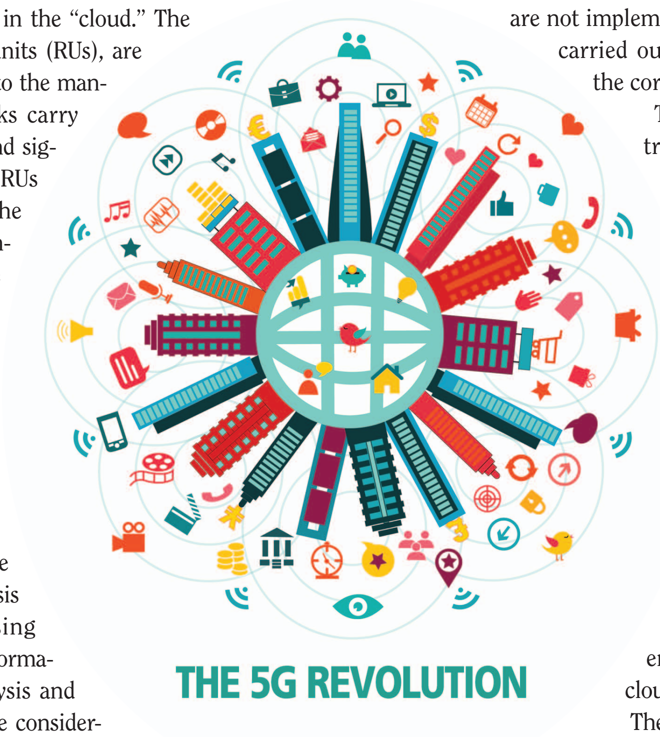
Seok-Hwan Park, Osvaldo Simeone, Onur Sahin, and Shlomo Shamai (Shitz)

Fronthaul Compression for Cloud Radio Access Networks

Signal processing advances inspired by network information theory

Cloud radio access networks (C-RANs) provide a novel architecture for next-generation wireless cellular systems whereby the baseband processing is migrated from the base stations (BSs) to a control unit (CU) in the "cloud." The BSs, which operate as radio units (RUs), are connected via fronthaul links to the managing CU. The fronthaul links carry information about the baseband signals—in the uplink from the RUs to the CU and vice versa in the downlink—in the form of quantized in-phase and quadrature (IQ) samples. Due to the large bit rate produced by the quantized IQ signals, compression prior to transmission on the fronthaul links is deemed to be of critical importance and is receiving considerable attention. This article provides a survey of the work in this area with emphasis on advanced signal processing solutions based on network information theoretic concepts. Analysis and numerical results illustrate the considerable performance gains to be expected for standard cellular models.



©ISTOCKPHOTO.COM/ZONADEARTE

INTRODUCTION

C-RANs provide a promising architecture for next-generation wireless cellular systems that is based on the separation of distributed RUs and centralized information processing nodes [1], [2]. Unlike current cellular systems, in C-RANs, the functionalities needed to process the baseband complex or IQ envelopes of the radio signals received/transmitted by the RUs are not implemented at the RUs; instead, they are carried out remotely within the "cloud" of the core network.

To this end, the baseband signals are transferred between the cloud and the RUs on a network of fronthaul links. As an example, Figure 1 illustrates the uplink of a C-RAN in a heterogeneous cellular network with RUs consisting of macro-BSs and pico-BSs and a multihop fronthaul topology between the RUs and the cloud (see, e.g., [3]). Note that on the used nomenclature, fronthaul links are often distinguished from backhaul links in that they have more stringent requirements on latency and synchronization to enable baseband processing in the cloud [3].

The centralization of information processing made possible by C-RANs enables interference management at the geographical scale covered by the distributed RUs (see, e.g., [4]). In fact, C-RANs provide an effective means to implement network multiple-input, multiple-output (MIMO) [5], [6] in heterogeneous wireless networks via the joint processing of the baseband signals at a CU, also known as baseband unit, in the cloud.

As discussed, the key feature of C-RANs is the use of a fronthaul network for the transfer of baseband information to

Digital Object Identifier 10.1109/MSP.2014.2330031
Date of publication: 15 October 2014

and from the cloud. Current solutions, which are the object of various standardization efforts [3], prescribe the use of conventional scalar quantizers for this purpose. However, with this approach, fronthaul capacity limitations are known to impose a formidable bottleneck to the system performance.

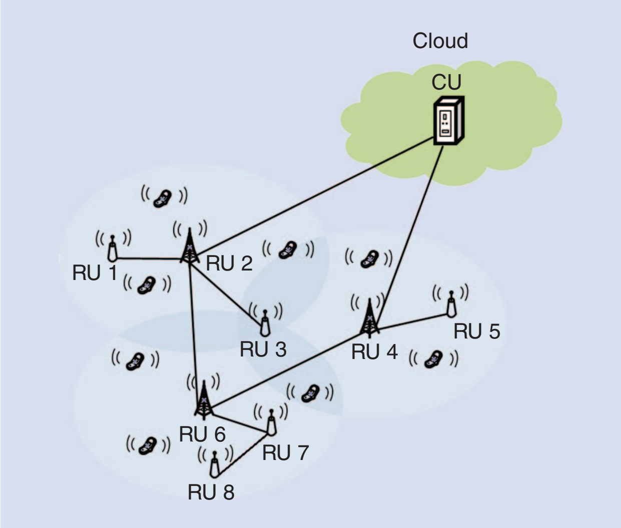
EXAMPLE

Consider an RU consisting of an long-term evolution (LTE) macro-BS that serves three cell sectors with five carriers and two receive antennas. As summarized in [7], it can be calculated that, using standard scalar quantization techniques with 15 bits/baseband IQ sample, the throughput required on the fronthaul links exceeds even the 10 Gbit/s provided by standard fiber optics links. The problem is even more pronounced for smaller RUs, e.g., pico-BSs or home-BSs, that, while operating with fewer antennas, channels, and sectors, are typically connected to fronthaul links of lower capacity, such as DSL-based wireline or millimeter-wave channels. ■

To alleviate the performance bottleneck identified above, recent efforts have targeted the design of more advanced fronthaul compression schemes. These schemes are based on point-to-point compression algorithms (see, e.g., [1], [7], and [8]). However, as is well known from network information theory, point-to-point techniques fail to achieve the optimal performance in the context of even the simplest networks, such as star, or single-hop, topologies [9].

Motivated by the previous discussion, this article aims at providing a survey of the work in the area of fronthaul compression with emphasis on advanced signal processing solutions based on network information theoretic concepts. Specifically, the main ideas that are brought to bear from network information theory are:

1) *Multiterminal compression*: In contrast to point-to-point compression, multiterminal compression allows for the joint processing of the compressed IQ samples of different RUs at the CU. Specifically, in the uplink, joint decompression enables the CU to leverage the correlation among the signals received by neighboring RUs. The key technique that makes this possible is distributed compression or Wyner–Ziv coding [10]. Instead, in the downlink, joint compression allows the CU to correlate the quantization noises of the baseband signals transmitted by neighboring RUs. This can be done via the information-theoretic technique of multivariate compression [9, Ch. 9].



[FIG1] 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.

2) *Structured coding*: Point-to-point and multiterminal compression employ unstructured quantization codebooks that are designed independently of the channel codebooks used for transmission on the wireless channels. As a conceptually different alternative, structured codes that are matched to the channel codebooks may instead be used. This leads to new strategies for C-RANs based on the framework of compute-and-forward [11].

In the following, we review point-to-point/multiterminal fronthaul compression and structured coding for the uplink and downlink of a C-RAN. Throughout, we provide numerical results to illustrate the key concepts. We also provide simulation results over standard cellular models to substantiate the gains that are expected from the implementation of multiterminal fronthaul compression in real-world systems. See "Information Theoretic Measures" for a brief review of the standard information theoretic notations used in the article.

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 N_U multiantenna user equipment (UE) and N_R multiantenna 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.

INFORMATION THEORETIC MEASURES

Throughout the article, we adopt standard information-theoretic definitions for the mutual information $I(X;Y)$, conditional mutual information $I(X;Y|Z)$, differential entropy $h(X)$ and conditional differential entropy $h(X|Y)$ [9]. For jointly complex Gaussian variables $(\mathbf{x},\mathbf{y})\sim\mathcal{CN}(0,\Omega_{\mathbf{x},\mathbf{y}})$, we define the conditional covariance matrix as $\Omega_{\mathbf{x}|\mathbf{y}}=\mathbb{E}[\mathbf{x}\mathbf{x}^{\dagger}|\mathbf{y}]= \ \Omega_{\mathbf{x}}-\Omega_{\mathbf{x},\mathbf{y}}\Omega_{\mathbf{y}}^{-1}\Omega_{\mathbf{x},\mathbf{y}}^{\dagger}$, where $\Omega_{\mathbf{x}}=\mathbb{E}[\mathbf{x}\mathbf{x}^{\dagger}]$, $\Omega_{\mathbf{y}}=\mathbb{E}[\mathbf{y}\mathbf{y}^{\dagger}]$, $\Omega_{\mathbf{x},\mathbf{y}}=\mathbb{E}[\mathbf{x}\mathbf{y}^{\dagger}]$ and the operation $(\cdot)^{\dagger}$ denotes the Hermitian transpose of a matrix or vector. Then, for joint complex Gaussian vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, the quantities $I(\mathbf{x};\mathbf{y})$ and $I(\mathbf{x};\mathbf{y}|\mathbf{z})$ are computed as $I(\mathbf{x};\mathbf{y})=h(\mathbf{x})-h(\mathbf{x}|\mathbf{y})=\text{logdet}(\Omega_{\mathbf{x}})-\text{logdet}(\Omega_{\mathbf{x}|\mathbf{y}})$ and $I(\mathbf{x};\mathbf{y}|\mathbf{z})=h(\mathbf{x}|\mathbf{z})-h(\mathbf{x}|\mathbf{y},\mathbf{z})=\text{logdet}(\Omega_{\mathbf{x}|\mathbf{z}})-\text{logdet}(\Omega_{\mathbf{x}|\mathbf{y},\mathbf{z}})$, respectively.

The fronthaul network connecting the RUs to the CU may have a single-hop topology, in which all RUs are directly connected to the CU or, more generally, a multihop topology. We first concentrate on the single-hop topology and then discuss the multihop case. An example of a single-hop C-RAN is the network shown in Figure 1 when restricted to RU 2 and RU 4.

Assuming flat-fading channels, the discrete-time pulse-matched baseband or IQ signal $\mathbf{y}_i^{\mathrm{ul}}$ received by the ith RU at any given time sample can be written using the standard linear model

$$\mathbf{y}_i^{\mathrm{ul}} = \mathbf{H}_i^{\mathrm{ul}}\mathbf{x}^{\mathrm{ul}} + \mathbf{z}_i^{\mathrm{ul}}, \tag{1}$$

where $\mathbf{H}_i^{\mathrm{ul}}$ represents the channel matrix from all the UE in the cluster toward the ith RU; $\mathbf{x}^{\mathrm{ul}}$ is the vector of IQ symbols transmitted by all the UE in the cluster; and $\mathbf{z}_i^{\mathrm{ul}} \sim \mathcal{CN}(\mathbf{0}, \Omega_{\mathbf{z}_i^{\mathrm{ul}}})$ models thermal noise and the interference arising from the other clusters. The signals $\mathbf{x}^{\mathrm{ul}}$ transmitted by the UE are assumed to be jointly complex Gaussian and independent across the UE. This corresponds to assuming standard point-to-point channel codes at the UE (see, e.g., [9, Ch. 3]). The channel matrices are assumed to be fixed and to remain constant during a coding block, which is of size n samples. Note that in (1) and in the following, we do not denote explicitly the dependence of the signals on the sample index to simplify the notation.

In the single-hop topology under study, each RU i is connected to the CU via a fronthaul link of capacity C_i bits/s/Hz. The fronthaul capacity is normalized to the bandwidth of the uplink channel. This implies that for any uplink coding block of n symbols, nC_i bits can be transmitted on the ith fronthaul link.

REMARK 1

Model (1), which is typically used in related literature, assumes implicitly that the RUs perform time and frequency synchronization locally. In fact, signal (1) is free of frequency drift and is sampled at the baud rate. It is noted that, if time synchronization is not carried out at the RUs, then the RUs need to oversample the baseband signals prior to transferring them on the fronthaul links. This is, for instance, prescribed in the CPRI standard [3].

REMARK 2

Following Remark 1, while model (1) assumes that time and frequency synchronization is done locally, the optimal allocation of layer 1 functionalities, such as synchronization and channel estimation, between the RUs and the CU is a subject of ongoing investigations (see, e.g., [12] for a related discussion).

POINT-TO-POINT FRONTHAUL COMPRESSION

In baseline C-RAN systems, each ith RU uses conventional point-to-point compression strategies to process the n samples of the received IQ signal $\mathbf{y}_i^{\mathrm{ul}}$, as illustrated in Figure 2.

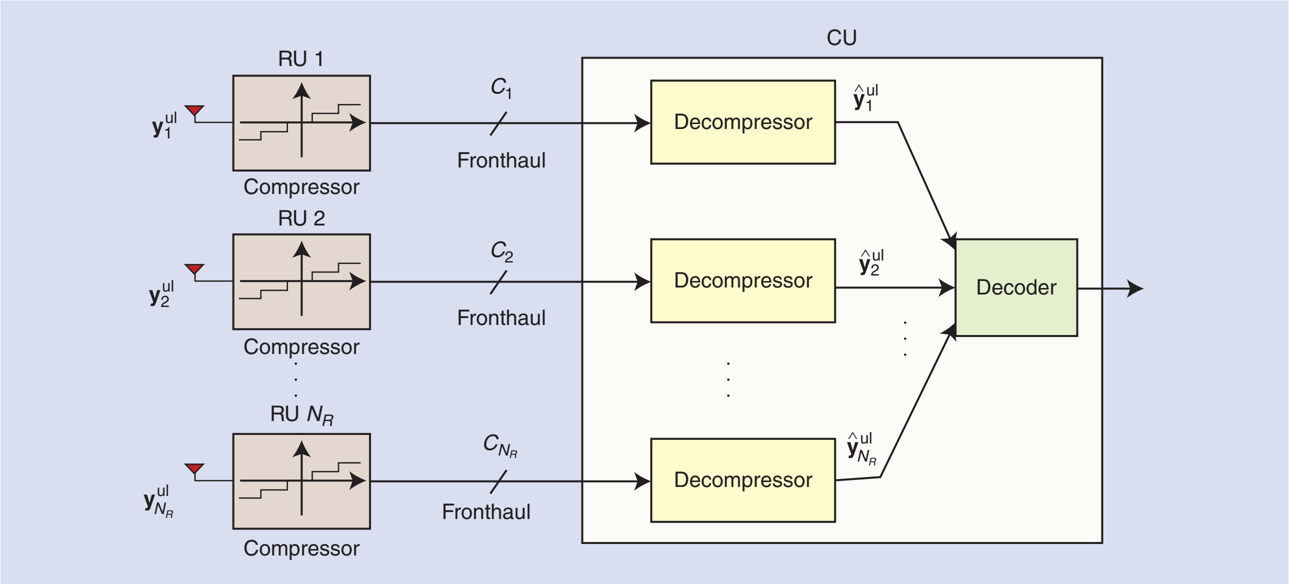
POINT-TO-POINT COMPRESSION

As a result of compression, each ith RU produces a binary string of (at most) nC_i 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 $\hat{\mathbf{y}}_i^{\mathrm{ul}}$, and is selected by the ith RU from a quantization codebook of 2^{nC_i} codewords (see, e.g., [13]). The example of a scalar quantizer ($n = 1$) at each RU is illustrated in Figure 2 for either the I or the Q component of the IQ sample.

KEY INFORMATION-THEORETIC RESULTS

A standard way of modeling the relationship between the received baseband signal $\mathbf{y}_i^{\mathrm{ul}}$ and its compressed version $\hat{\mathbf{y}}_i^{\mathrm{ul}}$ at RU i is to follow information-theoretic considerations (see, e.g., [9, Ch. 3]) and adopt the Gaussian "test channel"

$$\hat{\mathbf{y}}_i^{\mathrm{ul}} = \mathbf{y}_i^{\mathrm{ul}} + \mathbf{q}_i^{\mathrm{ul}}, \tag{2}$$



CU
RU 1
$\mathbf{y}_1^{\mathrm{ul}}$
Compressor
C_1
Fronthaul
Decompressor
$\hat{\mathbf{y}}_1^{\mathrm{ul}}$
RU 2
$\mathbf{y}_2^{\mathrm{ul}}$
Compressor
C_2
Fronthaul
Decompressor
$\hat{\mathbf{y}}_2^{\mathrm{ul}}$
Decoder
RU N_R
$\mathbf{y}_{N_R}^{\mathrm{ul}}$
Compressor
C_{N_R}
Fronthaul
Decompressor
$\hat{\mathbf{y}}_{N_R}^{\mathrm{ul}}$

[FIG2] Point-to-point fronthaul compression for the uplink of C-RANs.

where the quantization noise $\mathbf{q}_i^{\text{ul}}$ is independent of the signal $\mathbf{y}_i^{\text{ul}}$ and distributed as $\mathbf{q}_i^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i^{\text{ul}})$. The quantization noise statistics are thus defined by the covariance matrix $\mathbf{\Omega}_i^{\text{ul}}$. Connecting the information-theoretic viewpoint with classical vector quantization, the covariance matrix $\mathbf{\Omega}_i^{\text{ul}}$ can be thought of defining the shape of the quantization regions of the compressor.

Information theory provides analytical conditions that relate the quantization noise statistics $\mathbf{\Omega}_i^{\text{ul}}$ to the size of the quantization codebooks, and hence to the fronthaul capacity C_i, needed to satisfy the condition (2). More precisely, under these conditions (and for n large enough), the theory guarantees that a quantization codebook exists that contains a codeword of n samples $\hat{\mathbf{y}}_i^{\text{ul}}$ for any input sequence of n samples $\mathbf{y}_i^{\text{ul}}$, such that the joint empirical statistic of the two sequences is "close" to the joint distribution implied by (2) [9, Ch. 3].

Specifically, a standard result in information theory states that, if the fronthaul capacity C_i satisfies the condition

$$I(\mathbf{y}_i^{\text{ul}}; \hat{\mathbf{y}}_i^{\text{ul}}) \leq C_i, \tag{3}$$

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 $\mathbf{\Omega}_i^{\text{ul}}$ in the sense discussed above (see, e.g., [9, Ch. 3]). At an intuitive level, condition (3) says that a "smaller" covariance matrix $\mathbf{\Omega}_i^{\text{ul}}$, and hence a larger mutual information $I(\mathbf{y}_i^{\text{ul}}; \hat{\mathbf{y}}_i^{\text{ul}})$, calls for a larger required fronthaul capacity C_i.

SYSTEM DESIGN

Assuming that the condition (3) is satisfied for all RUs, the quantized IQ signals $\hat{\mathbf{y}}_1^{\text{ul}}, \ldots, \hat{\mathbf{y}}_{N_R}^{\text{ul}}$ 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 $\mathbf{x}^{\text{ul}}$. As a result, the uplink sum-rate

$$R_{\text{sum}}^{\text{ul}} = I(\mathbf{x}^{\text{ul}}; \hat{\mathbf{y}}_1^{\text{ul}}, \ldots, \hat{\mathbf{y}}_{N_R}^{\text{ul}}), \tag{4}$$

where the mutual information can be calculated from (1) and (2), is achievable (see, e.g., [9, Ch. 4]). Note that individual rates could also be similarly calculated using standard results on the capacity region of multiple access channels, and so could rates achievable with suboptimal decoding strategies such as treating interference as noise (see [9, Ch. 4]).

The sum-rate (4) depends on the compression strategies used by the RUs through the covariance matrices $\mathbf{\Omega}_i^{\text{ul}}$, $i = 1, \ldots, N_R$. The sum-rate can then be maximized with respect to these matrices to identify the optimal compression strategies to be used at the RUs. The nonconvex problem of maximizing the sum-rate under the fronthaul constraints (3), for $i = 1, \ldots, N_R$, over the matrices $\mathbf{\Omega}_i^{\text{ul}}$, $i = 1, \ldots, N_R$, falls in the category of difference-of-convex problems and can be tackled by using the so-called majorization minimization (MM) algorithm [14].

REMARK 3

To compress its received signal $\mathbf{y}_i^{\text{ul}}$, each RU i must only be informed about the quantization codebook to be used. Furthermore, the achievability of the sum-rate (4) hinges on the assumption that the CU is aware of the channel matrices of all the active UE. Each ith RU may estimate the channel matrix $\mathbf{H}_i^{\text{ul}}$ based on standard uplink training and then forward the estimated matrix to the CU on the fronthaul links. The CU can then optimize the compression strategies as discussed above and inform accordingly the RUs. We refer to [15] and [16] for an analysis of the overhead associated with the transfer of channel state information on the fronthaul links for ergodic fading channels.

DISTRIBUTED FRONTHAUL COMPRESSION

As seen in Figure 2, with standard point-to-point compression, compression and decompression across different RUs take place in parallel. This separate processing across the RUs neglects the key fact that the baseband signals $\mathbf{y}_i^{\text{ul}}$ in (1) are correlated across the RU index i, since they are noisy observations of the same transmitted signals $\mathbf{x}^{\text{ul}}$. Based on this fact, the joint processing of the signals received on the fronthaul links at the CU via distributed compression is expected to be advantageous, as first proposed in [17].

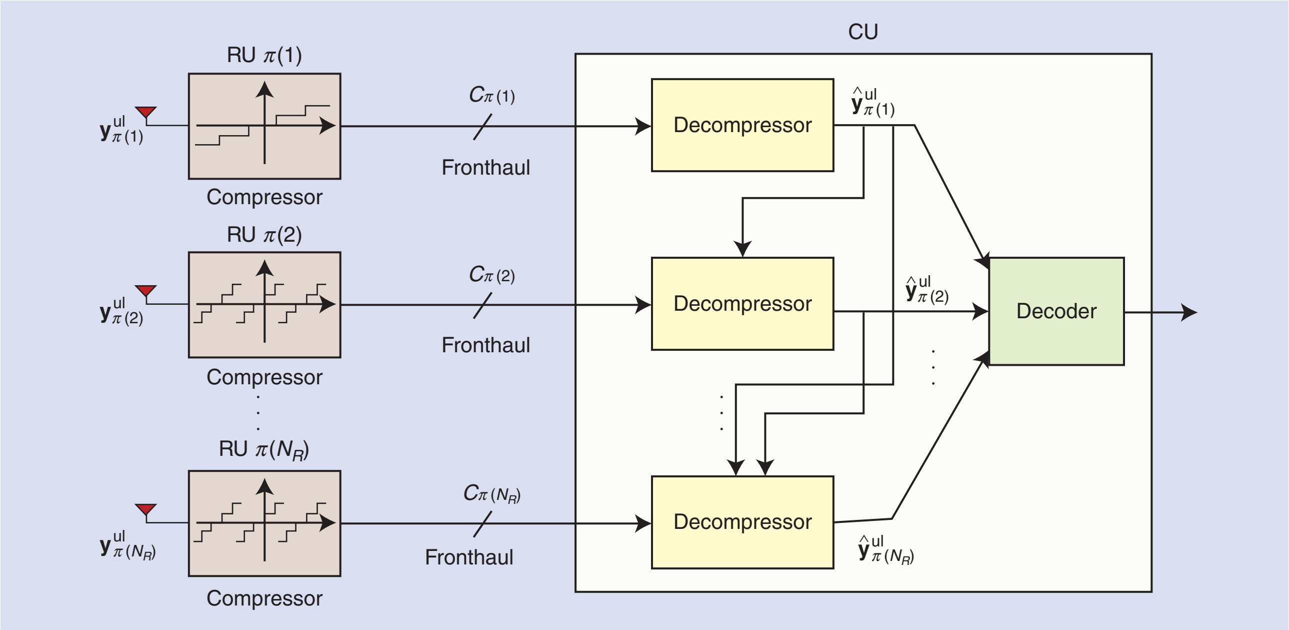
DISTRIBUTED COMPRESSION

To explain distributed compression, here we concentrate on the practical implementation that uses sequential decompression (see [9, Ch. 10] and also [18] and [19]). To this end, we fix an ordering π on the RU indices $\{1, \ldots, N_R\}$. As shown in Figure 3, the CU first decompresses the signal $\hat{\mathbf{y}}_{\pi(1)}^{\text{ul}}$, then $\hat{\mathbf{y}}_{\pi(2)}^{\text{ul}}$, and so on until $\hat{\mathbf{y}}_{\pi(N_R)}^{\text{ul}}$. Therefore, when decompressing $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$, the CU has already retrieved the signals $\{\hat{\mathbf{y}}_{\pi(1)}^{\text{ul}}, \ldots \hat{\mathbf{y}}_{\pi(i-1)}^{\text{ul}}\}$, which are correlated with the signal of interest $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$.

Wyner–Ziv compression offers the information-theoretically optimal approach to leverage side information available at the decompressor to improve the quality of the description $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$. Specifically, Wyner–Ziv compression enables the compressor to use a finer quantizer and hence to obtain a better description $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$, as compared to conventional point-to-point compression, for the same fronthaul capacity $C_{\pi(i)}$.

The approach works as follows. Since a finer quantizer has more codewords than the $2^{nC_{\pi(i)}}$ binary strings that can be supported on the fronthaul link, Wyner–Ziv compression associates the same binary string of $nC_{\pi(i)}$ bits to a subset of codewords. This is in contrast to point-to-point compression in which a distinct binary string is associated with each codeword in the quantization codebook. This subset is known as *bin*, and the *binning* step can be, in practice, realized by using a coset of linear codes or hashing (see, e.g., [10]). Therefore, the complexity of compression is not significantly increased as compared to the point-to-point approach. An example is shown in Figure 3, where RUs $\pi(i)$ with $i > 1$ use a scalar quantizer $(n = 1)$ that assigns the same quantization level to multiple regions of the real line (for the I and Q components).

When using Wyner–Ziv compression, the decompressor is thus faced with the problem of having to distinguish among all codewords $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$ that belong to the bin indexed by the binary string received on the fronthaul link. As long as the bins are not


CU
RU $\pi(1)$
$\mathbf{y}^{ul}_{\pi(1)}$
Compressor
$C_{\pi(1)}$
Fronthaul
Decompressor
$\hat{\mathbf{y}}^{ul}_{\pi(1)}$
RU $\pi(2)$
$\mathbf{y}^{ul}_{\pi(2)}$
Compressor
$C_{\pi(2)}$
Fronthaul
Decompressor
$\hat{\mathbf{y}}^{ul}_{\pi(2)}$
Decoder
RU $\pi(N_R)$
$\mathbf{y}^{ul}_{\pi(N_R)}$
Compressor
$C_{\pi(N_R)}$
Fronthaul
Decompressor
$\hat{\mathbf{y}}^{ul}_{\pi(N_R)}$

[FIG3] Multiterminal fronthaul compression for the uplink of C-RANs.

too large, this can be done by leveraging the available correlated side information $\{\hat{\mathbf{y}}^{ul}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{ul}_{\pi(i-1)}\}$. In fact, because of their statistical dependence, the real codeword $\hat{\mathbf{y}}^{ul}_{\pi(i)}$ is expected to be "closer" to the side information sequences. This detection step can be in practice performed by using channel decoding algorithms such as message passing or trellis search (see, e.g., [10]).

KEY INFORMATION-THEORETIC RESULTS

A classical information-theoretic result states that, using Wyner–Ziv compression, a given quantization error matrix $\boldsymbol{\Omega}^{ul}_{\pi(i)}$ in (2) is attainable if the fronthaul capacity $C_{\pi(i)}$ satisfies the inequality

$$I(\mathbf{y}^{ul}_{\pi(i)}; \hat{\mathbf{y}}^{ul}_{\pi(i)} \,|\, \hat{\mathbf{y}}^{ul}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{ul}_{\pi(i-1)}) \leq C_{\pi(i)}. \tag{5}$$

It is observed that, by standard properties of the mutual information [9, Ch. 2], the constraint (5) imposed on the quantization covariances $\boldsymbol{\Omega}^{ul}_{\pi(i)}$ is weaker than the constraint (3) corresponding to point-to-point compression. Specifically, the gap between the two mutual information quantities on the left-hand sides of (3) and (5) increases as the correlation between the useful signal $\hat{\mathbf{y}}^{ul}_{\pi(i)}$ and the side information $\{\hat{\mathbf{y}}^{ul}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{ul}_{\pi(i-1)}\}$ grows and vanishes if signal and side information are independent.

SYSTEM DESIGN

We are now interested in maximizing the achievable sum-rate (4) with respect to the quantization noise covariances $\boldsymbol{\Omega}^{ul}_{i}$, for $i = 1, \ldots, N_R$ under the fronthaul constraints (5) imposed by distributed compression for a fixed decompression order π. This order can be also optimized upon, as further discussed in Remark 4.

The optimization problem at hand is generally challenging. In [18, Sec. III], a (suboptimal) block-coordinate optimization approach was proposed that leverages a key result in [20]. Accordingly, one optimizes the covariance matrices following the same order π that is employed for decompression. In particular, at the ith step, for fixed (already optimized upon) covariances $\boldsymbol{\Omega}^{ul}_{\pi(1)}, \ldots, \boldsymbol{\Omega}^{ul}_{\pi(i-1)}$, the covariance $\boldsymbol{\Omega}^{ul}_{\pi(i)}$ is obtained by solving the following problem:

$$\begin{aligned} &\underset{\boldsymbol{\Omega}^{ul}_{\pi(i)} \succeq 0}{\text{maximize}}\; I(\mathbf{x}^{ul}; \hat{\mathbf{y}}^{ul}_{\pi(i)} \,|\, \hat{\mathbf{y}}^{ul}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{ul}_{\pi(i-1)}) \\ &\text{s.t.}\; I(\mathbf{y}^{ul}_{\pi(i)}; \hat{\mathbf{y}}^{ul}_{\pi(i)} \,|\, \hat{\mathbf{y}}^{ul}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{ul}_{\pi(i-1)}) \leq C_{\pi(i)}. \end{aligned} \tag{6}$$

In (6), by the chain rule of mutual information (see [9, Ch. 2]), the objective function measures the sum-rate increase obtained by transmitting the signal $\hat{\mathbf{y}}^{ul}_{\pi(i)}$ to the CU that already has the knowledge of the signals $\{\hat{\mathbf{y}}^{ul}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{ul}_{\pi(i-1)}\}$.

It was shown in [20] that an optimal covariance $\boldsymbol{\Omega}^{ul}_{\pi(i)}$ of this problem is given as

$$\boldsymbol{\Omega}^{ul}_{\pi(i)} = \mathbf{U}_{\pi(i)} \mathbf{D}_{\pi(i)} \mathbf{U}^{\dagger}_{\pi(i)}, \tag{7}$$

where $\mathbf{U}_{\pi(i)}$ is a unitary matrix whose columns are the orthonormal eigenvectors of the covariance matrix of the signal $\mathbf{y}^{ul}_{\pi(i)}$ conditioned on the signals $\{\hat{\mathbf{y}}^{ul}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{ul}_{\pi(i-1)}\}$, and $\mathbf{D}_{\pi(i)}$ is a diagonal matrix whose diagonal elements are obtained following a procedure similar to conventional reverse waterfilling [20, Th. 1].

The compression strategy described by the test channel (2) with the derived covariance matrix $\boldsymbol{\Omega}^{ul}_{\pi(i)}$ in (7) can be implemented at the RU $\pi(i)$ using classical transform coding [13] as discussed in [20, Sec. III-A]. Accordingly, the RU $\pi(i)$ first applies the linear preprocessing matrix $\mathbf{U}^{\dagger}_{\pi(i)} \mathbf{D}^{-1/2}_{\pi(i)}$ to the received signal vector $\mathbf{y}^{ul}_{\pi(i)}$ and then independently compresses the resulting signal streams using a Gaussian test channel with noise of unit variance. It can be proved that multiplication by the unitary transform $\mathbf{U}^{\dagger}_{\pi(i)}$, also referred to as conditional Karhunen–Loeve transform (KLT) [21], decorrelates the received signal streams when conditioned on the side information signals $\{\hat{\mathbf{y}}^{\text{ul}}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{\text{ul}}_{\pi(i-1)}\}$.

REMARK 4

The decompression order π generally affects the achievable performance and should be optimized upon. A choice that is generally sensible, and close to optimal, is that of decompressing first the signals coming from macro-BSs and then those from pico- or femto-BSs in their vicinity. The rationale for this approach is that macro-BSs tend to have a larger fronthaul capacity and hence their decompressed signals provide relevant side information for the signals coming from smaller cells, which are typically connected with lower capacity fronthaul links.

REMARK 5

In the previous discussion, it was assumed that the CU first decompresses the signals and then decodes the messages of the UE based on the decompressed signals. The performance may be improved by performing joint decompression and decoding at the cost of an increased computational complexity [17].

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 approach proposed in [11] selects a (nested) lattice code. Lattice codes have the property that the weighted-sum—more precisely the modulo-sum with respect to the coarse lattice—of two codewords is also a codeword, as long as the weights are integer numbers. In the scheme of [11], each RU then decodes an appropriate (modulo-)sum, with integer weights, of the codewords transmitted by the UE. The bit stream sent on the fronthaul link identifies the decoded codeword within the lattice code. The idea is that, upon receiving a sufficient number of linear combinations of codewords from the RUs, the CU can invert the resulting linear system and recover the transmitted codewords.

The key potential advantage of the compute-and-forward strategy is that no quantization noise is introduced by the CU due to the fact that the channel and quantization codebooks are matched. On the flip side, the baseband signal (1) received at each RU is a sum with noninteger weights of the codewords transmitted by the UE. Therefore, the difference between the decoded integer combination of codewords and the actual noninteger combination of codewords resulting from the channel affects decoding at each RU as an additional noise term.

NUMERICAL EXAMPLE

We now discuss the performance of point-to-point compression and of more advanced strategies in the context of a specific example. We focus on a standard three-cell circulant Wyner model (see, e.g., [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). This implies that the received signal (1) is given as $y^{\text{ul}}_i = x^{\text{ul}}_i + gx^{\text{ul}}_{[i-1]_3} + gx^{\text{ul}}_{[i+1]_3} + z^{\text{ul}}_i$, where x^{ul}_j is the signal sent by the UE in cell j and $[\cdot]_3$ represents the modulo-3 operation. The intercell channel gain is equal to $g = 0.4$. Moreover, every RU has the same fronthaul capacity of 3 bits/s/Hz.

Figure 4 plots the achievable per-cell sum-rate for point-to-point compression, distributed compression, and compute-and-forward versus the transmitted UE power P, which can be taken as a measure of signal-to-noise ratio (SNR). For reference, we also show the per-cell sum-rate achievable with single-cell processing, whereby each RU decodes the signal of the in-cell UE by treating all other UE signals as noise, and the cut-set upper bound [6]. 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. As for compute-and-forward, its performance at low SNR coincides with single-cell processing, as the RUs tend to decode trivial combinations consisting only of the signals of the local UE. 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. Further discussion can be found in the "Downlink" section.

MULTIHOP FRONTHAUL TOPOLOGY

In this subsection, we study the case in which the fronthaul network has a general multihop topology. As an example, in Figure 1, RU 6 communicates to the CU via a two-hop fronthaul connection that passes through RU 2 and RU 4. Note that each RU may have multiple incoming and outgoing fronthaul links.

ROUTING

To convey the quantized IQ samples from the RUs to the CU through multiple hops, each RU must decide on the information to be transmitted on each outgoing fronthaul link based on the information received on the incoming fronthaul links. A first option is to use routing: the bits received on the incoming links are simply forwarded on the outgoing links without any additional processing. This approach requires the optimization of standard flow variables that define the allocation of fronthaul capacity to the different bit streams. The problem is formulated and addressed via the MM algorithm in [22].

IN-NETWORK PROCESSING

Routing may be highly inefficient in the presence of a dense deployment of RUs. In fact, in this case, an RU may be connected to a large number of nearby RUs, all of which receive correlated baseband signals. In this case, it is wasteful of the fronthaul capacity to merely forward all the bit streams received from the connected RUs. Instead, it is possible to combine the correlated baseband signals at the RU to reduce redundancy. We refer to this processing of incoming signals as *in-network processing*.

To allow for in-network processing, the RU at hand must first decompress the received bit streams from the connected RUs to recover the baseband signals. The decompressed baseband signals are then linearly processed, along with the IQ signal received locally by the RU. After in-network processing, the obtained signals must be recompressed before they can be sent on the outgoing fronthaul links. The effect of the resulting quantization noise must thus be counterbalanced by the advantages of in-network processing to make the strategy preferable to routing. The optimal design of in-network processing is addressed in [22] using the MM algorithm.

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 5 shows the average sum-rate versus the number N of RUs in layer 1 with $N_U = 4$ UE and average received per-antenna SNR of 20 dB at all RUs. It is observed that the performance gain of in-network processing over routing becomes more pronounced as the number N of RUs in the first layer increases. This suggests that, as the density of the RUs' deployment increases, it is desirable for each RU in layer 2 to perform in-network processing of the signals received from layer 1.

DOWNLINK

SYSTEM MODEL

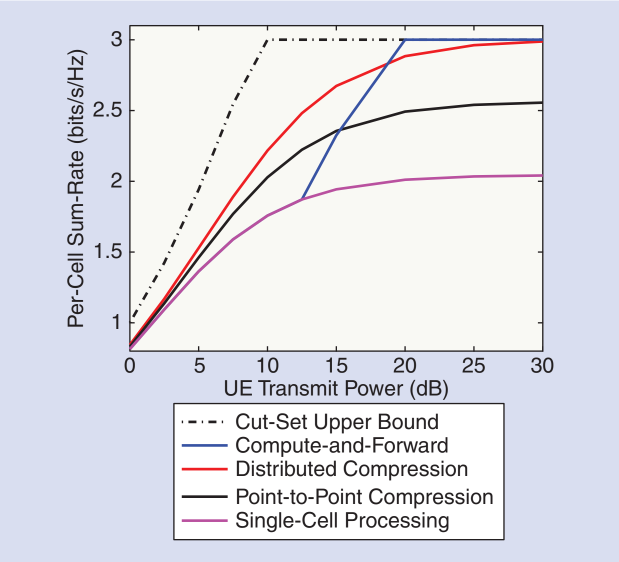
In the downlink, the CU that manages a given cluster processes the information messages of the UE within the cluster by performing channel coding and precoding on behalf of the RUs. As seen in Figure 6, the precoded baseband signals are then compressed by the CU, which finally forwards the compressed 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. Note that we concentrate here on a single-hop fronthaul topology. The multihop case can be addressed following the analysis for the uplink, but this is not further detailed here and is left as an interesting future work.

Similar to the uplink, assuming flat-fading channels, each UE k in the cluster under study receives a discrete-time baseband signal given as

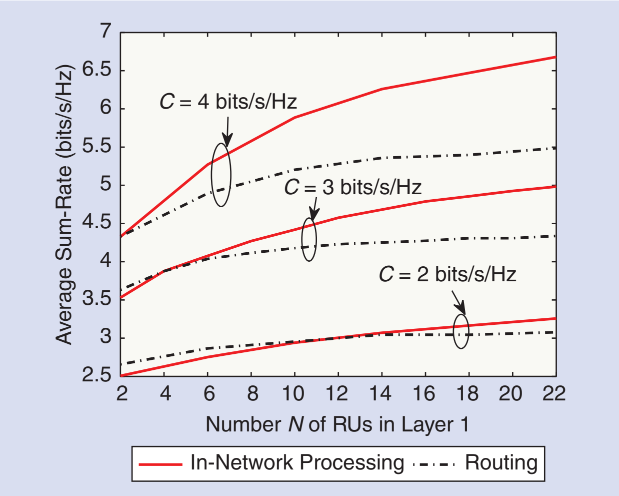
$$\mathbf{y}_k^{\mathrm{dl}} = \mathbf{H}_k^{\mathrm{dl}}\mathbf{x}^{\mathrm{dl}} + \mathbf{z}_k^{\mathrm{dl}}, \qquad (8)$$

where $\mathbf{x}^{\mathrm{dl}}$ is the aggregate baseband signal vector transmitted by all the RUs in the cluster; the additive noise $\mathbf{z}_k^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \Omega_{\mathbf{z}_k^{\mathrm{dl}}})$ accounts for thermal noise and interference from the other clusters; and the matrix $\mathbf{H}_k^{\mathrm{dl}}$ denotes the channel response matrix from all the RUs in the cluster toward UE k.

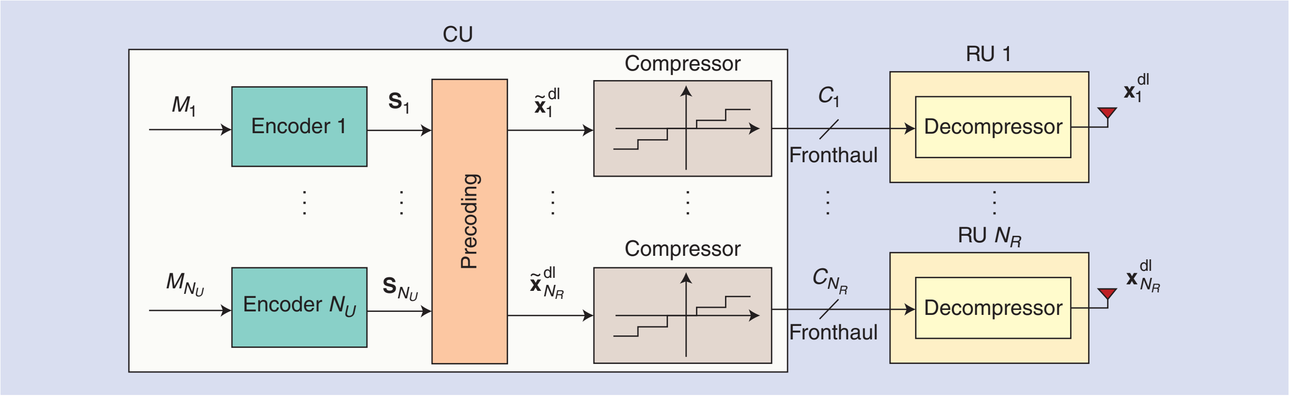
As mentioned, the transmitted signals $\mathbf{x}^{\mathrm{dl}}$ are quantized versions of the baseband signals produced by the CU that manages



[FIG4] 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 $g = 0.4$.



[FIG5] Average uplink sum-rate versus the number of RUs in layer 1 with $N_U = 4$ UEs, average received per-antenna SNR of 20 dB and fronthaul capacity of 2–4 bits/s/Hz.



[FIG6] Point-to-point fronthaul compression for the downlink of C-RANs.

the cluster. As shown in Figure 6, to obtain x^{dl}, the CU first performs channel encoding separately for different UE. This produces the IQ samples $s=[s_1;\ldots;s_{N_U}]$, with s_k representing the signal intended for UE k. The CU then performs linear precoding of the channel-encoded baseband signals s. We observe that non-linear precoding via "dirty-paper" coding can also be considered with minor modifications. The precoded IQ signals $\tilde{x}^{dl}$ produced by the CU can be written as

$$\tilde{x}^{dl}=[\tilde{x}_1^{dl};\ldots;\tilde{x}_{N_R}^{dl}]=As, \tag{9}$$

where $\tilde{x}_i^{dl}$ is the precoded signal intended for transmission by RU i and $A=[A_1,\ldots,A_{N_U}]$ is the precoding matrix with the submatrix A_k multiplied to the signal s_k. The compression of the signals $\tilde{x}_i^{dl}$, with $i=1,\ldots,N_R$ to produce x^{dl} is discussed next.

POINT-TO-POINT FRONTHAUL COMPRESSION

Similar to the uplink, in the conventional C-RAN implementation, the CU compresses separately the precoded IQ signals $\tilde{x}_i^{dl}$ intended for transmission by different RUs i using point-to-point compression, as shown in Figure 6. The index describing the compressed signal x_i^{dl} is sent to the ith RU via the corresponding fronthaul link of capacity C_i. Using compression with a Gaussian test channel, the compressed signal x_i^{dl} is given as

$$x_i^{dl}=\tilde{x}_i^{dl}+q_i^{dl}, \tag{10}$$

where the compression noise q_i^{dl} is distributed as $q_i^{dl}\sim\mathcal{CN}(\mathbf{0},\Omega_i^{dl})$. The quantization noises are independent across the RU index i due to the separate compression of the RUs' IQ signals.

Using the information theoretic results reviewed in the previous section, the quantization error matrix Ω_i^{dl} can be realized if the fronthaul link capacity C_i satisfies the inequality

$$I(\tilde{x}_i^{dl};x_i^{dl})\le C_i. \tag{11}$$

Moreover, assuming that each kth UE treats the signals intended for other UE as noise, the information rate

$$R_k^{dl}=I(s_k;y_k^{dl}) \tag{12}$$

can be achieved for UE k. The optimization of the weighted-sum-rate $R_{\text{sum}}^{dl}=\sum_{k=1}^{N_U}w_kR_k^{dl}$ subject to per-RU power constraints and to the constraints (11), for $i=1,\ldots,N_R$, with respect to the variables A and Ω_i^{dl} for $i=1,\ldots,N_R$ was tackled in [23, Sec. V-C] by using the MM algorithm.

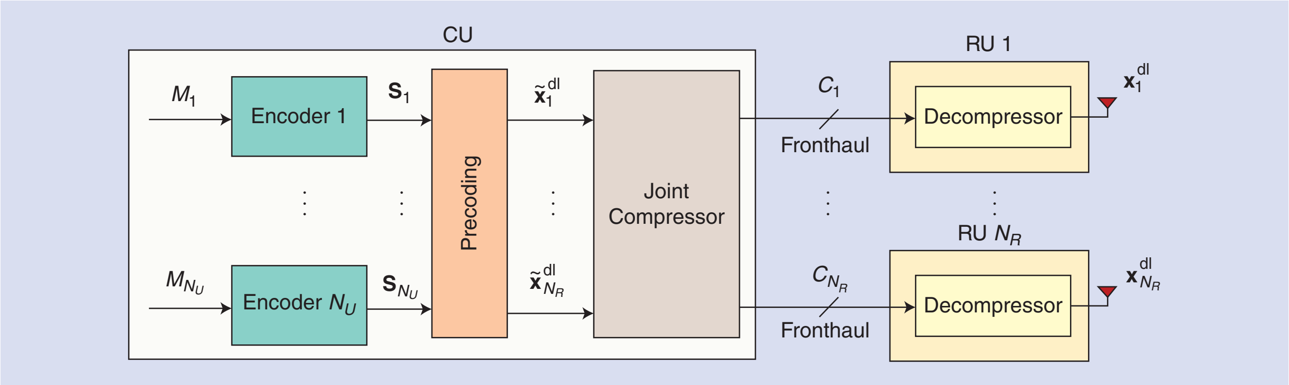
MULTIVARIATE FRONTHAUL COMPRESSION

We now investigate possible improvements to point-to-point compression based on multiterminal compression principles. Our starting observation is that point-to-point compression yields quantization errors that are independent across the RUs. In contrast, multivariate compression [9, Ch. 7] allows correlated quantization noises to be produced, at the expense of a joint, rather than separate, compression of the baseband signals $\tilde{x}_i^{dl}$ for $i=1,\ldots,N_R$ at the CU.

MULTIVARIATE COMPRESSION

The block diagram of the CU and RUs in a cluster operating with multivariate fronthaul compression is shown in Figure 7. As for the conventional point-to-point case of Figure 6, the CU performs channel encoding separately for each UE and applies precoding, hence obtaining the baseband signals $\tilde{x}_i^{dl}$ in (9) for $i=1,\ldots,N_R$. However, unlike point-to-point compression, the signals $\tilde{x}_i^{dl}$ are jointly compressed to select the quantized signals x_i^{dl} from the corresponding quantization codebooks $i=1,\ldots,N_R$.

Before providing some details on multivariate compression, we observe that correlating the quantization noises can be beneficial to control the effect of the additive quantization noises on the reception of the UE. To see this, assume that the quantization noise vector $q^{dl}=[q_1^{dl};\ldots;q_{N_R}^{dl}]$ in (10) are distributed as $\mathcal{CN}(\mathbf{0},\Omega^{dl})$, where the covariance matrix Ω^{dl} is a block matrix whose (i,j)th block $\Omega_{i,j}^{dl}=\mathbb{E}[q_i^{dl}q_j^{dl\dagger}]$ defines the correlation between the quantization noises of RU i and RU j. By using (8) and (10), the effective noise at the kth UE is given by $z_k^{dl}+H_k^{dl}q^{dl}$. The covariance matrix of the effective noise is then given as $\Omega_{z_k^{dl}}+H_k^{dl}\Omega^{dl}H_k^{dl}$, and can hence be controlled by designing the quantization error covariance matrix Ω^{dl}. As a result, one can reduce the impact of the effective noise on the reception of the useful signal and enhance the achievable rates (12).



[FIG7] Multiterminal fronthaul compression for the downlink of C-RANs.

With reference to vector quantization concepts, one can think of the matrix Ω^{dl} as defining the shape of the quantization regions in the space of the baseband signals of all RUs. Specifically, while point-to-point compression leads to regions that are merely the Cartesian product of the quantization regions of the separate quantizers, multivariate compression allows for more general shapes.

KEY INFORMATION-THEORETIC RESULTS

The multivariate compression lemma in [9, Ch. 9] provides sufficient conditions on the fronthaul capacities under which a given joint quantization error matrix Ω^{dl} can be realized. It is recalled that, if the error matrix Ω^{dl} is block diagonal, i.e., if the submatrices $\Omega_{i,j}^{\mathrm{dl}}$ have all zero entries for $i \neq j$, then the conditions at hand reduce to (11) for $i = 1, \ldots, N_R$. Instead, for a general covariance matrix Ω^{dl}, the multivariate compression lemma requires that the following inequality

$$\sum_{i \in \mathcal{S}} h(\mathrm{x}_i^{\mathrm{dl}}) - h(\mathrm{x}_{\mathcal{S}}^{\mathrm{dl}} \,|\, \tilde{\mathrm{x}}^{\mathrm{dl}}) \leq \sum_{i \in \mathcal{S}} C_i \qquad (13)$$

be satisfied for all subsets $\mathcal{S} \subseteq \{1, \ldots, N_R\}$. Using standard properties of the mutual information, it can be seen that if Ω^{dl} is block diagonal, then the system of conditions (13) for all subsets $\mathcal{S}$ is equivalent to the system of inequalities (11) for $i = 1, \ldots, N_R$. Otherwise, the inequalities (13) provide more stringent constraints on the fronthaul capacities than (11). The optimization over the precoding matrix A and the compression noise covariance Ω^{dl} was tackled by using the MM algorithm in [23].

REMARK 6

Similar to Figure 3 for the uplink, multivariate compression can be implemented using a sequential architecture, whereby the baseband signals of different RUs are sequentially, rather than jointly, compressed [23, Sec. IV-D].

COMPUTE-AND-FORWARD

Similar to the "Uplink" section, we now observe that the schemes discussed so far for the downlink employ quantization codebooks that are designed separately from the channel codebooks used for encoding the messages of the UE. An alternative approach, which is dual to the one studied for the uplink, leverages instead the same (nested) lattice code for both channel coding and quantization.

Specifically, according to the approach introduced in [24], the CU employs the same lattice code to perform channel encoding for all UE. Then, it performs precoding using only integer (modulo-)sum operations. In this fashion, the resulting precoded baseband signals are still codewords of the same lattice code. Finally, the CU transmits on the fronthaul links directly the index of the obtained precoded codewords.

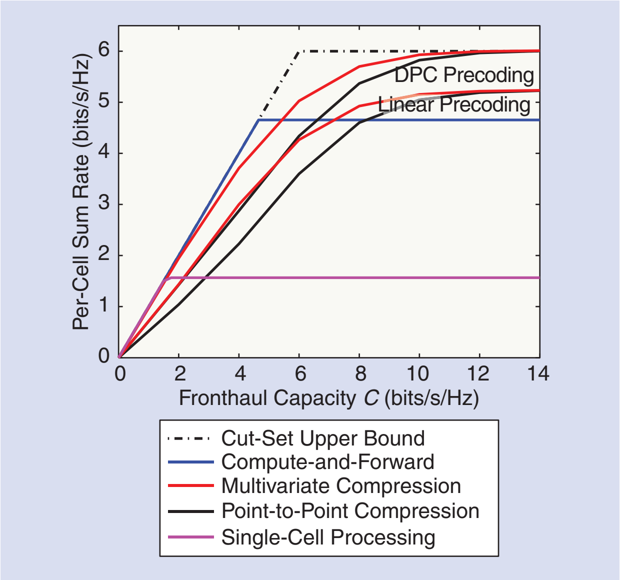
The scheme at hand has similar advantages and disadvantages as compared to its counterpart for the uplink. Specifically, while not adding any quantization noise, it is limited by the integrality constraints on the coefficients of the precoding matrix.

NUMERICAL EXAMPLE

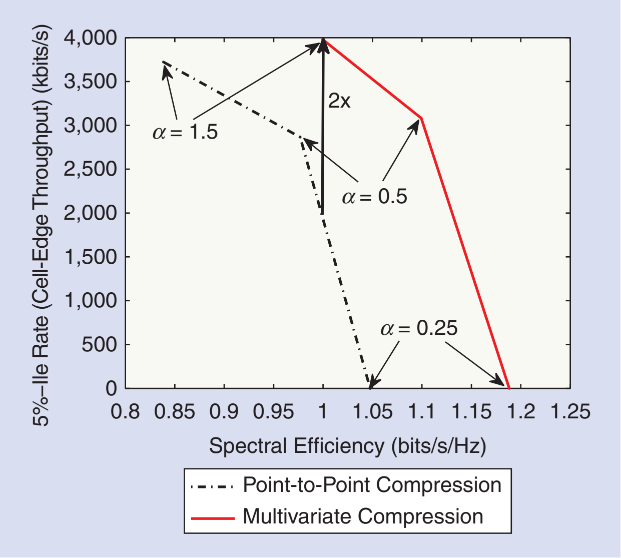
We consider here the same three-cell circulant Wyner model used in Figure 4 for the uplink, where the intercell channel gain is equal to $g = 0.5$, and every RU uses the same transmit power of 20 dB and has the same fronthaul capacity C. Figure 8 shows the per-cell sum-rate of point-to-point compression and multivariate compression, as applied to both linear precoding and "dirty paper" nonlinear precoding [25], and also of compute-and-forward. For reference, we also show the cut-set upper bound and the performance with single-cell processing, whereby each RU transmits only the signal of the in-cell UE. 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.

PERFORMANCE EVALUATION

This section provides a performance evaluation of the discussed fronthaul compression techniques using the standard cellular topology and channel models of [26]. We focus on the performance of the macrocell located at the center of a



[FIG8] 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 $g = 0.5$.



[FIG9] 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, $(C_{\text{macro}}, C_{\text{pico}}) = (3, 1)$ bits/s/Hz, $T = 5$ and $\beta = 0.5$.

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. All interference signals from other macrocells are treated as independent noise signals. The system parameters are as indicated in [26]. We focus here on the downlink, but comparable results were observed also for the uplink [27].

We adopt the conventional metric of cell-edge throughput versus the average per-UE spectral efficiency (see, e.g., [8, Fig. 5]). This is obtained by running a proportional fairness scheduler on a sequence of T time-slots with independent fading realizations, and by then evaluating the cell-edge throughput as the fifth percentile rate and the average spectral efficiency as the average sum-rate normalized by the number of UE. We recall that the proportional fairness scheduler maximizes at each time-slot the weighted-sum-rate $R_{\text{sum}}^{\text{fair}} = \sum_{k=1}^{N_U} R_k^{\text{dl}}/\bar{R}_k^{\alpha}$, with $\alpha \geq 0$ being a fairness constant, R_k^{dl} in (12), and $\bar{R}_k$ being the average data rate accrued by UE k so far. After each time-slot, the rate $\bar{R}_k$ is updated as $\bar{R}_k \leftarrow \beta\bar{R}_k + (1-\beta)R_k^{\text{dl}}$ where $\beta \in [0,1]$ is a forgetting factor. Increasing α leads to a more fair rate allocation among the UE.

Figure 9 plots the cell-edge throughput versus the average spectral efficiency for $K = 4$ UE, $(C_{\text{macro}}, C_{\text{pico}}) = (3,1)$ bits/s/Hz, $T = 5$ and $\beta = 0.5$. The curve is obtained by varying the fairness constant α in the utility function $R_{\text{sum}}^{\text{fair}}$. 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 2 × gain in terms of cell-edge throughput for a spectral efficiency of 1 bits/s/Hz.

CONCLUSIONS AND OUTLOOK

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. Among the many open issues, we mention here the investigation of structured coding schemes that are robust to nonintegrality limitations (see [11] and [24]); the performance analysis for limited frame lengths; the optimal allocation of layer-1 functionalities between the RUs and the CU [12]; the study of the impact of the fronthaul latency on higher-layer performance metrics; and the analysis of the interplay of the considered techniques with multiuser scheduling and limited-feedback channel state information.

AUTHORS

Seok-Hwan Park (seokhwan81@gmail.com) received the B.Sc. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea in 2005 and 2011, respectively. From 2012 to 2014, he was a postdoctoral research associate with the Center for Wireless Communication and Signal Processing Research, New Jersey Institute of Technology, Newark. Since March 2014, he has been with Samsung Electronics, Suwon, South Korea, as a senior engineer. His research interests include communication, information, and optimization theories with applications to various multiple-input, multiple-output wireless systems. He received the Best Paper Award at the 2006 Asia-Pacific Conference on Communications and an Excellent Paper Award at the IEEE Student Paper Contest in 2006.

Osvaldo Simeone (osvaldo.simeone@njit.edu) received the M.Sc. degree (with honors) and the Ph.D. degree in information engineering from Politecnico di Milano, Italy, in 2001 and 2005, respectively. He is currently with the Center for Wireless Communications and Signal Processing Research, New Jersey Institute of Technology, Newark, where he is an associate professor. His research interests include wireless communications, information theory, data compression, and machine learning. He is a corecipient of Best Paper Awards of the 2007 IEEE International Workshop on Signal Processing Advances in Wireless Communications and 2007 IEEE Conference on Wireless Rural and Emergency Communications. He is an editor of *IEEE Transactions on Information Theory*.

Onur Sahin (Onur.Sahin@interdigital.com) received the B.S. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2003 and the M.Sc. and Ph.D. degrees in electrical engineering from the Polytechnic Institute of New York University, Brooklyn, in 2005 and 2009, respectively. Since November 2009, he has been with the Advance Air Interface Group, InterDigital Inc., Melville, New York, as a staff engineer. He conducts research on the cross-layer design of the next-generation cellular and Wi-Fi systems such as ultradense networks, long-term evolution advanced, and beyond. He is the author of more than 20 publications and 12 international patent applications on next-generation wireless communication techniques and system design. His recent research interest includes small cell networks with the utilization of millimeter-wave (60 GHz) link technology to achieve multi-Gigabit/s throughput experience. He received the 2012 InterDigital Innovation Award.

Shlomo Shamai (Shitz) (sshlomo@ee.technion.ac.il) is with the Department of Electrical Engineering, Technion, Israel Institute of Technology, where he is now the William Fondiller Technion Distinguished Professor. He is an IEEE Fellow, a member of the Israeli Academy of Sciences and Humanities, and a foreign member of the U.S. National Academy of Engineering. He received the 2011 Claude E. Shannon Award and the 2014 Rothschild Prize in Mathematics/Computer Sciences and Engineering. He received the URSI 1999 van der Pol Gold Medal, the 2000 IEEE Donald G. Fink Prize Paper Award, the 2003 and 2004 Joint IEEE Information Theory Society and IEEE Communications Society Paper Award, the 2007 IEEE Information Theory Society Paper Award, the 2009 European Commission FP7, the 2014 EURASIP Best Paper Award, and the 2010 Thomson Reuters International Excellence in Scientific Research Award. He served on the executive editorial board of *IEEE Transactions on Information Theory.*

REFERENCES

[1] J. Segel and M. Weldon, "Lightradio portfolio—Technical overview," Technology White Paper 1, Alcatel-Lucent.

[2] China Mobile, "C-RAN: The road towards green RAN," White Paper, ver. 2.5, China Mobile Research Institute, Oct. 2011.

[3] Ericsson AB, Huawei Technologies, NEC Corporation, Alcatel Lucent and Nokia Siemens Networks, "Common public radio interface (CPRI); interface specification," CPRI specification v5.0, Sept. 2011.

[4] Z. Ding and H. V. Poor, "The use of spatially random base stations in cloud radio access networks," *IEEE Signal Processing Lett.*, vol. 20, no. 11, pp. 1138–1141, Nov. 2013.

[5] G. J. Foschini, K. Karakayali, and R. Valenzuela, "Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency," *IEE Proc. Commun.*, vol. 153, no. 4, pp. 548–555, Aug. 2006.

[6] O. Simeone, N. Levy, A. Sanderovich, O. Somekh, B. M. Zaidel, H. V. Poor, and S. Shamai (Shitz), "Cooperative wireless cellular systems: an information-theoretic view," *Found. Trends Commun. Inform. Theory*, vol. 8, no. 1–2, pp. 1–177, 2012.

[7] Integrated Device Technology, Inc., "Front-haul compression for emerging C-RAN and small cell networks," Apr. 2013.

[8] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[9] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[10] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Mag.,* vol. 21, no. 5, pp. 80–94, Sept. 2004.

[11] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inform. Theory,* vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[12] C. J. Bernardos, A. De Domenice, J. Ortin, P. Rost, and D. Wubben, "Challenges of designing jointly the backhaul and radio access network in a cloud-based mobile network," in *Proc. Future Networks Mobile Summit 2013*, Lisbon, Portugal, pp. 1–10.

[13] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

[14] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, 2004.

[15] J. Hoydis, M. Kobayashi, and M. Debbah, "Optimal channel training in uplink network MIMO systems," *IEEE Trans. Signal Processing,* vol. 59, no. 6, pp. 2824–2833, June 2011.

[16] J. Kang, O. Simeone, J. Kang, and S. Shamai (Shitz), "Joint signal and channel state information compression for the backhaul of uplink network MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1555–1567, Mar. 2014.

[17] A. Sanderovich, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. Inform. Theory*, vol. 55, no. 8, pp. 3457–3478, Aug. 2009.

[18] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[19] L. Zhou and W. Yu, "Uplink multicell processing with limited backhaul via per-base-station successive interference cancellation," *IEEE J. Select. Areas Commun.*, vol. 31, no. 10, pp. 1981–1993, Oct. 2013.

[20] A. del Coso and S. Simoens, "Distributed compression for MIMO coordinated networks with a backhaul constraint," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4698–4709, Sept. 2009.

[21] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed Karhunen-Loeve transform," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5177–5196, Dec. 2006.

[22] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Multi-hop backhaul compression for the uplink of cloud radio access networks," arXiv:1312.7135. [Online]. Available: http://arxiv.org/abs/1312.7135

[23] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Processing*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[24] S.-N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 5227–5243, Sept. 2013.

[25] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Downlink multicell processing with limited backhaul capacity," *EURASIP J. Adv. Signal Process.,* vol. 2009, 2009.

[26] 3GPP, TR 136.931 ver. 9.0.0 Rel. 9, May 2011.

[27] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Performance evaluation of multiterminal backhaul compression for cloud radio access networks," in *Proc. CISS*, Princeton, NJ, Mar. 19–21, 2014.

[SP]