Cloud Radio Access Networks

Principles, Technologies, and Applications

Edited by
Tony Q. S. Quek, Mugen Peng, Osvaldo Simone, and Wei Yu

List of contributors

Wei Yu
Electrical and Computer Engineering Department
University of Toronto

Pratik Patil
Electrical and Computer Engineering Department
University of Toronto

Binbin Dai
Electrical and Computer Engineering Department
University of Toronto

Yuhan Zhou
Electrical and Computer Engineering Department
University of Toronto

1 Cooperative Beamforming and Resource Optimization in C-RAN

Cloud radio access network (C-RAN) architecture offers two key advantages as compared to traditional radio access network (RAN) from physical-layer transmission point of view. First, the centralization and virtualization of RAN allow coordination of base-stations (BSs) across a large geographic area, thereby enabling coordinated physical-layer resource allocation across the BSs. The physical-layer resources here refer to frequency, time, and spatial dimensions that can be utilized by radio transmission. Second and more importantly, the C-RAN architecture also opens up the possibility of joint transmission and joint reception of user signals across multiple BSs, thereby fundamentally addressing the issue of inter-cell interference. As interference is the main bottleneck in modern densely deployed wireless networks, the C-RAN architecture offers significant advantage in that it provides the possibility of interference mitigation leading to performance enhancement without the need for additional site and bandwidth acquisition.

This chapter provides an optimization framework for cooperative beamforming and resource allocation in C-RANs. The chapter begins by identifying frequency, time, and spatial resources in wireless cellular networks, and defining the overall spectrum allocation, scheduling, and beamforming problem in a cooperative network. This chapter then provides a network model for the C-RAN architecture, and illustrates typical network objective functions and constraints for network utility maximization. A key characteristic of the C-RAN architecture is that the fronthaul connections between the cloud and the BSs may have limited capacities. One of the main goals of this chapter is to illustrate the impact of limited fronthaul capacity on the cooperative beamforming and resource allocation in C-RANs.

The chapter explores the optimization of design variables associated with C-RANs, depending on the transmission strategies at the cooperative BSs. For the uplink C-RAN, we illustrate compress-forward as the main strategy at the BSs, and focus on the impact of the choice of quantization noise levels at the BSs and possible joint transmit optimization strategies. For the downlink C-RAN, we compare the compression-based strategy and the data-sharing strategy, and illustrate the problem formulation and solution strategy in both cases. Throughout the chapter, key optimization techniques for solving resource allocations problems in C-RANs are presented.

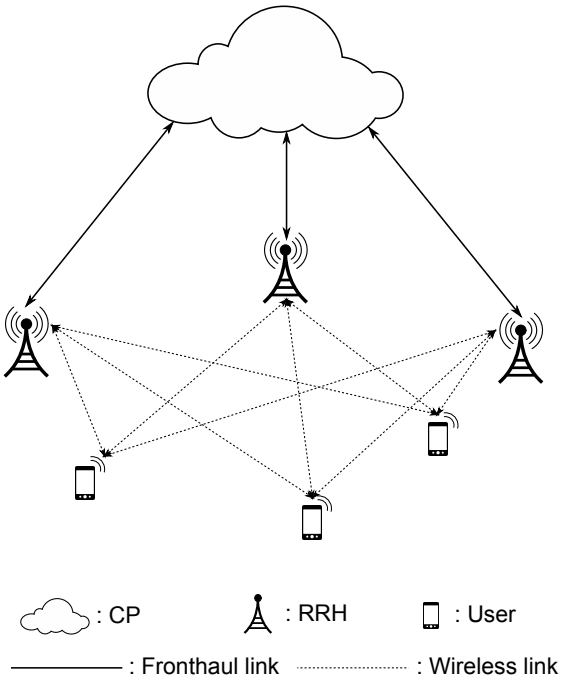

Figure 1.1 C-RAN system model.

1.1 C-RAN Model

In the C-RAN architecture, the baseband processing, traditionally performed locally at each BS, is aggregated and performed centrally at a cloud computing center. This is enabled by high-speed connections, referred to as *fronthaul* links, between the BSs and the cloud. Such centralized signal processing allows for the possibility of interference cancellation and interference pre-compensation across all the users in the uplink and downlink, respectively. The C-RAN architecture thus facilitates the implementation of network multiple-input multiple-output (network MIMO) [12], also known as coordinated multi-point (CoMP) or multi-cell processing (MCP) in the literature [6, 26]. The main focus of this chapter is on the interference mitigation capability enabled by C-RAN architecture. Toward this end, we abstract a physical-layer channel model in order to allow an information-theoretic understanding of the capacity limits of the C-RAN model as compared to traditional RAN.

1.1.1 System Model

To highlight key benefits of the C-RAN architecture, we focus on the network topology of one central processor (CP) coordinating a cluster of BSs serving users over a certain geographic area as illustrated in Fig. 1.1. The BSs in the C-RAN

architecture are also termed remote radio heads (RRHs) as their functionality is often restricted to transmission and reception of radio signals. These RRHs are managed by the cloud-computing based CP that communicates with RRHs via fronthaul links. The fronthaul connections can be dedicated fiber optic cables, or they can be wireless links. Although analog transport is a possible option, this chapter models the fronthaul links as finite-capacity noiseless digital links. Our aim is to understand the impact of limited fronthaul capacity on the overall system performance, and subsequently to design efficient transmission and relaying strategies that account for the limited available fronthaul capacity.

As a concrete setup, this chapter considers a C-RAN model consisting of a CP coordinating a total of L RRHs each equipped with M antennas serving K users each equipped with N antennas. The analysis developed in this chapter can be easily extended to the case with unequal number of antennas at different terminals. The main resources in the system are the fronthaul link capacities, and the power budgets at the users and at the RRHs. We denote the capacity of the fronthaul link connecting the RRH l to the CP by C_l. The power spectrum density constraint at the user k in the uplink is denoted as P_k^{ul}, and at the RRH l in the downlink as P_l^{dl}. The precise uplink and downlink channel models are abstracted out in the next section for an information-theoretic study of the C-RAN architecture.

To enable signal level cooperation for joint signal processing, it is crucial to be able to precisely synchronize the signals of different users. In this chapter, perfect synchronization among the RRHs in the downlink and among the users in the uplink is assumed. The impact of synchronization error in the context of uplink C-RAN is considered in [7]. In addition, instantaneous and perfect channel state information (CSI) is assumed to be available to all the RRHs and the users, and also at the CP. In practice, the amount of CSI available is limited by the coherence time of the channel and the overhead of communicating CSI to the CP. The effect of partial CSI and channel estimation errors are taken in account in [22]. The main focus on this chapter is to illustrate different fundamental transmission strategies in C-RAN and their interference mitigation capabilities.

1.1.2 Information Theoretical Model

From an information theoretical point of view, the C-RAN model is best understood as a relay network. The RRHs can be thought of as relays that facilitate the communication between the CP and the mobile users. In the uplink, multiple users communicate with the CP through the RRHs, and thus can be modeled as an instance of a multiple-access relay channel. In the downlink, the CP communicates with multiple users through RRHs. The downlink C-RAN can thus be modeled as an instance of a broadcast relay channel. We assume frequency-flat channels for now. The difference with the frequency-selective channels is discussed later in the section.

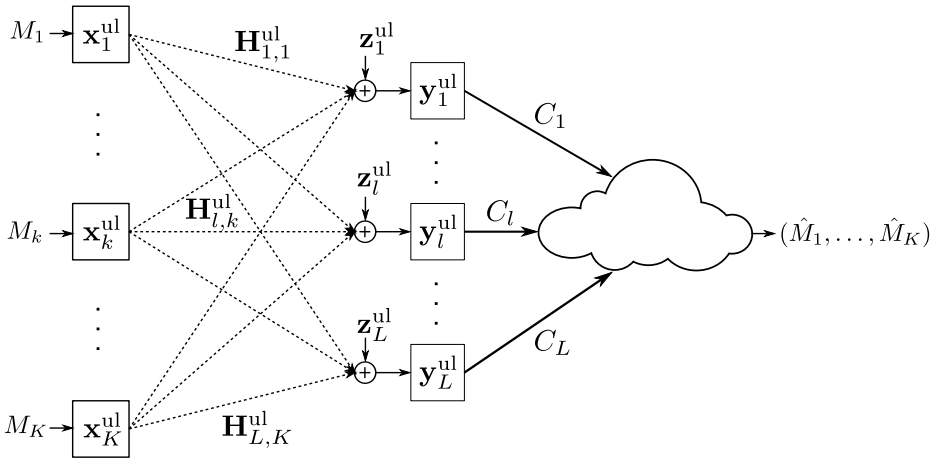

Figure 1.2 Information-theoretic uplink C-RAN model.

Uplink C-RAN

Let $\mathbf{x}_k^{\text{ul}} \in \mathbb{C}^{N\times 1}$ be the transmit signal from user k, and $\mathbf{y}_l^{\text{ul}} \in \mathbb{C}^{M\times 1}$ be the received radio signal at RRH l. Assuming additive Gaussian noises at the RRH receivers, the channel response at RRH l can be modeled as:

$$\mathbf{y}_l^{\text{ul}} = \sum_k \mathbf{H}_{l,k}^{\text{ul}} \mathbf{x}_k^{\text{ul}} + \mathbf{z}_l^{\text{ul}}, \tag{1.1}$$

where $\mathbf{H}_{l,k}^{\text{ul}} \in \mathbb{C}^{M\times N}$ is the channel from user k to RRH l, and $\mathbf{z}_l^{\text{ul}} \in \mathbb{C}^{M\times 1} \sim \mathcal{CN}(\mathbf{0}, \sigma_{\text{ul}}^2\mathbf{I})$ is the additive Gaussian noise. Fig. 1.2 illustrates the uplink system model.

In traditional RAN, after receiving the radio signals, each BS independently decodes the messages of its scheduled users, treating the combined signal from all other users as interference. In the C-RAN architecture, however, the RRHs have the flexibility to relay some information about its observed signals to the CP, which can then jointly process the information from all the RRHs for decoding. Joint processing has the advantage that the effect of inter-user interference can be mitigated. There are various different possible relaying strategies, depending on the information the RRHs relay to the CP and the eventual decoding strategy. These strategies are discussed in detail Section 1.2.

Downlink C-RAN

Let $\mathbf{x}_l^{\text{dl}} \in \mathbb{C}^{M\times 1}$ be the transmitted signal from RRH l. Assuming additive Gaussian noise, the received signal at user k, $\mathbf{y}_k^{\text{dl}} \in \mathbb{C}^{N\times 1}$, is represented as:

$$\mathbf{y}_k^{\text{dl}} = \sum_l \mathbf{H}_{k,l}^{\text{dl}} \mathbf{x}_l^{\text{dl}} + \mathbf{z}_k^{\text{dl}}, \tag{1.2}$$

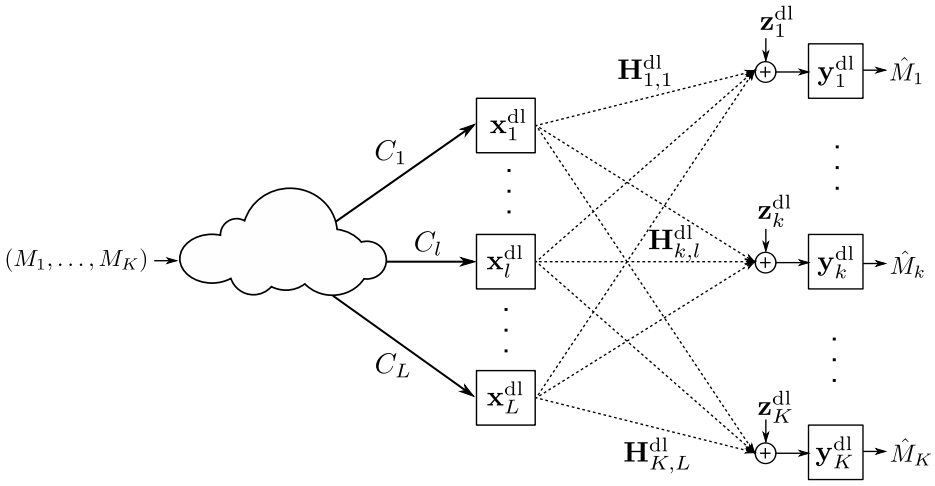

Figure 1.3 Information-theoretic downlink C-RAN model.

where $\mathbf{H}_{k,l}^{\mathrm{dl}} \in \mathbb{C}^{N \times M}$ is the channel from RRH l to user k, and $\mathbf{z}_k^{\mathrm{dl}} \in \mathbb{C}^{N \times 1} \sim \mathcal{CN}(\mathbf{0}, \sigma_{\mathrm{dl}}^2 \mathbf{I})$ is the additive Gaussian noise. Fig. 1.3 illustrates the downlink system model.

In traditional RAN, the user messages from the core network are sent directly to the BSs, which independently encode the messages for users within the cells. As consequence, the transmit signals from the neighboring BSs interfere with each other. In contrast, in C-RAN architecture, the fact that the CP has access to the messages of all the users enables joint encoding across the cooperating cluster, thereby allowing inter-cell interference to be mitigated. Depending on the specific ways that the CP utilizes the capacity-limited fronthaul to enable joint encoding, different downlink strategies are possible. These strategies are discussed in more detail in Section 1.3.

1.1.3 Achievable Rate Region

The different transmission, relaying, and decoding strategies for both uplink and downlink result in different achievable rate-tuples for the users. As multiple users share radio resources, an increase in user rate for one user usually comes at the cost of rates of other users. The concept of rate region captures this tradeoff. The rate region is the set of all the achievable user rates, $\mathcal{R} = \{R_1, \ldots, R_K\}$, for a particular channel model. Given a transmission and relaying strategy in C-RAN, the rate region $\mathcal{R}$ is a function of the underlying channels $\mathbf{H}_{l,k}^{\mathrm{ul}}$, the fronthaul capacities C_l, and the power constraints P_k^{ul} in the uplink, and similarly in the downlink. In allocating these resources to different users, a desirable operating point is to be chosen depending on the overall system objective. With that in mind, the overall goal of this chapter is to provide an optimization framework to maximize certain system objective under rate regions for different strategies,

and subsequently to point out the overall design insight obtained from such resource allocation perspective. Towards this end, we describe below the widely used system objective based on network utility considered in this chapter.

1.1.4 Network Utility Maximization

Network utility maximization is an optimization framework that takes into account the physical layer tradeoffs in terms of the rate region as well as the application layer tradeoffs in terms of varying usefulness of rates for different users (e.g., the value of additional rate increase for video application might be very different from file transfer). In the network utility maximization framework, each user has an utility $U_k(\bar{R}_k)$ as a function of its average user rate $\bar{R}_k$ that captures the value of having such a rate for user k. Most common utility functions are concave increasing functions. The overall network utility maximization is the problem of maximizing the sum utility over all the users in the system over operating parameters such as scheduling, beamforming, and quantization.

More specifically, the network utility maximization problem considered in this chapter aims to solve the following problem in every time slot:

$$\text{maximize} \quad \sum_{k=1}^{K} U_k(\bar{R}_k) \tag{1.3a}$$

$$\text{subject to} \quad (R_1, \ldots, R_K) \in \mathcal{R}, \tag{1.3b}$$

where the objective function above depends on the average user rates $\bar{R}_k$, while the optimization parameters affect the instantaneous rate R_k. The average rate is usually computed in a windowed fashion. For example, with exponential weighting the average rate is obtained as

$$\bar{R}_k^{\text{updated}} = (1-\alpha)\bar{R}_k^{\text{prior}} + \alpha R_k, \tag{1.4}$$

where $\bar{R}_k^{\text{prior}}$ is the average rate prior to the present time slot, and R_k is the instantaneous rate of the current time slot. The above optimization problem is repeatedly solved for each time slot under the rate-region constraint on the instantaneous rates, until the average user rates eventually converge.

A common user utility function U_k is the logarithm function, i.e., $U_k(\bar{R}_k) = \log(\bar{R}_k)$. Under the log utility and exponentially weighted rate averaging, and assuming that the new contribution of the instantaneous rate αR_k is small, the optimization of the network utility objective function can be solved approximately by a maximization of the instantaneous *weighted* sum rate, where the weights are inverses of average rates, as follows:

$$\text{maximize} \quad \sum_{k=1}^{K} w_k R_k \tag{1.5a}$$

$$\text{subject to} \quad (R_1, \ldots, R_K) \in \mathcal{R}, \tag{1.5b}$$

where $w_k = \frac{1}{\bar{R}_k^{\text{prior}}}$. The above weighted sum-rate maximization problem is solved

for each time slot over the transmission strategies with weights updated after each iteration. This transmit optimization problem under the logarithmic utility is known as the proportionally fair resource allocation problem. The rest of this chapter focuses on this weighted rate-sum maximization problem for C-RAN.

We remark that the log-utility is not the only possible choice of utility function. For delay sensitive applications, it is often desirable to maximize the minimum rate, or to guarantee a minimum rate while maximizing the sum rate. Different choices of utility functions would lead to different optimization formulations.

1.1.5 Resource Allocation Problem

The resource allocation problem for C-RAN consists of solving the above optimization problem over the operating parameters and under the system constraints. The operating parameters to be optimized can include not only cellular transmission parameters such as scheduling (i.e., which users to assign non-zero rate), beamforming, bandwidth and power allocation, but also relay strategies such as quantization levels in the context of C-RAN. The system constraints are the fronthaul link capacities, and the transmit power spectral density constraints at the users for the uplink and at the RRHs for the downlink.

1.1.6 Disjoint versus User-Centric Clustering

While defining the system model for C-RAN, we have implicitly assumed that RRHs are clustered into disjoint clusters, and RRHs within each cluster cooperatively serve the users in the cluster. Such model has explicit cluster boundaries, and the users near the cluster boundaries still suffer from inter-cluster interference. One way to further reduce inter-cluster interference is to let each user form a user-centric cluster of RRHs. Different clusters for different users may overlap in this case, and there are no explicit cluster boundaries. Such user-centric clustering typically improves the fairness in the system.

1.1.7 Frequency-Selective Channels

The chapter mainly considers frequency-flat channel model. But wireless channels are often frequency selective. In this case, one can employ orthogonal frequency division multiplex (OFDM) to divide the total bandwidth into a number of flat subchannels. Then each subchannel can independently employ the relay strategies for the frequency-flat channel model considered in this chapter.

The OFDMA-based C-RAN presents an additional dimension for resource allocation, namely among the frequency subchannels. This includes the assignment of the subchannels to the different users, and the allocation of fronthaul capacities as well as transmit power among the different subchannels. Some initial work on resource allocation for C-RAN employing OFDM has been carried out in [16] under certain simplifying modeling assumptions.

1.2 Uplink C-RAN

The ability to manage interference is one of the main benefits of C-RAN. In the uplink, different users in the cluster communicate their messages to the CP through RRHs. The RRHs, instead of decoding the messages locally, can relay information about their observations to the CP for centralized processing. This enables the interference mitigation capability for the uplink C-RAN.

In the ideal case where the fronthaul links have infinite capacities, the RRHs can convey its exact observations to the CP. The resulting channel model reduces to a MIMO multiple-access channel. Practical systems, however, have capacity-constrained fronthaul links. This limits the amount of information that the RRHs can relay. A key question is then to decide what information about the observed signals is the most useful at the CP so as to enable as much interference cancellation as possible.

This section discusses different strategies for relaying and centralized processing in the uplink C-RAN, then formulates their respective resource optimization problems, and indicates methods to solve these problems. We provide key insights obtained from such optimization throughout the chapter.

1.2.1 Compress, Decode, vs. Compute-Forward

From the perspective of maximizing the overall capacity of the network, the aim of the RRH should be to preserve as much information as possible in relaying its observation to the CP under the finite fronthaul capacity constraint. A natural strategy is for the RRHs to describe the observed signals by compressing the received analog signals, and relaying their digital representations to the CP [21, 25, 29]. The resolution of compression determines the amount of fronthaul capacity needed. Higher fronthaul capacity leads to lower quantization noise, which in turn leads to higher achievable user rates. At the CP, the user messages can be jointly decoded based on the compression indices received from all the RRHs in the cluster. Such joint processing at the CP enables effective interference cancellation. This relaying strategy is known as *compress-forward* in the literature. Note that the compress-forward strategy also inevitably forwards some part of the receiver noise at the RRHs to the CP.

There are different ways of performing compression and decompression, depending on whether some side information is utilized in the compression process, leading to either independent or Wyner-Ziv compression strategies. There are also different ways of performing decoding at the CP, depending on how the user messages and the compression codewords are decoded successively. These possibilities are discussed in detail in the next section. We mention here that in theory, there is also the possibility of performing decompression and message decoding at the CP jointly [6]. Doing so is in fact information theoretically more justified, but it also has very high complexity and is impractical to implement. For this reason, this chapter restricts to successive decoding type of strategies.

As an alternative to compress-forward, some of the RRHs can attempt to decode messages of users closest to them, and relay the messages themselves (rather than the compressed version of their observations) to the CP. The users being decoded at the RRHs cannot benefit from the joint processing capabilities of C-RAN, but these decoded messages can help the decoding of other users at the CP. This type of relaying strategy can be broadly referred to as a version of *decode-forward*.

Finally, the RRHs may opt to decode some linear combination of user messages, or more generally some function of user messages, and forward it to the CP. This is called the *compute-forward* strategy [17]. In compute-forward, the users choose the transmit codewords from a structured lattice codebook. The benefit of using structured codebook is that linear integer combinations of different codewords are still codewords. After receiving the signals, the RRHs compute functions of the user codewords from the received signal. Typically, functions that closely mimic the channel output at the RRHs are the ones that give the best computation rate. The indices corresponding to the function values are sent over the fronthaul links. After receiving all such function values, the CP inverts the functions to recover the original user messages.

The main advantage of decode-forward and compute-forward is that they eliminate noises at the RRHs. But in practice, there are only limited number of scenarios in which they outperform compress-forward. Further, compute-forward is quite sensitive to channel estimation error [18]. With this in mind, this chapter mostly focuses on the compress-forward strategy. We refer the reader to [9] for details regarding the achievable rate region and network optimization for the compute-forward strategy in the context of uplink C-RAN.

The use of compress-forward for C-RAN can also be justified from information theoretic consideration. For a Gaussian multi-message multicast network, it can be shown that compress-forward (and its variations called quantize-map-forward [1] and noisy network coding [15]) can achieve the information theoretic capacity of the network to within a constant gap, which only depends on the network topology, but is independent of other channel parameters.

The rest of this section focuses on compress-forward as the main relaying strategy for uplink C-RAN, and discusses different variants and their corresponding achievable rates and resource optimization.

1.2.2 Compress-Forward Strategy

In the compress-forward strategy, the received signals $\mathbf{y}_l^{\mathrm{ul}}$ are compressed at the RRHs, and the compression indices are sent to the CP. The CP then decodes the original user signals $\mathbf{x}_k^{\mathrm{ul}}$ from these indices.

There are different ways of performing compression at the RRHs and different ways of decoding the user messages at the CP, leading to different variations of the compress-forward strategy. The two main compression methods are *independent compression* and *Wyner-Ziv compression*. In independent compression,

the observations at the RRHs are compressed and decompressed independently. In Wyner-Ziv compression, it is possible to take advantage of the fact that the observed signals at the RRHs are correlated in order to reduce the amount of fronthaul capacity needed.

The processing at the CP can also take different forms. For example, after decoding the compression codewords, the CP may perform linear beamforming across the RRH signals for independent decoding of user messages, or the CP may perform successive interference cancellation (SIC). Alternatively, the CP may even interleave the decoding of user messages and compression codewords, using the decoded user messages as side information in subsequent processing.

To characterize the achievable rates and the required fronthaul capacities for the different compress-forward strategies, we model the user transmission and the compression process below. These models are based on information theoretical considerations; they provide accurate, yet simplified rate expressions for different variants of the compress-forward strategy.

We assume that the input signals $\mathbf{x}_k^{\mathrm{ul}}$ at the users are chosen according to a Gaussian codebook. While the choice of Gaussian-like input is not necessarily optimal for the compress-forward strategy [25], it makes the evaluation of rate region tractable. Let $\mathbf{U}_k \in \mathbb{C}^{N \times d_k}$ denote the transmit beamformer that user k utilizes to transmit the message signal $\mathbf{s}_k^{\mathrm{ul}} \in \mathbb{C}^{d_k \times 1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ to the CP. Here d_k denotes the number of data streams per user k. The transmit signal at user k is then given by $\mathbf{x}_k^{\mathrm{ul}} = \mathbf{U}_k \mathbf{s}_k^{\mathrm{ul}}$ with covariance matrix $\mathbb{E}[\mathbf{x}_k^{\mathrm{ul}} (\mathbf{x}_k^{\mathrm{ul}})^H] = \mathbf{U}_k \mathbf{U}_k^H$. The total transmission power consumed at user k is expressed as $\mathrm{Tr}(\mathbf{U}_k \mathbf{U}_k^H)$. With the linear Gaussian channel model as described earlier in the chapter, the received signal at RRH l in the uplink can thus be expressed as

$$\mathbf{y}_l^{\mathrm{ul}} = \sum_k \mathbf{H}_{l,k}^{\mathrm{ul}} \mathbf{U}_k \mathbf{s}_k^{\mathrm{ul}} + \mathbf{z}_l^{\mathrm{ul}}. \tag{1.6}$$

For the compression process, we again assume a Gaussian codebook. Let $\hat{\mathbf{y}}_l^{\mathrm{ul}}$ denote the compressed signal for RRH l. Then, the quantization process at RRH l is modeled as the addition of independent Gaussian quantization noises as follows:

$$\hat{\mathbf{y}}_l^{\mathrm{ul}} = \mathbf{y}_l^{\mathrm{ul}} + \mathbf{q}_l^{\mathrm{ul}}, \tag{1.7}$$

where $\mathbf{q}_l^{\mathrm{ul}} \in \mathbb{C}^{M \times 1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_l^{\mathrm{ul}})$ and $\mathbf{Q}_l^{\mathrm{ul}}$ is the covariance matrix of the quantization noise in the compressed signal corresponding to the RRH l. We point out that, even though it may seem that a more general linear additive model for compression is to first process the received signal $\mathbf{y}_l^{\mathrm{ul}}$ using a transformation matrix $\mathbf{A}_l$ and then compress the resulting transformed output $\mathbf{A}_l \mathbf{y}_l^{\mathrm{ul}}$ (perhaps even of lower dimension than $\mathbf{y}_l^{\mathrm{ul}}$), with appropriate choice of $\mathbf{Q}_l^{\mathrm{ul}}$, the model in (1.7) can be shown to be equivalent to such a linear model and is therefore without loss of generality.

1.2.3 Compression Strategies

Full benefit of the joint processing, in terms of its interference cancellation capability, would be achieved if each RRH is able to convey the exact $\mathbf{y}_l^{\mathrm{ul}}$ to the CP. In practice, the more accurately the compressed signal $\hat{\mathbf{y}}_l^{\mathrm{ul}}$ resembles the actual received signals $\mathbf{y}_l^{\mathrm{ul}}$ at the RRHs, higher the achievable rate would be for the overall network. There is, however, a cost for transmitting high fidelity version of $\mathbf{y}_l^{\mathrm{ul}}$ through the digital fronthaul link. This cost can be modeled using the information theoretical rate-distortion theory.

The rate-distortion tradeoff can be most easily understood in terms of the quantization noise $\mathbf{q}_l^{\mathrm{ul}}$ introduced in the compression process. On one hand, the quantization noise level directly provides an indication of the accuracy of $\hat{\mathbf{y}}_l^{\mathrm{ul}}$; it enters the achievable rate expression as additional noise introduced by the quantization process. On the other hand, the level of the quantization noise indicates the amount of fronthaul capacity needed for compression. Higher fronthaul capacity leads to better compression resolution and smaller quantization noises. The precise relationship between the fronthaul capacity and the quantization noise can be understood via rate-distortion theory as follows. Consider a single RRH l. In order to keep the statistical variance of the quantization noise to a certain level $\mathbf{Q}_l^{\mathrm{ul}}$, the amount of fronthaul capacity needed must satisfy:

$$C_l^{\mathrm{indep,ul}} \geq I\left(\mathbf{y}_l^{\mathrm{ul}}; \hat{\mathbf{y}}_l^{\mathrm{ul}}\right) \tag{1.8}$$

$$= \log \frac{\left|\sum_{k=1}^{K} \mathbf{H}_{l,k}^{\mathrm{ul}} \mathbf{U}_k \mathbf{U}_k^H (\mathbf{H}_{l,k}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_l^{\mathrm{ul}}\right|}{\left|\mathbf{Q}_l^{\mathrm{ul}}\right|}. \tag{1.9}$$

As expected, the above is a decreasing function of $\mathbf{Q}_l^{\mathrm{ul}}$. The superscript 'indep' refers to the fact that the quantization process is done independently for each RRH without utilizing any potential side information at the CP.

The above fronthaul rate can be improved using a more sophisticated compression technique that utilizes the fact that signals received at different RRHs are often highly correlated as they come from the same set of user messages. Thus, once some of the quantization codewords are decoded, they can serve as side information in subsequent decoding of other quantization codewords. As result, the fronthaul capacity needed for compression can be reduced. This compression technique is referred to as Wyner-Ziv compression. Assuming that the compressed signals from RRHs are recovered in the order of 1 to L, the fronthaul capacity required for Wyner-Ziv compression for compressing received signal at RRH l is given as follows:

$$C_l^{\mathrm{WZ,ul}} \geq I\left(\mathbf{y}_l^{\mathrm{ul}}; \hat{\mathbf{y}}_l^{\mathrm{ul}} | \hat{\mathbf{y}}_1^{\mathrm{ul}}, \ldots, \hat{\mathbf{y}}_{l-1}^{\mathrm{ul}}\right) \tag{1.10}$$

$$= \log \frac{\left|\sum_{k=1}^{K} \mathbf{H}_{1:l,k}^{\mathrm{ul}} \mathbf{U}_k \mathbf{U}_k^H (\mathbf{H}_{1:l,k}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{1:l} + \mathbf{Q}_{1:l}^{\mathrm{ul}}\right|}{\left|\sum_{k=1}^{K} \mathbf{H}_{1:l-1,k}^{\mathrm{ul}} \mathbf{U}_k \mathbf{U}_k^H (\mathbf{H}_{1:l-1,k}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{1:l-1} + \mathbf{Q}_{1:l-1}^{\mathrm{ul}}\right|} - \log |\mathbf{Q}_l^{\mathrm{ul}}|. \tag{1.11}$$

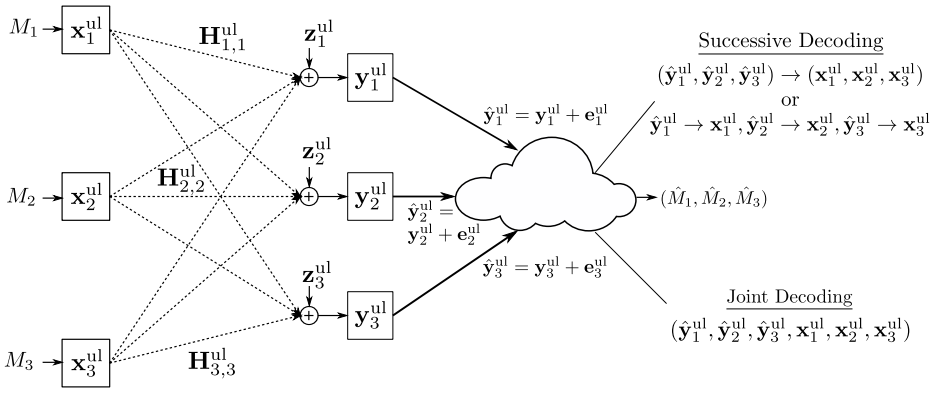

Figure 1.4 Illustration of compress-forward strategies for uplink C-RAN.

Here and throughout the rest of this section, the notation $\mathbf{H}_{\mathcal{S},\mathcal{T}}^{\text{ul}}$ donates the channel matrix from the users in the set $\mathcal{T}$ to the RRHs in the set $\mathcal{S}$, $\mathbf{Q}_{\mathcal{S}}^{\text{ul}}$ denotes the the block diagonal matrix formed with the quantization covariance matrices of the RRHs belonging to the set $\mathcal{S}$, and $1:l$ denotes the set $\{1,\ldots,l\}$. In the mutual information expression above, because the signals already recovered at the CP $\hat{\mathbf{y}}_1^{\text{ul}},\ldots,\hat{\mathbf{y}}_{l-1}^{\text{ul}}$ can serve as side information, they can be included in the conditioning in order to reduce the fronthaul rate for the compression at RRH l. We remark that the above compression rates are the information theoretical limit for compression with side information. Practical implementation of Wyner-Ziv compression is not trivial.

We further remark that, in the case where some of the user messages are decoded before the compressed signals for some other subset of RHH signals are recovered, we can include the decoded user messages as side information in the decompression process as well in order to further lower the fronthaul capacity requirements for these RRHs.

1.2.4 Decoding Strategies

The goal of the CP in compress-forward in the uplink C-RAN is to decode the user messages based on the compression indices sent from RRHs. The CP has various options for decoding user messages. The CP can choose to first recover all the compressed signals at the RRHs, then subsequently decode the user messages based on the compressed versions of the received signals. Alternatively, the CP can arbitrarily interleave the decoding of the message messages and the compression codewords. Doing so can benefit the users decoded later in the process at the expense of earlier users. The achievable rates of these various options are discussed in detail in this section.

The CP can first recover the compressed signals $\hat{\mathbf{y}}_l^{\text{ul}}$ from all the RRHs, then use these compressed signals to decode the user messages, which are encoded in

$\mathbf{x}_k^{\mathrm{ul}}$. Such a successive decoding strategy essentially converts the uplink C-RAN setup into a virtual multiple-access model with the CP receiving $(\hat{\mathbf{y}}_1^{\mathrm{ul}}, \ldots, \hat{\mathbf{y}}_L^{\mathrm{ul}})$ for decoding the user messages. The achievable rate region of this successive decoding strategy thus resembles the rate region of a multiple access channel with additional quantization noises.

For example, with all the compression codewords $\hat{\mathbf{y}}_l^{\mathrm{ul}}$ already decoded, the decoding of $\hat{\mathbf{x}}_k^{\mathrm{ul}}$ can be done independently for each user, resulting in the following achievable rate region:

$$R_k^{\mathrm{linear,ul}} \leq I\left(\mathbf{x}_k^{\mathrm{ul}}; \hat{\mathbf{y}}_1^{\mathrm{ul}}, \ldots, \hat{\mathbf{y}}_L^{\mathrm{ul}}\right) \tag{1.12}$$

$$= \log \frac{\left|\sum_{j=1}^{K} \mathbf{H}_{1:L,j}^{\mathrm{ul}} \mathbf{U}_j \mathbf{U}_j^H (\mathbf{H}_{1:L,j}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_{1:L}^{\mathrm{ul}}\right|}{\left|\sum_{j\neq k} \mathbf{H}_{1:L,j}^{\mathrm{ul}} \mathbf{U}_j \mathbf{U}_j^H (\mathbf{H}_{1:L,j}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_{1:L}^{\mathrm{ul}}\right|} \tag{1.13}$$

In writing down the above achievable rate region, we have implicitly assumed that a linear minimum-mean-squared-error (MMSE) network-wide beamforming is performed across the signals received from RRHs. The above rate region therefore already includes the capability of inter-RRH interference cancellation to certain extend.

The above rate region can be improved if SIC is implemented across the users. In particular, assuming that user messages are decoded in the order 1 to K, the SIC achievable rate for user k can be written as:

$$R_k^{\mathrm{SIC,ul}} \leq I\left(\mathbf{x}_k^{\mathrm{ul}}; \hat{\mathbf{y}}_1^{\mathrm{ul}}, \ldots, \hat{\mathbf{y}}_L^{\mathrm{ul}} | \mathbf{x}_1^{\mathrm{ul}}, \ldots, \mathbf{x}_{k-1}^{\mathrm{ul}}\right) \tag{1.14}$$

$$= \log \frac{\left|\sum_{j=k}^{K} \mathbf{H}_{1:L,j}^{\mathrm{ul}} \mathbf{U}_j \mathbf{U}_j^H (\mathbf{H}_{1:L,j}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_{1:L}^{\mathrm{ul}}\right|}{\left|\sum_{j>k} \mathbf{H}_{1:L,j}^{\mathrm{ul}} \mathbf{U}_j \mathbf{U}_j^H (\mathbf{H}_{1:L,j}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_{1:L}^{\mathrm{ul}}\right|} \tag{1.15}$$

Note that the achievable rate above reduces to the successive decoding rate region of a multiple-access channel, if the quantization noises are ignored.

Alternatively, instead of recovering all the compressed signals before decoding any user messages, the decoding can also be done on a per-RRH basis [28]. More specifically, once the compressed signals from RRH l, $\hat{\mathbf{y}}_l^{\mathrm{ul}}$, are recovered, the messages of the users associated with that RRH can be decoded immediately. Such decoding resembles the traditional per-BS decoding, except that since the decoding of all users is done centrally at the CP, previously decoded user messages can serve as side information in subsequent decoding, so their interference can be subtracted. Assuming $K = L$ and that user k is associated with RRH k, the achievable rate for user k in this case can be written as:

$$R_k^{\mathrm{perRRH,ul}} \leq I\left(\mathbf{x}_k^{\mathrm{ul}}; \hat{\mathbf{y}}_k^{\mathrm{ul}} | \mathbf{x}_1^{\mathrm{ul}}, \ldots, \mathbf{x}_{k-1}^{\mathrm{ul}}\right) \tag{1.16}$$

$$= \log \frac{\left|\sum_{j=k}^{K} \mathbf{H}_{k,j}^{\mathrm{ul}} \mathbf{U}_j \mathbf{U}_j^H (\mathbf{H}_{k,j}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_k^{\mathrm{ul}}\right|}{\left|\sum_{j>k} \mathbf{H}_{k,j}^{\mathrm{ul}} \mathbf{U}_j \mathbf{U}_j^H (\mathbf{H}_{k,j}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_k^{\mathrm{ul}}\right|} \tag{1.17}$$

Note that the above rate expression for per-RRH decoding can be further im-

proved by including the compressed signals of the RRHs recovered before the RRH k in the conditioning in the mutual information expression. Moreover, the Wyner-Ziv compression rate (1.11) can also benefit from the conditioning of the already decoded users signals before user k in per-RRH decoding.

We remark that the main benefit of C-RAN, namely inter-RRH interference mitigation, is achieved in the uplink via two mechanism: either through beamforming, i.e., the decoding of user message based on the received signals across multiple RRHs, or through SIC, i.e., the previously decoded user messages serve as side information for subsequent decoding, or both. In general, the benefit of network-wide beamforming is more important than successive decoding alone as in per-RRH SIC. This is because per-RRH SIC necessarily requires some of the users to be decoded first; these users therefore cannot benefit from centralized processing. The largest achievable rates are obtained if both beamforming and SIC are implemented. With this in mind, the rest of this section focuses on the achievable rates involving network-wide beamforming, i.e., either $R_k^{\text{linear,ul}}$ in (1.13) or $R_k^{\text{SIC,ul}}$ in (1.15).

1.2.5 Optimization Framework for Compress-Forward

Within the framework of network utility maximization, the optimization of the compress-forward strategy for uplink C-RAN is essentially a problem of solving a weighted sum rate maximization problem (1.5a) over the transmission and relaying strategies. The underlying optimization variables are the user scheduling, user transmit power and beamformers, and the quantization codebook — constrained by the input power and fronthaul capacity constraints.

User scheduling is usually determined by network layer protocols as function of user priorities, traffic delay constraints, and also physical layer channel conditions. While in theory user scheduling should be included in the weighted sum rate maximization, doing so rigorously is often difficult, especially when there are a large number of potential users in the system. In practice, it is often desirable to use heuristics that combine user traffic demand with channel strength considerations to schedule users. For example, users with longer queues of data to transmit should be scheduled first; users with stronger channels should be given priority; grouping users with near orthogonal channels to the cluster of RRHs is a sensible strategy.

When successive decoding of the user signals and the compressed signals (in the case of SIC and Wyner-Ziv compression, respectively) are implemented, decoding orders are additional variables to be optimized. Exhaustive searches for a C-RAN cluster of K users and L RRHs would involve $K!$ user orderings and $L!$ RRH orderings, respectively, and are clearly impractical, but sensible heuristic strategies often exist. For example, for SIC, users with strong channels should usually be decoded first in order to help the weak users and to improve fairness. For maximizing weighted sum rate, the SIC user decoding order normally should be chosen to be in the ascending order of the user priority weights. Likewise, good

heuristic ordering for Wyner-Ziv compression is also possible. For example, [30] proposes to decompress first the signals from those RRHs with either the higher value of the fronthaul capacity or the lower value of the average received signal power. The rationale here is that the already decompressed signals can serve as side information for subsequent decompression, so this ordering helps balance the effective quantization noise levels across the RRHs.

To simplify the problem, we now fix the set of users to be scheduled, and fix the orders in which decoding is performed. Without loss of generality, assume that the user signals are decoded in an order from 1 to K. Similarly, in the case of Wyner-Ziv compression, assume that the signals from RRHs are decompressed in an order from 1 to L. In this case, the joint transmitter and quantization noise covariance optimization problem can be formulated as follows:

$$\begin{aligned}
\underset{\mathbf{U}_k, \mathbf{Q}_l^{\mathrm{ul}}}{\text{maximize}} \quad & \sum_{k=1}^{K} w_k R_k^{\mathrm{ul}} && \text{(1.18a)}\\
\text{subject to} \quad & R_k^{\mathrm{ul}} = (1.13) \quad \text{or} \quad R_k^{\mathrm{ul}} = (1.15), \quad \forall k, && \text{(1.18b)}\\
& (1.9) \quad \text{or} \quad (1.11) \quad \leq C_l, \quad \forall l, && \text{(1.18c)}\\
& \mathrm{Tr}\left(\mathbf{U}_k \mathbf{U}_k^H\right) \leq P_k^{\mathrm{ul}}, \quad \forall k, && \text{(1.18d)}\\
& \mathbf{Q}_l^{\mathrm{ul}} \succeq \mathbf{0}, \quad \forall l, && \text{(1.18e)}
\end{aligned}$$

where w_k's are the priority weights in the weighted sum-rate maximization framework. The optimization has two sets of design variables, the transmit beamformer for user k, $\mathbf{U}_k$, which is constrained by the power budget, and the quantization covariance matrix for RRH l, $\mathbf{Q}_l^{\mathrm{ul}}$, which is constrained by the fronthaul capacity. This optimization problem is nonconvex; it is in general challenging to find its global optimum solution.

In formulating the above optimization problem, we have implicitly assumed that both the transmit strategy at the user side and the compression process at RRHs can be done adaptively, in the sense that the users can adaptively choose their transmit power level, beamformers, and rate, and the RRHs can adaptively choose different quantization codebooks, according to the network condition. While transmit optimization is invariably included in modern cellular network, adaptive quantization may not be. The analysis below discusses the issue of adaptive quantization noise optimization first, followed by transmit beamforming and power optimization.

1.2.6 Optimization of Quantization at RRHs

In this section, we analyze the quantization noise optimization component of (1.18). To illustrate the key ideas, we first consider one instance of the optimization problem (1.18) with independent compression and successive decoding of user messages ordered according to the user priority weights (i.e., we assume $w_1 \leq \cdots \leq w_K$). Similar analysis can be obtained under Wyner-Ziv coding and

with linear MMSE beamforming. Denote $\mathbf{\Sigma}_k = \mathbf{U}_k\mathbf{U}_k^H$ as the transmit signal covariance matrix for user k. The weighted sum rate maximization problem thus becomes:

$$\underset{\mathbf{\Sigma}_k, \mathbf{Q}_l^{\mathrm{ul}}}{\text{maximize}} \quad \sum_{k=1}^{K} w_k \log \frac{\left|\sum_{k=1}^{K} \mathbf{H}_{1:L,k}^{\mathrm{ul}} \mathbf{\Sigma}_k (\mathbf{H}_{1:L,k}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_{1:L}^{\mathrm{ul}}\right|}{\left|\sum_{j>k} \mathbf{H}_{1:L,j}^{\mathrm{ul}} \mathbf{\Sigma}_j (\mathbf{H}_{1:L,j}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_{1:L}^{\mathrm{ul}}\right|} \tag{1.19a}$$

$$\text{subject to} \quad \log \frac{\left|\sum_{k=1}^{K} \mathbf{H}_{l,k}^{\mathrm{ul}} \mathbf{\Sigma}_k (\mathbf{H}_{l,k}^{\mathrm{ul}})^H + \sigma_{\mathrm{ul}}^2 \mathbf{I} + \mathbf{Q}_l^{\mathrm{ul}}\right|}{\left|\mathbf{Q}_l^{\mathrm{ul}}\right|} \leq C_l, \quad \forall l, \tag{1.19b}$$

$$\mathrm{Tr}\left(\mathbf{\Sigma}_k\right) \leq P_k^{\mathrm{ul}}, \quad \forall k, \tag{1.19c}$$

$$\mathbf{Q}_l^{\mathrm{ul}} \succeq \mathbf{0}, \quad \forall l. \tag{1.19d}$$

First focus on the optimization over $\mathbf{Q}_l^{\mathrm{ul}}$ with fixed $\mathbf{\Sigma}_k$. The main difficulty in solving the above optimization problem stems from the fact that the objective function is not a concave function and the fronthaul capacity constraints are not convex functions of $\mathbf{Q}_l^{\mathrm{ul}}$. A method based on successive convex approximation (SCA) is proposed in [30] to solve this problem. The basic idea behind SCA is to first approximate the original problem into a convex program by linearizing the nonconvex parts in the objective function and the constraints at a suitable starting point. Then after solving the convex program, a new convex approximation is made around the updated solution from the previous iteration. This procedure is iterated until convergence and can be shown to reach the local optimum of the original optimization problem.

The optimal solution $\mathbf{Q}_l^{\mathrm{ul}}$ obtained from the procedure above is a set of positive semi-definite matrices. These optimized quantization noise covariance matrices can be implemented using an architecture where the received vector signal at the RRH is first beamformed, followed by compression across the components of the resulting signal. Assuming the eigenvalue decomposition of $\mathbf{Q}_l^{\mathrm{ul}} = \mathbf{A}_l^H \mathbf{\Lambda}_l \mathbf{A}_l$, where $\mathbf{A}_l$ is a unitary matrix and $\mathbf{\Lambda}_l$ is a diagonal matrix, the quantization process with $\mathbf{Q}_l^{\mathrm{ul}}$ is equivalent to first beamforming $\mathbf{y}_l^{\mathrm{ul}}$ with $\mathbf{A}_l$, then performing compression across each element of the newly beamformed vector $\mathbf{A}_l\mathbf{y}_l^{\mathrm{ul}}$. The diagonal entries in $\mathbf{\Lambda}_l$ represent the quantization noise levels in each of the resulting components. If some of these noise levels in the optimal $\mathbf{\Lambda}_l$ are nearly infinite, this implies that those corresponding components are not useful for decoding at the CP, in which case the effective beamforming matrix essentially projects the received signal at the RRH into a lower dimensional space.

We remark that the optimized beamformers $\mathbf{A}_l$ and the quantization noise levels $\mathbf{\Lambda}_l$ depend on the channels $\mathbf{H}_{l,k}^{\mathrm{ul}}$ and the transmit beamformers $\mathbf{U}_k$, which often change as the user scheduling, user locations, etc., change. To implement jointly optimized transmission and quantization therefore requires an adaptive compression architecture at RRHs that dynamically adapts to the changing transmission and channel parameters. There is, however, a special case where such adaptive design is not necessary. Under a high signal-to-quantization-noise-ratio condition and assuming that as many users as total number of RRH an-

tennas are scheduled, uniform quantization noise levels across the antennas, i.e., setting $\mathbf{Q}_l^{\text{ul}} = \gamma_l \mathbf{I}$, can be shown to be a reasonable strategy for maximizing the sum rate [30], where the proportionality constant γ_l is chosen to satisfy the fronthaul capacity constraint at RRH l. Thus, under this special condition, adaptive quantization at RRHs is not needed; independent quantization on a per-antenna basis is already an approximately optimal design.

1.2.7 Fronthaul-Aware Transmit Beamforming

We now address the optimization of transmit beamforming in fronthaul capacity-limited uplink C-RAN. Consider again the optimization problem (1.19), but over the transmit covariance matrices $\mathbf{\Sigma}_k$. If we assume that the quantization noise covariance matrices $\mathbf{Q}_l^{\text{ul}}$ are fixed, then the maximization of the weighted sum-rate subject to the input power constraints resembles a conventional MIMO multiple-access channel input optimization problem, but with additional quantization noise $\mathbf{Q}_l^{\text{ul}}$.

The optimization problem (1.19) assumes that SIC is implemented. The objective function in this case is concave in the transmit covariance matrices $\mathbf{\Sigma}_k$, and the problem can be solved using efficient convex optimization methods. When linear MMSE receive beamforming is implemented, the optimization problem is nonconvex, but a class of algorithms known as weighted minimum mean-square error (WMMSE) algorithms [27] are well suited for this scenario. The WMMSE algorithm is capable of reaching a locally optimal transmit beamforming solution for the problem.

The above discussion assumes that the quantization noise covariance matrices $\mathbf{Q}_l^{\text{ul}}$ are fixed. In the general case, where the transmit covariance matrices $\mathbf{\Sigma}_k$ and the quantization noise covariance matrices $\mathbf{Q}_l^{\text{ul}}$ are optimized jointly, a method called WMMSE-SCA, which incorporates SCA into the WMMSE algorithm, can be used to arrive at a stationary point of the weighted sum rate maximization problem [30].

We conclude this section by pointing out the importance of being fronthaul aware when designing transmit beamformers, particularly for the heterogeneous C-RAN architecture, where the fronthaul capacities of different RRHs can be quite different. Transmit beamforming serves to steer the radio transmission in certain spatial directions. Intuitively, if certain RRHs have more limited fronthaul capacities than others, the beamformers should steer away from them and instead point toward RRHs with higher fronthaul capacities.

In fact, as the joint optimization framework of the transmit covariance and quantization noise covariance matrices for the uplink C-RAN model shows, for optimized performance, the transmit beamformers should adapt to the quantization noise levels, and conversely the quantization noise levels should also adapt to the transmit beamforming.

1.3 Downlink C-RAN

In the downlink C-RAN, messages intended for different users in the cluster originate from the CP. Since the CP has access to all the user data, it can send useful information about the user messages to multiple RRHs in order to facilitate cooperation among different RRHs so as to minimize the unwanted interference seen by the users.

In the ideal case with infinite fronthaul capacities, the data of all the users in the entire cluster can be provided to all the RRHs. This reduces the downlink model to a MIMO broadcast channel with distributed antennas. However, the practical case with finite fronthaul capacities allows for limited information transfer. As in the uplink, a key question is to decide the most useful information about the user messages to be sent to the RRHs in order to enable as much interference pre-subtraction as possible.

This section discusses various relaying strategies that utilize the limited fronthaul capacities in different manners for the downlink C-RAN, along with their corresponding optimization frameworks and methods for finding the solutions. We conclude by providing design insights learned from such optimization.

1.3.1 Data-Sharing, Compression, vs. Reverse Compute-Forward

In the downlink, the benefit of the C-RAN architecture arises from the ability to cooperatively transmit signals from RRHs to minimize the effect of unwanted interference at users. Cooperative transmission from multiple RRHs takes the form of network-wide beamforming. A straightforward way for the CP to enable cooperation is to simply share each user message with multiple RRHs, which can then form cooperative cluster to serve the users. Ideally, to enable full cooperation, message of each user needs to be shared with all the RRHs in the entire network. However, such full cooperation may not be feasible due to the corresponding fronthaul capacity constraints. One way to reduce the fronthaul consumption is to share each user's message with only a subset of RRHs which then locally form beamformed signals to serve the users. This strategy is termed the *data-sharing* strategy.

Another way to achieve cooperation is to centrally compute the beamformed signals to be transmitted by the RRHs at the CP. These signals are then compressed and sent to the individual RRH for transmission to the users. Since the CP has the messages of all the users, the signals computed at the CP can mimic full cooperation. However, since these signals are analog, they need to be compressed before they can be sent to the RRHs. This introduces quantization noises that limit the system performance. Such a strategy is termed the *compression-based* strategy in this chapter.

Instead of sharing direct user messages or sharing the beamformed signals, there is also a possibility of sharing some function of user messages to the RRHs. In *reverse compute-forward* strategy [8], linear functions of user signals, cho-

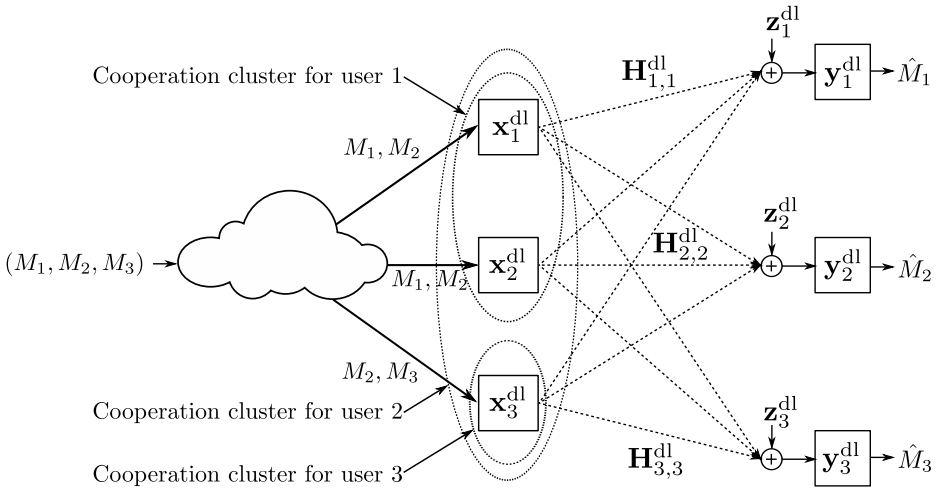

Figure 1.5 Illustration of data-sharing strategy for downlink C-RAN.

sen from a structured lattice codebook, are sent to the RRHs. These functions are computed in a way that after passing through the channels, each user can effectively retrieve its own message.

In the data-sharing strategy, the finite fronthaul capacity limits the size of cooperation cluster; while in the compression-based strategy, the limited fronthaul capacity adds additional quantization noises. Further, as with compute-forward in the uplink, the performance of reverse compute-forward strategy in the downlink is quite sensitive to the channel gains. With this in mind, in this chapter, we focus on data-sharing and compression-based strategies. Readers are referred to [9] for details about optimization in reverse compute-forward strategy.

From an information theoretic perspective, the downlink C-RAN is an instance of the broadcast-relay channel. While it reduces to a broadcast channel if the fronthaul links have infinite capacities, the capacity characterization for the practical case with finite fronthaul capacities is very challenging. Approximate capacity and approximately optimal relaying strategies for the general broadcast-relay network have been studied in [11, 14], but the exact characterization of capacity for the downlink C-RAN remains an open problem.

1.3.2 Data-Sharing Strategy

In traditional RAN, each BS receives raw data for users in its cell, and computes the transmit signal based on that data independently of other BSs. From a user's perspective, it receives useful signal from its serving BS and overhears interference from other nearby BSs. In C-RAN, the fronthaul connections from the CP to RRHs open up the possibility of signal level cooperative transmission. Since the CP has access to the data of all the users in its cluster, a straightforward

way to enable such cooperative transmission is to share data of each user to all the RRHs. This essentially coverts the overall C-RAN downlink setup into a large antenna array with the antennas distributed over the network, or equivalently as a broadcast channel. However, sharing data of each user to all the RRHs requires very high fronthaul capacity. In the more practical case where the fronthaul links have limited capacities, each RRH can only receive data for a subset of users, or equivalently each user gets served by only a subset of RRHs as illustrated in Fig. 1.5. The effect of such limited cooperation is characterized below.

To illustrate the key ideas, we assume Gaussian signaling and use linear beamforming. Let $\mathbf{V}_{k,l} \in \mathbb{C}^{M\times d_k}$ denote the matrix of transmit beamformers that convey d_k data streams from RRH l to user k. The transmit signal from RRH l is given by $\mathbf{x}_l^{\mathrm{dl}} = \sum_k \mathbf{V}_{k,l}\mathbf{s}_k^{\mathrm{dl}}$, where $\mathbf{s}_k^{\mathrm{dl}} \in \mathbb{C}^{d_k\times 1} \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$ is the message signal for user k. The covariance matrix of the signal transmitted by RRH l is given by $\mathbb{E}\left[\mathbf{x}_l^{\mathrm{dl}}(\mathbf{x}_l^{\mathrm{dl}})^H\right] = \sum_k \mathbf{V}_{k,l}\mathbf{V}_{k,l}^H$ with total transmit power $\sum_k \mathrm{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right)$. Note that if user k's message $\mathbf{s}_k$ is not available at RRH l, then the corresponding beamformer $\mathbf{V}_{k,l}$ is zero. Finally, with the linear Gaussian channel model described earlier in the chapter, the received signal at user k can be written as

$$\mathbf{y}_k^{\mathrm{dl}} = \sum_l \mathbf{H}_{k,l}^{\mathrm{dl}}\mathbf{V}_{k,l}\mathbf{s}_k^{\mathrm{dl}} + \sum_l \sum_{j\neq k} \mathbf{H}_{k,l}^{\mathrm{dl}}\mathbf{V}_{j,l}\mathbf{s}_j^{\mathrm{dl}} + \mathbf{z}_k^{\mathrm{dl}}. \tag{1.20}$$

Given (1.20), the achievable rate for user k under data-sharing strategy, treating inference as noise, can be expressed as

$$R_k^{\mathrm{data,dl}} = I(\mathbf{s}_k^{\mathrm{dl}}, \mathbf{y}_k^{\mathrm{dl}}) \tag{1.21}$$

$$= \log \frac{\left|\sum_j \mathbf{H}_k^{\mathrm{dl}}\mathbf{V}_j\mathbf{V}_j^H\left(\mathbf{H}_k^{\mathrm{dl}}\right)^H + \sigma_{\mathrm{dl}}^2\mathbf{I}\right|}{\left|\sum_{j\neq k} \mathbf{H}_k^{\mathrm{dl}}\mathbf{V}_j\mathbf{V}_j^H\left(\mathbf{H}_k^{\mathrm{dl}}\right)^H + \sigma_{\mathrm{dl}}^2\mathbf{I}\right|} \tag{1.22}$$

where $\mathbf{H}_k^{\mathrm{dl}} \in \mathbb{C}^{N\times LM} = \left[\mathbf{H}_{k,1}^{\mathrm{dl}},\ldots,\mathbf{H}_{k,L}^{\mathrm{dl}}\right]$ and $\mathbf{V}_k \in \mathbb{C}^{LM\times d_k} = \left[\mathbf{V}_{k,1}^T,\ldots,\mathbf{V}_{k,L}^T\right]^T$ are the combined channel gains and transmit beamformers from all the RRHs to user k.

To support these user rates, the fronthaul capacity must support the aggregate data of users that each RRH participates in beamforming to. The fronthaul capacity required to send data to RRH l is thus simply the sum of rates of users that are served by RRH l. To write this mathematically, we make use of the fact that the transmit beamformer from RRH l to user k is zero, i.e. $\mathbf{V}_{k,l} = \mathbf{0}$, if RRH does not serve user k, or equivalently $\mathrm{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right) = 0$. Writing it in this way is useful for the optimization of the data-sharing strategy later on. The total fronthaul required for RRH l can now be written as $\sum_k \mathbb{1}\left\{\mathrm{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right)\right\} R_k^{\mathrm{data,dl}}$, where $\mathbb{1}\left\{\mathrm{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right)\right\}$ is the indicating function defined as

$$\mathbb{1}\left\{\mathrm{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right)\right\} = \begin{cases} 0, & \text{if } \mathrm{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right) = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{1.23}$$

It determines whether or not user k's message is revealed to RRH l.

Note that to participate in beamforming to user k, there is also the overhead of transmitting the beamformer coefficients of the user to all the RRHs involved in order for them to combine with the user data. In practice, sending the beamforming coefficients usually requires much less fronthaul capacity than sending user messages, especially in a slow varying environment as beamforming coefficients typically only need to be updated as the user channels vary.

We further remark that the fronthaul consumption model (1.3.2) assumes that all the data streams of user k are either completely available or not at all at RRH l and ignores the possibility that only part of the data stream is revealed to a RRH. If such possibility is considered, then a user may receive different data streams from different serving RRHs and the fronthaul consumption model (1.3.2) needs to be adjusted by using the indicator function and the rate expression for each individual data stream instead.

Finally, we point out that, instead of linear beamforming, a non-linear precoding technique (e.g. dirty paper coding) can also be utilized to improve the achievable user rates. The optimization framework developed in the next section can be easily extended to such case.

1.3.3 Optimization Framework for Data-Sharing

Given (1.22) and (1.3.2), the weighted sum-rate maximization problem for data-sharing strategy can be formulated as

$$\underset{\mathbf{V}_{k,l}}{\text{maximize}} \quad \sum_k w_k R_k^{\text{data,dl}} \tag{1.24a}$$

$$\text{subject to} \quad \sum_k \mathbb{1}\left\{\text{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right)\right\} R_k^{\text{data,dl}} \le C_l, \quad \forall l, \tag{1.24b}$$

$$\sum_k \text{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right) \le P_l^{\text{dl}}, \quad \forall l, \tag{1.24c}$$

where w_k in (1.24a) is the priority weight associated with user k.

The above optimization problem is nonconvex, so finding its globally optimal solution is challenging. One source of nonconvexity arises from the indicator function in (1.24b). One way to tackle this issue is to recast the indicator function into an expression involving an ℓ_0-norm, which can be further approximated as a convex weighted ℓ_1-norm using the compressive sensing idea [2]. Another source of nonconvexity is the rate $R_k^{\text{data,dl}}$ expressed in (1.22). To resolve this difficulty, $R_k^{\text{data,dl}}$ in (1.24b) can be fixed as a constant, then updated iteratively. This turns the fronthaul constraint into a convex constraint for a given iteration. Then, the WMMSE algorithm [3, 27] can be applied to reach a stationary point solution of the beamforming problem. The details of such an approach can be found in [4]. Although this algorithm does not have theoretical convergence proof, it is numerically observed to converge, and performs as well as other algorithms with theoretical convergence guarantees [5].

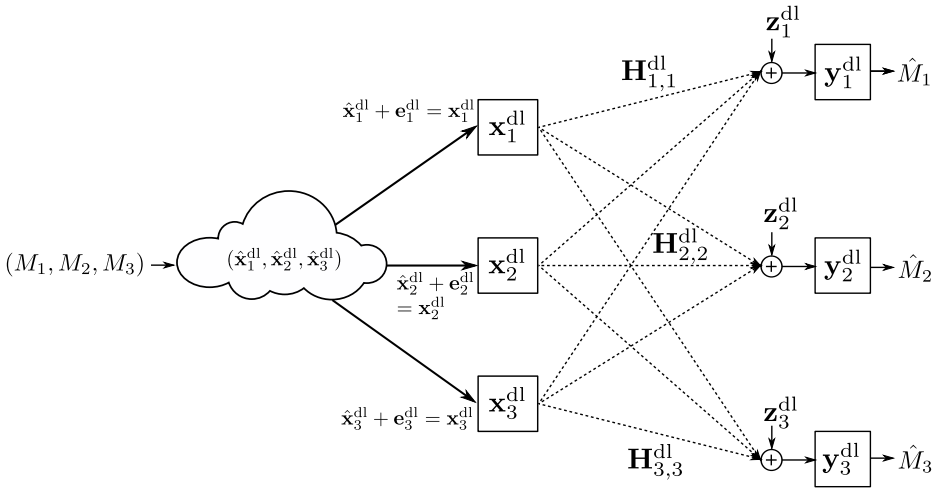

Figure 1.6 Illustration of compression-based strategy for downlink C-RAN.

In the problem formulation (1.24), it is assumed that the RRH cluster for each user can be updated dynamically in each time slot. In the case where the RRH clustering is static and is only updated when the user locations change, the compressive sensing idea can still be applied to address the fronthaul constraints [4]. But in this case, the optimal static RRH clustering design problem needs to be formulated based on loading considerations and is not trivial to solve.

One way to form such static RRH clusters is simply to partition the entire set of RRHs geographically into different groups. RRHs within the same group form a cooperative array of antennas and jointly serve the users that fall in that geographic area [10]. In such a user-RRH association, however, users near the boundary of the partitions still suffer from considerable interference.

In an alternate way, each individual user can decide on a static and fixed set of serving RRHs. The criteria to select the best RRHs need to be based on both the channel strengths as well as the loading at the RRHs. We refer to [4,5,13,19] for details on possible ways to form such static user-centric RRH associations.

1.3.4 Compression-Based Strategy

In the data-sharing strategy, the limited fronthaul capacities restrict the cooperation size of the RRH cluster in serving a user. However, since the all the user data are available at the CP, the CP can centrally compute the beamformed signals that the RRHs should transmit. Such signals computed at the CP can in principle mimic the effect of full cooperation. The downside to such an approach is that the beamformed signals are no longer discrete (unlike the raw data in the data-sharing strategy), but instead are analog in nature. So these signals need to be compressed before they can be sent over the digital fronthaul links of

finite capacities. The process of compression introduces compression noise. The amount of such noise is determined by the available fronthaul capacities. Higher fronthaul capacity leads to finer compression and less quantization noise. Fig. 1.6 illustrates the compression-based strategy. In the following, we characterize the effect of such quantization noises on the performance of the downlink C-RAN system.

We make similar transmission assumptions as in the case of the data-sharing strategy. With $\mathbf{V}_{k,l}$ as the matrix of beamforming vectors for user k from RRH l, we can write the precoded signal computed at the CP and intended for transmission by RRH l as

$$\hat{\mathbf{x}}_l^{\mathrm{dl}} = \sum_k \mathbf{V}_{k,l}\mathbf{s}_k^{\mathrm{dl}}. \tag{1.25}$$

These signals are then compressed and sent to the RRHs. As with the compress-forward strategy in the uplink, we model the compression process mathematically as an additive process

$$\mathbf{x}_l^{\mathrm{dl}} = \hat{\mathbf{x}}_l^{\mathrm{dl}} + \mathbf{e}_l^{\mathrm{dl}}, \tag{1.26}$$

where $\mathbf{x}_l^{\mathrm{dl}}$ is the reconstructed signal that RRH l actually transmits to the users, and the additional noise $\mathbf{e}_l^{\mathrm{dl}} \in \mathbb{C}^{M\times 1}$ (assumed to be independent of the signals to be compressed) captures the effect of quantization. We assume a Gaussian quantization model with $\mathbf{e}_l^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_{l,l}^{\mathrm{dl}})$. We remark that, similar to the uplink, the additive model for the compression process above is without loss of generality and includes the possibility of processing $\hat{\mathbf{x}}_l^{\mathrm{dl}}$ with a beamformer $\mathbf{B}_l$ (possibly to reduce the rank) prior to quantization. Note that the transmit power at RRH l can be represented as $\sum_k \mathrm{Tr}\left(\mathbf{V}_{k,l}\mathbf{V}_{k,l}^H\right) + \mathrm{Tr}\left(\mathbf{Q}_{l,l}^{\mathrm{dl}}\right)$; it accounts for the contribution due to the quantization noises. It is also worth noting that the quantization noises of different RRHs are not necessarily independent of each other as the signals for all the RRHs are compressed jointly at the CP.

Let $\mathbf{Q}^{\mathrm{dl}} \in \mathbb{C}^{LM\times LM}$ denote the covariance matrix of the jointly Gaussian quantization noises of all the RRH signals with $\mathbf{Q}_{l,l}^{\mathrm{dl}}$ being the lth diagonal block submatrix in $\mathbf{Q}^{\mathrm{dl}}$. The received signal at user k under the compression strategy can be expressed as

$$\mathbf{y}_k^{\mathrm{dl}} = \sum_l \mathbf{H}_{k,l}^{\mathrm{dl}}\mathbf{V}_{k,l}\mathbf{s}_k^{\mathrm{dl}} + \sum_l \sum_{j\neq k} \mathbf{H}_{k,l}^{\mathrm{dl}}\mathbf{V}_{j,l}\mathbf{s}_j^{\mathrm{dl}} + \mathbf{H}_k^{\mathrm{dl}}\mathbf{e}^{\mathrm{dl}} + \mathbf{z}_k^{\mathrm{dl}}, \tag{1.27}$$

where $\mathbf{e}^{\mathrm{dl}} = \left[\mathbf{e}_1^{\mathrm{dl}}, \ldots, \mathbf{e}_L^{\mathrm{dl}}\right]$. As can be seen from (1.27), the received signal in the compression strategy has an additional noise term due to the quantization noises in the signals transmitted to the RRHs.

Given (1.27), the achievable rate for user k under the compression strategy,

again treating interference as noise, can be expressed as

$$R_k^{\text{comp,dl}} = I(\mathbf{s}_k^{\text{dl}}; \mathbf{y}_k^{\text{dl}}) \tag{1.28}$$

$$= \log \frac{\left| \sum_j \mathbf{H}_k^{\text{dl}} \mathbf{V}_j \mathbf{V}_j^H \left(\mathbf{H}_k^{\text{dl}}\right)^H + \mathbf{H}_k^{\text{dl}} \mathbf{Q}^{\text{dl}} \left(\mathbf{H}_k^{\text{dl}}\right)^H + \sigma_{\text{dl}}^2 \mathbf{I} \right|}{\left| \sum_{j \neq k} \mathbf{H}_k^{\text{dl}} \mathbf{V}_j \mathbf{V}_j^H \left(\mathbf{H}_k^{\text{dl}}\right)^H + \mathbf{H}_k^{\text{dl}} \mathbf{Q}^{\text{dl}} \left(\mathbf{H}_k^{\text{dl}}\right)^H + \sigma_{\text{dl}}^2 \mathbf{I} \right|}. \tag{1.29}$$

As compared to the rate in the data-sharing strategy (1.22), the rate (1.29) in the compression-based strategy has an additional term that represents the combined quantization noise after it passes through the channel. This quantization noise lowers the achievable rate.

On the plus side, since the beamformers are computed at the CP, there are no specific constraints on $\mathbf{V}_{k,l}$ that limit the participation of RRHs in serving the users. So long as the CSI from the serving RRHs to the users is available at the CP, the CP can pre-compute all the beamformers and describe the beamformed signals to the RRHs in an efficient way.

We now look at the relationship between the quantization noise levels and the fronthaul capacities. The precise relationship depends on the compression technique used at the CP. We start with the case where the signals of different RRHs are compressed independently. In such a scenario, the quantization noises at different RRHs are uncorrelated, and the quantization noise covariance matrix $\mathbf{Q}^{\text{dl}}$ is a block-diagonal matrix with $\mathbf{Q}_{l,l}^{\text{dl}}$ on the diagonal blocks. Using results from rate-distortion theory, similar to the case of independent compression in the uplink, the fronthaul capacity required for independent compression at RRH l is given by

$$C_l^{\text{indep,ul}} \geq I(\mathbf{x}_l^{\text{dl}}; \hat{\mathbf{x}}_l^{\text{dl}}) \tag{1.30}$$

$$= \log \left| \sum_k \mathbf{V}_{k,l} \mathbf{V}_{k,l}^H + \mathbf{Q}_{l,l}^{\text{dl}} \right| - \log \left| \mathbf{Q}_{l,l}^{\text{dl}} \right|. \tag{1.31}$$

Note that when independent compression is performed across signals of different RRHs, i.e., with block-diagonal $\mathbf{Q}^{\text{dl}}$, the aggregated effect of the quantization noises at the users, $\mathbf{H}_k^{\text{dl}} \mathbf{Q}^{\text{dl}} \left(\mathbf{H}_k^{\text{dl}}\right)^H$, is just the sum of contributions $\mathbf{H}_{k,l}^{\text{dl}} \mathbf{Q}_{l,l}^{\text{dl}} (\mathbf{H}_{k,l}^{\text{dl}})^H$ from each RRH. However, it is possible to improve the achievable rates by considering a more general compression scheme that allows for arbitrary correlation among quantization noises in the signals of different RRHs. Such correlation allows the possibility of nonzero off-diagonal block matrices in $\mathbf{Q}^{\text{dl}}$ that can potentially lead to terms that eventually cancel each other at the user side. This type of compression is termed multivariate compression, as first proposed in [20], and is discussed below.

Assuming a compression order from RRH 1 to L, the fronthaul required to compress the signals for RRH l for multivariate compression can be expressed

as:

$$C_l^{\text{mult,dl}} \geq I(\mathbf{x}_l^{\text{dl}}; \hat{\mathbf{x}}_l^{\text{dl}}) + I(\mathbf{e}_l^{\text{dl}}; \mathbf{e}_1^{\text{dl}}, \ldots, \mathbf{e}_{l-1}^{\text{dl}}) \tag{1.32}$$

$$= \log \left| \sum_k \mathbf{V}_{k,l} \mathbf{V}_{k,l}^H + \mathbf{Q}_{l,l}^{\text{dl}} \right| - \log \left| \mathbf{Q}_{l,l}^{\text{dl}} - \mathbf{Q}_{l,1:l-1}^{\text{dl}} \left(\mathbf{Q}_{1:l-1,1:l-1}^{\text{dl}} \right)^{-1} \left(\mathbf{Q}_{l,1:l-1}^{\text{dl}} \right)^H \right|. \tag{1.33}$$

Here, $\mathbf{Q}_{\mathcal{A},\mathcal{B}}^{\text{dl}}$ denotes the covariance submatrix of $\mathbf{Q}^{\text{dl}}$ indexed by the RRHs in the sets $\mathcal{A}$, and $\mathcal{B}$ and $1:l$ denotes the set $\{1, \ldots, l\}$. As can be seen from the expression above, introducing correlation between the quantization noises of different RRHs actually costs more fronthaul capacity as compared with independent compression. The benefit of such correlation is that since these quantization noises pass through the channel and add up at the end users, we can potentially design the noise correlations in such a way as to aligning them appropriately in order to make the noises cancel each other at the user side, thereby improving the overall system performance.

As with the Wyner-Ziv compression in the uplink, different ordering of the RRHs results in different fronthaul requirements and quantization noise covariance matrices. For a fixed order, a practical implementation of the multivariate compression has been proposed in [20].

1.3.5 Optimization Framework for Compression

Under the different compression strategies described above, the weighted sum rate maximization problem for compression-based strategy in the downlink C-RAN can be formulated differently as follows:

$$\underset{\mathbf{V}_{k,l}, \mathbf{Q}}{\text{maximize}} \quad \sum_k w_k R_k^{\text{comp,dl}} \tag{1.34a}$$

$$\text{subject to} \quad (1.31) \quad \text{or} \quad (1.33) \quad \leq C_l, \quad \forall l, \tag{1.34b}$$

$$\sum_k \text{Tr}\left(\mathbf{V}_{k,l} \mathbf{V}_{k,l}^H\right) + \text{Tr}\left(\mathbf{Q}_{l,l}^{\text{dl}}\right) \leq P_l^{\text{dl}}, \quad \forall l, \tag{1.34c}$$

where $R_k^{\text{comp,dl}}$ in (1.34a) is defined in (1.29). Note that additional constraints on the format of the covariance matrix $\mathbf{Q}^{\text{dl}}$ are to be imposed depending on the compression strategy. For example, in (1.31), $\mathbf{Q}^{\text{dl}}$ needs to be a block-diagonal matrix with the diagonal matrices $\mathbf{Q}_{l,l}^{\text{dl}}$ being positive semi-definite, i.e. $\mathbf{Q}_{l,l}^{\text{dl}} \succeq \mathbf{0}, \forall l$; in (1.33), $\mathbf{Q}^{\text{dl}}$ needs to be a positive semi-definite matrix, i.e. $\mathbf{Q}^{\text{dl}} \succeq \mathbf{0}$.

Unfortunately, none of the above optimization problems is a convex optimization program. In [20], the optimization problems (1.34) under (1.31) and (1.33) are solved through the majorize-minimization (MM) method. The main observation that allows such a method is that both the nonconvex objective and the fronthaul relation can be represented as a difference of convex functions. To implement the MM-based method proposed in [20], first the transmit beamforming variables are converted into transmit covariance matrices and the rank

constraints on the covariance matrices are relaxed in subsequent optimization. Then a sequence of convex programs are solved over the covariance matrices by repeatedly linearizing the convex parts in the objective function and the concave parts in the fronthaul constraints until some convergence criterion is met. Such a method can be shown to reach a local optimum of the rank-relaxed problem. In the end, to get back the appropriate beamformers, the eigenvectors corresponding to the largest eigenvalues of the final transmit covariance matrices are selected.

1.3.6 Hybrid Strategy

The data-sharing and compression-based strategies utilize the fronthaul capacity in two distinct ways. In data-sharing, the fronthaul links carry raw user messages for RRHs to compute the beamformed signals, while in compression-based strategy, the fronthaul links carry compressed bits of the already computed beamformed signals. The advantage of data-sharing approach is that the RRHs receive clean messages to be used for joint transmission. However, the fronthaul capacity constraint limits the cooperation cluster size. The main advantage of the compression-based approach is that the fronthaul capacity is more efficiently utilized when beamformed signals of multiple user messages are transmitted through the fronthaul. However, it pays a price in the extra quantization noise term in the resulting rate expression.

Based on the above comparison, a hybrid compression and data-sharing strategy is proposed in [23] to obtain the benefit of both strategies. In the hybrid strategy, a part of the fronthaul capacity is used to carry direct messages for some users and the remaining is used to carry the compressed beamformed signal of the rest of the users.

The rationale behind such an approach is the following. The desired precoded signal typically consists of both strong and weak signals and both high-rate and low-rate data streams. It would be beneficial to directly carry clean messages for the relatively strong signal with relatively low rate, because in this case it is typically more efficient to send the information bits themselves than to do compression on such signals. With these strong signals separated out, the amplitude of the rest of the signal is now lower. It would therefore require fewer bits to compress.

From the RRH's perspective, each RRH receives the direct messages for the strong users and the compressed precoded signals for the rest of the weak users in the network. It can compute a beamformed signal based on the direct messages and the decompressed signal, and transmit the result on its antennas. An optimization framework to design such a hybrid strategy is discussed in [23]. The key design parameters in such a hybrid approach are the selection of users that are suitable for direct data-sharing, in addition to the beamforming and quantization noise variables.

1.3.7 Data-Sharing versus Compression

Two fundamentally distinct strategies of data-sharing and compression are presented in this chapter for the downlink C-RAN. A natural question to ask is which one performs better in a realistic wireless network? The answer to this question depends on the amount of fronthaul capacity available.

In theory, to achieve full cooperation across the cluster managed by the CP, the amount of fronthaul capacity required for data-sharing strategy at each RRH is simply the sum of the achievable rates of all the users across the cluster, which is finite. However, for the compression-based strategy to achieve full cooperation, infinite fronthaul capacity would be needed in order to bring the the quantization noises to zero. Thus at extremely high fronthaul capacities, data-sharing has an advantage as compared to compression.

At extremely low fronthaul capacities, data-sharing also has an advantage. This is because this case reduces to traditional single-cell processing, where each user's data is sent to one RRH only. Since the user data is discrete, it is more efficient to send messages rather than the compressed version of the analog signal.

However, for most realistic network settings, where the fronthaul capacity is moderately high, the compression-based strategy almost always outperforms the data-sharing strategy. This is because the effect of quantization noises is usually quite small. Further, compression is a more efficient utilization of the fronthaul capacity than data-sharing, because the latter essentially replicates the same user message across multiple fronthaul links, which is inefficient. Numerical comparison of the two strategies has been investigated in [24] under a realistic network topology under different fronthaul capacities. When the fronthaul capacity is moderate and the two strategies are comparable, the hybrid of the two can bring additional gains [23].

In the downlink C-RAN, the gains due to cooperation depends crucially on the ability of the CP to obtain CSI of the users in its cluster. The discussion so far assumes that CSI of all users in the cluster is available at the CP. But in practice, CSI acquisition and sharing consume significant fronthaul capacity, and are expected to be major factors in limiting the size of cooperation cluster in the C-RAN architecture. Note that at the same cluster size data-sharing strategy achieves higher rate than the compression strategy due to the additional quantization noise in compression. So, to achieve the same rate, the compression strategy requires larger cluster size, hence more CSI. In a typical deployment, the cooperation cluster size under the compression strategy is mostly limited by CSI availability, while for data-sharing it is mostly limited by the fronthaul capacity.

As a concluding remark, we note that the implementations of the data-sharing and compression strategies have key differences in that the RRHs need to have knowledge of the modulation and coding format for implementing data-sharing, but such codebook knowledge is not needed for compression. Thus, the RRHs for implementing the compression strategy can be made much simpler.

1.4 Summary

This chapter illustrates cooperative beamforming and relaying strategies and the associated resource allocation for both uplink and downlink C-RAN. In the uplink, we show compress-forward as the fundamental strategy and provide an optimization framework for transmit beamforming at the users and quantization at the RRHs. In the downlink, we demonstrate data-sharing and compression as two competing and fundamentally different strategies. The data-sharing optimization framework for RRH clustering and transmit beamforming and the compression optimization framework for cooperative beamforming and quantization at the CP are discussed. In all cases, the finite fronthaul capacity has major impact on the analysis and design of different transmission and relaying strategies in the C-RAN architecture.

The achievable user rate and the fronthaul rate expressions used throughout the chapter are based on information theoretic analysis and assume the use of capacity-achieving and rate-distortion achieving codes. The codes used in practice usually operate below the information theoretical limit. However, to a good approximation, the performance due to such practical codes can be captured by incorporating gap factors in the respective user rate and fronthaul rate expressions. The optimization algorithms developed in this chapter can be easily extended with such factors taken into account.

Notes

References

[1] A. Avestimehr, S. Diggavi, and D. Tse, "Wireless network information flow: A deterministic approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1872–1905, Apr. 2011.

[2] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[3] S. Christensen, R. Agarwal, E. de Carvalho, and J. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792 –4799, Dec. 2008.

[4] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, 2014.

[5] ——, "Backhaul-aware multicell beamforming for downlink cloud radio access network," in *Proc. IEEE Int. Commun. Conf. (ICC) Workshop*, June 2015, pp. 2689–2694.

[6] D. Gesbert, S. Hanly, H. Huang, S. Shamai (Shitz), O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[7] E. Heo, O. Simeone, and H. Park, "Optimal fronthaul compression for synchronization in the uplink of cloud radio access networks," *CoRR*, vol. abs/1510.01545, 2015. [Online]. Available: http://arxiv.org/abs/1510.01545

[8] S.-N. Hong and G. Caire, "Reverse compute and forward: A low-complexity architecture for downlink distributed antenna systems," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2012, pp. 1147–1151.

[9] ——, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5227–5243, Sept. 2013.

[10] H. Huang, M. Trivellato, A. Hottinen, M. Shafi, P. Smith, and R. Valenzuela, "Increasing downlink cellular throughput with limited network MIMO coordination," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 2983–2989, June 2009.

[11] S. Kannan, A. Raja, and P. Viswanath, "Approximately optimal wireless broadcasting," *IEEE Trans. Inf. Theory*, vol. 58, no. 12, pp. 7154–7167, 2012.

[12] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Commun. Mag.*, vol. 13, no. 4, pp. 56–61, Aug. 2006.

[13] S. Kaviani, O. Simeone, W. Krzymien, and S. Shamai, "Linear precoding and equalization for network MIMO with partial cooperation," *IEEE Trans. Veh. Technol.*, vol. 61, no. 5, pp. 2083–2096, Jun. 2012.

[14] S. H. Lim, K. T. Kim, and Y. Kim, "Distributed decode-forward for relay networks," *CoRR*, vol. abs/1510.00832, 2015. [Online]. Available: http://arxiv.org/abs/1510.00832

[15] S. H. Lim, Y.-H. Kim, A. El Gamal, and S.-Y. Chung, "Noisy network coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 3132–3152, May 2011.

[16] L. Liu, S. Bi, and R. Zhang, "Joint power control and fronthaul rate allocation for throughput maximization in ofdma-based cloud radio access network," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4097–4110, Nov. 2015.

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

[18] B. Nazer, A. Sanderovich, M. Gastpar, and S. Shamai, "Structured superposition for backhaul constrained cellular uplink," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2009, pp. 1530–1534.

[19] C. T. K. Ng and H. Huang, "Linear precoding in cooperative MIMO cellular networks with limited coordination clusters," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1446–1454, Dec. 2010.

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

[21] ——, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[22] ——, "Robust layered transmission and compression for distributed uplink reception in cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 204–216, Jan. 2014.

[23] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," in *Proc. Inf. Theory Applicat. Workshop (ITA)*, Feb. 2014, pp. 1–6.

[24] P. Patil, B. Dai, and W. Yu, "Performance comparison of data-sharing and compression strategies for cloud radio access networks," in *Proc. European Signal Process. Conf. (EUSIPCO)*, July 2015, pp. 2456–2460.

[25] A. Sanderovich, S. Shamai, Y. Steinberg, and G. Kramer, "Communication via decentralized processing," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3008–3023, Jul. 2008.

[26] S. Shamai and B. M. Zaidel, "Enhancing the cellular downlink capacity via coprocessing at the transmitting end," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, vol. 3, May 2001, pp. 1745–1749.

[27] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

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

[29] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.

[30] ——, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," *IEEE Trans. Signal Process.*, submitted, 2015.