**Fronthaul Compression for Cloud Radio Access Networks**

Seok-Hwan Park, Osvaldo Simeone, Onur Sahin, and Shlomo Shamai

IEEE Signal Procession Magazine, Nov 2014

Summary by: Zafar Iqbal, INFONET Lab, GIST

**Introduction**

Cloud Radio Access Networks (C-RANs) are novel next-generation wireless cellular network architectures where the baseband processing is shifted from the base station (BS) to the control unit (CU) in the “cloud” of the operator. In C-RANs [1], [2], unlike the current cellular systems, the functionalities needed to process the IQ samples of the radio signals received/transmitted by the RUs are not implemented in the RUs but in the CU within the cloud of the core network. The BSs which function as radio units (RUs) are connected to the CU via fronthaul links. The fronthaul links carry information to the CU from the RU and vice versa in the form of quantized in-phase and quadrature (IQ) samples. This approach has the advantages of enabling interference management at the geographical scale covered by the distributed RUs [4] and also an effective means to implement network MIMO [5], [6] in heterogeneous wireless networks via joint processing of the baseband signals at a CU. However, the current solutions for standardization efforts [3] use conventional scalar quantization techniques which impose a bottleneck to the system performance due to limitations in the fronthaul capacity. Figure shows a typical C-RAN heterogeneous cellular network.

For example a 3-sector LTE-A system with five 20 MHz carriers (Bandwidth: 100 MHz, 2x2 MIMO, 15 bits per I/Q sample) translates to 13.8 Gbps throughput on the fronthaul [7], which is very hard to support with standard fiber optic links of about 10 Gbps capacity. Due the large amount of IQ data generated by the network and limited capacity of the fronthaul links, compression prior to transmission on the fronthaul is very necessary and is of significant importance in the design of next generation wireless cellular networks. This article provides a summary of the recent work in this area with emphasis on the advanced signal processing solutions.



**The uplink of a C-RAN with a multihop fronthaul topology between the RUs and the cloud, which contains the CU. The solid lines represent the fronthaul links.**

1. **Uplink**

**System Model**

In a C-RAN, the RUs are partitioned into clusters, such that all RUs within a cluster are managed by a single CU. Within the area covered by a given cluster, there are *NU* multi-antenna user equipment (UE) and *NR* multi-antenna RUs. In the uplink, the UE transmits wirelessly to the RUs. In turn, the RUs compress the received baseband signals and transmit the compressed signals on the fronthaul network toward the managing CU.

Assuming flat-fading channels, the discrete-time pulse-matched baseband or IQ signal **y***i*ul received by the *i*th RU at any given time sample can be written using the standard linear model,

 (1)

where **H***i*ul represents the channel matrix from all the UE in the cluster toward the *i*th RU; **x**ul is the vector of IQ symbols transmitted by all the UE in the cluster; and models thermal noise and the interference arising from the other clusters.

**Point-to-point Fronthaul Compression**

Each *i*th RU uses conventional point-to-point strategies to process the n samples of the received IQ signal **y***i*ul as shown in figure. As a result of compression, each *i*th RU produces a binary string of (at most) *nCi* bits, which allows the corresponding decompressor at the CU to identify the quantized signal within the quantization codebook. The quantized signal consists of n samples and is selected by the *i*th RU from a quantization codebook of 2*nCi* codewords [13].



**Point-to-point fronthaul compression for the uplink of C-RANs.**

**Information-theoretic Results**

The relationship between the received baseband signal **y***i*ul and its compressed version at RU *i* is given as follows,

 (2)

where the quantization noise **q***i*ul is independent of the signal **y***i*ul and distributed as , and is the covariance matrix that define the quantization noise statistics. The covariance matrix can be thought of as defining the shape of the quantization regions of the compressor. For *n* large enough, information theory guarantees that a quantization codebook exists that contains a codeword of *n* samples for any input sequence of *n* samples **y***i*ul, such that the joint empirical statistics of the two sequences is “close” to the joint distribution implied by (2). Specificall, if the fronthaul capacity *Ci* satisfies the condition

 (3)

Where the mutual information is calculated using (2), then it is possible to design a compression strategy that realizes the given quantization error covariance matrix [9].

**System Design**

Assuming that condition (3) is satisfied for all RUs, the quantized IQ signals are successfully recovered at the CU. The CU then performs joint decoding of the messages sent by all UE, which are encoded in the signals **x**ul. As a result, the uplink sum-rate

 (4)

where the mutual information can be calculated from (1) and (2), is achievable [9]. The sum-rate depends on the compression startegies used by the RUs through the covariance matrices, , which can then be maximized with respect to these matrices to identify optimal compression strategy to be used at the RUs.

**Distributed Fronthaul Compression**

As the point-to-point compression and decompression takes place in parallel, the separate processing across all the RUs neglects the fact that the baseband signals **y***i*ul are correlated across all the RUs since they are noisy observations of the same transmitted signal **x**ul. Due to this fact, the joint processing of the signals received at the CU by applying distributed compression is expected to be advantageous.

The distributed compression is done using sequential decompression [9], [18], [19], for which Wyner-Ziv compression provides information-theoretically optimal approach to leverage side information at the decompressor. This approach is shown in the figure below.



**Multiterminal fronthaul compression for the uplink of C-RANs.**

In sequential decompression, an ordering *π* is fixed on the RU indices {1,…,*NR*}. As shown in the figure, the CU performs decompression sequentially and when it decompresses the signal , it has already retrieved the signals , which are correlated with the signal of interest . Wyner-Ziv compression enables the compressor to use a finer quantizer and hence to obtain a better description of , as compared to conventional point-to-point compression, for the same fronthaul capacity *C(π)i.* As opposed to quantization codebook used in point-to-point compression, distributed compression uses binning technique which does not significantly increase the complexity of compression.

**Information-theoretic Results**

Using Wyner-Ziv compression, a given quantization error matrix in (2) is attainable if the fronthaul capacity *C(π)i* satisfies the inequality,

 (5)

It is observed that, by standard properties of mutual information [9], constraint (5) imposed on the quantization covariances is weaker than the constraint (3) corresponding to point-to-point compression.

**System Design**

Now, we want to maximize the achievable sum-rate (4) with respect to the quantization noise covariances under the fronthaul constraints (5), imposed by distributed compression for a fixed decompression order *π*. This order can be optimized and is a challenging problem. In [18], a block-coordinate optimization approach was proposed, which optimizes the covariance matrices following the same order *π* that is employed for decompression. In particular, at the ith step, for fixed (already optimized) covariances , the covariance  is obtained by solving the following,

 (6)

In (6), by the chain rule of mutual information [9], the objective function measures the sum-rate increase obtained by transmitting the signal  to the CU that already has the knowledge of the signals . An optimal covariance  of this problem is given as,

 (7)

Where **U***π(i)* is a unitary matrix whose columns are the orthonormal eigenvectors of the covariance matrix of the signal **y***𝛑i*ul conditioned on the signals , and **D***π(i)* is a diagonal matrix whose diagonal elements are obtained following a procedure similar to conventional reverse waterfilling [20]. The compression strategy described by the test channel (2) with the derived covariance matrix in (7) can be implemented at the RU *π(i)* using classical transform coding [13], as discussed in [20].

**Compute-and-forward**

In the schemes discussed so far, the quantization codebooks used by the RUs are designed separately from the channel codebooks used by the UE for transmission in the uplink. A conceptually different approach was instead proposed in [11] based on the principle of compute-and-forward. Accordingly, the same codebook is used both for channel encoding at all the UE and for quantization at the RUs.

The main advantage of this strategy is that no quantization noise is introduced by the CU because the channel and quantization codebooks are matched. On the other hand, the baseband signal (1) received at each RU is a sum with non-integer weights of the codewords transmitted by the UE. Therefore, the difference between the decoded integer combination of codewords and the actual non-integer combination of codewords resulting from the channel, affects decoding at each RU as an additional noise term.

**Numerical Example**

We focus on a standard three-cell circulant Wyner model [6], where each cell contains a single-antenna UE and a single-antenna RU, and intercell interference takes place only between adjacent cells (the first and third cell are considered to be adjacent). The intercell channel gain is equal to *g* = 0.4. Moreover, every RU has the same fronthaul capacity of 3 bits/s/Hz. The given figure plots the achievable per-cell sum-rate for point-topoint compression, distributed compression, and compute-andforward versus the transmitted UE power *P*, which can be taken as a measure of signal-to-noise ratio (SNR). It can be seen that the performance advantage of distributed compression over point-to-point compression increases as the SNR grows larger. This is because the correlation of the received signals in (1) at the RUs becomes more pronounced as the SNR increases. On the other hand, compute-and-forward outperforms all the other schemes as the SNR increases, i.e., in the regime where the fronthaul capacity is the main performance bottleneck.



**Per-cell uplink sum-rate versus the transmitted UE power *p* for the circulant Wyner model with *C*** = **3 bits/s/Hz dB and**

**intercell channel gain equal to *g*** = **0.4.**

**Multihop Fronthaul Topology**

To convery the quantized IQ samples from the RUs to the CU via multiple hops, each RU must decide on how to forward the information to the CU. One option is to use routing, which is inefficient in the presence of a dense deployment of RUs because it is wasteful of the fronthaul capacity to merely forward all the bit streams received from the connected RUs. Therefore, the correlated baseband signals are combined at the RU to reduce redundancy. This process is known as *in-network processing*.

**Numerical Example**

We now compare the sum-rates achievable with routing and with in-network processing for the uplink of a C-RAN with a two-hop fronthaul network. Specifically, there are *N* RUs in the first layer and two RUs in the second layer, all receiving in the uplink. The RUs in the first layer do not have direct fronthaul links to the CU, while the RUs in the second layer do. Half of the RUs in the first layer is connected to one RU in the second layer, and half to the other RU in the second layer. We assume that all fronthaul links have capacity equal to 2–4 bits/s/Hz and all channel matrices have identically and independently distributed (i.i.d.) complex Gaussian entries with unit power (Rayleigh fading).

Figure shows the average sum-rate versus the number *N* of RUs in layer 1 with *NU* =4 UE and average received per-antenna SNR of 20 dB at all RUs. As the number of RUs in the first layer increases, the performance of in-network processing over routing becomes more pronounced. Therefore, in-network processing is desirable for each RU in layer 2 in case of dense deployment of RUs.



**Average uplink sum-rate versus the number of RUs in layer 1 with *NU*** = **4 UEs, average received per-antenna SNR of**

**20 dB and fronthaul capacity of 2–4 bits/s/Hz.**

1. **Downlink**

**System Model**

In the downlink, the CU that manages a given cluster processes the information messages of the UE within the cluster by performing channel coding and precoding on behalf of the RUs. As seen in Figure 6, the precoded baseband signals are then compressed by the CU, which finally forwards the compressed IQ signals to the RUs on the fronthaul links. Each RU decompresses the signal received on the fronthaul link (by looking up the corresponding quantization codebook), performs pulse shaping, upconverts the resulting signal, and transmits it to the UE on the wireless downlink channel.

**Point-to-point Fronthaul Compression**

Similar to the uplink, in the conventional C-RAN implementation, the CU compresses separately, the pre-coded IQ signals intended for transmission by different RUs, using point-to-point compression, as shown in the figure.



**Point-to-point fronthaul compression for the downlink of C-RANs.**

**Multivariate Fronthaul Compression**

Unlike point-to-point compression, the IQ signals are jointly compressed to select the quantized signals from the corresponding quantization codebooks. Correlating the quantization noises are beneficial to control the effect of the additive quantization noises on the reception of the UE and enhance the achievable rates. Multivariate fronthaul compression block diagram is shown in the figure.



**Multiterminal fronthaul compression for the downlink of C-RANs.**

**Compute-and-forward**

Similar to the uplink, a compute-and-forward strategy is used for the downlink in which the CU employs the same lattice code to perform channel encoding for all the UE, as the one used for quantization [24]. While not adding any quantization noise, this scheme also has limited performance due to constraints on the coefficients of the precoding matrix.

**Numerical Example**

Figure shows the per-cell sum-rate of compression schemes as applied to both linear precoding and dirty paper nonlinear precoding with the same model as used before. It is observed that multivariate compression significantly outperforms point-to-point compression for both linear precoding and “dirty paper” nonlinear precoding. Moreover, compute-and-forward is the most effective strategy in the regime of moderate fronthaul capacity *C* in which the limitations imposed by integer precoding are not dominant. In contrast, for sufficiently large fronthaul capacity *C*, both compression based schemes attain the upper bound, while compute-and-forward is limited by the mentioned integrality constraints.



**Per-cell downlink sum-rate versus the fronthaul capacity *C* for the circulant Wyner model with transmitted UE power *p*** = **20 dB and intercell channel gain equal to *g*** = **0.5.**

**Performance Evaluation**

Using the standard cellular topology channel models of [26], a performance evaluation of the discussed fronthaul compression techniques is done by using the convetional metric of cell-edge throughput versus the average per-UE spectral efficiency [8]. Consider a macro-cell located at the center of a two-dimensional 19-cell hexagonal cellular layout. In each macrocell, there are *K* randomly and uniformly located UE; a macro-BS with three sectorized antennas placed in the center; and a single randomly and uniformly located single-antenna pico-BS. A single-hop fronthaul topology is assumed, where each macrocell is a cluster served by a CU that is connected directly to the macro-BS and the pico-BS in the macrocell. Specifically, the fronthaul links to each macro-BS antenna and to each pico-BS have capacities *C*macro and *C*pico, respectively.

Figure shows the cell-edge throughput versus the average spectral efficiency for *K*=4 UE. It is observed that spectral efficiencies larger than 1.05 bits/s/Hz are not achievable with point-to-point compression, while they can be obtained with multivariate compression. Moreover, it is seen that multivariate compression provides 2x gain in terms of cell-edge throughput for a spectral efficiency of 1 bits/s/Hz.



**Cell-edge throughput, i.e., fifth percentile rate, versus the average per-UE spectral efficiency for various fairness constants**

*𝛼* **in the downlink of a C-RAN with *K*** = **4 UEs, (*C*macro,*C*pico)** = **(3,1) bits/s/Hz, *T*** = **5 and *𝛃***= **0.5.**

**Conclusions**

The design of C-RANs poses a host of new research challenges to the signal processing community. One key problem is that of devising effective compression algorithms for the fronthaul links connecting the RUs with the CU that resides within the “cloud” of the operator’s core network. As reviewed in this article, the performance of conventional point-to-point compression strategies can be substantially improved by leveraging techniques inspired by network information theory. Most notably, we have emphasized the potential gains of multiterminal compression—distributed compression for the uplink and multivariate compression for the downlink—and of structured coding via compute-and-forward.

**References**

Same as in the original paper.