Cloud Radio Access Network: Challenges and (Some) Solutions

Osvaldo Simeone
New Jersey Institute of Technology
Cloud Radio Access Networks

- Heterogeneous dense cellular network
- Macro, femto, pico-BSs
- C-RAN: Joint baseband processing takes place in the “cloud”
- Instance of network function virtualization
Cloud Radio Access Networks

- Fronthaul links carry “radio” signals to/from control unit (CU) or baseband unit (BBU)
- Base stations act as radio units (RUs) or remote radio heads (RRHs)
- Cloud resources shared by all connected RUs unlike D-RAN
Cloud Radio Access Networks

• Analog (e.g., radio-over fiber) vs digital (e.g., CPRI) fronthaul transmission
• Digital transmission: digitized complex (IQ) baseband signals
Advantages:
• Dense deployment with low-cost “green” BSs (RUs)
• Flexible radio and computing resource allocation (statistical multiplexing)
• Effective interference mitigation via joint baseband processing (e.g., eICIC and CoMP in LTE-A)
• Easier network upgrades and maintenance

Key challenge:
Capacity and latency bottleneck imposed by fronthaul network limitations
Fronthauling

• Fiber-optic communication is the preferred choice.

• Mmwave front/backhauling [Ghosh ‘13] [Checko et al ‘15][Fujitsu ‘15]

• Copper (LAN cable) for indoor coverage [Lu et al ‘14]
Fronthaul Capacity Limitations

- Common Public Radio Interface (CPRI) specifies the fronthaul interface between CU and RUs
- Sampling and scalar quantization

<table>
<thead>
<tr>
<th>Cell configuration</th>
<th>Bit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 MHz LTE, 15+1 CPRI IQ Sample width, 10/8 line coding, 2x2 MIMO</td>
<td>2.5 Gbps</td>
</tr>
<tr>
<td>5x20 MHz LTE-A, 15 CPRI IQ Sample width, 2x2 MIMO, 3 sectors</td>
<td>13.8 Gbps</td>
</tr>
<tr>
<td>20 MHz LTE, 4x2 MIMO, 3 sectors</td>
<td>16.6 Gbps</td>
</tr>
<tr>
<td>TD-LTE, 3 sectors</td>
<td>30 Gbps</td>
</tr>
<tr>
<td>1.6 MHz TD-SCDMA, 8Tx/8Rx antennas, 4 times sampling rate</td>
<td>330 Mbps</td>
</tr>
<tr>
<td>TD-SCDMA S444, 3 sectors</td>
<td>6 Gbps</td>
</tr>
<tr>
<td>200 kHz GSM, 2Tx/2Rx antennas, 4x sampling rate</td>
<td>25.6 Mbps</td>
</tr>
</tbody>
</table>

[Checko et al ‘15]
Fronthaul Latency Limitations

- Latency associated with two-way mobile-CU transmission:
  - Air interface
  - Fronthaul transport ($\propto$ fronthaul length; single vs. multi-hop; <250 μs [NGMN ’15])
  - CU processing (≈ ms)

- Fronthaul transport and CU processing depends on transport technology (wireless vs. wired).
Overview

- Fronthaul capacity limitations
  - Point-to-point fronthaul compression (compressed CPRI)
  - Multivariate fronthaul compression
  - Alternative functional splits

- Fronthaul latency limitations
  - HARQ with local feedback

- Conclusions
Overview

o Fronthaul capacity limitations
  - Point-to-point fronthaul compression (compressed CPRI)
  - Multivariate fronthaul compression
  - Alternative functional splits

o Fronthaul latency limitations
  - HARQ with local feedback

o Conclusions
System Model

Data streams → CU → RU

CSI → C_{1} → RU

\[ \begin{align*}
C_{1} & \quad \vdots
C_{N_{B}} & \quad x_{1} \quad \vdots
\end{align*} \]

\[ \begin{align*}
x_{1} & \quad \vdots
x_{N_{B}} & \quad y_{1} \quad \vdots
\end{align*} \]

fronthaul → wireless downlink channel

MS 1 → y_{1} → MS N_{B}
CPRI-Based Implementation
Compressed CPRI
Compressed CPRI

• Filtering [Samardzija et al ‘12] [Guo et al ‘12]: Remove redundancies due to oversampling

• Per-block scaling [Samardzija et al ‘12] [Guo et al ‘12]: Mitigate large peak-to-peak variations

• Optimized non-uniform quantization [Samardzija et al ‘12] [Guo et al ‘12]: Lloyd-Max or companding

• Noise shaping [Nieman and Evans ‘13]: predictive quantization

• Lossless compression [Vosoughi et al ‘12]: entropy coding

• Compressed CPRI reduces the fronthaul rate by a factors around 3 [Dotsch et al ‘13]
Overview

- Fronthaul capacity limitations
  - Point-to-point fronthaul compression (compressed CPRI)
  - Multivariate fronthaul compression
  - Alternative functional splits

- Fronthaul latency limitations
  - HARQ with local feedback

- Conclusions
Is Point-to-Point Quantization Optimal?

![Diagram showing the process of point-to-point quantization involving channel encoders, precoding, quantization, and fronthaul.]
Multivariate Fronthaul Quantization?

Central Unit

Channel encoder 1

Precoding

Multivariate quantization

Channel encoder $N_M$

CSI

fronthaul

$M_1$

$M_{N_M}$

$s_1$

$s_{N_M}$

$\tilde{x}_1$

$\tilde{x}_{N_B}$

$C_1$

$C_{N_B}$

$x_1$

$x_{N_B}$
Point-to-Point Quantization

$x_1$  
$x_2$
Point-to-Point Quantization

$x_2$

$x_1$
Point-to-Point Quantization

\[ x_1 \]

\[ x_2 \]
Point-to-Point Quantization
Point-to-Point Quantization

quantization noise = interference
**Point-to-Point Quantization**

- The impact of interference depends on CSI.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>h</td>
</tr>
</tbody>
</table>

**signal space**
Point-to-Point Quantization

signal space

h

affects MS

does not affect MS
Multivariate Fronthaul Quantization

Multivariate Fronthaul Quantization

signal space
Multivariate Fronthaul Quantization
Multivariate Fronthaul Quantization

- Dual role of precoding and multivariate quantization

![Diagram showing the process of multivariate quantization in a signal processing system.]
Multivariate Fronthaul Compression

- Multivariate quantization can be combined with (block) compression.

Multivariate Fronthaul Compression

- Multivariate compression can be analyzed using information-theoretic methods [Park et al ‘13].

- An application of multivariate compression from network information theory (used, e.g., for multi-description coding)
Simulation Set-up

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

- LTE-recommended simulation set-up [3GPP-TR-136942]
- Proportional fairness scheduler
- Information-theoretic performance metrics (with clipping)
Numerical Results

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

- Quantization vs compression

![Graph showing the comparison between Ideal fronthaul, Multivariate compression, and Multivariate quantization rates against fronthaul capacity B. The graphs show different cases for P = 20 dB and P = 0 dB.]
Overview

- Fronthaul capacity limitations
  - Point-to-point fronthaul compression (compressed CPRI)
  - Multivariate fronthaul compression
  - Alternative functional splits

- Fronthaul latency limitations
  - HARQ with local feedback

- Conclusions
Alternative Functional Splits

[Dötsch et al '13]
[Wübben et al '14]
Alternative Functional Splits

Compression in the frequency domain
[Dötsch et al ’13]: lower PAPR but minor gains

[Dötsch et al ’13]
[Wübben et al ’14]
Alternative Functional Splits

Compression of only used subcarriers and adaptive quantization
[Dötsch et al ‘13] [Lorca and Cucala ‘13]
[Grieger et al ‘12]: significant gains in lightly loaded systems
Example

- Alternative functional split with resource demapping at RU: differentiated pilot-data fronthaul compression

Overview

- Fronthaul capacity limitations
  - Point-to-point fronthaul compression (compressed CPRI)
  - Multivariate fronthaul compression
  - Alternative functional splits

- Fronthaul latency limitations
  - HARQ with local feedback

- Conclusions
Fronthaul Latency Constraints

- Fronthaul latency (transport and processing) affects higher layer protocols, most notably ARQ and HARQ

Air interface

Fronthaul transport ($\propto$ fronthaul length; single vs. multi-hop; $<250$ μs [NGMN '15])

CU processing (~ms)

Fronthaul transport

Air interface
Uplink (H)ARQ in D/C-RAN
Uplink (H)ARQ in D/C-RAN

MS 1

RU 1

RU 2

CU

BBU(s)

2 Baseband fronthaul transfer

43
Uplink (H)ARQ in D/C-RAN
Uplink (H)ARQ in D/C-RAN

MS 1

MS 2

RU 1

RU 2

CU

BBU(s)

ACK/NAK

fronthaul transfer
Uplink (H)ARQ in D/C-RAN

- MS 1
- MS 2
- RU 1
- RU 2
- CU
- BBU(s)

5 ACK/NAK feedback
Fronthaul Latency and HARQ

- Ex.: FD LTE
Low-Latency Local Feedback

- With alternative functional splits, low-latency local control
- Low-latency local control vs. high-latency global control at the cloud
Low-Latency Local Feedback

- With alternative functional splits, low-latency local control
- Low-latency local control vs. high-latency global control at the cloud
Low-Latency Local Feedback

- Low-latency local feedback based on alternative functional split for D-RAN [Dotsch et al ’13]
Low-Latency Local Feedback

- Low-latency local feedback based on alternative functional split for D-RAN [Dotsch et al ’13]

\[ P_e(MCS, CSI) \geq P_{th} \]

NACK

estimate based on look-up table or error exponents [Rost and Prasad ‘14] or finite-blocklength bounds [Khalili and Simeone ‘15]
Low-Latency Local Feedback

- Low-latency local feedback based on alternative functional split for D-RAN [Dotsch et al ’13]

\[ P_e(MCS, CSI) \geq P_{th} \] 

ACK/NACK local feedback

estimate based on look-up table or error exponents [Rost and Prasad ‘14] or finite-blocklength bounds [Khalili and Simeone ‘15]
## Low-Latency Local Feedback

<table>
<thead>
<tr>
<th>Local feedback by RU</th>
<th>Decoding outcome at CU</th>
<th>Action</th>
<th>Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACK/NAK</td>
<td>Correct/Incorrect</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>ACK</td>
<td>Incorrect</td>
<td>ARQ retransmission (higher layer)</td>
<td>Delay</td>
</tr>
<tr>
<td>NAK</td>
<td>Correct</td>
<td>None</td>
<td>Delay</td>
</tr>
</tbody>
</table>
Performance Evaluation

• Throughput: number of correctly delivered bits/ total transmission time

• Probability of success: probability of ACK transmission and successful CU decoding by the maximum number of allowed transmission attempts

• Analysis based on finite-blocklength bounds [Polyanskiy et al ‘10]

Numerical Results

SISO, $N_{max} = 5$, $s = 3$ dB, $P_S = 0.99$

Throughput loss [%] vs. Blocklength for $r = 1$ bit/s/Hz and $r = 3$ bit/s/Hz.
Numerical Results

$N_{\text{max}} = 5$, $s = 3$ dB, $k = 50$ and $r = 3$ bit/symbol
D-RAN

RU 1

Fronthaul links

BBU 1
BBU 2

CU

RU 2
D-RAN vs. C-RAN

- Multi-bit local feedback from RUs to user + user-centric decision
Overview

- Fronthaul capacity limitations
  - Point-to-point fronthaul compression (compressed CPRI)
  - Multivariate fronthaul compression
  - Alternative functional splits

- Fronthaul latency limitations
  - HARQ with local feedback

- Conclusions
Conclusions

• Key advantages of C-RAN derive from the virtualization of base station functionalities.

• Main challenges: Capacity and latency limitations of fronthaul network

• Some solutions:
  – Quantization and compression inspired by network information theoretic principles (see also uplink, multi-hop,...)
  – Alternative functional splits
  – Local low-latency feedback

• Open issues: random access, interplay with caching,...
Extra Slides
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

- Cut-set upper bound
- Point-to-point compression
- Single-cell processing

Graph showing per-cell sum-rate [bits/s/Hz] vs. MS transmit power [dB].
Numerical Results

- Cut-set upper bound
- Distributed compression
- Point-to-point compression
- Single-cell processing
- Compute-and-forward

Graph showing per-cell sum-rate [bits/s/Hz] vs. MS transmit power [dB] with various techniques.
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 $\alpha = 0.5$)
Numerical Results

$K=5$, $N=2$, SNR = 0 dB
Network-Aware Fronthaul Compression

• Block implementation: Successive estimation-compression architecture
## 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(dB) = 128.1 + 37.6 log_{10} R (R in km)</td>
</tr>
<tr>
<td>Path-loss (pico-BS - MS)</td>
<td>PL(dB) = 38 + 30 log_{10} R (R in m)</td>
</tr>
<tr>
<td>Antenna pattern for sectorized macro-BS antennas</td>
<td>$A(\theta) = -\min[12(\theta / \theta_{3dB})^2, A_m]$</td>
</tr>
<tr>
<td></td>
<td>($\theta_{3dB} = 65^\circ, A_m = 20$ 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]

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

where

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

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

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

\[ s = 3 \text{ dB}, \quad n_{\max} = 5, \quad r = 2 \text{ bits/symbol}, \quad k = 50, \quad m_t = 1 \text{ and } m_r = 1 \]
Numerical Results

$s = 3 \text{ dB}, \ n_{\text{max}} = 5, \ r = 2 \text{ bits/symbol}, \ k = 50, \ m_t = 1 \text{ and } m_r = 1$
HARQ-TI

$n_{\text{max}}$  $n_{\text{th}}$  3rd  2nd  1st
HARQ-TI
HARQ-TI

\[ n_{\text{max}} \rightarrow n_{\text{th}} \rightarrow 3rd \rightarrow 2nd \rightarrow 1st \]
HARQ-TI

$n_{\text{max}}$ $n_{\text{th}}$ 3rd 2nd 1st
HARQ-CC

\[ n_{\text{max}} + n_{\text{th}} = n_{\text{th}} \text{ max} \]

CU

\[ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \]
\[ n_{\text{max}} + n_{\text{th}} + 3\text{rd} + 2\text{nd} + 1\text{st} = \text{max} \]
HARQ-CC

\[ n_{\text{max}} + \text{nth} + 3\text{rd} + 2\text{nd} + 1\text{st} = \text{red squares} \]
HARQ-IR

$n_{\text{max}}$  $n_{\text{th}}$  3rd  2nd  1st

CU
HARQ-IR

$n_{\text{max}}$ \quad \ldots \quad nth \quad 3rd \quad 2nd \quad 1st
Multivariate Quantization: Design

- Information-theoretic analysis assumes long encoding blocks
- Optimization via successive convex approximation
- Symbol-by-symbol quantization: multivariate quantization as vector quantization with product codebook -- Lloyd-Max-like design [Gersho and Gray ’92]
Design Constraint

- Add complexity only at the cloud and not at the RUs
Design Constraint

- Add complexity only at the cloud and not at the RUs