1 Fronthaul-Aware Design for Cloud Radio-Access Networks

Liang Liu, Wei Yu, and Osvaldo Simeone

1.1 Introduction

Cloud radio-access network (C-RAN) is an emerging paradigm for the fifth-generation (5G) wireless cellular network, where the traditional physical-layer base-station (BS) transmission and reception infrastructure is being virtualized using cloud computing techniques. The virtualization of wireless access also enables centralized control and management of wireless access-points, which further provides significant benefit from a transmission spectral efficiency perspective. In the current 3G/4G cellular network, each user (also called UE) is solely served by its own BS. This traditional single-cell processing paradigm shown in Fig. 1.1 suffers from severe inter-cell interference especially for cell-edge users. In the C-RAN paradigm shown in Fig. 1.2, as the BSs are coordinated centrally from the cloud, they can potentially transmit and receive radio signals to/from the users jointly, thereby creating the possibility of interference cancellation, which can significantly improve the overall network throughput.

In a C-RAN architecture, the traditional BSs essentially become remote radio heads (RRHs) that serve to relay information between the mobile users and the central processor (CP) in the cloud. Baseband processing together with its associated decoding/encoding complexities is implemented in the cloud rather than taking place locally at each BS as in traditional 3G/4G networks. As the RRHs in C-RAN require only rudimentary wireless access capabilities, they are much more cost effective to deploy, therefore allowing the C-RAN architecture to be more easily scaled geographically, leading to denser deployment of remote antennas and the ability for the network to support many more users. Furthermore, as baseband units (BBUs) are now implemented centrally at the CP, the C-RAN architecture allows the pooling of computational resources across the entire network, leading to better utilization of the computational units and higher energy efficiency for the network.

Because of both the distributed nature of wireless antenna placement and the centralized nature of cloud computing resources in C-RAN, the communication links between the RRHs and the CP are of central importance in C-RAN design.

Liang Liu and Wei Yu are with the University of Toronto, Canada. Osvaldo Simeone is with New Jersey Institute of Technology, Newark, New Jersey, USA.

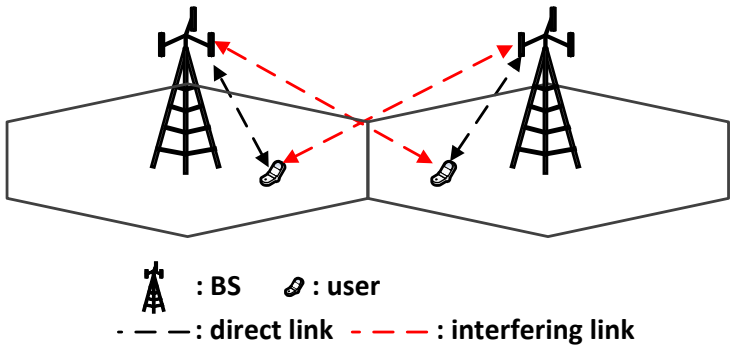

Figure 1.1 Traditional 3G/4G cellular network: each BS serves its associated users in each cell; cell-edge users suffer from severe interference.

These links are often referred to as *fronthaul* links, as they connect the radio front-end with the BBUs implemented in the cloud (in contrast to the *backhaul* links between the traditional 3G/4G BSs and the backbone network). The fronthaul links are typically implemented with fibre-optics, but they can also be implemented as wireless links, especially for pico- and femto-BSs in heterogeneous networks (HetNets) where self-backhauling is increasingly desirable.

The capacity and latency performance of the fronthaul links have significant impact on the design of C-RAN. For example, the current standardized common public radio interface (CPRI) defining the communication protocol between RRH and BBU specifies fronthaul rates ranging from 100's Mbps to 10's Gbps. When multiple RRH's are aggregated, the deluge of data required to be transported between the RRHs and the cloud can easily overwhelm the physical limitation of practical fronthaul implementations. Furthermore, as the C-RAN architecture now allows the BBUs to be physically located much further away from RRHs, the ensuing latency would have a significant impact on the overall delay performance of the network.

This chapter aims to illustrate that the physical and data link layer design of a C-RAN system must adapt to the capacity and latency limitations of fronthaul links. The first part of the chapter provides an information theoretical evaluation of achievable rates of C-RAN with the impact of finite-capacity fronthaul links taken into account. Toward this end, a practical user-RRH clustering strategy as well as various fronthaul techniques for implementing uplink and downlink beamforming in C-RAN are considered; their achievable rates subject to the fronthaul capacity constraints are evaluated. In the second part of the chapter, the effect of fronthaul latency on the throughput and efficiency of data link layer is discussed. Toward this end, a novel design of Hybrid Automatic Repeat Request (hybrid ARQ or HARQ) protocol is proposed to circumvent the additional delay caused by the multihop topology of C-RAN.

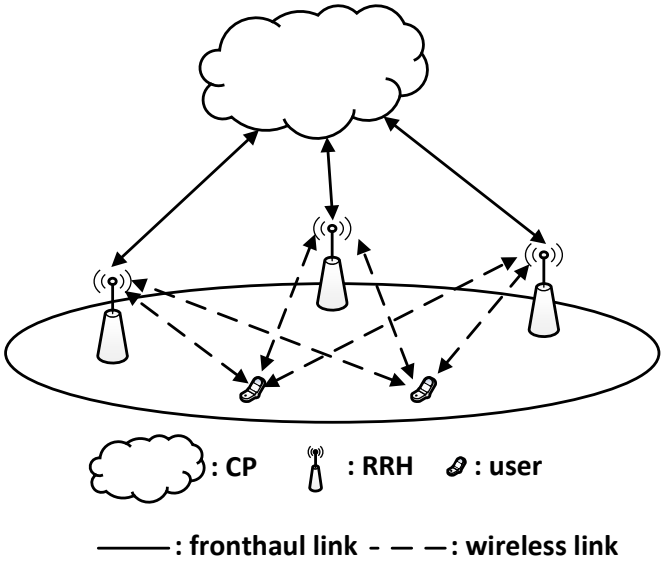

Figure 1.2 C-RAN: the RRHs serve the users under the coordination of the CP via finite-capacity fronthaul links; intercell interference can be effectively mitigated.

1.2 Fronthaul-Aware Cooperative Transmission and Reception

A key benefit of the C-RAN architecture as compared to traditional single-cell processing is that it enables cooperative transmission and reception across multiple RRHs via *beamforming*. This section illustrates beamforming design techniques for both uplink and downlink C-RAN and characterizes the theoretical achievable rates of a C-RAN deployment as functions of the fronthaul capacity constraints.

Consider a typical C-RAN deployment as depicted in Fig. 1.2, where a cluster of RRHs each equipped with multiple antennas cooperatively serve multiple single-antenna users under the coordination of the CP via finite-capacity fronthaul links. The fronthaul links are modeled as finite-capacity noiseless digital links with some fixed capacity for each link. We remark that although the analog optical modulation using the radio-over-fiber technique is also an alternative, the digital fronthaul model is adopted here, because it is considerably easier to implement in practice.

To illustrate the benefit of cooperation in C-RAN, the following network model is adopted in this section. Assuming a network with N RRHs each equipped with M antennas, each user is associated with its strongest RRH. Among its associated users, each RRH schedules $K < M$ users at each time slot for service. Furthermore, to achieve a cooperation gain, each scheduled user is jointly served by a cooperative cluster of RRHs. Specifically, in the uplink, each RRH forwards its received signal to the CP over its fronthaul link. The CP then decodes each

user's message based on the signals received from the RRHs in the cooperative cluster of the user. In the downlink, with coordination from the CP, the transmit signal of each RRH is designed as function of the messages of users whose cooperation cluster includes the RRH. This enables joint transmission across the cluster of RRHs to each user. Note that the size of the cooperation cluster depends on the ability of acquiring the channel state information (CSI) between the RRHs and the users. For simplicity, the cluster size is assumed to be fixed in this section.

This section assumes a *user-centric* clustering strategy in which each user is always placed at the center of its cooperation cluster, but the clusters for different users may overlap. As compared to disjoint clustering (which partitions the entire network into disjoint sets of cooperating RRHs), user-centric clustering has the advantage that it completely eliminates cluster edges, hence resulting in better fairness in rate distribution across the users [1].

The remainder of this section describes a particular zero-forcing (ZF) beamforming strategy across the user-centric clusters using various fronthaul techniques for both uplink and downlink C-RAN. Notationally, let $\mathcal{K}$ denote the set of scheduled users in a particular timeslot, and let Θ_k denote the cooperative cluster of RRHs for user $k \in \mathcal{K}$, with $D_k = |\Theta_k|$ being the cluster size. As each RRH in Θ_k schedules K active users, a sensible strategy is to zero-force all the interference due to the signals of the scheduled users in the cluster. We use Ω_k to denote the set of users scheduled by all the RRHs in Θ_k (i.e., the cluster for user k), so that $|\Omega_k| = KD_k$. Finally, from the RRHs' perspective, it is convenient to define the set of all users whose signals are zero-forced by RRH n as Φ_n, i.e., $\Phi_n = \{k : n \in \Theta_k, k = 1, \cdots, K\}$. Throughout the section, the wireless channels between the RRHs and the users are assumed to be quasi-static flat-fading channels over a fixed bandwidth of B Hz. The fronthaul capacity between each RRH and the CP is assumed to be C bits per second (bps).

1.2.1 Uplink

In the uplink C-RAN, at each time slot each RRH's observed signal is a superposition of the signals sent from all the scheduled users in the set $\mathcal{K}$. Specifically, let x_k^{ul} denote the transmit signal of user k, and $\mathbf{y}_n^{\text{ul}} \in \mathbb{C}^{M\times 1}$ denote the received signal at RRH n, then

$$\mathbf{y}_n^{\text{ul}} = \sum_{k\in\mathcal{K}} \mathbf{h}_{n,k}^{\text{ul}} x_k^{\text{ul}} + \mathbf{z}_n^{\text{ul}}, \quad \forall n, \tag{1.1}$$

where $\mathbf{h}_{n,k}^{\text{ul}} \in \mathbb{C}^{M\times 1}$ is the uplink channel from user k to RRH n, and $\mathbf{z}_n^{\text{ul}} \in \mathbb{C}^{M\times 1} \sim \mathcal{CN}(\mathbf{0}, \sigma_{\text{ul}}^2\mathbf{I})$ denotes the additive white Gaussian noise (AWGN).

Observe that the received signal of each RRH contains useful information even for the users that are not associated with it. However, the single-cell processing mechanism of the current 3G/4G cellular networks cannot take advantage of this information, as it restricts decoding to be done locally at each BS for its own

associated users. To make the best utilization of the received signals in the entire network, the RRHs in C-RAN should relay their observed signals to the CP over the fronthaul links, so that the message of each user can be decoded by the CP based on the observations of all the RRHs serving this user.

If the fronthaul links have infinite capacities, each RRH can perfectly convey its observed signal to the CP, and the full joint decoding gain can be achieved. In practice, however, the fronthaul links have finite capacities, thus each RRH can only convey an approximate version of its received signal. An interesting but essential question arises: What is the appropriate way for the RRHs to preserve as much information as possible in relaying their observations to the CP, while satisfying the finite-capacity constraints of the fronthaul links?

The basic strategy is that the RRHs should compress its received signal. Below we describe two uplink fronthaul compression techniques. When the C-RAN system is fully loaded, i.e., when the spatial dimensions at the RRHs are fully occupied by all the users, a *compress-forward* strategy [2, 3] works very well. When the C-RAN is lightly loaded, i.e., when there are excess spatial dimensions as compared to number of active users, it may be advantageous to employ a *beamform-compress-forward* strategy [4, 3].

Compress-Forward Strategy: In the compress-forward strategy, the RRHs first downconvert their received RF signals to the baseband signals, which are analog in nature, then compress the baseband signals and send the corresponding compression indices, which are represented by digital codewords, to the CP. After receiving the compression indices, the CP first decompresses these quantized signals in order to recover a distorted version of the received signal across all the RRHs, then decodes the user messages based on the entire decompressed signals.

Intuitively, the compression resolution is determined by the available fronthaul capacity, i.e., a more stringent fronthaul capacity constraint would imply "coarser compression", which in rate-distortion theory is reflected as larger quantization noise. We note that the optimal compression in the C-RAN setting would involve *vector quantization* across the antennas and *Wyner-Ziv compression* across the RRHs, which are techniques capable of taking advantage of the fact that the received signals across the antennas and across the RRHs are correlated. But for simplicity, the model in the rest of this section assumes scalar quantization modeled by independent additive Gaussian quantization noise.

With independent compression across RRHs and scalar quantization across antennas at each RRH, we can now describe the compress-forward strategy in uplink C-RAN as follows. For simplicity, we assume that all the users transmit with an identical power denoted by p_u so that the transmit signal of user k is expressed as $x_k^{\text{ul}} = \sqrt{p_u} s_k^{\text{ul}}$, where $s_k^{\text{ul}} \sim \mathcal{CN}(0,1)$ denotes the message for user k chosen from a Gaussian codebook. With the channel model as described in (1.1), the discrete-time baseband signal received by RRH n is given by

$$\mathbf{y}_n^{\text{ul}} = \sum_{k \in \mathcal{K}} \sqrt{p_u} \mathbf{h}_{n,k}^{\text{ul}} s_k^{\text{ul}} + \mathbf{z}_n^{\text{ul}}, \quad \forall n. \tag{1.2}$$

The scalar quantization process at the mth antenna of the nth RRH is modeled as a Gaussian test channel with the uncompressed signal as the input and compressed signal as the output, i.e.,

$$\tilde{y}^{\mathrm{ul}}_{n,m} = y^{\mathrm{ul}}_{n,m} + e^{\mathrm{ul}}_{n,m} = \sum_{k\in\mathcal{K}} \sqrt{p_u} h^{\mathrm{ul}}_{n,m,k} s^{\mathrm{ul}}_k + z^{\mathrm{ul}}_{n,m} + e^{\mathrm{ul}}_{n,m}, \quad \forall n, m, \tag{1.3}$$

where $h^{\mathrm{ul}}_{n,m,k}$ is the channel from user k to the mth antenna of RRH n, and $z^{\mathrm{ul}}_{n,m}$ is the Gaussian noise at the mth antenna of RRH n, further $e^{\mathrm{ul}}_{n,m} \sim \mathcal{CN}(0, q^{\mathrm{ul}}_{n,m})$ denotes the quantization noise in compressing $y^{\mathrm{ul}}_{n,m}$, and $q^{\mathrm{ul}}_{n,m}$ denotes its variance. Note that since independent compression is employed across RRHs and scalar quantization is employed at each RRH, the quantization noises $e^{\mathrm{ul}}_{n,m}$ are independent over the RRHs and the antennas.

With the above Gaussian test channel model, the design of the compression codebook is equivalent to setting the variances of the compression noise. To achieve higher compression resolution, which leads to higher achievable rates in the uplink C-RAN, the quantization noise should be made as small as possible at each RRH. However, the minimum amount of quantization noise is also limited by the fronthaul capacity, as given by rate-distortion theory. In practice, assuming a gap Γ_q to the rate-distortion limit, the fronthaul capacity in bps required to transmit $\tilde{y}^{\mathrm{ul}}_{n,m}$ to the CP can be expressed as

$$C^{\mathrm{ul}}_{n,m} = B \log_2 \left(\frac{\Gamma_q \left(\sum_{k\in\mathcal{K}} p_u |h^{\mathrm{ul}}_{n,m,k}|^2 + \sigma^2_{\mathrm{ul}} \right) + q^{\mathrm{ul}}_{n,m}}{q^{\mathrm{ul}}_{n,m}} \right), \ \forall m, n. \tag{1.4}$$

For simplicity, the total fronthaul capacity of RRH n is assumed to be equally allocated to its M antennas, i.e., $C^{\mathrm{ul}}_{n,m} = C/M$, $\forall m$. From (1.4), the variance of the quantization noise for compression $y^{\mathrm{ul}}_{n,m}$ is then given by

$$q^{\mathrm{ul}}_{n,m} = \frac{\Gamma_q \left(\sum_{k\in\mathcal{K}} p_u |h^{\mathrm{ul}}_{n,m,k}|^2 + \sigma^2_{\mathrm{ul}} \right)}{2^{\frac{C}{BM}} - 1}, \quad \forall n, m. \tag{1.5}$$

This allows us to derive the achievable rate of each user as follows. The CP decodes the message of user k based on the signals sent from its serving RRHs in the cooperation cluster, i.e., the set Θ_k. Denote the received signal across Θ_k as $\tilde{\mathbf{y}}^{\mathrm{ul},k} = [\cdots, \tilde{y}^{\mathrm{ul}}_{n,1}, \cdots, \tilde{y}^{\mathrm{ul}}_{n,M}, \cdots]^T_{n\in\Theta_k} \in \mathbb{C}^{MD_k\times 1}$, $\forall k$. For convenience, define $\mathbf{g}^{\mathrm{ul}}_{k,i} = [\cdots, (\mathbf{h}^{\mathrm{ul}}_{n,i})^T, \cdots]^T_{n\in\Theta_k} \in \mathbb{C}^{MD_k\times 1}$ as the collective channel vector from user i to the RRHs in Θ_k. Then,

$$\tilde{\mathbf{y}}^{\mathrm{ul},k} = \underbrace{\sqrt{p_u}\mathbf{g}^{\mathrm{ul}}_{k,k} s^{\mathrm{ul}}_k}_{\text{desired signal}} + \underbrace{\sum_{i\neq k, i\in\Omega_k} \sqrt{p_u}\mathbf{g}^{\mathrm{ul}}_{k,i} s^{\mathrm{ul}}_i}_{\text{intra-cluster interference}} + \underbrace{\sum_{j\notin\Omega_k} \sqrt{p_u}\mathbf{g}^{\mathrm{ul}}_{k,j} s^{\mathrm{ul}}_j}_{\text{inter-cluster interference}} + \bar{\mathbf{z}}^{\mathrm{ul}}_k + \bar{\mathbf{e}}^{\mathrm{ul}}_k, \ \forall k \tag{1.6}$$

where $\bar{\mathbf{z}}_k^{\mathrm{ul}}$ and $\bar{\mathbf{e}}_k^{\mathrm{ul}}$ are the collective AWGN and quantization noises across the RRHs in Θ_k, with covariances $\mathbf{S}_{\bar{\mathbf{z}},k}^{\mathrm{ul}} = \mathbb{E}[\bar{\mathbf{z}}_k^{\mathrm{ul}}(\bar{\mathbf{z}}_k^{\mathrm{ul}})^H] = \sigma_{\mathrm{ul}}^2\mathbf{I}$ and $\mathbf{S}_{\bar{\mathbf{e}},k}^{\mathrm{ul}} = \mathbb{E}[\bar{\mathbf{e}}_k^{\mathrm{ul}}(\bar{\mathbf{e}}_k^{\mathrm{ul}})^H] = \mathrm{diag}([\cdots, q_{n,1}^{\mathrm{ul}}, \cdots, q_{n,M}^{\mathrm{ul}}, \cdots]_{n\in\Theta_k}^T)$, respectively.

The CP applies a linear beamformer $\mathbf{w}_k^{\mathrm{ul}} \in \mathbb{C}^{MD_k\times 1}$ with unit norm to $\tilde{\mathbf{y}}^{\mathrm{ul},k}$ for decoding s_k^{ul}:

$$\begin{aligned}\hat{s}_k^{\mathrm{ul}} =& \sqrt{p_u}(\mathbf{w}_k^{\mathrm{ul}})^H\mathbf{g}_{k,k}^{\mathrm{ul}}s_k^{\mathrm{ul}} + \sum_{i\neq k, i\in\Omega_k}\sqrt{p_u}(\mathbf{w}_k^{\mathrm{ul}})^H\mathbf{g}_{k,i}^{\mathrm{ul}}s_i^{\mathrm{ul}} \\ &+ \sum_{j\notin\Omega_k}\sqrt{p_u}(\mathbf{w}_k^{\mathrm{ul}})^H\mathbf{g}_{k,j}^{\mathrm{ul}}s_j^{\mathrm{ul}} + (\mathbf{w}_k^{\mathrm{ul}})^H\bar{\mathbf{z}}_k^{\mathrm{ul}} + (\mathbf{w}_k^{\mathrm{ul}})^H\bar{\mathbf{e}}_k^{\mathrm{ul}}, \quad \forall k. \end{aligned} \tag{1.7}$$

As a practical choice of $\mathbf{w}_k^{\mathrm{ul}}$, this section considers the ZF beamforming technique, where the intra-cluster interference due to users in Ω_k is completely eliminated, i.e., $(\mathbf{w}_k^{\mathrm{ul}})^T\mathbf{g}_{k,i}^{\mathrm{ul}} = 0$, $\forall i \in \Omega_k$ and $i \neq k$. To achieve this goal, the following ZF beamforming vectors can be utilized:

$$\mathbf{w}_k^{\mathrm{ul}} = \frac{(\mathbf{I} - \mathbf{G}_{-k}^{\mathrm{ul}}(\mathbf{G}_{-k}^{\mathrm{ul}})^\dagger)\mathbf{g}_{k,k}^{\mathrm{ul}}}{\|(\mathbf{I} - \mathbf{G}_{-k}^{\mathrm{ul}}(\mathbf{G}_{-k}^{\mathrm{ul}})^\dagger)\mathbf{g}_{k,k}^{\mathrm{ul}}\|_2}, \quad \forall k, \tag{1.8}$$

where $\mathbf{G}_{-k}^{\mathrm{ul}} = [\cdots, \mathbf{g}_{k,i}^{\mathrm{ul}}, \cdots]_{i\neq k, i\in\Omega_k} \in \mathbb{C}^{MD_k\times(KD_k-1)}$ denotes the collection of channels from the intra-cluster users (excluding user k) to the RRHs serving user k, and $(\mathbf{G}_{-k}^{\mathrm{ul}})^\dagger$ denotes the pseudo-inverse of $\mathbf{G}_{-k}^{\mathrm{ul}}$.

The achievable rate of user k under the above compress-forward scheme can now be characterized as

$$r_k^{\mathrm{ul,CF}} = B\log_2\left(1 + \frac{p_u\left|(\mathbf{w}_k^{\mathrm{ul}})^H\mathbf{g}_{k,k}^{\mathrm{ul}}\right|^2}{\left(\sum_{j\neq\Omega_k} p_u\left|(\mathbf{w}_k^{\mathrm{ul}})^H\mathbf{g}_{k,j}^{\mathrm{ul}}\right|^2 + \sigma_{\mathrm{ul}}^2 + (\mathbf{w}_k^{\mathrm{ul}})^H\mathbf{S}_{\bar{\mathbf{e}},k}^{\mathrm{ul}}\mathbf{w}_k^{\mathrm{ul}}\right)\Gamma_s}\right), \tag{1.9}$$

where Γ_s is the signal-to-interference-plus-noise ratio (SINR) gap due to the coding and modulation scheme used in practice.

Beamform-Compress-Forward Strategy: The compress-forward strategy discussed above compresses the received signals at each antenna independently; the fronthaul capacity is shared across all the antennas. This compression strategy can be shown to be close to optimal when the system is fully loaded (i.e. it schedules as many users as there are antennas), and operates at high signal-to-quantization-noise ratio (SQNR), and further if equal quantization noise level, rather than equal allocation of quantization bits, is applied across the antennas [3]. In general, however, it may not be the most efficient use of the limited fronthaul especially for systems with many more antennas at the RRHs than the number of scheduled users. This is an increasingly possible scenario with the emerging massive multiple-input multiple-output (MIMO) technology. To address this issue, in the following we propose a beamform-compress-forward

scheme, where each RRH first performs beamforming of its received signals across its antennas followed by compression in a reduced dimensional space.

Specifically, the beamforming operation applied by RRH n to its received signal $\mathbf{y}_n^{\text{ul}}$ given in (1.2) can be modeled as

$$\hat{\mathbf{y}}_n^{\text{ul}} = \mathbf{V}_n^{\text{ul}} \mathbf{y}_n^{\text{ul}} = \sum_{k \in \mathcal{K}} \sqrt{p_u} \mathbf{V}_n^{\text{ul}} \mathbf{h}_{n,k}^{\text{ul}} s_k^{\text{ul}} + \mathbf{V}_n^{\text{ul}} \mathbf{z}_n^{\text{ul}}, \quad \forall n, \tag{1.10}$$

where $\mathbf{V}_n^{\text{ul}} = [\mathbf{v}_{n,1}^{\text{ul}}, \cdots, \mathbf{v}_{n,L_n}^{\text{ul}}]^T \in \mathbb{C}^{L_n \times M}$ denotes the beamforming matrix at RRH n with $L_n \leq M$ denoting the (reduced) dimension of the output signal $\hat{\mathbf{y}}_n^{\text{ul}}$ after beamforming. The beamformers $\mathbf{V}_n^{\text{ul}}$ at the RRHs essentially transform the effective channel between the users and the RRHs from $\mathbf{h}_{n,k}^{\text{ul}} \in \mathbb{C}^{M \times 1}$ to $\mathbf{V}_n^{\text{ul}} \mathbf{h}_{n,k}^{\text{ul}} \in \mathbb{C}^{L_n \times 1}$, $\forall k, n$, while transforming the effective noise at the RRHs from $\mathbf{z}_n^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \sigma_{\text{ul}}^2 \mathbf{I})$ to $\mathbf{V}_n^{\text{ul}} \mathbf{z}_n^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \sigma_{\text{ul}}^2 \mathbf{V}_n^{\text{ul}} (\mathbf{V}_n^{\text{ul}})^H)$, $\forall n$. As a result, the aforementioned compress-forward strategy can be applied to the new uplink C-RAN channel model given in (1.10), and the achievable rate of user k under the beamform-compress-forward scheme can be similarly derived as in (1.9) with the new effective channels and noises. The remaining question is how to determine the beamforming vectors at each RRH to maximize the achievable rates.

The above question can be answered through an optimization framework involving quantization noise covariance matrices across the RRHs [3], but such an optimization is complex and it assumes vector quantization across the antennas. Below, we offer an heuristic approach of first determining L_n, the dimension of the beamforming matrix at each RRH n, then finding a beamformer through identifying the principal component of the received signal. The proposed heuristic is to set

$$L_n = \min(|\Phi_n|, M) \tag{1.11}$$

at each RRH n. Recall that Φ_n is the set of users served by each RRH. Thus, for a lightly loaded C-RAN system in which the number of users being served is less than the number of antennas, each RRH n compacts its received signal $\mathbf{y}_n^{\text{ul}}$ into $|\Phi_n|$ dimensions to preserve the useful information for its served users.

Next, to extract the L_n most informative dimensions, we perform a singular-value decomposition on the covariance matrix of the received signal of RRH n as $\mathbf{S}_{\mathbf{y},n}^{\text{ul}} = \mathbf{U}_n \Lambda_n \mathbf{U}_n^H$, where

$$\mathbf{S}_{\mathbf{y},n}^{\text{ul}} = \mathbb{E}[\mathbf{y}_n^{\text{ul}} (\mathbf{y}_n^{\text{ul}})^H] = \sum_{k \in \mathcal{K}} p_u \mathbf{h}_{n,k}^{\text{ul}} (\mathbf{h}_{n,k}^{\text{ul}})^H + \sigma_{\text{ul}}^2 \mathbf{I}, \quad \forall n. \tag{1.12}$$

To access information in the strongest subspace associated with the largest eigenvalues, the beamforming matrix adopted by RRH n is chosen as a collection of the first L_n dominant eigenvectors of $\mathbf{S}_{\mathbf{y},n}^{\text{ul}}$ in $\mathbf{U}_n$, i.e.,

$$\mathbf{V}_n^{\text{ul}} = [\mathbf{u}_{n,1}, \cdots, \mathbf{u}_{n,L_n}]^H, \quad \forall n. \tag{1.13}$$

Table 1.1. System Parameters of the Numerical Example

Channel Bandwidth	20 MHz
Distance between Cell Sites	0.8 km
Number of RRHs per Cell	3
Number of Antennas per RRH (M)	12
Number of Scheduled Users per RRH	2
User Transmit Power (p_u)	23 dBm
Antenna Gain	15 dBi
Path Loss Model	$140.7 + 36.7 \log_{10}(d)$ dB
Log-Normal Shadowing	8 dB
Rayleigh Small Scale Fading	0 dB
SINR Gap (Γ_s)	6 dB
Rate-Distortion Gap (Γ_q)	4.3 dB
AWGN Power Spectrum Density	-169 dBm/Hz

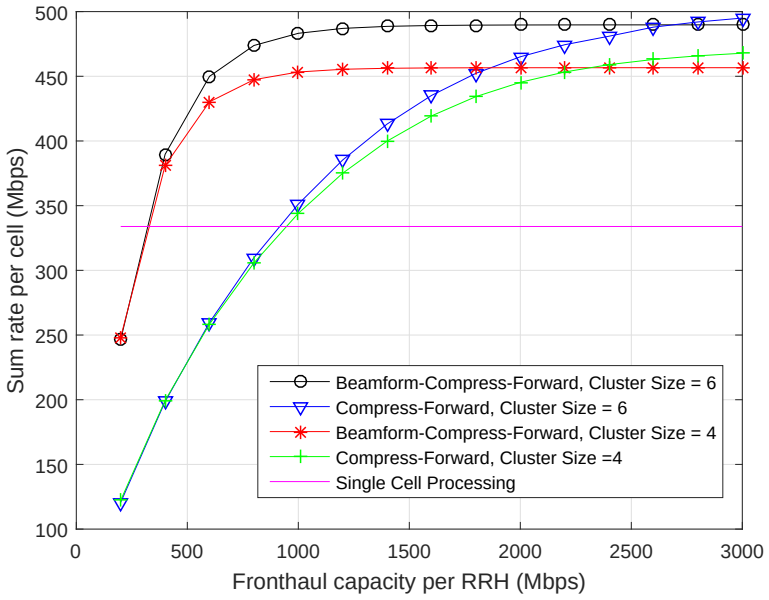

Figure 1.3 Performance comparison between the compress-forward and beamform-compress-forward strategies in the uplink C-RAN.

As mentioned above, the achievable rates of the beamform-compress-forward scheme can now be obtained in the same manner as for the compress-forward scheme, but with the new effective channels $\mathbf{V}_n^{\mathrm{ul}}\mathbf{h}_{n,k}^{\mathrm{ul}}$ and effective noises $\mathbf{V}_n^{\mathrm{ul}}\mathbf{z}_n^{\mathrm{ul}}$.
Performance Evaluation: To compare the performance of the beamform-compress-forward strategy and the compress-forward strategy, we present a numerical example of a 19-cell wrapped-around cellular network simulation

topology with the parameters listed in Table 1.1. Each cell is sectorized using 3 RRHs. The per-cell sum rate as a function of per-RRH fronthaul capacities with different cluster sizes is shown in Fig. 1.3.

In this example, each RRH is equipped with 12 antennas but serves only 2 users in each timeslot. Thus, the system is not fully loaded; beamform-compress-forward is expected to show advantage as compared to compress-forward. Indeed, at moderate fronthaul capacity, a sum-rate gain of $20\% - 50\%$ is observed, due to the fact that beamform-compress provides better utilization of the fronthaul. The difference between the two strategies diminishes for large fronthaul capacities, because the quantization noise is no longer the limiting factor.

It is of interest to note that the C-RAN system significantly outperforms the single-cell processing baseline. Furthermore, the amount of fronthaul capacity required in order to reap the full benefit of uplink C-RAN is about six times the access rate using beamform-compress-forward, making the practical implementation of C-RAN feasible.

1.2.2 Downlink

In the downlink C-RAN, each user's observed signal is a superposition of the signals sent from all the RRHs. Specifically, let $\mathbf{x}_n^{\mathrm{dl}} \in \mathbb{C}^{M\times 1}$ denote the transmit signal of RRH n, and y_n^{dl} denote the received signal at RRH n. Then, the received signal at user k can be modeled as

$$y_k^{\mathrm{dl}} = \sum_{n=1}^{N} (\mathbf{h}_{k,n}^{\mathrm{dl}})^H \mathbf{x}_n^{\mathrm{dl}} + z_k^{\mathrm{dl}}, \quad \forall k, \tag{1.14}$$

where $\mathbf{h}_{k,n}^{\mathrm{dl}} \in \mathbb{C}^{M\times 1}$ is the downlink channel from RRH n to user k, and $z_k^{\mathrm{dl}} \sim \mathcal{CN}(0, \sigma_{\mathrm{dl}}^2)$ denotes the AWGN at user k.

In the current 3G/4G cellular network, each scheduled user is served by one BS and sees interference from all neighboring BSs. The benefit of the C-RAN architecture arises from the ability of multiple RRHs to cooperatively serve users, thereby minimizing the effect of undesired interference. As the messages intended for different users in C-RAN all originate from the CP, the CP can relay useful information about the user messages to the RRHs via the fronthaul links, thus allowing the RRHs to perform network-wide beamforming in order to achieve cooperative transmission.

If the fronthaul links have infinite capacities, the CP can perfectly convey the data of all the users in the C-RAN to each RRH, thus achieving full cooperation. With finite-capacity fronthaul links, however, the CP can only send a limited amount of information to each RRH. As a result, a key task for the CP is to convey information about the user messages to the RRHs in the most succinct form in order to enable as much interference cancellation as possible.

One possibility is to use a *compression-based* strategy [5], which can be thought of as the dual operation of the compress-forward strategy used in the uplink. The

idea is to pre-form the cooperative beamformed signals to be transmitted by the RRHs at the CP. The analog signals to be transmitted on the antennas are then compressed and sent digitally to the corresponding RRHs via the fronthaul links for cooperative transmission. As a simpler alternative, the CP may opt to share the user messages directly with the RRHs via fronthaul links, leading to the *data-sharing* strategy [6, 7]. With user messages at hand, the RRHs can beamform their transmit signals on their own then transmit to the users. In the following, we quantify the achievable rates and the fronthaul requirements in the downlink C-RAN using the compression-based and data-sharing strategies, respectively.

Compress-Forward Strategy: Under the compression-based strategy, the transmit signal of each RRH n, i.e., $\mathbf{x}_n^{\mathrm{dl}}$, is a compressed version of the beamformed signal at the CP, denoted by $\tilde{\mathbf{x}}_n^{\mathrm{dl}}$. Specifically, the CP first forms the beamformed transmit signal for each RRH n as follows:

$$\tilde{\mathbf{x}}_n^{\mathrm{dl}} = \sum_{k\in\Phi_n} \mathbf{w}_{n,k}^{\mathrm{dl}} \sqrt{p_b} s_k^{\mathrm{dl}}, \quad \forall n, \tag{1.15}$$

where $s_k^{\mathrm{dl}} \sim \mathcal{CN}(0,1)$ denotes the message intended for user k chosen from the Gaussian codebook, $\mathbf{w}_{n,k}^{\mathrm{dl}} = [w_{n,1,k}^{\mathrm{dl}}, \cdots, w_{n,M,k}^{\mathrm{dl}}]^T \in \mathbb{C}^{M\times 1}$ denotes the beamforming vector at RRH n for user k so that the overall unit-norm beamformer for user k across all of its serving RRHs is $\mathbf{w}_k^{\mathrm{dl}} = [\cdots, (\mathbf{w}_{n,k}^{\mathrm{dl}})^T, \cdots]_{n\in\Theta_k}^T$, and p_b is the power of the beamformers assumed to be identical across all the $\mathbf{w}_k^{\mathrm{dl}}$'s.

As in the uplink, we assume here that the CP applies scalar quantization to compress each component of the beamformed signals independently, (although we note here multivariate compression across the RRHs is also possible [5] although with much higher complexity). The compression process is modeled as a Gaussian test channel with independent additive Gaussian quantization noise. As a result, the transmit signal from the mth antenna of RRH n is given by:

$$x_{n,m}^{\mathrm{dl}} = \tilde{x}_{n,m}^{\mathrm{dl}} + e_{n,m}^{\mathrm{dl}} = \sum_{k\in\Phi_n} w_{n,m,k}^{\mathrm{dl}} \sqrt{p_b} s_k^{\mathrm{dl}} + e_{n,m}^{\mathrm{dl}}, \quad \forall n, m, \tag{1.16}$$

where $\tilde{x}_{n,m}^{\mathrm{dl}}$ denotes the mth component of $\tilde{\mathbf{x}}_n^{\mathrm{dl}}$, $e_{n,m}^{\mathrm{dl}} \sim \mathcal{CN}(0, q_{n,m}^{\mathrm{dl}})$ denotes the quantization noise for quantizing $\tilde{x}_{n,m}^{\mathrm{dl}}$, and $q_{n,m}^{\mathrm{dl}}$ denotes the variance of the quantization noise. Note that the $x_{n,m}^{\mathrm{dl}}$ are transmitted from the CP to the corresponding RRHs as quantization indices in digital format. By rate-distortion theory, the fronthaul capacity in terms of bps required for sending $x_{n,m}^{\mathrm{dl}}$ is given by

$$C_{n,m}^{\mathrm{dl}} = B \log_2 \left(\frac{\Gamma_q \sum\limits_{k\in\Phi_n} p_b |w_{n,m,k}^{\mathrm{dl}}|^2 + q_{n,m}^{\mathrm{dl}}}{q_{n,m}^{\mathrm{dl}}} \right), \quad \forall n, m. \tag{1.17}$$

If, for simplicity, we further assume that the fronthaul capacity is evenly allocated to all the antennas of each RRH, i.e., $C_{n,m}^{\mathrm{dl}} = C/M$, $\forall n, m$, the quantization noise resulting from compressing the transmit signal of the mth antenna of RRH n is

given by

$$q_{n,m}^{\mathrm{dl}} = \frac{\Gamma_q \sum\limits_{k\in\Phi_n} p_b |w_{n,m,k}^{\mathrm{dl}}|^2}{2^{\frac{C}{MB}} - 1}, \quad \forall n, m. \tag{1.18}$$

Next, we discuss the choice of the per-beam transmit power p_b so that the transmit power constraint is satisfied. For convenience, we assume a sum-power constraint for all the RRHs, denoted by P_R. According to (1.16) and (1.18), the total transmit power across all the RRHs is given by

$$p^{\mathrm{dl}} = \sum_{n=1}^{N} \mathbb{E}\|\mathbf{x}_n^{\mathrm{dl}}\|^2 = \frac{2^{\frac{C}{MB}} - 1 + \Gamma_q}{2^{\frac{C}{MB}} - 1} \cdot \sum_{k\in\mathcal{K}} p_b. \tag{1.19}$$

By setting the total transmit power to be equal to the sum-power constraint, i.e., $p^{\mathrm{dl}} = P_R$, the per-beam transmit power p_b is given by

$$p_b = \frac{(2^{\frac{C}{MB}} - 1)P_R}{(2^{\frac{C}{MB}} - 1 + \Gamma_q)|\mathcal{K}|}. \tag{1.20}$$

Lastly, we quantify the achievable rates of the downlink C-RAN under the above compression-based strategy. For convenience, define $\mathbf{g}_{k,i}^{\mathrm{dl}} = [\cdots, (\mathbf{h}_{k,n}^{\mathrm{dl}})^T, \cdots]_{n\in\Theta_i}^T \in \mathbb{C}^{MD_k\times 1}$ as the collective channel from user i's serving set of RRHs (i.e., Θ_i) to user k. Then, the received signal at user k given in (1.14) can be expressed as

$$\begin{aligned} y_k^{\mathrm{dl}} = &\underbrace{(\mathbf{g}_{k,k}^{\mathrm{dl}})^H \mathbf{w}_k^{\mathrm{dl}} \sqrt{p_b} s_k^{\mathrm{dl}}}_{\text{desired signal}} + \underbrace{\sum_{i\neq k, i\in\Omega_k} (\mathbf{g}_{k,i}^{\mathrm{dl}})^H \mathbf{w}_i^{\mathrm{dl}} \sqrt{p_b} s_i^{\mathrm{dl}}}_{\text{intra-cluster interference}} \\ &+ \underbrace{\sum_{j\notin\Omega_k} (\mathbf{g}_{k,j}^{\mathrm{dl}})^H \mathbf{w}_j^{\mathrm{dl}} \sqrt{p_b} s_j^{\mathrm{dl}}}_{\text{inter-cluster interference}} + z_k^{\mathrm{dl}} + \sum_{n=1}^{N} (\mathbf{h}_{k,n}^{\mathrm{dl}})^H \mathbf{e}_n^{\mathrm{dl}}, \quad \forall k, \end{aligned} \tag{1.21}$$

where $\mathbf{e}_n^{\mathrm{dl}} = [e_{n,1}^{\mathrm{dl}}, \cdots, e_{n,M}^{\mathrm{dl}}]^T$ denotes the collective quantization noise for compressing the signal at RRH n across its M antennas. Since scalar quantization is applied, the covariance matrix of $\mathbf{e}_n^{\mathrm{dl}}$ is diagonal, i.e., $\mathbf{S}_{\mathbf{e},n}^{\mathrm{dl}} = \mathbb{E}[\mathbf{e}_n^{\mathrm{dl}}(\mathbf{e}_n^{\mathrm{dl}})^H] = \mathrm{diag}(q_{n,1}^{\mathrm{dl}}, \cdots, q_{n,M}^{\mathrm{dl}})$.

As in the uplink, we apply ZF beamforming so that the downlink transmit beamforming vectors $\mathbf{w}_k^{\mathrm{dl}}$ are designed to completely cancel the intra-cluster interference term in (1.21), i.e., $(\mathbf{g}_{i,k}^{\mathrm{dl}})^H \mathbf{w}_k^{\mathrm{dl}} = 0$, $\forall k \in \Omega_i$. Specifically, define $\mathbf{G}_{-k}^{\mathrm{dl}} = [\cdots, \mathbf{g}_{i,k}^{\mathrm{dl}}, \cdots]_{i\neq k, k\in\Omega_i}$ as the collection of channel vectors from all the RRHs serving user k to its intra-cluster users (excluding user k). Similar to the uplink ZF beamforming design given in (1.8), the downlink beamforming vectors to zero-force the intra-cluster interference can be obtained as follows:

$$\mathbf{w}_k^{\mathrm{dl}} = \frac{(\mathbf{I} - \mathbf{G}_{-k}^{\mathrm{dl}}(\mathbf{G}_{-k}^{\mathrm{dl}})^\dagger)\mathbf{g}_{k,k}^{\mathrm{dl}}}{\|(\mathbf{I} - \mathbf{G}_{-k}^{\mathrm{dl}}(\mathbf{G}_{-k}^{\mathrm{dl}})^\dagger)\mathbf{g}_{k,k}^{\mathrm{dl}}\|_2}, \quad \forall k, \tag{1.22}$$

where $(\mathbf{G}_{-k}^{\mathrm{dl}})^{\dagger}$ denotes the pseudo-inverse of $\mathbf{G}_{-k}^{\mathrm{dl}}$. To this end, the achievable rate of user k with the compression-based transmission strategy is then given by

$$r_k^{\mathrm{dl,CP}} = B\log_2\left(1+\frac{p_b\left|(\mathbf{g}_{k,k}^{\mathrm{dl}})^H\mathbf{w}_k^{\mathrm{dl}}\right|^2}{\left(\sum\limits_{j\neq\Omega_k}p_b\left|(\mathbf{g}_{k,j}^{\mathrm{dl}})^H\mathbf{w}_j^{\mathrm{dl}}\right|^2+\sigma_{\mathrm{ul}}^2+\sum\limits_{n=1}^{N}(\mathbf{h}_{k,n}^{\mathrm{dl}})^H\mathbf{S}_{\mathbf{e},n}^{\mathrm{dl}}\mathbf{h}_{n,k}^{\mathrm{dl}}\right)\Gamma_s}\right). \tag{1.23}$$

Data-Sharing Strategy: An alternative to the compression strategy for the downlink C-RAN is that instead of sending the compressed version of the beamformed signals, CP can directly share user messages with the RRHs, which then perform beamforming locally and cooperatively transmit the beamformed signals to the users.

Specifically, the message of each user k, i.e., s_k^{dl}, is sent from the CP to all the RRHs serving this user, i.e., Θ_k, via the fronthaul links. In this case, the transmit signal of RRH n is given by

$$\mathbf{x}_n^{\mathrm{dl}} = \sum_{k\in\Phi_n}\mathbf{w}_{n,k}^{\mathrm{dl}}\sqrt{p_b}s_k^{\mathrm{dl}}, \quad \forall n. \tag{1.24}$$

Observe that there is no quantization noise term in the above. As a consequence, if the ZF beamforming design given in (1.22) is applied for all the users, the achievable rate of user k under the data-sharing strategy can be obtained similarly as in (1.23), but without the quantization noise, i.e.,

$$r_k^{\mathrm{dl,DS}} = B\log_2\left(1+\frac{p_b\left|(\mathbf{g}_{k,k}^{\mathrm{dl}})^H\mathbf{w}_k^{\mathrm{dl}}\right|^2}{\left(\sum\limits_{j\neq\Omega_k}p_b\left|(\mathbf{g}_{k,j}^{\mathrm{dl}})^H\mathbf{w}_j^{\mathrm{dl}}\right|^2+\sigma_{\mathrm{ul}}^2\right)\Gamma_s}\right), \ \forall k, \tag{1.25}$$

where the per-beam transmit power is now simply $p_b = P_R/|\mathcal{K}|$.

It is worth noting that although the data-sharing strategy does not suffer from quantization noise, the cluster size is severely limited by the finite-capacity fronthaul links. Specifically, if user k is served by RRH n, i.e., $k\in\Phi_n$, then s_k needs to be sent to RRH n at a rate of $r_k^{\mathrm{dl,DS}}$ bps. As a result, given a clustering strategy defined by Φ_n, the fronthaul capacity required for each RRH n is the sum of all the user rates served by it, i.e.,

$$C_n^{\mathrm{dl}} = \sum_{k\in\Phi_n} r_k^{\mathrm{dl,DS}}, \quad \forall n. \tag{1.26}$$

It is thus essential to design the cluster size carefully under the data-sharing strategy such that the fronthaul traffic does not exceed the fronthaul capacity at each link, i.e., $C_n^{\mathrm{dl}} \leq C$, $\forall n$. For example, compressed sensing techniques can be used to choose the serving cluster for each user in an intelligent fashion [6].

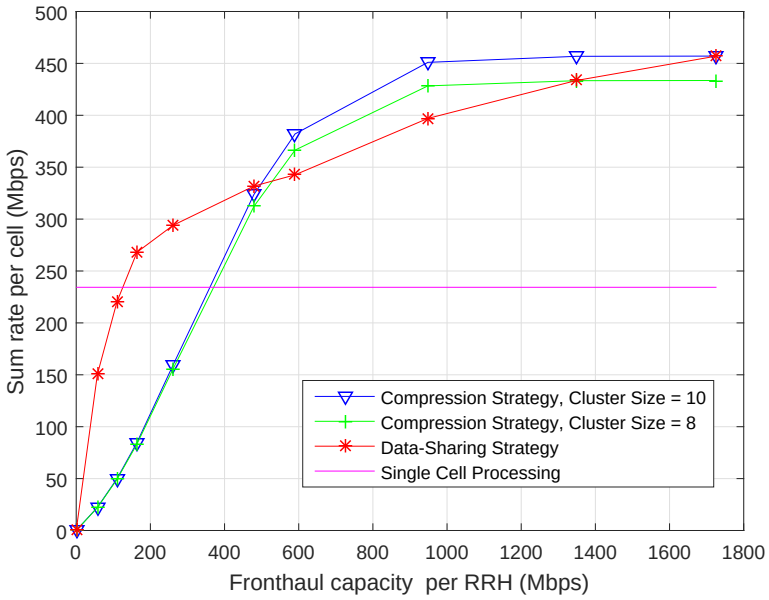

Figure 1.4 Performance comparison between compression-based and data-sharing strategies in the downlink C-RAN under user-centric clustering.

Performance Evaluation: A per-cell sum-rate comparison between the compression and data-sharing strategies in downlink C-RAN under user-centric clustering is shown in Fig. 1.4 under various per-RRH fronthaul capacities. The network setup is similar to the uplink as given in Table 1.1 with three RRHs per cell, except the number of antennas at each RRH is set to be $M = 4$, and the average transmit power of each RRH is set to be 43dBm. The user-centric cluster size for the compression strategy is fixed, while the cluster size for data-sharing ranges from 1 to 10. Observe that as in uplink, C-RAN brings considerable gain as compared to the single-cell baseline. Cooperative transmission is able to almost double the sum rate at fronthaul capacity cost of about six times the access rate.

Observe also that at low fronthaul capacity data-sharing outperforms compression, while at high fronthaul capacity compression outperforms data-sharing. The reason is that in the data-sharing strategy, the message of each user is repeatedly transmitted over different fronthaul links to its serving RRHs; this is not the most efficient use of the fronthaul when the cluster size is large.

We remark that this section assumes a single-hop C-RAN with direct links between the CP and RRHs. If the fronthaul network consists of edge routers and network processors over multiple hops, routing strategies can also play a significant role. In particular, as the data-sharing strategy amounts to multicast of user messages to multiple RRHs, network coding techniques can be applied to improve the efficiency of the fronthaul network [8].

1.3 Fronthaul-Aware Data Link and Physical Layers

So far we have discussed the impact of fronthaul capacity limitation on the spectral efficiency of C-RAN. This section addresses latency, which is an equally important system objective that affects the throughput of 5G deployment. An important case in point is the HARQ protocol, which runs across the data link and physical layers, and has the role of guaranteeing reliable communication over fading channels. HARQ accomplishes this goal via the transmission of additional information about data frames that have been previously transmitted but not correctly received and acknowledged by the receiver. Given that baseband processing and HARQ retransmission decisions are implemented at the BBU, delays in the communication between RRH and BBU due to fronthaul transmission entail an increased latency between successive retransmission attempts. This may disrupt the operation of existing HARQ protocols or cause excessive delays in latency-sensitive applications. As an example, in LTE a latency larger than 3 ms in the uplink is treated as system outage at the data link layer [9].

Fronthaul latency can be partly mitigated by deploying shorter, dedicated rather than multi-hop, fronthaul links between each RRH and the BBU, so as to reduce the transit time between RRH and BBU. Furthermore, one can enhance the computing power at the BBU, so as limit the time required to process signals at the BBU for fronthaul transmission and reception. To provide some reference values, two-way fronthaul transmission, excluding processing times, for single-hop fronthaul links may be of the order of 0.5 ms, while processing times at the BBU and UE can amount to a few milliseconds [10].

This section discusses solutions for scenarios in which the performance limitations incurred due to fronthaul latency constraints cannot be satisfactorily dealt with by means of the outlined approaches within the standard C-RAN architecture. Specifically, we consider the potential advantages that could be accrued by leveraging *alternative functional splits* between the BBU and the RRHs, whereby the RRHs implement some control functionalities for the HARQ protocol.

As discussed, the performance degradation of HARQ protocols in C-RAN is to be ascribed to the need to transfer baseband signals, as well as retransmission request (NAK) or positive acknowledgement (ACK) messages, between RRH and BBU, given that both control and data plane functionalities are implemented solely at the BBU. With the aim of reducing latency, the class of solutions investigated in this section allows for *HARQ control functions to be carried out at the RRH*.

In equipping the RRH with sufficient intelligence to perform some baseband functions as well as control decisions, the considered solutions deviate from the standard C-RAN architecture and are in line with the *alternative functional splits* being investigated in the literature and by the industry (see [12, 13]). In defining such alternative splits, emphasis will be given here to solutions that require RRHs with reduced complexity as compared to conventional base stations. This choice

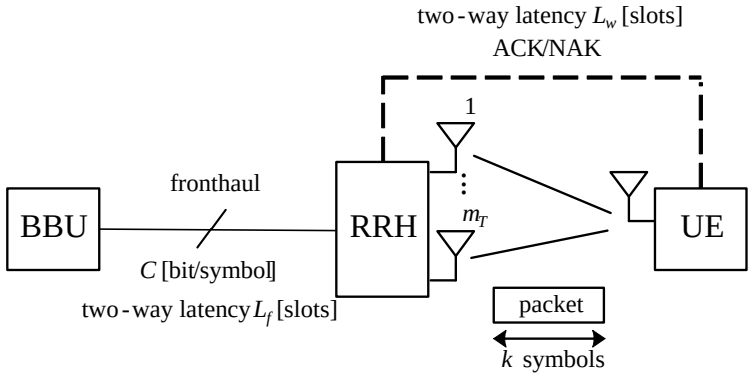

Figure 1.5 System model considered in Sec. 1.3 (with $m_R = 1$ receive antennas).

excludes, for instance, approaches that require full data decoding at the RRHs in the uplink or data encoding and precoding in the downlink.

The considered functional splits can be interpreted as implementing an instance of the *separation of control and data planes* that is currently advocated for next-generation wireless network architectures in a variety of guises (see [14] for a review). In particular, in the solutions at hand, the control functions associated with the HARQ protocol are carried out at the network edge, namely at the RRHs, while functionalities related to the data plane are still performed remotely at the BBU as in a conventional C-RAN system. The key advantages of this architecture are that: (*i*) retransmission control is not subject to the fronthaul latency constraints; (*ii*) the complexity of the RRHs can be kept in check given that data plane processing is performed at the BBU; (*iii*) joint baseband processing of data-plane information at the BBU yields spectral efficiency gains at the physical layer, as discussed in the previous section.

In the rest of this section, we describe HARQ protocols for both uplink and downlink based on the discussed separation of control and data planes. In order to simplify the discussion, instead of treating a C-RAN system with cooperative transmission and reception, this section focuses on the throughout and latency of a simplified Distributed-RAN (D-RAN) system in which each RRH serves its own set of users. Here, we use the term D-RAN to refer to the intermediate architecture between a standard cellular system and C-RAN in which the BBU of each base station is hosted at a remote site (see, e.g., [11, 12, 15]). In a D-RAN, unlike a C-RAN, the BBUs of different RRHs are hence physically distinct. Note that joint baseband decoding is generally not feasible in a D-RAN, since this would require the exchange of baseband signals among BBUs, rather than user-plane data as allowed by the X2 interface that may connect BBUs (see e.g., [15]). The focus on D-RAN allows us to concentrate on setups in which any given UE is assigned to a single RRH-BBU pair and to distill the essense of the effect of latency on the throughput performance.

1.3.1 Uplink

In this subsection, we consider the uplink, and we describe a solution, first proposed in [16, 17], whereby HARQ control functions are carried out at the RRH. The scheme works as follows: rather than waiting for an ACK or NAK message to be received from the BBU, the RRH assigned to an UE estimates the uplink channel based on the received signal, and it preemptively makes the control decision of sending a NAK message in case the signal-to-noise ratio (SNR) is found to be below a threshold and an ACK message otherwise. Importantly, the RRH does not perform data decoding and hence its complexity is kept significantly lower than that of a conventional base station. We now detail system model and performance analysis.

System Model: We concentrate on a D-RAN system in which an UE transmits on a dedicated spectral resource to an RRH, as illustrated in Fig. 1.5. The RRH is connected by means of a dedicated fronthaul link to a BBU. The BBU performs decoding, while the RRH is assumed to have limited baseband processing functionalities that allow for resource demapping and for the estimation of CSI. Note that different UEs are assumed to be served in distinct time-frequency resources, as done for instance in LTE, and hence we limit our attention to the performance of a given UE.

Each packet transmitted by the UE contains k encoded complex symbols and is transmitted within a coherence time/frequency interval of the channel, which is referred to as *slot*. The transmission rate of the first transmission of an information message is defined as r bits per symbol, so that kr is the number of information bits in the information message.

Each transmitted packet is acknowledged via the transmission of a feedback message by the RRH to the UE. We assume that these feedback messages are correctly decoded by the UE. The same information message may be transmitted for up to n_{max} successive slots using an HARQ protocol. Here, we adopt the Incremental Redundancy (IR) protocol, which operates across the physical and data link layers and is implemented in standards such as LTE [9]. With HARQ-IR, the UE transmits new encoded symbols at each transmission attempt and the BBU performs decoding based on all the received packets. Furthermore, we assume a backlogged UE that always has packets to transmit, and a selective repeat retransmission policy in which only frames from unsuccessfully received information messages are retransmitted.

In order to capture the impact of the fronthaul latency, we assume that a two-way delay of L_f slots is incurred for transmission between RRH and BBU. Furthermore, we assume that the round-trip transmission delay between UE and RRH, including the decoding delay at the UE, is given by L_w slots. As discussed, these two-way latencies may amount to a few milliseconds, which typically encompass multiple transmission intervals, e.g., multiple Transmission Time Intervals (TTIs) in LTE. The fronthaul capacity is denoted by C and is

measured in bits per symbol of the wireless channel or, equivalently, bits per second per Hz with respect to the wireless bandwidth.

The UE is equipped with m_T transmitting antennas, while m_R receiving antennas are available at the RRH. The received signal for any nth slot can be expressed as

$$\mathbf{y}_n = \sqrt{\frac{s}{m_T}}\mathbf{H}_n\mathbf{x}_n + \mathbf{z}_n, \tag{1.27}$$

where s measures the average SNR per receive antenna; $\mathbf{x}_n \in \mathbb{C}^{m_T \times 1}$ represents the symbols sent by the transmit antennas of the UE at a given channel use, whose average power is normalized as $\mathrm{E}[||\mathbf{x}_n||^2] = 1$; $\mathbf{H}_n \in \mathbb{C}^{m_R \times m_T}$ is the channel matrix, which is assumed to have independent identically distributed (i.i.d.) $\mathcal{CN}(0,1)$ entries (Rayleigh fading); and $\mathbf{z}_n \in \mathbb{C}^{m_R \times 1}$ is an i.i.d. Gaussian noise vector with $\mathcal{CN}(0,1)$ entries. The channel matrix $\mathbf{H}_n$ changes independently in each slot n. Moreover, it is assumed to be known to the RRH and to the BBU. We assume the use of Gaussian codebooks with an equal power allocation across the transmit antennas, although the analysis could be extended to arbitrary power allocation and antenna selection schemes.

The main performance metrics of interest are as follows.

- Throughput T: The throughput measures the average rate, in bits per symbol, at which information can be successfully delivered from the UE to the BBU;
- Probability $\mathrm{P_s}$ of success: The metric $\mathrm{P_s}$ measures the probability of a successful transmission within a given HARQ session, which is the event that, in one of the n_{max} allowed transmission attempts, the information message is decoded correctly at the BBU;
- Average latency D: The average latency D measures the average number N of transmission attempts per information message.

A few remarks are in place. First, the three metrics are interdependent. In particular, based on standard renewal theory arguments, the throughput can be calculated as [18]

$$T = \frac{r\mathrm{P_s}}{\mathrm{E}[N]}, \tag{1.28}$$

where we recall that r is the transmission rate, and the random variable N denotes the number of transmission attempts for a given information message. As it will be discussed, the average latency D is an increasing function of $\mathrm{E}[N]$. Therefore, given r, any two metrics defined above determine the third. Secondly, errors in the HARQ sessions, which occur with probability $1 - \mathrm{P_s}$ are typically dealt with by higher layers, as done by the RLC layer in LTE [16]. Finally, as it will be seen, the proposed schemes aim at trading a decrease in the throughput T for a reduction in the average latency D.

Conventional D-RAN: In a conventional D-RAN system, all processing and retransmission decisions are carried out at the BBU. Therefore, each transmission

requires a two-way latency of $L_w + L_f$ slots, due to the need to communicate in both directions on the wireless channel and on the fronthaul link. Assuming, as mentioned, a backlogged UE, the throughput can be written as (1.28), where the average number of transmissions can be computed as

$$\mathrm{E}[N] = \sum_{n=1}^{n_{max}-1} n\mathrm{P}(\mathrm{ACK}_n) + n_{max}\mathrm{P}(\mathrm{NAK}_{n_{max}-1}). \tag{1.29}$$

We denote as ACK_n the event that an ACK message is sent to the UE after exactly n transmission attempts, hence terminating the retransmission process. Note that, in a conventional D-RAN implementation, this implies that the BBU decodes successfully after exactly n transmission attempts. We also denote as NAK_n the event in which a NAK message is sent to the UE for all transmission attempts up to, and including, the nth one. In a conventional D-RAN implementation, this implies that the BBU does not decode successfully up to, and including, the nth transmission attempt. We observe that, by definition, we have the relationship $\mathrm{P}(\mathrm{ACK}_n) = \mathrm{P}(\mathrm{NAK}_{n-1}) - \mathrm{P}(\mathrm{NAK}_n)$. Furthermore, for a conventional D-RAN system, the probability of a successful transmission is given as

$$\mathrm{P_s} = 1 - \mathrm{P}(\mathrm{NAK}_{n_{max}}). \tag{1.30}$$

In order to evaluate probabilities of error, we will adopt the finite-blocklength Gaussian approximation proposed in [19], based on the work [20]. Accordingly, the probability $\mathrm{P_e}(r, k, \mathbf{H})$ of a decoding error for a transmission at rate r in a slot of k channel uses when the channel matrix is $\mathbf{H}$ is approximated as

$$\mathrm{P_e}(r, k, \mathbf{H}) = Q\left(\frac{C(\mathbf{H}) - r}{\sqrt{\frac{V(\mathbf{H})}{k}}}\right), \tag{1.31}$$

where we have defined

$$C(\mathbf{H}) = \sum_{j=1}^{m_{rt}} \log_2\left(1 + \frac{s\lambda_j}{m_T}\right) \text{ and } V(\mathbf{H}) = \left(m_{rt} - \sum_{j=1}^{m_{rt}} \frac{1}{\left(1 + \frac{s\lambda_j}{m_T}\right)^2}\right) \log_2^2 e, \tag{1.32}$$

with $m_{RT} = \min(m_R, m_T)$; $\{\lambda_j\}_{j=1,\ldots,m_{RT}}$ being the eigenvalues of the matrix $\mathbf{H}^H\mathbf{H}$; and $Q(\cdot)$ being the Gaussian complementary cumulative distribution function.

With this approximation, since with HARQ-IR, a set of n transmission attempts for a given information message can be treated as a transmission over n parallel channels (see, e.g., [18]), the error probability at the nth transmission can be computed as $\mathrm{E}[\mathrm{P_e}(r, k, \mathcal{H}_n)]$ where $\mathcal{H}_n = \mathrm{diag}([\mathbf{H}_1, ..., \mathbf{H}_n])$ and the expectation is taken with respect to the channel distribution [19]. Moreover, the probability of a decoding error up to, and including, the nth transmission can

be upper bounded by the probability of error *at* the n transmission as

$$\mathrm{P}(\mathrm{NAK}_n) \leq \mathrm{E}[\mathrm{P_e}(r, k, \mathcal{H}_n)]. \tag{1.33}$$

Using (1.33) in (1.28) and (1.30), we obtain lower bounds on the throughput and probability of success, respectively (within the approximation (1.32) of the probability of error). As for the average latency, given the discussion above, we can calculate

$$D = \mathrm{E}[N](L_w + L_f). \tag{1.34}$$

An upper bound on D can be computed using (1.29) and (1.33).

Edge-Based Retransmission: A low-latency edge-based HARQ control scheme was first proposed in [16, 17]. The approach assumes an RRH-BBU functional split whereby each RRH can perform synchronization and resource demapping, so as to be able to perform CSI estimation of the channel $\mathbf{H}_n$ at each transmission attempt n. The RRH is also assumed to have obtained the modulation and coding scheme (MCS) used for data transmission from the BBU, which selects the MCS during scheduling. The MCS information amounts here to the rate r and packet length k. Based on this information, the RRH can compute the probability of error for decoding at the BBU. This could be done, e.g., by using an analytical approximation, such as $\mathrm{P_e}(r, k, \{\mathbf{H}_i\}_{i\leq n})$ in (1.31), or a pre-computed look-up table. Note that this probability depends on all channel matrices $[\mathbf{H}_1, \cdots, \mathbf{H}_n]$ corresponding to prior and current transmission attempts. In the following, we will assume that the approximation $\mathrm{P_e}(r, k, \{\mathbf{H}_i\}_{i\leq n})$ is used by the RRH, although the discussion applies more generally.

The gist of the approach is to allow the RRH to make preemptive decisions regarding the feedback of ACK/NAK messages to the UE without waiting the L_f slots required for two-way communication on the fronthaul link. This is done as follows: if the decoding error probability $\mathrm{P_e}(r, k, \{\mathbf{H}_i\}_{i\leq n})$ is smaller than a given threshold $\mathrm{P_{th}}$, the RRH sends an ACK message to the UE, predicting a positive decoding event at the BBU; while, otherwise, a NAK message is transmitted, that is, the following rule is used by the RRH:

$$\mathrm{P_e}(r, k, \{\mathbf{H}_i\}_{i\leq n}) \underset{\mathrm{NAK}}{\overset{\mathrm{ACK}}{\lessgtr}} \mathrm{P_{th}}. \tag{1.35}$$

While reducing the average latency due to the implementation of control decisions at the RRH, the discussed edge-based HARQ scheme introduces a possible mismatch between the RRH's decisions and the actual decoding outcome at the BBU. In particular, there are two types of error. In the first type of error, the transmitted packet is not decodable at the BBU, but an ACK message is sent by the RRH. As seen, this type of mismatch needs to be dealt with by higher layers. In the second type of error, the received packet is decodable at the BBU, but a NAK message is sent by the RRH. In this case, the UE performs an unnecessary HARQ retransmission, unless the maximum number n_{max} of transmission attempts has already been performed. These errors generally cause a reduction of

throughput and probability of success, in return for which the edge-based scheme at hand promises significant gains in terms of delays. To see this, we note that the average latency until a packet is acknowledged, positively or negatively, to the UE is given by

$$D = \mathrm{E}[N]L_w, \tag{1.36}$$

since no use of the fronthaul link is required in order to complete the HARQ process. This may be significantly smaller than (1.34), depending on the relative value of L_w and L_f.

With regards to the optimization of the threshold $\mathrm{P_{th}}$ in (1.35), this needs to strike a balance between the probability of success $\mathrm{P_s}$, which would call for a smaller $\mathrm{P_{th}}$ and hence more retransmissions, and the throughput T, which may be generally improved by a larger $\mathrm{P_{th}}$, resulting in the transmission of new information.

Throughput and probability of success can be computed in a manner similar to the conventional D-RAN implementation. The main caveat is the definition of a successful event: a transmission is considered as successful here if an ACK message is sent to the UE within one of the n_{max} allowed transmissions attempts *and* if the BBU can correctly decode. Hence, by the law of total probability, the probability of success $\mathrm{P_s}$ can be written as

$$\mathrm{P_s} = \sum_{n=1}^{n_{max}} \mathrm{P}(\mathrm{S}_n|\mathrm{ACK}_n)\mathrm{P}(\mathrm{ACK}_n), \tag{1.37}$$

where S_n is the event that the BBU can succesfully decode at the nth transmission, while the event ACK_n is defined as above.

The probabilities needed to compute throughput (1.28) and probability of success (1.37) can be obtained, using the discussed Gaussian approximation, as follows. The probability of a NAK message being sent up to the n transmission attempt is

$$\mathrm{P}(\mathrm{NAK}_n) = \mathrm{P}\left(\mathrm{P_e}(r, k, \mathcal{H}_n) > \mathrm{P_{th}}\right). \tag{1.38}$$

Note that this is due to the monotonicity of the probability $\mathrm{P_e}(r, k, \mathcal{H}_n)$ as a function of each eigenvalue, so that the probability $\mathrm{P_e}(r, k, \mathcal{H}_n)$ is no larger than $\mathrm{P_e}(r, k, \mathcal{H}_{n-1})$. In a similar manner, we can also calculate

$$\mathrm{P}(\mathrm{S}_n|\mathrm{ACK}_n) = 1 - \mathrm{E}[\mathrm{P_e}(r, k, \mathcal{H}_n)|\mathcal{A}(\mathrm{P_{th}})] \tag{1.39}$$

where we have defined the event $\mathcal{A}(\mathrm{P_{th}}) = \{\{\mathrm{P_e}(r, k, \mathcal{H}_{n-1}) > \mathrm{P_{th}}\} \bigcap \{\mathrm{P_e}(r, k, \mathcal{H}_n) \leq \mathrm{P_{th}}\}\}$.

Numerical Example: We now corroborate the analysis presented in the previous sections by providing insights into the performance comparison of conventional D-RAN and edge-based scheme systems via a numerical example. Fig. 1.6 shows the throughput loss of the edge-based scheme as compared to the conventional D-RAN implementation, as a function of the blocklength k, for two

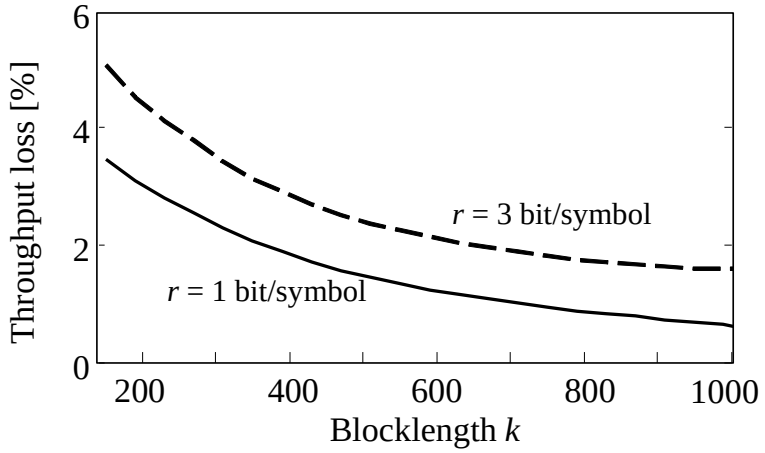

Figure 1.6 Throughput loss of the edge-based scheme with respect to the standard D-RAN implementation versus blocklength k system ($s = 4$ dB, $n_{max} = 10$ $m_t = 1$, $m_r = 1$, $\mathrm{P_s} > 0.99$ for $r = 1$ bit/symbol and $r = 3$ bit/symbol).

rates $r = 1$ bit/symbol and $r = 3$ bit/symbol. We set $s = 4$ dB, $n_{max} = 10$ and we focus on a single-antenna link, i.e., $m_T = m_R = 1$. For every value of k, the threshold $\mathrm{P_{th}}$ is optimized to maximize the throughput T under the constraint that the probability of success satisfies the requirement $\mathrm{P_s} > 0.99$. This is typically assumed to be acceptable in existing systems (see, e.g., [9]).

It can be seen that, as the blocklength increases, the throughput loss of local feedback decreases significantly. In this regime, the latency reduction afforded by edge-based control comes at a minor cost in terms of throughput loss. This reflects a fundamental insight: The performance loss of local feedback is due to the fact that the local decisions are taken by the RRH based only on channel state information, without reference to the specific channel noise realization that affects the received packet. Therefore, as the blocklength k increases, and hence as the errors due to atypical channel noise realizations become less likely, the local decisions tend to be consistent with the actual decoding outcomes at the BBU. In other words, as the blocklength k grows larger, it becomes easier for the RRH to predict the decoding outcome at the BBU: in the Shannon regime of infinite k, successful or unsuccessful decoding depends deterministically on whether the rate r is above or below capacity.

1.3.2 Downlink

In this section, we consider a low-latency HARQ protocol in which control is carried out at the RRH for the downlink of a D-RAN system. Similar to the uplink, the key idea is that of enabling the RRH to make low-latency retransmission control decisions, while still retaining all baseband encoding capabilities at the BBU so as to reduce the complexity of the RRH. According to the pro-

tocol, first proposed in [21], at the first transmission attempt, the RRH stores the transmitted baseband signal, which is encoded by the BBU and received by the RRH on the fronthaul link. In case a NAK is received, the RRH then retransmits the stored baseband signal without further baseband processing and without assistance from the BBU. We observe that, unlike the uplink mechanism discussed above, here no CSI estimation is needed at the RRH, but the RRH still needs to be equipped with sufficient baseband capabilities to enable the detection of ACK/NAK messages. Furthermore, the RRH is assumed to have enough memory to store previously transmitted baseband signals.

System Model: We consider downlink communication in the same D-RAN system studied above and shown in Fig. 1.5, consisting of a RRH that is connected to a BBU through a fronthaul link. We focus on the performance of a given single-antenna UE, i.e., $m_T = 1$, which is allocated dedicated spectral resources for downlink transmission. The RRH transmits a packet of length k symbols in each slot, and the UE sends an ACK or NAK message depending on the decoding outcome, which is assumed to be correctly decoded at the RRH or BBU. The key parameters r, L_w, and L_f are defined as for the uplink (see Fig. 1.5).

The received signal at the UE in a time slot t can be written as

$$y_t = \mathbf{h}_t^\dagger \mathbf{x}_t + z_t, \tag{1.40}$$

where $\mathbf{h}_t$ is the $m_R \times 1$ channel vector with m_R being the number of transmit antennas of the RRH; $\mathbf{x}_t$ is the $m_R \times 1$ signal transmitted by the RRH with power constraint $\mathrm{E}[\|\mathbf{x}_t\|^2] = P$; and z_t is complex Gaussian noise with unitary power, i.e., distributed as $\mathcal{CN}(0,1)$. Note that, unlike the uplink, we find it convenient here to express the received signal as a function of the slot index t rather than of the retransmission attempt index n. In particular, this allows us to keep track of the channel variations, which, as further elaborate in the rest of this subsection, play a key role in the downlink. To this end, we assume that the channel vector process $\mathbf{h}_t$ is correlated across two successive slots according to a stationary autoregressive model of order one, namely

$$\mathbf{h}_t = \rho \mathbf{h}_{t-1} + \mathbf{v}_t \tag{1.41}$$

for $t \in \{\dots, -2, -1, 0, 1, 2, \dots\}$ with correlation coefficient $\rho \in [0, 1)$, where $\mathbf{v}_t$ has independent $\mathcal{CN}(0, 1-\rho^2)$ entries. Full CSI is assumed at the UE.

Conventional D-RAN: In a conventional D-RAN implementation, as shown in Fig. 1.7, the RRH delivers the ACK/NAK feedback message, as well as updated CSI information, from the UE to the BBU, and the BBU carries out the encoding of the data-plane information of a previously transmitted packet in case a NAK is received or of a new packet in case an ACK is obtained. Therefore, a round trip delay of L_w slots on the wireless channel and a two-way fronthaul latency of L_f slots are elapsed between the transmission of a downlink packet and the time that packet may be retransmitted to the UE.

To elaborate, as for the uplink, we assume that the BBU implements the IR protocol. Accordingly, at any time t, the RRH sends a new part of an encoded

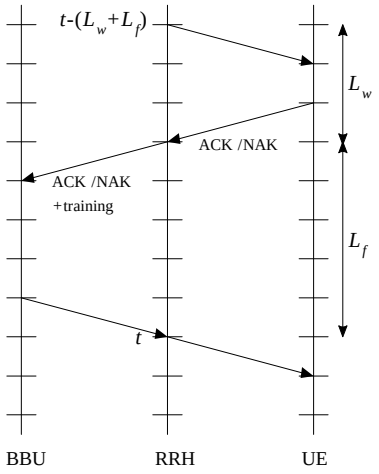

Figure 1.7 Illustration of the conventional implementation of downlink HARQ in D-RAN.

information frame that was last transmitted at slot $t-(L_w+L_f)$ if a NAK is received at the RRH at time $t-L_f$; otherwise, it transmits a packet from a new information frame.

In either case, the information-bearing symbol $s_t \sim \mathcal{CN}(0,1)$ to be transmitted at slot t is linearly precoded at the BBU by means of a $m_R \times 1$ beamforming vector $\mathbf{w}_t$, which is matched to the last channel realization that is available at the cloud, namely $\mathbf{h}_{t-L_f}$, that is $\mathbf{w}_t = \mathbf{h}_{t-L_f}/\|\mathbf{h}_{t-L_f}\|$. Then, the precoded signal $\tilde{\mathbf{x}}_t$ at the BBU is given as

$$\mathbf{x}_t = \sqrt{P}\frac{\mathbf{h}_{t-L_f}}{\|\mathbf{h}_{t-L_f}\|}s_t. \tag{1.42}$$

We refer to [21] for a discussion on how to account for fronthaul capacity limitations and for the corresponding performance degradation due to quantization noise.

Edge-based Retransmission: To potentially alleviate the performance limitations in terms of delay of the standard D-RAN system described above due to two-way fronthaul latency, we now consider a solution based on the implementation of the HARQ control function at the RRH. Accordingly, we allow for low-latency retransmissions by the RRH, under the working assumption that the RRH can store the previously transmitted baseband signals as well as decode the ACK/NAK feedback messages on the uplink.

As illustrated in Fig. 1.8, if a NAK is fed back by the UE regarding a packet previously sent at time $t-L_w$, the RRH autonomously retransmits the previously transmitted packet at time t without waiting for a newly encoded packet from the BBU. If an ACK is instead fed back by the UE, the RRH asks for an encoded packet associated to a new information frame from the BBU. As

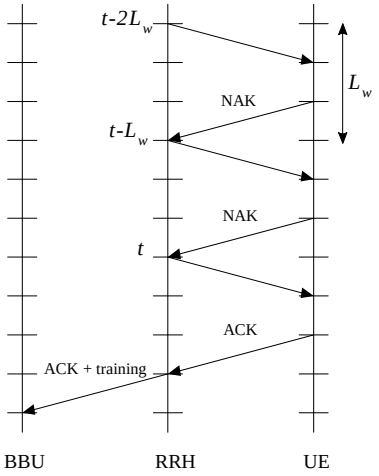

Figure 1.8 Illustration of edge-based HARQ for the downlink.

a result, for the first transmission of a packet, a signal (1.42) is transmitted, and, for each retransmission, the same signal is retransmitted by the RRH. Note that this strategy cannot adapt to channel variations and is implicitly based on HARQ Type I or Chase Combining [22].

Numerical Results and Discussion: In this section, we compare the throughput and latency performance of the conventional D-RAN and edge-based retransmission. Throughout this section, we assume that the number of transmit antennas at the RRH is $m_R = 4$, the length of a packet in each slot is $k = 100$ symbols, the transmit power P is 10 dB, the maximum number of transmission attempts is $n_{\max} = 10$, the two-way latencies are equal to $L_f = L_w = 2$ slots. We also consider, as discussed in [21], a fronthaul capacity C of $C = 3$ bit/symbol.

We first plot the throughput as a function of correlation coefficient ρ, which defines the time-variability of the channel, in Fig. 1.9 for $r = 2$ and $r = 3$ bit/symbol. The throughput loss of the edge-based scheme, which is caused by the lack of adaptation to the varying channel conditions in the retransmission attempts and to the simpler HARQ protocol, depends on the correlation coefficient ρ and on the transmission rate R. Specifically, for a lower transmission rate requiring with high probability no more than one retransmission, such as 2 bit/symbol, the throughput loss is minor. Instead, for a larger transmission rate, which calls for more retransmissions, the loss may be substantial, unless the correlation coefficient ρ is large enough. For instance, with $\rho = 0.8$, which corresponds to a speed of 60 km/h and a carrier frequency of 2.6 GHz for a slot duration of 1 ms according to Clarke's standard model, the loss is 2% for $r = 2$ bit/symbol and 18% for $r = 3$ bit/symbol.

The implementation of edge-based retransmission is justified if the discussed throughput loss is deemed to be acceptable when compared to the given reduction in latency. This is investigated by means of Fig. 1.10, which shows the latency

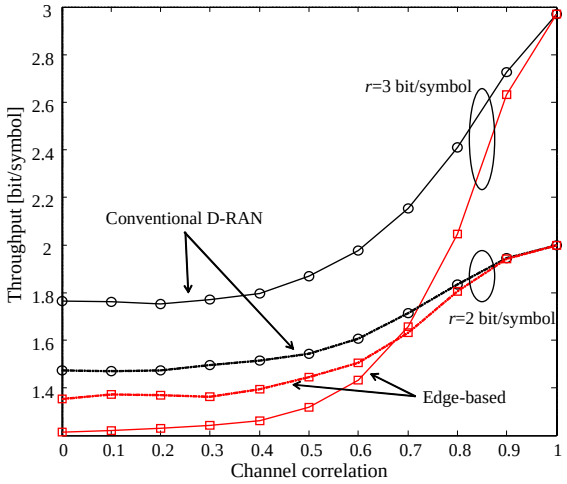

Figure 1.9 Throughput T versus the correlation coefficient ρ for $r = 2$ and $r = 3$ bit/symbol ($m_T = 4$, $k = 100$, $P = 10$ dB, $C = 3$ bits/symbol, $L_w = L_f = 2$ slots, and $n_{\max} = 10$).

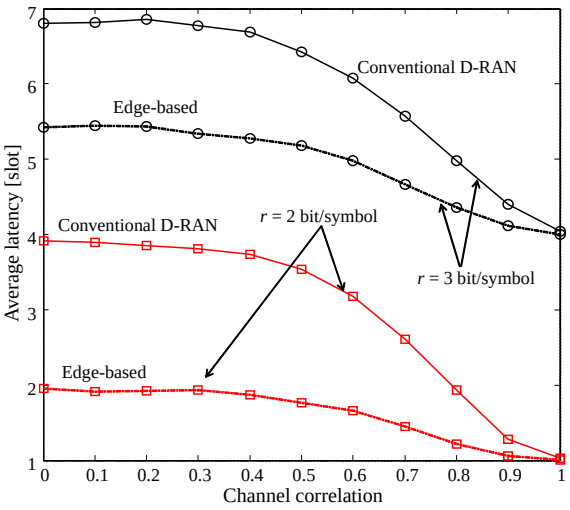

Figure 1.10 Latency D versus the correlation coefficient ρ for $r = 2$ and $r = 3$ bit/symbol ($m_T = 4$, $k = 100$, $P = 10$ dB, $C = 3$ bit/symbol, $L_w = L_f = 2$ slots, and $n_{\max} = 10$).

as a function of correlation coefficient ρ for the same parameters as above. It is seen that the reduction in latency can be very significant, particularly for sufficiently small rates and/or for slowly varying channels, i.e., for large enough ρ. As examples, for $\rho = 0.8$ as considered above, the latency can be reduced by 3.2 slots, at the cost of a throughput reduction of 0.05 bit/symbol, if $r = 2$

bit/symbol; while the latency reduction is 3 slots at the cost of a throughput loss of 0.35 bit/symbol if $r = 3$ bit/symbol.

1.4 Conclusions

The cloud radio-access network architecture enables significant increase in area spectral efficiency for 5G wireless cellular networks by allowing dense and distributed deployment of remote antennas together with centralized capability for joint baseband processing across the cooperative antenna clusters. The advent of such an architecture also necessitates re-thinking of both physical layer and data link layer operations. This chapter presents some of the challenges and opportunities in C-RAN design. The first part of the chapter illustrates that in the physical layer, coherent transmission and reception of user signals across multiple remote radio heads provide significant rate gain, but the designs of these cooperative communication strategies need to be adapted according to the capacity constraints of the fronthaul. We utilize the compression strategy as the fundamental technique for both uplink and downlink, and quantify the effect of the limited fronthaul capacity on the overall spectral efficiency in a zero-forcing based user-centric cooperative communication strategy. The second part of this chapter makes a further contribution in pointing out that the latency in C-RAN architecture can be managed and controlled in an intelligent way by re-thinking the design of the H-ARQ protocol in a multi-hop network. Together these techniques point to ways that would make radio access via cloud computing a reality.

1.5 Acknowledgments

Liang Liu and Wei Yu would like to acknowledge the support of Natural Sciences and Engineering Research Council (NSERC) of Canada. Osvaldo Simeone would like to thank Shahrouz Khalili (NJIT), Wonju Lee, and Joonhyuk Kang (KAIST) for their collaboration on the material covered in Sec. 1.3. The work of Osvaldo Simeone was partially supported by the U.S. NSF through grant 1525629.

References

[1] C. Zhu and W. Yu, "Stochastic snalysis of user-centric network MIMO," in *Proc. IEEE Inter. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, 2016.

[2] 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, June 2014.

[3] Y. Zhou and W. Yu, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," to appear in *IEEE Trans. Signal Process.*, 2016. [Online]. Available: http://arxiv.org/abs/1604.05001v1

[4] L. Liu and R. Zhang, "Optimized uplink transmission in multi-antenna C-RAN with spatial compression and forward," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5083-5095, Oct. 2015.

[5] 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.

[6] 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.

[7] 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)*, Sep. 2015.

[8] L. Liu and W. Yu, "Joint sparse beamforming and network coding for downlink multi-hop cloud radio access networks," submitted to *IEEE Global Commun. Conf. (Globecom)*, 2016.

[9] E. Dahlman, S. Parkvall, J. Skold, P. Bemin, *3G Evolution: HSPA and LTE for Mobile Broadband,* Academic Press, 2nd Ed., 2008.

[10] NGMN Alliance, "Further study on critical C-RAN technologies," 2015.

[11] China Mobile, "C-RAN: The road towards green RAN," White Paper, ver. 2011.

[12] A. Checko et al., "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tutorials*, vol. 17, no. 1, pp. 405-426, First quarter 2015.

[13] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *J. Commun. Networks*, vol. 18, no. 2, April 2016. [Online]. Available: http://arxiv.org/abs/1512.07743

[14] A. Mohamed, O. Onireti, M. Imran, A. Imran, and R. Tafazolli, "Control-data separation architecture for cellular radio access networks: A survey and outlook," in *IEEE Commun. Surveys Tutorials*, vol. 18, no. 1, pp. 446-465, First Quarter 2016.

[15] M. Nahas, A. Saadani, J. Charles, and Z. El-Bazzal, "Base stations evolution: Toward 4G technology," in *Proc. Int. Conf. Telecommun. (ICT)*, pp. 1-6, Aalborg, Denmark, Apr. 2012.

[16] 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 Tech. Journal*, vol. 18, no. 1, pp. 105128, Jun. 2013.

[17] P. Rost and A. Prasad, "Opportunistic hybrid ARQ-enabler of centralized-RAN over nonideal backhaul," *IEEE Wireless Commun. Letters*, vol. 3, no. 5, pp. 481484, Oct. 2014.

[18] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 19711988, Jul. 2001.

[19] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static MIMO fading channels at finite blocklength." [Online]. Available: http://arxiv.org/abs/1311.2012.

[20] Y. Polyanskiy, H. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307 2359, May 2010

[21] W. Lee, O. Simeone and J. Kang, "Edge-based HARQ for low-latency communications in downlink D-RAN," submitted, 2016.

[22] S. Khalili and O. Simeone, "Uplink HARQ for distributed and cloud RAN via separation of control and data planes," 2015. [Online]. Available: http://arxiv.org/abs/1508.06570