Cloud Radio Access Networks

Contents

List of illustrations *page* iv

1 Fronthaul Compression for C-RAN 1

1.1 Abstract 1

1.2 Introduction 1

1.3 State of the Art: Point-to-Point Fronthaul Processing 4

1.3.1 Scalar Quantization: CPRI 4

1.3.2 Compressed CPRI 5

1.3.3 Alternative functional splits 6

1.4 Network-Aware Fronthaul Processing: Uplink 7

1.4.1 Problem Setting 8

1.4.2 Distributed Quantization/ Compression (Wyner-Ziv Coding) 9

1.4.3 Examples 12

1.5 Network-Aware Fronthaul Processing: Downlink 14

1.5.1 Multivariate Quantization/ Compression 15

1.5.2 Example 16

1.6 Network-Aware Fronthaul Processing: In-Network Processing 17

1.6.1 Problem Setting 17

1.6.2 In-Network Fronthaul Processing 18

1.7 Concluding Remarks 20

Notes 21

References 22

Illustrations

1.1 A heterogeneous cellular network based on C-RAN with two clusters of RRHs managed by two BBUs and a multi-hop fronthaul network (fronthaul links are shown as solid lines). 2

1.2 Alternative functional splits of the physical layer between BBU and RRH. 7

1.3 Block diagram detailing baseband processing for the uplink. With conventional point-to-point solutions, the decompression block at the BBU performs separate decompression on each fronthaul link, while, with a network-aware solution based on distributed quantization/ compression, joint decompression across all connected RRHs is performed. 8

1.4 Block diagram detailing the process of network-aware decompression where signals that have been already decompressed can be used as side information to alleviate the fronthaul requirements for another RRH. 10

1.5 Left: Standard uplink quantization; Right: Quantization with side information (Wyner-Ziv coding). 11

1.6 Per-cell uplink sum-rate versus the uplink SNR for the circulant Wyner model with fronthaul capacity of 3 bits/s/Hz and inter-cell channel 5 dB smaller than the intra-cell channel gain. 12

1.7 Cell-edge throughput versus the average per-UE spectral efficiency for various fairness constants α in the uplink of a C-RAN with $N = 3$ pico-BSs, $K = 5$ UEs, $(C_{\mathrm{macro}}, C_{\mathrm{pico}})$=$(9, 3)$ bits/s/Hz, $T = 10$, $\beta = 0.5$ and a bandwidth of 10 MHz. 14

1.8 Block diagram detailing baseband processing for the downlink. With conventional point-to-point solutions, the compression block at the BBU performs separate decompression on each fronthaul link, while, with a network-aware solution based on multivariate quantization/ compression, joint compression across all connected RRHs is performed. 15

1.9 Left: Standard downlink quantization for two RRHs; Right: Multivariate quantization. 16

1.10 Cell-edge throughput, i.e., 5%-ile rate, versus the average per-UE spectral efficiency for various fairness constants α in the downlink of a C-RAN with $N = 1$ pico-BS, $K = 4$ UEs, $(C_{\mathrm{macro}}, C_{\mathrm{pico}})$=$(6, 2)$ bits/s/Hz, $T = 5$ and $\beta = 0.5$. 17

1.11 Block diagram detailing baseband processing at each RRH with (a) routing and (b) in-network processing strategies. 18

1.12 Average sum-rate versus the number of RRHs in the first layer of a two-hop topology with $N_M = 4$ UEs, average received per-antenna SNR of 20 dB and fronthaul capacity of 2, 3, 4 bits/s/Hz. 19

1 Fronthaul Compression for C-RAN

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

1.1 Abstract

This chapter covers fronthaul quantization and compression for C-RANs. First, the state-of-the-art is reviewed by discussing the CPRI specification, which prescribes scalar quantization of the baseband signals. Then, various improvements of CPRI are overviewed that apply to each fronthaul link operations such as filtering, scaling or lossless compression. These point-to-point, or per-fronthaul link, quantization/ compression solutions are generally oblivious to the network topology and state, e.g., density and channel conditions, and, as a result, they are suboptimal from a network information theoretic standpoint. Based on this observation, advanced fronthaul processing methods are presented that follow network information theoretic principles and leverage the joint processing capabilities of the baseband unit (BBU) along with information about network topology and state. It is demonstrated that the information theoretic concepts of distributed quantization and compression, multivariate quantization and compression, and in-network processing provide useful frameworks on which to base the design of fronthaul processing techniques that are able to improve network-wide, rather than per-link, performance criteria. In particular, distributed and multivariate quantization/ compression may enhance conventional point-to-point solutions by means of joint fronthaul processing at the BBU of the baseband signals corresponding to Remote Radio Heads (RRHs) with overlapping covering areas. Furthermore, in-network processing may overcome the limitations of standard multiplex-and-forward techniques in multi-hop fronthaul topologies by removing spatial redundancies across nearby RRHs. Numerical results are reported throughout the chapter in order to illustrate the relative merits of different solutions, and open problems are pointed out.

1.2 Introduction

The C-RAN architecture relies on fronthaul links to connect each Remote Radio Head (RRH) to the managing Baseband Unit (BBU). In particular, for the uplink, the fronthaul links allow the RRHs to convey their respective received signals, either in analog format or in the form of digitized baseband samples, to

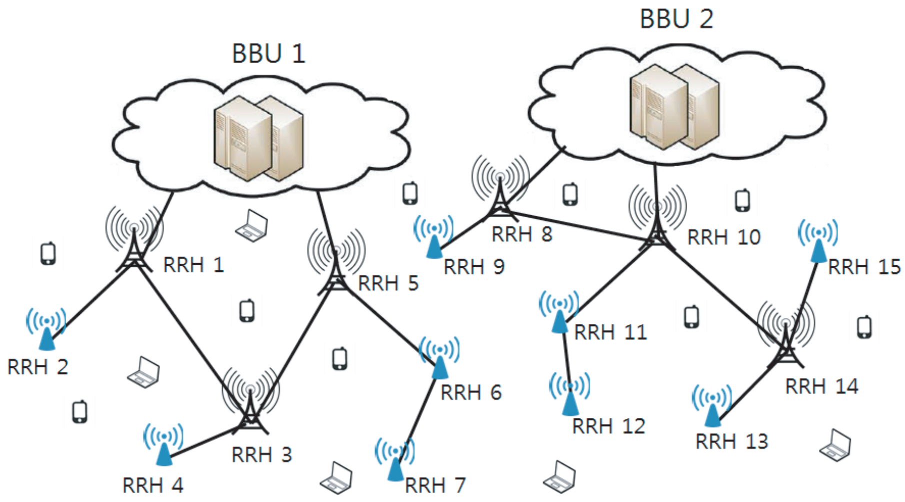

Figure 1.1 A heterogeneous cellular network based on C-RAN with two clusters of RRHs managed by two BBUs and a multi-hop fronthaul network (fronthaul links are shown as solid lines).

the BBU. For the downlink, instead, the BBU transfers the radio signal that each RRH is to transmit on the radio interface, in analog or digital format, on the fronthaul links to the RRHs. It is this transfer of radio or baseband signals that makes the virtualization of baseband and higher-layer functions of the RRHs at the BBU, which defines the C-RAN architecture, possible. The analog transport solution is typically implemented by means of radio-over-fiber (see, e.g., [1]) but solutions based on copper LAN cables are also available [2]. In contrast, the digital transmission of baseband, or IQ, samples is currently carried out by following the Common Public Radio Interface (CPRI) specification [3]. This ideally requires fiber optic fronthaul links, although practical constraints motivate the development of wireless-based digital fronthauling [4]. The digital approach seems to have attracted the most interest due to the traditional advantages of digital solutions, including resilience to noise and to hardware impairments as well as flexibility in the transport options (see, e.g., [5]). Furthermore, the connection between an RRH and the BBU may be direct, i.e., single-hop, or may take place over a cascade of fronthaul links, i.e., multi-hop, as illustrated in Fig. 1.1.

In this chapter, we provide an overview of the state of the art on the problem of transporting digitized IQ baseband signals on the fronthaul links. As mentioned, the current de facto standard that defines analog-to-digital processing and transport options is provided by the CPRI specification [3]. CPRI is widely understood to be unsuitable for the large-scale implementation of C-RAN owing to its significant fronthaul bit rate requirements under common operating conditions. As an example, as reported in [5, 6], the bit rate needed for an LTE base station that serves three cell sectors with carrier aggregation over five carriers and two receive antennas exceeds even the 10 Gbits/s provided by standard fiber optics links. The large bit rate is a consequence of the simple scalar quantization approach taken by CPRI, whereby each IQ sample is quantized using a given number - typically around 15 - of bits per I and Q sample. The rate requirements are even more problematic for network deployments in which fiber optic links are not available – a common occurrence due to the cost of installing or leasing fiber optic connections. Typical examples are heterogeneous dense networks with RRHs having small coverage, such as pico-base stations or home-base stations, for which wireless fronthauling is under study over mm-wave channels [4].

Motivated by the mentioned shortcomings of the CPRI specification in the presence of practical fronthaul capacity limitations, this chapter aims at providing a review of current and advanced solutions for the compression of baseband signals to be transmitted over digital fronthaul links. We observe that fronthaul links also impose constraints on the latency entailed by the transfer of information between BBU and RRHs, which have important consequences on the performance of protocols such as HARQ and random access; we refer to [5, 7] for discussions and references.

The content and organization of the chapter is as follows.

- **Point-to-Point Fronthaul Processing:** The state-of-the-art on fronthaul quantization/ compression of baseband signals is reviewed in Sec. 1.3. In particular, in this section, we discuss the CPRI specification and various improvements thereof that apply solutions such as filtering, scaling or lossless compression to each fronthaul link.
- **Network-Aware Fronthaul Processing:** The point-to-point, or per-fronthaul link, quantization/ compression solutions reviewed in Sec. 1.3 are generally oblivious to the network topology and state, e.g., density and channel conditions. As a result, they are generally suboptimal from a network information theoretic standpoint. Based on this observation, we then overview advanced fronthaul processing methods that follow network information theoretic principles and leverage the joint processing capabilities of the BBU along with information about network topology and state. We refer to this class of techniques as being *network-aware*. Specifically, in Sec. 1.4, *distributed quantization/ compression* is discussed for the uplink of C-RAN systems; in Sec. 1.5, *multivariate quantization/ compression* is presented for the downlink; and Sec. 1.6 elaborates on the use of *in-network processing* for multi-hop network topologies. In each section, the information theoretic principles underlying each solution are explained by using intuitive arguments and illustrations.

Network-aware fronthaul processing techniques operate across multiple fronthaul links and RRHs, and hence their benefits should be measured at a system level rather than merely in terms of rate reduction on each fronthaul link. Therefore, numerical results are reported throughout the chapter in order to illustrate the relative merits of different solutions in terms of network-wide criteria such as sum-rate or edge-cell rate.

We end this introduction by emphasizing two important themes that are recurring in the chapter. The first is the fact that, in a C-RAN, significant gains can be accrued by the joint optimization of the operation of the system across the wireless channels and the fronthaul network. The second, broader, theme is the important role that network information theory can play in guiding the design of practical solutions for complex systems such as C-RAN and, more generally, 5G systems and beyond.

1.3 State of the Art: Point-to-Point Fronthaul Processing

In this section, we first review the basics of the CPRI specification in Sec. 1.3.1. Then, having identified the limitations of the scalar quantization approach prescribed by CPRI, Sec. 1.3.2 presents techniques that have been proposed to reduce the fronthaul bit rate by means of more advanced quantization and compression solutions applied separately on each fronthaul link, i.e., via *point-to-point* fronthaul processing.

1.3.1 Scalar Quantization: CPRI

The CPRI specification was issued by a consortium of radio equipment manufacturers with the aim of standardizing the communication interface between BBU and RRHs[1] on the fronthaul network. CPRI prescribes, on the one hand, the use of sampling and scalar quantization for the digitization of the baseband signals, and, on the other, a constant bit rate serial interface for the transmission of the resulting bit rate. Note that the baseband signals are either obtained from downconversion in the uplink or produced by the BBU after baseband processing in the downlink.

The CPRI interface specifies a frame structure that is designed to carry user-plane data, namely the quantized IQ samples, along with the control and management plane, for, e.g., error detection and correction, and the synchronization plane data. It supports 3GPP GSM/EDGE, 3GPP UTRA and LTE, and allows for star, chain, tree, ring and multihop fronthaul topologies. CPRI signals are defined at different bit rates up to 9.8 Gbps and are constrained by strict requirements in terms of probability of error (10^{-12}), timing accuracy (0.002 ppm) and delay (5 μs excluding propagation).

The line rates are proportional to the bandwidth of the signal to be digitized, to the number of receive antennas and to the number of bits per sample. Specifically, the bit rate can be calculated as [7]

$$R_{\text{CPRI}} = 2N_{\text{ant}}R_{\text{s}}N_{\text{res}}N_{\text{ov}}, \tag{1.1}$$

[1] The terminology used in CPRI is Radio Equipment Control (REC) and Radio Equipment (RE), respectively.

where N_{ant} is the number of receive antennas at the RRH; R_{s} is the sampling rate, which depends on the signal bandwidth according to a specified table [3]; N_{res} is the number of bits per I or Q sample, so that $2N_{\text{res}}$ is the number of bits for each complex sample; and N_{ov} accounts for the overhead of the management, control plane and synchronization planes. The parameter N_{res} ranges in the interval from 8 to 20 bits for LTE in both the uplink and the downlink. It is noted that, using (1.1), it is easy to identify common scenarios, such as the one discussed at the beginning of this section, in which the maximum CPRI rate of 9.8 Gbs is violated, particularly in the presence of carrier aggregation and/or large-array MIMO systems [5, 6].

As discussed, the basic approach prescribed by CPRI, which is based on sampling and scalar quantization, is bound to produce bit rates that are difficult to accommodate within the available fronthaul capacities – most notably for small cells with wireless fronthauling and for larger cells with optical fronthaul links in the presence of carrier aggregation and large-array MIMO transceivers. This has motivated the design of strategies that reduce the bit rate of the CPRI data stream, while limiting the distortion incurred on the quantized signal. In the following, we provide an overview of these schemes by differentiating between techniques that adhere to the standard C-RAN implementation, characterized by the full migration of baseband processing to the BBU, and solutions that explore different functional splits between RRHs and BBU. We refer to the former class as compressed CPRI, and review both classes separately in the next subsections.

1.3.2 Compressed CPRI

The full separation of baseband processing at the BBU from the radio functionalities implemented at the RRHs is made possible by the fact that CPRI performs quantization of the time-domain baseband signals. We recall that these signals are either received by the RRHs in the uplink or produced by means of baseband processing at the BBU for the downlink. The separation at hand can be maintained, while reducing the required fronthaul rate, by compressing the time-domain baseband samples rather than simply performing scalar quantization. We refer to this class of approaches as compressed CPRI. Compressed CPRI is based on a number of principles, which are briefly discussed in the following.

1) *Filtering* [8, 9]: As per the CPRI standard, the time-domain signal is oversampled. For instance, for a 10 MHz LTE signal a sampling frequency of 15.36 MHz is adopted. Therefore, a low-pass filter can be applied to the signal without affecting the information content.

2) *Per-block scaling* [8, 9]: The dynamic range of the quantizer needs to be selected to accommodate the peak-to-peak variations of the time-domain signal. Given the generally large peak-to-average power ratio (PAPR), this calls for quantization with a large number of bits in order to maintain a small quantization noise over the entire dynamic range (e.g., typically 15 bits in CPRI). In LTE,

the problem is particularly relevant for the OFDM-based downlink due to the large PAPR of OFDM signals. This limitation can be mitigated by dividing the signal into subblocks of small size (e.g., 32 samples in [8]) and rescaling the signal in each subblock so that the peak-to-peak variations within the block fit the dynamic range of the quantizer. In this fashion, the relevant peak-to-peak amplitude is not that measured across the entire block of samples but only within each subblock. Note that this approach entails some overhead as the receiver needs to be informed regarding the scaling factor applied to each block – there is hence a tension between the effectiveness of this solution and the required fronthaul overhead.

3) *Optimized non-uniform quantization* [8, 9]: Rather than adopting uniform scalar quantization, the quantization levels can be optimized as a function of the statistics of the baseband signal by means of standard strategies such as the Lloyd-Max algorithm.

4) *Noise shaping* [10]: Due to the correlation of successive baseband samples, predictive, or noise shaping, quantization techniques based on a feedback filter can be beneficial to reduce the rate of optimized quantization.

5) *Lossless compression* [11][2]: Any residual correlation among successive quantized baseband samples, possibly after predictive quantization, can be further leveraged by entropy coding techniques that aim at reducing the rate down to the entropy of the digitized signal.

As a rule of thumb, compressed CPRI techniques are seen to reduce the fronthaul rate by a factors around 2-3 [7].

1.3.3 Alternative functional splits

In order to obtain further fronthaul rate reductions by means of point-to-point compression techniques, alternative functional splits to the conventional C-RAN implementation need to be explored [7, 12, **?**]. Accordingly, some baseband functionalities are implemented at the RRH, such as frame synchronization, FFT/ IFFT or resource demapping. The rationale is that, by keeping some baseband functionalities at the RRHs, one can potentially reduce the fronthaul overhead.

A first solution prescribes the implementation of frame synchronization and FFT in the uplink and of the IFFT in the downlink at the RRH (see demarcation point "A" in Fig. 1.2). The rest of the baseband functionalities, such as channel decoding/encoding, are instead performed at the BBU. This functional split enables the signal to be quantized in the frequency domain, that is, after the FFT in the uplink and prior to the IFFT in the downlink. Given that the signal has a lower PAPR in the frequency domain, particularly in the LTE downlink, the number of bits per sample can be reduced at a minor cost in terms of signal-to-quantization-noise ratio. The experiments in [7] do not demonstrate, however, very significant fronthaul rate gains with this approach.

[2] Reference [11] in fact considers time-domain modulation and not OFDM but the principle is the same as discussed here.

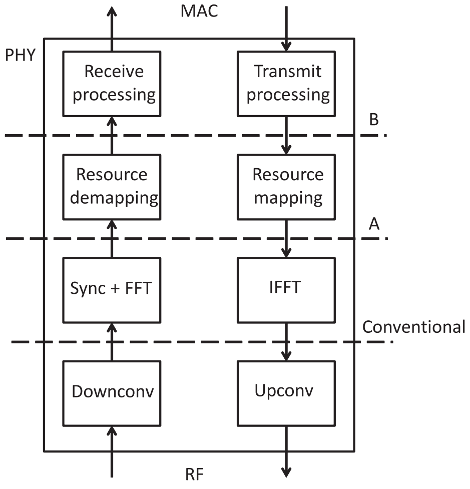

Figure 1.2 Alternative functional splits of the physical layer between BBU and RRH.

A more promising approach implements also resource demapping for the uplink and resource mapping for the downlink at the RRH (see demarcation point "B" in Fig. 1.2). For the uplink, this implies that the RRH can deconstruct the frame structure and distinguish among the different physical channels multiplexed in the resource blocks. As a result, the RRH can apply different quantization strategies to distinct physical channels, e.g., by quantizing more finely channels carrying higher-order modulations. More importantly, in the case of lightly loaded frames, unused resource blocks can be neglected. This approach was shown in [7, 13] to lead to compression ratios of the order of up to 30 – an order of magnitude larger than with compressed CPRI – in the regime of small system loads. A similar approach was also implemented in the field trials and reported in [14].

1.4 Network-Aware Fronthaul Processing: Uplink

The solutions explored so far to reduce the fronthaul capacity requirements of the C-RAN architecture have been based on point-to-point quantization and compression algorithms. Here we revisit the problem of fronthaul compression by taking a more fundamental viewpoint grounded in network information theory. Accordingly, we look at the problem at the network level rather than at the granularity of each individual fronthaul link. As the rest of this chapter illustrates, this network-aware perspective on the design of fronthaul processing has the potential to move significantly beyond the limitations of point-to-point

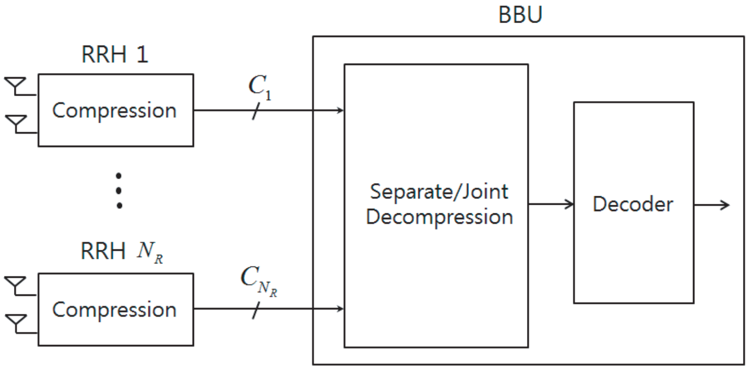

Figure 1.3 Block diagram detailing baseband processing for the uplink. With conventional point-to-point solutions, the decompression block at the BBU performs separate decompression on each fronthaul link, while, with a network-aware solution based on distributed quantization/ compression, joint decompression across all connected RRHs is performed.

approaches towards the network information theoretic optimal performance. We start by analyzing the uplink with a single-hop fronthaul topology in this section, while the downlink, under the same fronthaul topology is treated in the next section. Multi-hop fronthauling is finally considered in Sec. 1.6.

1.4.1 Problem Setting

The block diagram of an uplink C-RAN system characterized by a single cluster of RRHs with a single-hop fronthaul topology is shown in Fig. 1.3. Note that the fronthaul links between the RRHs and the BBU may have significantly different capacity limitations, as indicated by the parameters C_i in the figure. For instance, some RRHs may be endowed with optical fronthaul links while others with wireless fronthauling. Depending on the available fronthaul link budget, each RRH quantizes/ compresses the locally received baseband signal to convey it on the corresponding fronthaul link to the BBU. Note that we mark the fronthaul processing block at the RRHs as "Compression", although this block also includes quantization, and, furthermore, no compression may take place following quantization as in a standard CPRI implementation. An analogous discussion applies to the "Decompression" block at the BBU.

In a conventional implementation based on the point-to-point solutions reviewed in the previous section, the decompression block at the BBU involves separate decompression operations for each fronthaul link. In contrast, with a network-aware solution based on distributed source coding, joint decompression across all connected RRHs is performed. The rationale for joint decompression is that the signals received by different RRHs are correlated as they represent noisy versions of the same transmitted signals. This correlation is expected to be particularly significant for dense networks – an important use case for the C-RAN architecture. To realize the performance advantages of joint decompression, the RRHs implement the network information theoretic technique of *distributed source coding* (see, e.g., [15] for an introduction), as discussed next.

1.4.2 Distributed Quantization/ Compression (Wyner-Ziv Coding)

The impact of distributed quantization/ compression, or source coding, on the performance of C-RAN systems has been investigated from an information theoretic viewpoint in a number of papers starting from the original work [16], including [17, 18, 19, 20]. The key idea of distributed source coding can be easily explained with reference to the problem of quantization or compression with side information at the receiver's side. Specifically, given that the signals received by different RRHs are correlated, once the BBU has recovered the signal of one RRH, that signal can be used as *side information* for the decompression of the signal of another RRH. As illustrated in Fig. 1.4, this process can be iterated in a decision-feedback-type loop, whereby signals that have been already decompressed can be used as side information to alleviate the fronthaul requirements for the RRHs whose signals have yet to be decompressed. As we will see, the availability of side information at the decompressor allows the required fronthaul rate to be reduced with no penalty on the accuracy of the quantized signal; or, in a dual fashion, to enhance the quantization accuracy at the same fronthaul rate.

As a practical remark, we note that the process at hand requires some decompression order across the RRHs to be established. In particular, as argued in [21], a choice that is generally sensible, and close to optimal, is that of decompressing first the signals coming from macro-base stations (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.

The coding strategy to be implemented at the RRHs in order to leverage the side information at the receiver is known in information theory as *Wyner-Ziv coding* [22]. Note that Wyner-Ziv coding does not require the RRHs to be aware of the side information available at the BBU but only of the correlation between the received signal and the side information. More discussion on this point will be provided below. Rather than offering a technical description of Wyner-Ziv coding, we provide here a simple example for the most basic case of scalar quantization. Improvements based on compression are possible and follow in a manner similar to the discussion above.

Consider Fig. 1.5. On the left, a portion of a standard uniform quantizer with five levels is shown. Note that we focus for simplicity of illustration on real samples. The black box represent the quantization levels and the vertical axis

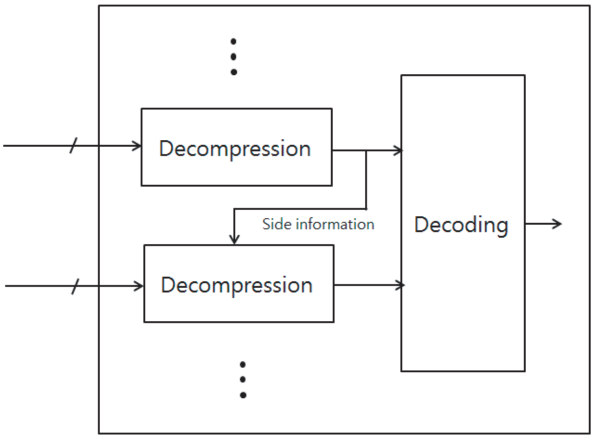

Figure 1.4 Block diagram detailing the process of network-aware decompression where signals that have been already decompressed can be used as side information to alleviate the fronthaul requirements for another RRH.

reports the corresponding binary labels[3]. Now, assume that the receiver has some side information that is correlated with the sample to be quantized. Wyner-Ziv quantization enables the RRH to use a finer quantizer, and hence to achieve an enhanced resolution, without increasing the fronthaul rate or, conversely, to keep the same resolution while reducing the fronthaul rate. The first effect is illustrated in the right part of Fig. 1.5, where the same number of binary labels, and hence the same fronthaul rate, is used to support a finer subdivision of the dynamic range of the received uplink signal. Note that, with Wyner-Ziv quantization, the same binary label is assigned to multiple quantization levels.

The BBU can distinguish between quantization levels that are assigned the same binary label – known collectively as "bin" in information theory – by leveraging the side information: The quantization level that is "closer" to the side information sample is likely to be the correct one. "Closeness" generally depends on the correlation between the signal to be decompressed and the side information. For instance, a standard minimum-distance decoder may be adopted to decode within a bin under the assumption that the received signals can be approximately described as jointly Gaussian random variables – a typical assumption in the presence of channel state information at the BBU. It is emphasized that decompressing a Wyner-Ziv quantized sample hence entails the additional decoding step of selecting the correct level within the bin indicated by the quantizer. Alternatively, one can design the quantizers jointly with the demapping function as in [23].

The description above refers to an implementation of Wyner-Ziv coding based on sample-by-sample scalar quantization. In order to reduce the probability of

[3] An odd number of levels is considered here for simplicity of illustration.

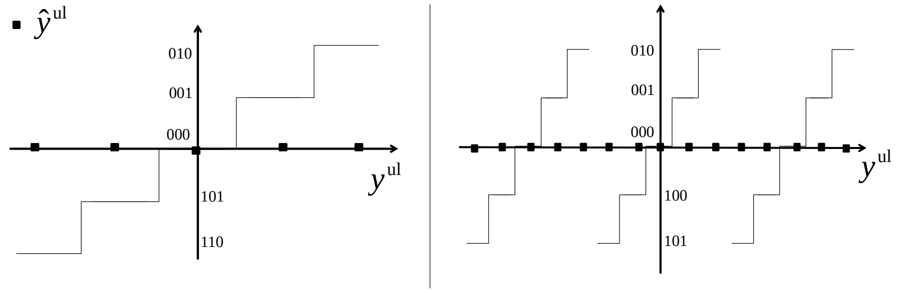

Figure 1.5 Left: Standard uplink quantization; Right: Quantization with side information (Wyner-Ziv coding).

error for the in-bin decoding step described above and/or to complement quantization with compression, in practice, block processing that leverages the mature state of the art on modern source and channel coding is desirable. A first solution is to follow standard scalar quantization, as used in CPRI, with a block that computes the syndrome of a binary linear block code on the resulting bit stream [24][25]. In this fashion, bins are implemented as cosets of a linear binary block code. Hence, by selecting codes that admit efficient decoding, such as LDPC or turbo codes, in-bin decoding becomes feasible, particularly considering that the computational burden is on the BBU. A second alternative is that of using nested codes with the property that the finer code is a good source code, or quantization codebook while the coarser code, and its cosets, are good channel codes that play the role of bins. Examples of such codes include trellis codes [26], polar codes [27], compound LDGM-LDPC codes [28].

Another key practical issue is the need to inform each RRH about the correlation between the received signal and the side information corresponding to the signals that are decompressed by the BBU before that of the RRH at hand. This correlation is essential for the RRH to select a quantization/ compression scheme that may allow the BBU to decode within a bin based on the side information with acceptable reliability. Moreover, it depends on the channel state information of the involved RRHs and it amounts to a covariance matrix of the size of the number of receive antennas at the RRH (see, e.g., [18]). Therefore, the BBU may convey this correlation matrix on the fronthaul link to the RRH or, more conventionally, the BBU may inform the RRH about which particular quantizer/ compressor to apply among the available algorithms in a codebook of possible choices. The design of such codebook and of rules for the selection of specific quantizers/compressors is an interesting open problem.

We finally observe that the discussion above assumes that the BBU first decompresses the quantized signals and then decodes the UEs' messages based on the decompressed signals. It is known that the performance may be potentially improved by performing joint decompression and decoding at the cost of an increased computational complexity [29].

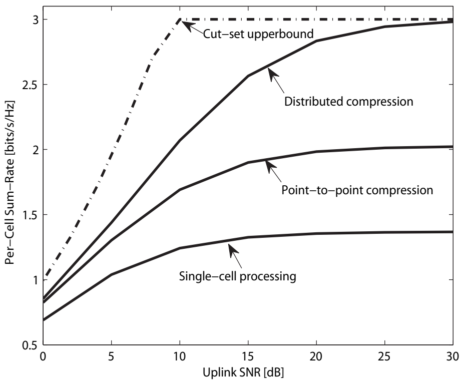

Figure 1.6 Per-cell uplink sum-rate versus the uplink SNR for the circulant Wyner model with fronthaul capacity of 3 bits/s/Hz and inter-cell channel 5 dB smaller than the intra-cell channel gain.

1.4.3 Examples

Distributed quantization/ compression, or Wyner-Ziv coding, was demonstrated in a number of theoretical papers, including [16, 18, 19, 20], to offer significant potential performance gains for the C-RAN uplink. Here we first describe one such result and then provide some discussion about performance under a more complete scenario of relevance for LTE systems. Throughout, achievable rates are computed by using standard information theoretic characterizations (i.e., by evaluating appropriate mutual information terms). The relationship between the fronthaul overhead and the accuracy of the quantized and compressed baseband signals is also modeled by using information theoretic arguments, namely from rate-distortion theory. The use of information theoretic metrics implies that the displayed performance of point-to-point solutions corresponds to the maximum theoretically achievable rate with optimal point-to-point techniques that leverage all of the ideas described in the previous section. The performance of distributed source coding also reflects that of an optimal block-based implementation that uses state-of-the-art codes and perfect information about the correlation between RRHs as introduced above. In this regard, we note that this information, in this example, reduces to a scalar since we consider each RRH to have a single antenna.

Fig. 1.6 plots the achievable per-cell uplink sum-rate for point-to-point and distributed compression versus the uplink signal-to-noise ratio (SNR) in a standard three-cell circulant Wyner model (see, e.g., [30]), where each cell contains a single-antenna UE and single-antenna RRH, and inter-cell interference takes place only between adjacent cells (the first and third cell are considered to be adjacent). Note that the sum-rate is calculated here under the assumption of joint decoding of the signals of all users at the CU for both point-to-point and distributed compression. The inter-cell channel gains are set to be 5 dB smaller than the intra-cell channel gain, and every RRH has the same fronthaul capacity of 3 bits/s/Hz, where normalization is with respect to the uplink bandwidth, or, equivalently, 3 bits for each sample of the received signal if sampled at the baud rate. For reference, we also show the per-cell sum-rate achievable with single-cell processing, whereby each RRH decodes the signal of the in-cell UE by treating all other UE signals as noise, and the cut-set upper bound [30]. It can be seen that the performance advantage of distributed compression over point-to-point compression increases as the SNR grows larger, since the correlation of the received signals at the RRHs becomes more pronounced when the effect of noise is less relevant. We also note that, with distributed compression, the uplink SNR at which the quantization noise becomes the dominant factor limiting the performance, hence causing a floor on the achievable sum-rate, increases significantly.

We now provide a performance evaluation using the standard cellular topology and channel models described in the LTE document [31]. We focus on the performance of the macro-cell located at the center of a two-dimensional 19-cell hexagonal cellular layout. In each macro-cell, there are K randomly and uniformly located single-antenna UEs and a number of RRHs as follows: a macro-base station (BS) with three sectorized antennas placed in the center and N randomly and uniformly located single-antenna pico-BSs. A single-hop fronthaul topology is assumed, where a BBU is connected directly to the macro-BS and the pico-BS in the macro-cell. 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 macro-cells are treated as independent noise signals. More details on the system parameters can be found in [32].

We adopt the conventional metric of cell-edge throughput versus the average per-UE spectral efficiency (see, e.g., [33, Fig. 5]). The achievable rates are evaluated using the rate functions in [31] that account for the smallest and largest allowed spectral efficiencies allowed by existing modulation and coding schemes. Moreover, the rates are computed 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 5%-ile rate and the average spectral efficiency as the average sum-rate normalized by the number of UEs. We recall that the proportional fairness scheduler maximizes at each time-slot the weighted sum $R_{\mathrm{sum}}^{\mathrm{fair}} = \sum_{k=1}^{K} R_k^{\mathrm{dl}}/\bar{R}_k^{\alpha}$ of per-UE rates R_k^{dl} with $\alpha \geq 0$ being a fairness constant 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 $\bar{R}_k \leftarrow \beta\bar{R}_k + (1-\beta)R_k^{\mathrm{dl}}$ where $\beta \in [0,1]$ is a forgetting factor. Increasing α leads to a more fair rate allocation among the UEs.

Fig. 1.7 plots the cell-edge throughput versus the average spectral efficiency for $N = 3$ pico-BSs, $K = 5$ UEs, $(C_{\mathrm{macro}}, C_{\mathrm{pico}})$=$(9,3)$ bits/s/Hz, $T = 10$, $\beta = 0.5$ and a bandwidth of 10 MHz. The curve is obtained by varying the fairness constant α in the utility function $R_{\mathrm{sum}}^{\mathrm{fair}}$. It is observed that spectral efficiencies larger than 1.01 bits/s/Hz are not achievable with point-to-point compression, while they can be obtained with multivariate compression. Moreover, it is seen that

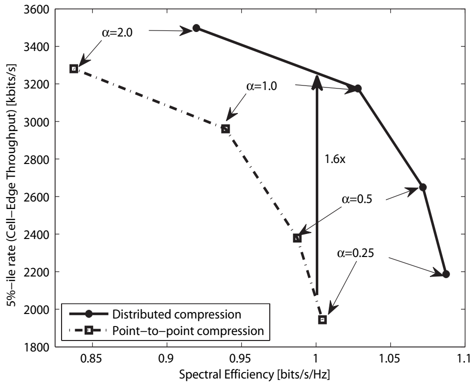

Figure 1.7 Cell-edge throughput versus the average per-UE spectral efficiency for various fairness constants α in the uplink of a C-RAN with $N = 3$ pico-BSs, $K = 5$ UEs, $(C_{\text{macro}}, C_{\text{pico}})$=$(9, 3)$ bits/s/Hz, $T = 10$, $\beta = 0.5$ and a bandwidth of 10 MHz.

distributed compression provides a gain of 1.6× in terms of cell-edge throughput for a spectral efficiency of 1 bits/s/Hz.

1.5 Network-Aware Fronthaul Processing: Downlink

In this section, we consider the downlink of a C-RAN with a single-hop fronthaul topology. As seen in the previous section, in the uplink, the traditional solution consisting of separate fronthaul quantizers/ compressors is suboptimal, from a network information theoretic viewpoint, based on the principle of distributed source coding. In this section, we will see that a different principle, namely that of multivariate quantization/ compression (see, e.g., [15]), is relevant for the downlink.

The block diagram of a downlink C-RAN system with a single cluster of RRHs connected by means of a single-hop fronthaul topology to a BBU is shown in Fig. 1.8. The BBU performs baseband processing to encode the downlink data streams to be delivered to the UEs. Specifically, the BBU first carries out channel encoding on each data stream and then applies linear precoding, which is computed from the available channel state information, in order to enable multi-user MIMO transmission. The resulting baseband signals are then quantized and compressed before being transmitted on the fronthaul links to the RRHs. As for the uplink, we mark as “Compression” and “Decompression” blocks that also include quantization.

In a conventional implementation based on the point-to-point solutions reviewed in Sec. 1.3, the compression block at the BBU involves separate compression for each fronthaul link. In contrast, with a network-aware solution based on multivariate quantization/ compression, joint compression across all connected

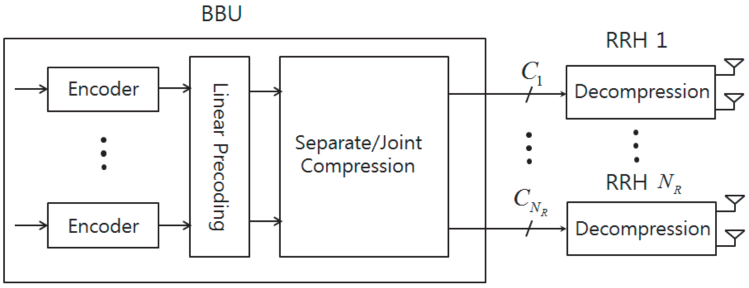

Figure 1.8 Block diagram detailing baseband processing for the downlink. With conventional point-to-point solutions, the compression block at the BBU performs separate decompression on each fronthaul link, while, with a network-aware solution based on multivariate quantization/ compression, joint compression across all connected RRHs is performed.

RRHs is performed. As discussed below, joint compression allows the BBU to shape the distribution of the quantization noise across nearby RRHs and, in so doing, to control the impact of the quantization noise on the downlink reception at the UEs.

1.5.1 Multivariate Quantization/ Compression

To understand the key ideas in the simplest terms, let us start by assessing the impact of scalar quantization on the downlink performance. We focus, as done above, on the quantization of real samples for simplicity of illustration. In the left-hand side of Fig. 1.9, we illustrate the quantization regions resulting from standard scalar quantization of the signals $\tilde{x}_i^{\mathrm{dl}}$ with $i = 1, 2$ to be transmitted by two RRHs. The black squares represent the quantization levels ${x_i}^{\mathrm{dl}}$ with $i = 1, 2$ used by the two quantizers and the black balls denote the corresponding quantization levels on the plane. Given the current channel state information, an UE may be more sensitive to quantization noise in a certain spatial dimension. In particular, if the UE has a single antenna, the channel vector from the two RRHs to the UE defines this signal direction. However, as seen in the figure, with traditional separate quantization, no control of the shape of the quantization regions is possible. Therefore, one cannot leverage the channel state information at BBU to implement a more effective quantizer that reduces the impact of the quantization noise along the signal direction.

As proposed in [34], the limitation identified above can be alleviated by multivariate compression whereby the vector $\tilde{\mathbf{x}}^{\mathrm{dl}} = [\tilde{x}_1^{\mathrm{dl}} \tilde{x}_2^{\mathrm{dl}}]^T$ is jointly, rather than separately, quantized at the BBU. Note that, unlike vector quantization, multivariate compression is constrained by the numbers of levels for each axis, which yield the fronthaul rates for each RRH, rather than on the total number of levels on the plane. As illustrated in the right-hand side of Fig. 1.9, multivariate compression enables the shaping of the quantization regions on the plane, hence

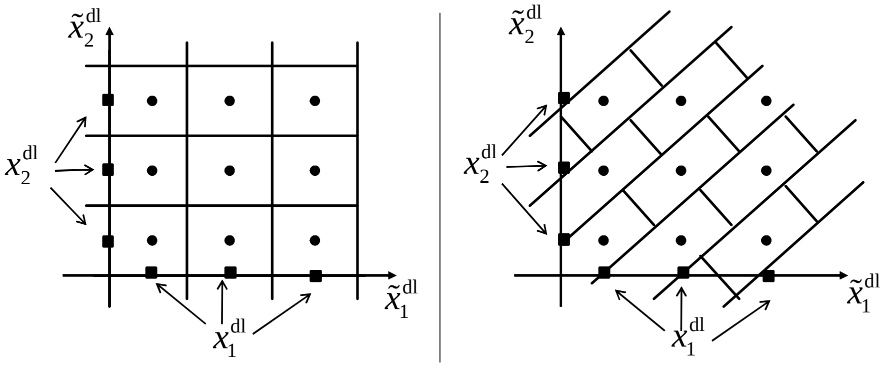
$\tilde{x}_2^{\text{dl}}$
x_2^{dl}
$\tilde{x}_1^{\text{dl}}$
x_1^{dl}
$\tilde{x}_2^{\text{dl}}$
x_2^{dl}
$\tilde{x}_1^{\text{dl}}$
x_1^{dl}

Figure 1.9 Left: Standard downlink quantization for two RRHs; Right: Multivariate quantization.

allowing a finer control of the impact of the quantization noise on the received signals.

As mentioned, the implementation of multivariate compression hinges on the availability of channel state information at the BBU, which is to be expected. Moreover, it requires the BBU to inform each RRH about the quantization levels, i.e., the quantization codebook, to be used. Importantly, however, the RRHs need not be informed about the specific mapping carried out by joint compression (e.g., about the shapes of the quantization regions on the plane in Fig. 1.9) as a function of the channel state information. In practice, therefore, the selection of the quantization codebooks should be performed at a coarse time scale, based only on long-term channel state information, while the specific mapping carried out by joint compression should be adapted on the basis of current channel state information at the BBU.

1.5.2 Example

This section provides a performance evaluation of the discussed fronthaul compression techniques considering the downlink of the LTE-based scenario adopted for Fig. 1.7. Fig. 1.10 plots the cell-edge throughput versus the average spectral efficiency for $N = 1$ pico-BS, $K = 4$ UEs, $(C_{\text{macro}}, C_{\text{pico}})=(6, 2)$ bits/s/Hz, $T = 5$ and $\beta = 0.5$. We recall that the curve is obtained by varying the fairness constant α in the utility function $R_{\text{sum}}^{\text{fair}}$. The rates are evaluated here under the assumption of an optimized linear precoder based on channel state information at the CU for both point-to-point and multivariate compression. As a side remark, we note that, in [34], the performance is also evaluated for non-linear precoders following the "dirty paper coding" principle. Similar to the uplink, it is seen that spectral efficiencies larger than 3.12 bps/Hz are not achievable with point-to-point compression, while they can be obtained with multivariate com-

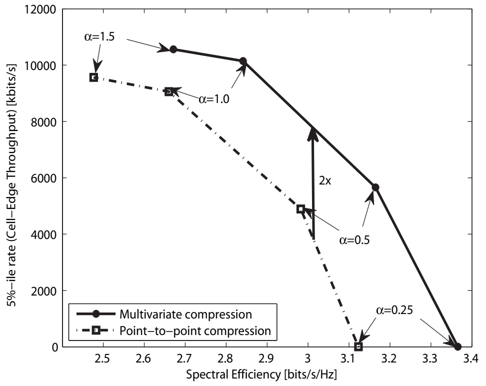

Figure 1.10 Cell-edge throughput, i.e., 5%-ile rate, versus the average per-UE spectral efficiency for various fairness constants α in the downlink of a C-RAN with $N = 1$ pico-BS, $K = 4$ UEs, $(C_{\text{macro}}, C_{\text{pico}})$=$(6, 2)$ bits/s/Hz, $T = 5$ and $\beta = 0.5$.

pression. Furthermore multivariate compression provides about a twofold gain in terms of cell-edge throughput for spectral efficiency of 3 bps/Hz.

1.6 Network-Aware Fronthaul Processing: In-Network Processing

In this section, we study the case in which the fronthaul network has a general multi-hop topology. As an example, in Fig. 1.1, RRH 7 communicates to the BBU via a two-hop fronthaul connection that passes through RRH 6 and RRH 5. Note that each RRH may have multiple incoming and outgoing fronthaul links. As it will be discussed, the information theoretic idea of in-network processing plays a key role in this scenario.

1.6.1 Problem Setting

In order to convey the quantized IQ samples from the RRHs to the BBU through multiple hops, each RRH forwards, on each outgoing fronthaul link, some information about the signals received on the wireless channel and the incoming fronthaul links. A first standard option based on point-to-point fronthaul processing is to use routing: The bits received on the incoming links are simply forwarded, along with the bit stream produced by the local quantizer/ compressor, on the outgoing links without any additional processing as illustrated in Fig. 1.11(a). 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 in [35].

Routing may be highly inefficient in the presence of a dense deployment of RRHs. In fact, under this assumption, an RRH may be close to a large number

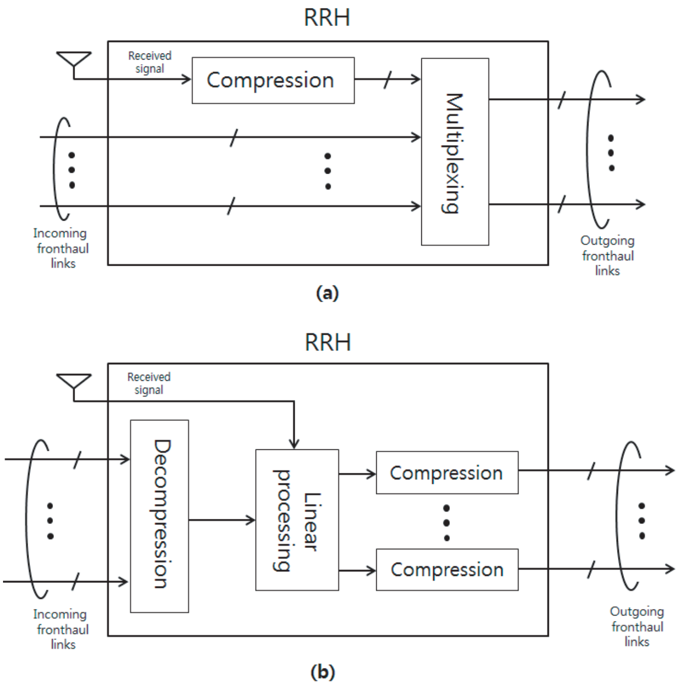

Figure 1.11 Block diagram detailing baseband processing at each RRH with (a) routing and (b) in-network processing strategies.

of other RRHs, all of which receive correlated baseband signals. In this case, it is wasteful of the available fronthaul capacity to merely forward all the bit streams received from the connected RRHs. Instead, it is possible to combine the correlated baseband signals at the RRH prior to forwarding in order to reduce redundancy. We refer to this processing of incoming signals as *in-network processing*.

1.6.2 In-Network Fronthaul Processing

A possible implementation of in-network processing based on linear operations is shown in Fig. 1.11(b). Accordingly, in order to allow for in-network processing, each RRH first decompresses the received bit streams from the connected RRHs so as to recover the baseband signals. The decompressed baseband signals are then linearly processed, along with the IQ signal received locally by the RRH. After in-network processing, the obtained signals must be recompressed before they can be sent on the outgoing fronthaul links. The effect of the quantization noise resulting from this second quantization step must thus be counterbalanced by the advantages of in-network processing in order to make the strategy prefer-

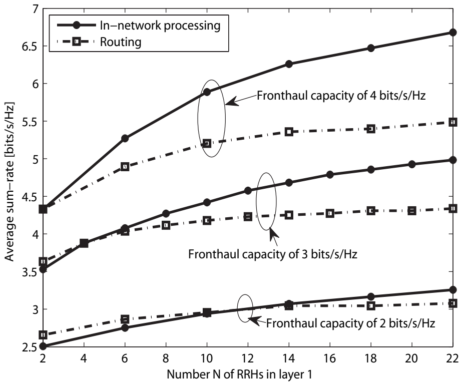

Figure 1.12 Average sum-rate versus the number of RRHs in the first layer of a two-hop topology with $N_M = 4$ UEs, average received per-antenna SNR of 20 dB and fronthaul capacity of 2, 3, 4 bits/s/Hz.

able to routing. The optimal design of in-network processing is addressed in [35].

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 RRHs in the first layer and two RRHs in the second layer. The RRHs in the first layer do not have direct fronthaul links to the BBU, while the RRHs in the second layer do. Half of the RRHs in the first layer is connected to one RRH in the second layer, and half to the other RRH 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). Fig. 1.12 shows the average sum-rate versus the number N of RRHs in the first layer with $N_M = 4$ UEs and average received per-antenna SNR of 20 dB at all RRHs. It is first observed that the performance gain of in-network processing over routing becomes more pronounced as the number N of RRHs in the first layer increases. This suggests that, as the RRHs' deployment becomes more dense, it is desirable for each RRH in the second layer to perform in-network processing of the signals received from the first layer. Moreover, in-network processing is more advantageous when the fronthaul links have a larger capacity, as the distortion introduced by the recompression step discussed above becomes smaller.

1.7 Concluding Remarks

In this chapter, we have provided an overview of the state of the art on fronthaul quantization and compression for C-RAN. We have differentiated between point-to-point, or per-fronthaul link, quantization/ compression solutions, which are generally oblivious to the network topology and state, and network-aware approaches, which instead follow network information theoretic principles and leverage the joint processing capabilities of the BBU. It was demonstrated, via various examples, that the information theoretic concepts of distributed quantization and compression, multivariate quantization and compression, and in-network processing provide useful frameworks on which to base the design of fronthaul processing techniques that are able to significantly outperform point-to-point solutions in terms of network-wide performance criteria. Interesting open problems concerning the implementation of network-aware fronthaul compression include the design of efficient feedback mechanisms on the fronthaul network aimed at satisfying the discussed channel state information requirements of this class of techniques.

In closing, we would like to mention a related technique that is also inspired by network information theory and that may play a role in the design of next-generation cellular systems based on generalizations of the C-RAN architecture, namely compute-and-forward [36]. Compute-and-forward relies on the use of nested lattice codes at the UEs, whose structure guarantees that any integer (modulo-) sum of codewords is a codeword in the same lattice codebook. Thanks to this property, in the uplink, the RRHs can decode a linear function of the uplink codewords with the aim of providing the CU with enough linear equations to recover all transmitted messages. Since the size of codebook of the possible functions to be decoded can be adapted to the fronthaul capacity, compute-and-forward does not require any quantization at the RRHs. Drawbacks of the method include an increased complexity of the RRH, which need to operate as full-fledged base stations. A version of this technique also exists for the downlink as proposed in [37]. A discussion on compute-and-forward in the context of the C-RAN architecture can be also found in [21].

Finally, we observe that many of the network information theoretic problems underlying the design of C-RANs are still open (see, e.g., [38]) and hence advances in this domain may lead to progress in the C-RAN technology.

Notes

References

[1] H. Al-Raweshidy and S. Komaki, "Radio over fiber technologies for mobile communications networks," *Artech House*, 2002.

[2] C. Lu, H. Almeida, E. Trojer, K. Laraqui, M. Berg, O. V. Tidblad, and P.-E. Eriksson, "Connecting the dots: small cells shape up for high-performance indoor radio," *Ericsson Review*, Dec. 2014.

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

[4] Fujitsu, "The benefits of cloud-ran architecture in mobile network expansion," White Paper, Fujitsu 2015.

[5] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 1, pp. 405–426, First quarter 2015.

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

[7] U. Dotsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for lte," *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, 2013.

[8] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, "Compressed transport of baseband signals in radio access networks," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 9, pp. 3216–3225, 2012.

[9] B. Guo, W. Cao, A. Tao, and D. Samardzija, "Lte/lte-a signal compression on the cpri interface," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 117–133, 2013.

[10] K. F. Nieman and B. L. Evans, "Time-domain compression of complex-baseband lte signals for cloud radio access networks," *Proc. IEEE Glob. Conf. on Sig. and Inf. Proc.*, pp. 1198–1201, Dec. 2013.

[11] M. W. A. Vosoughi and J. R. Cavallaro, "Baseband signal compression in wireless base stations," *Proc. IEEE Glob. Comm. Conf.*, Dec. 2012.

[12] D. Wubben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5g signal processing: Flexible centralization through cloud-ran," *IEEE Sig. Proc. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[13] J. Lorca and L. Cucala, "Lossless compression technique for the fronthaul of lte/lte-advanced cloud-ran architectures," *Proc. IEEE Int. Symp. World of Wireless, Mobile and Multimedia Networks*, Jun. 2013.

[14] S. B. S. Grieger and G. Fettweis, "Large scale field trial results on frequency domain compression for uplink joint detection," *Proc. IEEE Glob. Comm. Conf.*, Dec. 2012.

[15] A. E. Gamal and Y.-H. Kim, *Network Information Theory.* Cambridge University Press, 2011.

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

[17] A. Sanderovich, S. Shamai, and Y. Steinberg, "Distributed mimo receiver - achievable rates and upper bounds," *IEEE Trans. Info. Th.*, vol. 55, no. 10, p. 4419, Oct. 2009.

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

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

[20] L. Zhou and W. Yu, "Uplink multicell processing with limited backhaul via per-base-station successive interference cancellation," *IEEE Jour. Select. Areas in Comm.*, vol. 31, no. 10, pp. 2246–2254, Oct. 2013.

[21] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Fronthaul compression for cloud radio access networks," *IEEE Signal Processing Magazine, Special Issue on Signal Processing for the 5G Revolution*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[22] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Info. Th.*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[23] A. Saxena, J. Nayak, and K. Rose, "On efficient quantizer design for robust distributed source coding," in *Data Compression Conference, 2006. DCC 2006. Proceedings.* IEEE, 2006, pp. 63–72.

[24] Z. Liu, S. Cheng, A. D. Liveris, and Z. Xiong, "Slepian-wolf coded nested lattice quantization for wyner-ziv coding: High-rate performance analysis and code design," *IEEE Trans. Info. Th.*, vol. 52, no. 10, pp. 4358–4379, Oct. 2006.

[25] A. Aaron and B. Girod, "Compression with side information using turbo codes," *Proc. IEEE Data Compression Conf.*, Apr. 2002.

[26] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (discus): Design and construction," *IEEE Trans. Info. Th.*, vol. 49, no. 3, pp. 626–643, Mar. 2010.

[27] S. B. Korada and R. L. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Trans. Info. Th.*, vol. 56, no. 4, pp. 1751–1768, Apr. 2010.

[28] E. Martinian and M. J. Wainwright, "Analysis of ldgm and compound codes for lossy compression and binning," *arXiv:0602.046.*

[29] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint decompression and decoding for cloud radio access networks," *IEEE Sig. Proc. Lett.*, vol. 20, no. 5, pp. 503–506, May 2013.

[30] O. Simeone, N. Levy, A. Sanderovich, O. Somekh, B. M. Zaidel, H. V. Poor, and S. Shamai, *Cooperative wireless cellular systems: An information-theoretic view.* Foundations and Trends in Commun. Inf. Theory, 2011.

[31] "3gpp tr 36.931 ver. 9.0.0 rel. 9," no. 6, May 2011.

[32] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Performance evaluation of multiterminal backhaul compression for cloud radio access networks," *Proc. of Conf. on Info. Scien. and Systems*, Mar. 2014.

[33] R. Irmer, H. Droste, P. March, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Comm. Mag.*, Feb. 2011.

[34] 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. Sig. Proc.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[35] ——, "Multihop backhaul compression for the uplink of cloud radio access networks," *arXiv:1312.7135.*

[36] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *Information Theory, IEEE Transactions on*, vol. 57, no. 10, pp. 6463–6486, 2011.

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

[38] N. Liu and W. Kang, "A new achievability scheme for downlink multicell processing with finite backhaul capacity," in *Information Theory (ISIT), 2014 IEEE International Symposium on.* IEEE, 2014, pp. 1006–1010.