<td>($\theta_{3\text{dB}} = 65^\circ, A_m = 20 \text{ dB}$)</td>
</tr>
<tr>
<td>Lognormal shadowing (macro-BS - MS)</td>
<td>10 dB standard deviation</td>
</tr>
<tr>
<td>Lognormal shadowing (pico-BS - MS)</td>
<td>6 dB standard deviation</td>
</tr>
<tr>
<td>Antenna gain after cable loss (macro-BS)</td>
<td>15 dBi</td>
</tr>
<tr>
<td>Antenna gain after cable loss (pico-BS, MS)</td>
<td>0 dBi</td>
</tr>
<tr>
<td>Noise figure</td>
<td>5 dB (macro-BS), 6 dB (pico-BS), 9 dB (MS)</td>
</tr>
<tr>
<td>Transmit power</td>
<td>46 dBm (macro-BS), 24 dBm (pico-BS), 23 dBm (MS)</td>
</tr>
<tr>
<td>Small-scale fading model</td>
<td>Rayleigh-fading</td>
</tr>
<tr>
<td>Synchronization</td>
<td>Perfect synchronization</td>
</tr>
<tr>
<td>Inter-site distance (site: macro-BS)</td>
<td>750 m</td>
</tr>
<tr>
<td>Frequency reuse factor</td>
<td>$F=1/3$</td>
</tr>
<tr>
<td>Number of antennas</td>
<td>Single antenna at each macro/pico-BS and MS</td>
</tr>
<tr>
<td>Channel state information (CSI)</td>
<td>Full CSI at control units about BSs in the cluster</td>
</tr>
</tbody>
</table>
Simulation Set-up

- LTE rate model [3GPP-TR-136942]

\[ \tilde{R}_k(\gamma_k) = \begin{cases} 
0, & \text{if } \gamma_k \leq \gamma_{\text{min}} \\
\alpha_{\text{attenuate}} S(\gamma_k), & \text{if } \gamma_{\text{min}} < \gamma_k \leq \gamma_{\text{max}} \\
R_{\text{max}}, & \text{if } \gamma_k > \gamma_{\text{max}}
\end{cases} \]

where

- \( \gamma_k \): SINR at MS \( k \); \( S(\gamma) = \log_2(1+\gamma) \);
- \( \gamma_{\text{max}} = S^{-1}(R_{\text{max}} / \alpha_{\text{attenuate}}) \);
- \( \alpha_{\text{attenuate}} \): attenuation factor representing implementation losses;
- \( R_{\text{max}} \): Maximum and minimum throughput of the codeset, bps/Hz;
- \( \gamma_{\text{min}} \): Minimum SINR of the codeset.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>UL</th>
<th>DL</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_{\text{max}} )</td>
<td>2.0</td>
<td>4.4</td>
<td>Based on 16-QAM 3/4 (UL) &amp; 64-QAM 4/5 (DL)</td>
</tr>
<tr>
<td>( \gamma_{\text{min}} )</td>
<td>-10 dB</td>
<td>-10 dB</td>
<td>Based on QPSK with 1/5 (UL) &amp; 1/8 (DL)</td>
</tr>
<tr>
<td>( \alpha_{\text{attenuate}} )</td>
<td>0.4</td>
<td>0.6</td>
<td>Representing implementation losses</td>
</tr>
</tbody>
</table>

[3GPP-TR-136942, Annex A]
Simulation Set-up

• Proportional fairness metric

\[ R_{\text{sum-PF}}(t) = \sum_{k=1}^{K} \frac{R_k(t)}{\bar{R}_k^\alpha} \quad \cdots \text{(P1)} \]

where \( \alpha \): fairness constant;
\( R_k(t) \): instantaneous rate for MS \( k \) at time \( t \);
\( \bar{R}_k \): historical data rate for MS \( k \) until time \( t-1 \).

– At each time \( t \), the rate \( \bar{R}_k \) is updated as

\[ \bar{R}_k \leftarrow \beta \bar{R}_k + (1 - \beta) R_k(t) \]

where \( \beta \in [0,1] \): the forgetting factor.
Numerical Results

- Cell-edge throughput versus average spectral efficiency
  - Uplink, 1-cell cluster, $N = 3$ pico-BS, $K = 5$ MSs, $(C_{\text{macro}}, C_{\text{pico}}) = (9, 3)$ bps/Hz, $T_{\text{max}} = 10$, $\beta = 0.5$, $F = 1/3$
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
    - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
    - Compute-and-Forward
    - Time-Varying Channels
    - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Joint Decompression and Decoding

- An achievable rate with JDD was derived in [Sanderovich et al '09]

\[
R_{\text{sum}} = \min_{s \subseteq \mathcal{N}_B} \left\{ \sum_{j \in s} \left( C_j - I(y_j; \hat{y}_j \mid x) \right) - I(x; \hat{y}_s) \right\},
\]
Joint Decompression and Decoding

- The problem of maximizing the sum-rate is a Difference of Convex (DC) problem [Park et al '14].

![Graph showing average per-cell sum-rate vs. inter-cell channel gain with different methods: cutset upper bound, JDD with MM algorithm, SDD with exhaustive ordering, and SDD with greedy ordering. The graph includes annotations for joint decompression and decoding, separate decompression and decoding, and parameters $N_B = 3, n_M = 1$, $n_B = 4$, $C = 12$ bits/s/Hz, $P = 20$ dB.]
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - **Compute-and-Forward**
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - **Compute-and-Forward**
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Compute-and-Forward
[Nazer et al ’09] [Hong and Caire ’11]

• The MSs use nested lattice codes:
Compute-and-Forward

[Nazer et al ’09] [Hong and Caire ’11]
Numerical Results

• Three-cell SISO circular Wyner model
Numerical Results

$C = 3 \text{ bit/s/Hz and } \alpha = 0.4$
Numerical Results

$C = 3 \text{ bit/s/Hz and } \alpha = 0.4$
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
So far, static channels
With static channels, the fronthaul overhead due to CSI transfer is negligible
Time-Varying Channels

- Time-varying ergodic channel with coherence blocks of length $T$ channel uses [Hassibi et al '03]

- RU-CU functional split: where to perform channel estimation?
Compress-Forward-Estimate (CFE) [Hoydis et al ‘11]

• The RUs compress both training and data fields
• The CU estimates the channel and performs coherent decoding
Estimate-Compress-Forward (ECF)

- Motivated by the information-theoretic argument of separate estimation and compression [Witsenhausen ‘80]
- The RUs estimate the channel and compress both CSI and data field
- The CU performs coherent decoding
Semi-Coherent Processing

[Kang et al ‘14]

- For low fronthaul capacities...
- BS estimates the CSI and performs local equalization
- BS compresses and forwards the equalized data signals to CU
- CU jointly decodes the data signal by mismatched decoding metric

[Weingarten et al ‘04]
Semi-Coherent Processing

[Kang et al ‘14]

- Mismatched decoding:
  \[
  \sum_{k=1}^{n} \gamma_k \| \hat{X}_{d,k} - X_{d,k} \|^2
  \]

  \(\gamma_k\) is a weight factor for each coherence block

- With one fronthaul bit per coherent block

  \[
  \gamma_k = \begin{cases} 
  \gamma_b, & \|\widetilde{H}\| < \omega \\ 
  \gamma_g, & \|\widetilde{H}\| \geq \omega
  \end{cases}
  \]

  weigh less blocks with poor CSI
Non-Coherent Capacity

[Kang et al ‘14]

- Do not impose training + data structure [Marzetta and Hochwald ‘99]
- The RUs quantize the received signal
Example

\[ N_B = N_M = 2, \quad N_t = N_r = 4, \quad C = 6, \quad T = 10, \quad K = 0 \]
Example

\[ N_t = N_r = 2, \ P = 10\text{dB}, \ T = 4, \ \text{and} \ K = 0 \]
Example

\[ N_t = N_r = 2, \quad P = 5 \text{dB}, \quad C = 5 \text{ bits/s/Hz}, \quad T = 4, \quad \text{and} \quad K = 0 \]
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions

69
System Model

Single-cluster single-hop fronthaul topology
Point-to-Point Fronthaul Compression
[Simeone et al ’09] [Patil and Yu ’14]

![Diagram of Point-to-Point Fronthaul Compression System]

- Central Unit
- Channel encoder 1
- Precoding
- Compressor 1
- RU 1
- Channel encoder $N_M$
- Precoding
- Compressor $N_B$
- RU $N_B$
Point-to-Point Fronthaul Compression

Ex.: Scalar quantization
Point-to-Point Fronthaul Compression

Ex.: Scalar quantization

... uncorrelated quantization noise
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Multivariate Fronthaul Compression

[Park et al ’14]
Multivariate Fronthaul Compression

- Multivariate compression produces compressed signals with correlated quantization noises: new design degree of freedom
Multivariate Fronthaul Compression

Ex.: Scalar quantization
Multivariate Fronthaul Compression

Ex.: Scalar quantization
Multivariate Compression Lemma

\[ p(\tilde{x}, x_1, \ldots, x_M) = p(\tilde{x}) p(x_1, \ldots, x_M | \tilde{x}) \]

\[ \sum_{i \in S} h(X_i) - h(X_S | \tilde{X}) \leq \sum_{i \in S} R_i, \text{ for all } S \subseteq \{1, \ldots, M\} \]
Multivariate Compression Lemma

• Multivariate compression lemma
  – Ex.: With $N_B = 2$, we have

$$h(x_1) - h(x_1 | \tilde{x}) = I(\tilde{x}_1; x_1) \leq C_1, \quad (A1)$$

$$h(x_2) - h(x_2 | \tilde{x}) = I(\tilde{x}_2; x_2) \leq C_2, \quad (A2)$$

$$h(x_1) + h(x_2) - h(x_1, x_2 | \tilde{x}) = I(\tilde{x}; x_1, x_2) + I(x_1; x_2) \leq C_1 + C_2. \quad (A3)$$

while with separate compression, we have

$$I(\tilde{x}_1; x_1) \leq C_1, \quad (B1)$$

$$I(\tilde{x}_2; x_2) \leq C_2. \quad (B2)$$
Multivariate Compression Lemma

(contrapolymatroid)
Multivariate Fronthaul Compression

- Linear precoding (DPC treated in a similar way)

- Gaussian test channel:

  \[ x_i = \tilde{x}_i + q_i, \quad q_i \sim \mathcal{CN}(0, \Omega_{i,i}), \quad i \in N_B \]

- The compressed signal

  \[
  \mathbf{x} = \left[ x_1^H, \ldots, x_{N_B}^H \right]^H
  \]

  is given as

  \[
  \mathbf{x} = \mathbf{A}s + \mathbf{q},
  \]

  with

  \[
  \mathbf{q} = \left[ q_1^H, \ldots, q_{N_B}^H \right]^H \sim \mathcal{CN}(0, \Omega) \quad \text{and}
  \]

  \[
  \Omega = \begin{bmatrix}
  \Omega_{1,1} & \Omega_{1,2} & \cdots & \Omega_{1,N_B} \\
  \Omega_{2,1} & \Omega_{2,2} & \cdots & \Omega_{2,N_B} \\
  \vdots & \vdots & \ddots & \vdots \\
  \Omega_{N_B,1} & \Omega_{N_B,2} & \cdots & \Omega_{N_B,N_B}
  \end{bmatrix}
  \]
Multivariate Fronthaul Compression

- Joint optimization of precoding and compression:

$$\text{maximize} \sum_{k=1}^{N_M} w_k f_k (A, \Omega)$$

$$\text{s.t.} \quad g_S (A, \Omega) \leq \sum_{i \in S} C_i, \text{ for all } S \subseteq \mathcal{N}_B,$$

$$\text{tr} \left( E_i^H A A^H E_i + \Omega_{i,i} \right) \leq P_i, \text{ for all } i \in \mathcal{N}_B.$$

where

$$f_k (A, \Omega) = I (s_k : y_k)$$

$$= \log \det \left( I + H_k (A A^H + \Omega) H_k^H \right) - \log \det \left( I + H_k \left( \sum_{l \neq k} A_l A_l^H + \Omega \right) H_k^H \right),$$

$$g_S (A, \Omega) = \sum_{i \in S} h (x_i) - h (x_S | \bar{x})$$

$$= \sum_{i \in S} \log \det \left( E_i^H A A^H E_i + \Omega_{i,i} \right) - \log \det \left( E_s^H \Omega E_s \right) \leq \sum_{i \in S} C_i.$$
Multivariate Fronthaul Compression

- The implementation of joint compression is complex
- A successive architecture with a given permutation $\pi : \mathcal{N}_B \rightarrow \mathcal{N}_B$.

Step 1: $\hat{X}_{\pi(1)} \rightarrow X_{\pi(1)}$

Step $i$: $u_{\pi(i)} = \begin{bmatrix} x_{\pi(1)} \\ \vdots \\ x_{\pi(i-1)} \\ \hat{x}_{\pi(i)} \end{bmatrix}$

MMSE estimation of $x_{\pi(i)}$ given $u_{\pi(i)}$

$\hat{x}_{\pi(i)} \rightarrow X_{\pi(i)}$

$\hat{q}_{\pi(i)} \sim \mathcal{CN}(0, \sum_{x_{\pi(i)}|u_{\pi(i)}})$

$\hat{q}_{\pi(i)} \sim \mathcal{CN}(0, \sum_{x_{\pi(i)}|u_{\pi(i)}})$

$\hat{q}_{\pi(i)} \sim \mathcal{CN}(0, \sum_{x_{\pi(i)}|u_{\pi(i)}})$

Compression

Computation rate at step $i$ ($i > 1$): $I(\hat{x}_{\pi(i)}; x_{\pi(i)})$
Multivariate Fronthaul Compression

- Successive estimation-compression architecture
Simulation Set-up

• In each macro-cell, $N$ pico-BSs and $K$ MSs are uniformly distributed.
Numerical Results

- Cell-edge throughput versus average spectral efficiency
  - Downlink, 1-cell cluster, $N = 1$ pico-BS, $K = 4$ MSs, $(C_{macro}, C_{pico}) = (3,1)$ bps/Hz, $T_{\text{max}} = 5$, $\beta = 0.5$, $F = 1/3$
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Compute-and-Forward

- Reverse compute-and-forward (RCoF) [Hong and Caire ’12]
Numerical Results

- Three-cell SISO circular Wyner model
Numerical Results

• Three-cell SISO circular Wyner model ($P = 20$ dB and $\alpha = 0.5$)
Numerical Results

- Three-cell SISO circular Wyner model ($P = 20$ dB and $\alpha = 0.5$)
Numerical Results

- Three-cell SISO circular Wyner model \( P = 20 \) dB and \( \alpha = 0.5 \)
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
• Ergodic channel with coherence blocks of length $T$ channel uses [Hassibi et al '03]
Time-Varying Channels

• RU-CU functional split: where to perform precoding?

• Instantaneous CSI: Design different precoders for each coherence period

• Stochastic CSI:
  o CU only knows the spatial correlation (e.g., [Adhikari et al ’14])

\[
H_{ji} = \Sigma_{R,ji}^{1/2} \hat{H}_{ji} \Sigma_{T,ji}^{1/2}, \quad [\Sigma_T(\theta_{ji}, \Delta_{ji})]_{m,n} = \frac{\alpha_{ji}}{2 \Delta_{ji}} \int_{\theta_{ji}-\Delta_{ji}}^{\theta_{ji}+\Delta_{ji}} \exp\left(-j\pi(m-n)\sin(\phi)\right) d\phi;
\]

  o Design a single precoder across all coherence periods
Compress-After-Precoding
Compress-Before-Precoding

[Park et al ‘13] [Kang et al ‘14] [Patil and Wu ‘14]

- Clustering: Send each user’s message to a subset of RUs
- Precoding performed at the RUs
Compress-Before-Precoding

[Park et al ‘13] [Kang et al ‘14] [Patil and Wu ‘14]

- Perfect CSI: Precoding overhead gets amortized over $T$
- Stochastic CSI: Since the precoding matrix is the same across all the coherence periods, negligible overhead
Example
Example

\[ N_R = 4, N_{t,i} = 2, N_{r,j} = 1, P = 10 \text{ dB}, C=4 \text{ bits/s/Hz}, T = 10 \]
Example

\[ N_R = N_M = 4, \ N_{t,i} = 2, \ N_{r,j} = 1, \ P = 20 \text{ dB}, \ C = 2 \text{ bits/s/Hz}, \]

instantaneous CSI
Example

\[ N_R = N_M = 4, \quad N_{t,i} = 2, \quad P = 10 \text{ dB}, \quad C = 3 \text{ bits/s/Hz}, \quad T = 10 \]
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Inter-Cluster Multivariate Compression Design

[Park et al ‘14]
Inter-Cluster Multivariate Compression Design

\[ N_C = 2, N_R = 2, \]
\[ N_M = 2, n_M = 1, \]
\[ n_R = 1, \]
\[ C = 2 \text{ bits/s/Hz} \]
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Multi-Hop Fronthaul Topology

[Park et al ‘14]
Multi-Hop Fronthaul Topology

- Multiplex-and-Forward (MF)
- Decompress-Process-and-Recompress (DPR): In-network processing
- Optimization of the uplink sum-rate
Related Works

• [Ni et al ‘13] Simulation-based study of in-network processing on the LLRs
• [Goela and Gastpar ‘12] Optimal linear in-network processing strategies in a multihop network for estimation
• [Avestimehr et al ‘08] Deterministic and Gaussian relay networks
• [Cuff et al ‘09] Cascade source coding
System Model

MS: mobile station
RU: radio unit

\[ y_i = H_i x + z_i \]

Capacitated directed acyclic graph (DAG)
RU Operation

- Multiplex-and-Forward (MF)
- Decompress-Process-and-Recompress (DPR): In-network processing
Multiplex-and-Forward
Multiplex-and-Forward

- Gaussian test channel

\[ \hat{y}_i = y_i + q_i, \quad q_i \sim \mathcal{CN}(0, \Omega_i). \]
Define as $f^i_e \geq 0$ the capacity used to convey the signal compressed at node $v_i$ on edge $e$ for $e \in \mathcal{E}_{\text{act}}, i \in \mathcal{M}$.

$$\sum_{i \in \mathcal{M}} f^i_e \leq \tilde{C}_e, \quad \forall e \in \mathcal{E}_{\text{act}}$$

Edge capacity constraint

$$\sum_{e \in \Gamma_I(j)} f^i_e \geq \sum_{e \in \Gamma_O(j)} f^i_e, \quad \forall j \in \mathcal{V} \setminus \mathcal{S} \setminus \{v_{M+1}\}, i \in \mathcal{M}.$$  

Flow conservation at each RU
Multiplex-and-Forward

• Sum-rate maximization:

\[
\begin{align*}
\text{maximize} \quad & \log \det(\mathbf{H}\Sigma_x\mathbf{H} + \mathbf{I} + \mathbf{\Omega}) - \log \det(\mathbf{I} + \mathbf{\Omega}) \\
\text{s.t.} \quad & \log \det(\mathbf{\Omega}_i + \Sigma_{y_i}) - \log \det(\mathbf{\Omega}_i) \leq \sum_{e \in \Gamma_1(M+1)} f^i_e, \quad i \in \mathcal{M} \\
& \sum_{i \in \mathcal{M}} f^i_e \leq \tilde{C}_e, \quad \forall e \in \mathcal{E}\text{act} \\
& \sum_{e \in \Gamma_1(j)} f^i_e \geq \sum_{e \in \Gamma_0(j)} f^i_e, \quad \forall j \in \mathcal{V} \setminus \mathcal{S} \setminus \{v_{M+1}\}, \quad i \in \mathcal{M}.
\end{align*}
\]

• Difference-of-convex (DC) problem: Majorization Minimization (MM) approach to find a locally optimal solution.
Decompress-Process-and-Recompress

\[ y_i \xrightarrow{\text{Decompression}} r_i \xrightarrow{\text{Linear Processing}} \{ u_e \}_{e \in \Gamma_I(i)} \xrightarrow{\text{Compress}} \{ u_e \}_{e \in \Gamma_O(i)} \xrightarrow{\text{Compress}} \text{To RUs and CU in next layers} \]

From RUs in previous layers

\[ \Gamma_I(i) \]

\[ \Gamma_O(i) \]
Decompress-Process-and-Recompress

- Process and decompress:
  \[
  u_e = L_e r_i + q_e,
  \]
  with \( q_e \sim \mathcal{CN}(0, \Omega_e) \)

- Characterization of the relation between input and output vector [Goela and Gastpar ‘12]
Decompress-Process-and-Recompress

- Sum-rate maximization

\[
\begin{align*}
\text{maximize} & \quad \left\{ \log \det \left( T H \Sigma \times H^H T^H + TT^H + \tilde{T} \Omega \tilde{T}^H \right) \right\} \\
\text{s.t.} & \quad \log \det \left( \Omega_e + L_e \Sigma_r L_e^H \right) - \log \det \left( \Omega_e \right) \leq \tilde{C}_e, \quad i = \text{tail}(e), \quad e \in E_{\text{act}}, \\
L_e^{\text{rx}} & \in \mathbb{C}^{d_e \times n_R, i}, \quad L_e^i \in \mathbb{C}^{d_e \times d_e^i}, \quad e \in E_{\text{act}}.
\end{align*}
\]
Decompress-Process-and-Recompress

• **Lemma.** We can set $L_e = I$ for $e \in \mathcal{E}_{\text{act}}$ without loss of optimality.

• The optimization over $\{\Omega_e\}_{e \in \mathcal{E}_{\text{act}}}$ is a DC problem: MM algorithm
Numerical Results

\[ M = 2N + 3 \]

- Each MS and RU have a single antenna
- All RUs have the same average SNR
- Rayleigh fading channels
Numerical Results

$K=4$, $SNR = 20$ dB

Numerical results for $K=4$, $SNR = 20$ dB are shown in the following graph. The average sum-rate is plotted against the number of RUs in layer 1 for different channel conditions.

- $\tilde{C}_e = 4$ bits/s/Hz
- $\tilde{C}_e = 3$ bits/s/Hz
- $\tilde{C}_e = 2$ bits/s/Hz

The graph shows the performance of DPR and MF quantization techniques.
Numerical Results

$K=5$, $N=2$, SNR = 0 dB
Overview

- Introduction and Motivation
- Uplink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Distributed Fronthaul Compression
  - Joint Decompression and Decoding
  - Compute-and-Forward
  - Time-Varying Channels
- Downlink with Single-Hop Fronthaul Topology
  - Point-to-Point Fronthaul Compression
  - Multivariate Fronthaul Compression
  - Compute-and-Forward
  - Time-Varying Channels
  - Inter-Cluster Multivariate Fronthaul Compression
- Uplink with Multi-Hop Fronthaul Topology
- Conclusions
Conclusions

- Information-theoretic view on C-RAN

- Novel approaches inspired by theory: distributed/multivariate fronthaul compression, compute-and-forward, joint decompression and decoding, estimate-compress-forward, semi-coherent processing, in-network processing,…

- Implementation: linear codes, scalar quantization, successive estimation and compression,…

- Performance of conventional techniques can be drastically improved by strategies inspired by information theory