Cloud Radio Access Networks: Uplink Channel Estimation and Downlink Precoding

Osvaldo Simeone, Jinkyu Kang, Joonhyuk Kang and Shlomo Shamai (Shitz)

I. Introduction

The gains afforded by cloud radio access network (C-RAN) in terms of savings in capital and operating expenses, flexibility, interference management and network densification rely on the presence of high-capacity low-latency fronthaul connectivity between remote radio heads (RRHs) and baseband unit (BBU). In light of the non-uniform and limited availability of fiber optics cables, the bandwidth constraints on the fronthaul network call, on the one hand, for the development of advanced baseband compression strategies and, on the other hand, for a closer investigation of the optimal functional split between RRHs and BBU. In this chapter, after a brief introduction to signal processing challenges in C-RAN, this optimal function split is studied at the physical (PHY) layer as it pertains to two key baseband signal processing steps, namely channel estimation in the uplink and channel encoding/ linear precoding in the downlink. Joint optimization of baseband fronthaul compression and of baseband signal processing is tackled under different PHY functional splits, whereby uplink channel estimation and downlink channel encoding/ linear precoding are carried out either at the RRHs or at the BBU. The analysis, based on information-theoretical arguments, and numerical results yields insight into the configurations of network architecture and fronthaul capacities

O. Simeone is with the Center for Wireless Information Processing (CWIP), ECE Department, New Jersey Institute of Technology (NJIT), Newark, NJ 07102, USA (Email: osvaldo.simeone@njit.edu).

Jinkyu Kang is with the School of Engineering and Applied Sciences (SEAS), Harvard University, Cambridge, MA 02138, USA (Email: jkkang@g.harvard.edu).

Joonhyuk Kang is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST) Daejeon, South Korea (Email: jhkang@ee.kaist.ac.kr).

S. Shamai (Shitz) is with the Department of Electrical Engineering, Technion, Haifa, 32000, Israel (Email: sshlomo@ee.technion.ac.il).

in which different functional splits are advantageous. The treatment also emphasizes the versatility of deterministic and stochastic successive convex approximation strategies for the optimization of C-RANs.

II. Technology Background

In a C-RAN architecture, the base station (BS) functionalities, from the PHY layer to higher layers, are implemented in a virtualized fashion on centralized general-purpose processors rather than on the local hardware of the base stations or access points. This results in a novel cellular architecture in which low-cost wireless access points−the RRHs−which retain only radio functionalities, are centrally managed by a reconfigurable centralized "cloud", the BBU. At a high level, the C-RAN concept can be seen as an instance of network function virtualization and hence as the RAN counterpart of the separation of control and data planes proposed for the core network in software-defined networking [1].

The C-RAN architecture has the following key advantages, which make it a key contender for inclusion in a 5G standard:

- Reduced capital expense due to the possibility to substitute full-fledged base stations with RRHs with reduced space and energy requirements;
- Statistical multiplexing gain thanks to the flexible allocation of radio and computing resources across all the connected RRHs;
- Easier implementation of coordinated and cooperative transmission/ reception strategies, such as Enhanced Inter-Cell Interference Coordination (eICIC) and Coordinated MultiPoint (CoMP) in Long Term Evolution Advanced (LTE-A), to mitigate multi-cell interference;
- Simplified network upgrades and maintenance owing to the centralization of RAN functionalities.

The C-RAN architecture depends on a network of so-called fronthaul links to enable the virtualization of BS functionalities at a BBU. This is because in the uplink, the RRHs are required to convey their respective received signals, either in analog format or in the form of digitized baseband samples, to the BBU for processing. Moreover, in a dual fashion, in a C-RAN downlink, each RRH needs to receive from the BBU either directly the analog radio signal to be transmitted on the radio interface, or a digitized version of the corresponding baseband samples. The RRH−BBU bidirectional links that carry such information are referred to as *fronthaul* links, in contrast to the backhaul links connecting the BBU to the core network.

The analog transport solution is typically implemented on fronthaul links by means of radio-over-fiber [2]. Instead, the digital transmission of baseband, or IQ, samples is currently carried out by following the Common Public Radio Interface (CPRI) standard [3], which most commonly requires fiber optic fronthaul links as well. The digital approach appears to be favored due to the traditional advantages of digital solutions, including resilience to noise and hardware impairments and flexibility in the transport options [4].

A. Signal Processing Challenges in C-RAN

The main roadblock to the realization of the mentioned promises of C-RAN hinges on the inherent restrictions on bandwidth and latency of the fronthaul links that may limit the advantages of centralized processing at the BBU.

1) Fronthaul capacity limitations: Implementing the CPRI standard, the bit rate required for base station that serve multiple cell sectors with carrier aggregation and with multiple antennas exceeds the 10 Gbit/s provided by standard fiber optics links [4], [5]. This problem is even more pronounced for networks in which fiber-optic links are not available due to the large expense required for their deployment or lease, as for heterogeneous networks with smaller RRHs [6]. The capacity limitations of the fronthaul link call for the development of compression strategies that reduce the fronthaul rate with minor or no degradation in the quality of the quantized baseband signal. Typical solutions are based on filtering, per-block scaling, lossless compression, predictive quantization, see [7]–[12]

When quantization and compression are not sufficient, as reported in [13], [14], the bottleneck on the performance of C-RANs due to the capacity limitations of the fronthaul links can be alleviated by implementing a more flexible separation of functionalities between RRHs and BBU, rather than performing all baseband processing at the BBU. Examples of baseband operations that can be carried out at the RRH include Fast Fourier Transform and Inverse Fast Fourier Transform (FFT and IFFT), demapping, synchronization, channel estimation, precoding and channel encoding. Note that [13] also investigates the possibility to implement functions at higher layers, such as error detection, at the RRHs. We will elaborate on important aspects of the functional split between RRH and BBU below.

2) Fronthaul latency limitations: Two of the communication protocols that are most affected by fronthaul delays are uplink hybrid automatic repeat request (HARQ) and random access [13]. For HARQ, the problem is that the outcome of decoding at the BBU may only become available at the RRH after the time required for

- the transfer of the baseband signals from the RRH to the BBU
- the processing at the BBU
- the transmission of the decoding outcome from the BBU to the RRH.

This delay may seriously affect the throughput achievable by uplink HARQ. For example, in LTE with frequency division multiplexing, the feedback latency should be less than 8 ms in order not to disrupt the operation of the system [13]. Similar issues impair the implementation of random access.

B. Chapter Overview

In this chapter, we explore the problem of optimal functional split between RRHs and BBU at the PHY layer by focusing on the two key baseband operations of channel encoding and channel encoding/ precoding. We recall that alternative functional splits are envisaged to be potentially advantageous in the presence of significant fronthaul capacity constraints.

For the uplink, we compare the standard implementation in which all baseband processing, including channel estimation, is performed at the BBU, with an alternative architecture in which channel estimation, along with the necessary frame synchronization and resource demapping, is instead implemented at the RRHs. This is discussed in Sec. III.

The downlink is discussed in Sec. IV, where we contrast the standard C-RAN implementation with an alternative one in which channel encoding and precoding are applied at the RRHs, while the BBU retains the function of designing the precoding matrices based on the available channel state information.

Throughout, we take an information-theoretic approach in order to evaluate analytical expressions for the achievable performance that illuminates the impact of different design choices. The analysis is corroborated by extensive numerical results that provide insight into the performance comparisons highlighted above. The chapter is concluded in Sec. V.

III. Uplink: Where to Perform Channel Estimation?

In this section, we study the uplink and address the potential advantages that could be accrued by performing channel estimation at the RRHs rather than at the BBU. The rationale for the exploration of this functional split is that communicating the digitized signal received within the training portion of the received signal, as done in the conventional implementation, may impose a more significant burden on the fronthaul network that communicating directly the estimated channel state information (CSI). This split is also supported by the known information-theoretic optimality of separate estimation and compression [15]. In particular, we compare two different approaches:

- the conventional approach, in which the RRHs quantize the training signals and CSI estimation takes place at the BBU;
- channel Estimation at the RRHs, in which the RRHs perform CSI estimation and forward a quantized version of the CSI to the BBU.

Note that the conventional approach was the subject of an earlier study [16] and that this section is adapted from our earlier work [17], to which we refer for proofs and additional considerations.

We start by discussing the system model in Sec. III-A and then elaborate on the two approaches in Sec. III-B and Sec. III-C. Finally, we present numerical results in Sec. III-D.

A. System Model

We study the uplink of a cellular system consisting of N_U User Equipments (UEs), N_R RRHs and a BBU, as shown in Fig. 1. We denote the set of all UEs, or mobile users, as $\mathcal{N}_U = \{1, \ldots, N_U\}$ and the set of all RRHs as $\mathcal{N}_R = \{1, \ldots, N_R\}$. Each i-th UE has $N_{t,i}$ transmit antennas, while each j-th RRH is

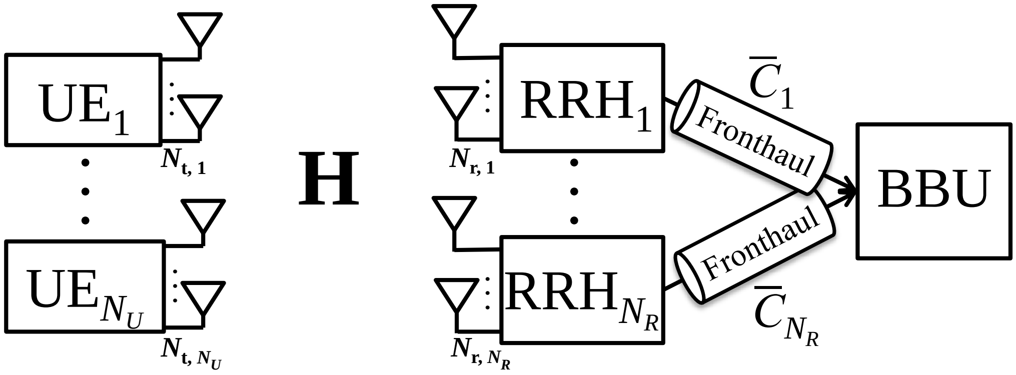

Fig. 1. Uplink of a C-RAN system consisting of N_U UEs and N_R RRHs. Each j-th RRH is connected to the BBU with a fronthaul link of capacity $\bar{C}_j$.

equipped with $N_{r,j}$ receive antennas. We define the number of total transmit antennas as $N_t = \sum_{i=1}^{N_U} N_{t,i}$. Each j-th RRH is connected to the BBU via a fronthaul link of capacity $\bar{C}_j$. All rates, including $\bar{C}_j$, are normalized to the bandwidth available on the uplink channel from the UEs to the RRHs and are measured in bits/s/Hz. We assume that coding is performed across a large number of channel coherence blocks, for example over many resource blocks of an LTE system operating on a channel with significant time-frequency diversity. This implies that the ergodic capacity describes the system performance in terms of achievable rates (see, e.g., [18]).

Each channel coherence block, of length T channel uses, is split into a phase for channel training of length T_p channel uses and a phase for data transmission of length T_d channel uses, with

$$T_p + T_d = T. \tag{1}$$

The signal transmitted by the i-th UE is given by a $N_{t,i} \times T$ complex matrix $\mathbf{X}_i$, where each column corresponds to the signal transmitted by the $N_{t,i}$ antennas in a channel use. This signal is divided into the $N_{t,i} \times T_p$ pilot signal $\mathbf{X}_{p,i}$ and the $N_{t,i} \times T_d$ data signal $\mathbf{X}_{d,i}$. We assume that the transmit signal $\mathbf{X}_i$ has a total per-block power constraint $T^{-1}E[||\mathbf{X}_i||^2] = \bar{P}_i$, and we define ${T_p}^{-1}E[||\mathbf{X}_{p,i}||^2] = P_{p,i}$ and ${T_d}^{-1}E[||\mathbf{X}_{d,i}||^2] = P_{d,i}$ as the powers used for training and data, respectively by the i-th UE. Note that $E[\cdot]$ refers throughout to the expectation operator. In terms of pilot and data signal powers, the power constraint is hence expressed as

$$\frac{T_p}{T}P_{p,i}+\frac{T_d}{T}P_{d,i}=\bar{P}_i. \tag{2}$$

For simplicity, we assume equal transmit power allocation for all UEs, and hence we have $\bar{P}_i=\bar{P}$, $P_{d,i}=P_d$ and $P_{p,i}=P_p$ for all $i\in\mathcal{N}_U$. Finally, we collect in matrices $\mathbf{X}_p$ and $\mathbf{X}_d$ all the pilot signals and the data signals transmitted by all UEs, respectively, i.e., $\mathbf{X}_p=[\mathbf{X}_{p,1}^T,\ldots,\mathbf{X}_{p,N_U}^T]^T$ and $\mathbf{X}_d=[\mathbf{X}_{d,1}^T,\ldots,\mathbf{X}_{d,N_U}^T]^T$.

The training signal is $\mathbf{X}_p=\sqrt{P_p/N_t}\mathbf{S}_p$, where $\mathbf{S}_p$ is a $N_t\times T_p$ matrix with orthogonal rows and unitary power entries corresponding to the orthogonal training sequences transmitted from each antenna by all UEs (as in, e.g., [16]). Note that this implies that each training sequence is transmitted with power P_p/N_t and that the condition $T_p\geq N_t$ holds. During the data phase, the UEs transmit independent space-time codewords without precoding. Using random coding arguments, we write $\mathbf{X}_d=\sqrt{P_d/N_t}\mathbf{S}_d$, where $\mathbf{S}_d$ is a $N_t\times T_d$ matrix of independent and identically distributed (i.i.d.) $\mathcal{CN}(0,1)$ variables.

The $N_{r,j}\times T$ signal $\mathbf{Y}_j$ received by the j-th RRH in a given coherence block, where each column corresponds to the signal received by the $N_{r,j}$ antennas in a channel use, can be split into the $N_{r,j}\times T_p$ received pilot signal $\mathbf{Y}_{p,j}$ and the $N_{r,j}\times T_d$ data signal $\mathbf{Y}_{d,j}$. The signal received at the j-th RRH is then given by

$$\begin{aligned}\mathbf{Y}_{p,j} &= \sqrt{\frac{P_p}{N_t}}\mathbf{H}_j\mathbf{S}_p+\mathbf{Z}_{p,j} &\text{(3a)}\\ \text{and}\quad \mathbf{Y}_{d,j} &= \sqrt{\frac{P_d}{N_t}}\mathbf{H}_j\mathbf{S}_d+\mathbf{Z}_{d,j}, &\text{(3b)}\end{aligned}$$

where $\mathbf{Z}_{p,j}$ and $\mathbf{Z}_{d,j}$ are respectively the $N_{r,j}\times T_p$ and $N_{r,j}\times T_d$ matrices of i.i.d. complex Gaussian noise variables with zero-mean and unit variance, i.e., $\mathcal{CN}(0,1)$. The $N_{r,j}\times N_t$ channel matrix $\mathbf{H}_j$ collects all the $N_{r,j}\times N_{t,i}$ channel matrices $\mathbf{H}_{ji}$ from the i-th UE to the j-th RRH as $\mathbf{H}_j=[\mathbf{H}_{j1},\ldots,\mathbf{H}_{jN_U}]$.

The channel matrix $\mathbf{H}_{ji}$ is modeled as having i.i.d. $\mathcal{CN}(0,\alpha_{ji})$ entries, where α_{ji} is the path loss coefficient between the i-th UE and the j-th RRH being given as

$$\alpha_{ji}=\frac{1}{1+\left(\frac{d_{ji}}{d_0}\right)^{\eta}}, \tag{4}$$

where d_{ji} is the distance between the i-th UE and the j-th RRH, d_0 is a reference distance, and η is the path loss exponent. The channel matrices are assumed to be constant during each channel coherence block and to change according to an ergodic process from block to block.

B. Conventional Approach

With the conventional approach, the RRH quantizes and compresses both its received pilot signal in Eq. (3a) and its received data signal in Eq. (3b), and forwards the compressed signals to the BBU on the fronthaul link. The BBU then estimates the CSI on the basis of the received quantized pilot signals and performs coherent decoding of the data signal. In the rest of Sec. III, we limit the analytical treatment to the case of a single UE and a single RRH, i.e., $N_U = 1$ and $N_R = 1$, for simplicity of presentation. We henceforth remove the subscripts indicating UE and RRH indices. A more general discussion can be found elsewhere [17].

1) Training Phase: During the training phase, the vector of received training signals $\mathbf{Y}_p$ in Eq. (3a) across all coherence times is quantized. In order to account for quantization and compression, throughout this chapter, we use the standard additive quantization noise model that follows conventional information-theoretical arguments based on random coding [19]. Accordingly, the quantized pilot signal can be written as

$$\widehat{\mathbf{Y}}_p = \mathbf{Y}_p + \mathbf{Q}_p, \tag{5}$$

where the compression noise matrix $\mathbf{Q}_p$ is assumed to have i.i.d. $\mathcal{CN}(0, \sigma_p^2)$ entries. Note that the assumption of Gaussian i.i.d. quantization noises is made here for simplicity of analysis without claim of optimality. On a practical note, Gaussian quantization noise can be realized by high-dimensional vector quantizers such as trellis-coded quantization [20]. The quantization noise variance σ_p^2 dictates the accuracy of the quantization and depends on the fronthaul capacity via standard information-theoretic identities [19], as further discussed below.

Based on Eq. (5), the channel matrix $\mathbf{H}$ from the UE to the RRH is estimated at the BBU by the minimum mean square error (MMSE) method. Hence, it can be expressed as

$$\mathbf{H} = \widehat{\mathbf{H}} + \mathbf{E}, \tag{6}$$

where the estimated channel $\widehat{\mathbf{H}}$ is a complex Gaussian matrix with i.i.d. $\mathcal{CN}(0, \sigma^2_{\widehat{h}})$ entries, and the estimation error $\mathbf{E}$ has i.i.d. $\mathcal{CN}(0, \sigma^2_e)$ entries. With $\sigma^2_{\widehat{h}} = \alpha - \sigma^2_e$ and $\sigma^2_e = \alpha N_t(1+\sigma^2_p)/(T_p P_p + N_t(1+\sigma^2_p))$, respectively [18], [21], where we recall that α is the power gain for the channel between UE and RRH.

2) Data Phase: The quantized data signal received at the BBU can be similarly expressed as $\widehat{\mathbf{Y}}_d = \mathbf{Y}_d + \mathbf{Q}_d$, where the quantization noise $\mathbf{Q}_d$ is assumed to have i.i.d. $\mathcal{CN}(0, \sigma^2_d)$ entries. Moreover, it can be written as the sum of a useful term $\widehat{\mathbf{H}}\mathbf{X}_d$ and of the equivalent noise $\mathbf{N}_d = \mathbf{E}\mathbf{X}_d + \mathbf{Z}_d + \mathbf{Q}_d$, namely

$$\widehat{\mathbf{Y}}_d = \widehat{\mathbf{H}}\mathbf{X}_d + \mathbf{N}_d, \tag{7}$$

where the equivalent noise $\mathbf{N}_d$ has i.i.d. entries with zero mean and power $1 + \sigma^2_d + P_d\sigma^2_e$. We observe that $\mathbf{N}_d$ is not Gaussian distributed and is not independent of $\mathbf{X}_d$. Further discussion can be found in the literature [17], [18].

3) Ergodic Rate: As mentioned, we adopt as the performance criterion of interest the ergodic rate, which, under the assumption of Gaussian codebooks, is given by the mutual information $T^{-1}I(\mathbf{X}_d; \widehat{\mathbf{Y}}_d|\widehat{\mathbf{H}})$ [bits/s/Hz] (see, e.g, [19, Ch. 3]). This quantity can be lower-bounded by the following expression [17]:

$$R = \frac{T_d}{T} E\left[\log_2 \det\left(\mathbf{I}_{N_r} + \rho_{\text{eff}}\widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger\right)\right], \tag{8}$$

with $\rho_{\text{eff}} = P_d/(N_t(1+\sigma^2_d + P_d\sigma^2_e))$ being the effective signal to noise ratio (SNR), which accounts for the effects of quantization and channel estimation, and $\widehat{\mathbf{H}}$ being distributed as in Eq. (6). The rate in Eq. (8) is hence an achievable ergodic rate [17]. Moreover, let us define as C_p the fronthaul rate allocated to transmit information about the pilot signals and as C_d the fronthaul rate for the data with $C_p + C_d = \bar{C}$. Then, if the conditions

$$C_p = \frac{T_p N_r}{T}\log_2\left(1 + \frac{P_p\alpha + 1}{\sigma^2_p}\right) \tag{9a}$$

$$\text{and } C_d = \frac{T_d N_r}{T}\log_2\left(1 + \frac{P_d\alpha + 1}{\sigma^2_d}\right) \tag{9b}$$

are satisfied, a quantization (and compression) scheme exists that guarantees the desired quantization errors (σ_d^2, σ_p^2) [17].

The ergodic achievable rate in Eq. (8) can now be optimized over the fronthaul allocation (C_p, C_d) under the fronthaul constraint $\bar{C} = C_p + C_d$, with C_p and C_d in Eq. (9), by maximizing the effective SNR ρ_{eff} in Eq. (8). This non-convex problem can be tackled using a line search method [22] in a bounded interval (e.g., over C_p in the interval $[0, \bar{C}]$).

C. Channel Estimation at the RRHs

With the mentioned alternative functional split, each RRH estimates the CSI on the basis of its received pilot signal in Eq. (3a), and then quantizes and compresses both its estimated CSI and its received data signal in Eq. (3b) for transmission on the fronthaul.

1) Training Phase: The RRH performs the MMSE estimate of the channel $\mathbf{H}$ given the observation $\mathbf{Y}_p$ in Eq. (3a). As a result, similar to Eq. (6), we can decompose the channel matrix $\mathbf{H}$ into the MMSE estimate $\widetilde{\mathbf{H}}$ and the independent estimation error $\mathbf{E}$, as

$$\mathbf{H} = \widetilde{\mathbf{H}} + \mathbf{E}, \tag{10}$$

where the error $\mathbf{E}$ has i.i.d. $\mathcal{CN}(0, \sigma_e^2)$ entries with $\sigma_e^2 = \alpha N_t/(T_p P_p + N_t)$ and $\widetilde{\mathbf{H}}$ has i.i.d. $\mathcal{CN}(0, \sigma_{\tilde{h}}^2)$ entries with $\sigma_{\tilde{h}}^2 = \alpha - \sigma_e^2$.

The sequence of channel estimates $\widetilde{\mathbf{H}}$ for all coherence times in the coding block is compressed by the RRH and forwarded to the BBU on the fronthaul link. The compressed channel $\widehat{\mathbf{H}}$ is related to the estimate $\widetilde{\mathbf{H}}$ as

$$\widetilde{\mathbf{H}} = \widehat{\mathbf{H}} + \mathbf{Q}_p, \tag{11}$$

where the $N_r \times N_t$ quantization noise matrix $\mathbf{Q}_p$ has i.i.d. $\mathcal{CN}(0, \sigma_p^2)$ entries.

2) Data Phase: During the data phase, the RRH quantizes the signal $\mathbf{Y}_d$ in Eq. (3b) and sends it to the BBU on the fronthaul link. The signal obtained at the BBU is related to $\mathbf{Y}_d$ as

$$\widehat{\mathbf{Y}}_d = \mathbf{Y}_d + \mathbf{Q}_d, \tag{12}$$

where $\mathbf{Q}_d$ is independent of $\mathbf{Y}_d$ and represents the quantization noise matrix with i.i.d. $\mathcal{CN}(0, \sigma_d^2)$ entries. Separating the desired signal and the noise in Eq. (12), the quantized signal $\widehat{\mathbf{Y}}_d$ can be expressed as

$$\widehat{\mathbf{Y}}_d = \widehat{\mathbf{H}}\mathbf{X}_d + \mathbf{N}_d, \tag{13}$$

where $\mathbf{N}_d$ denotes the equivalent noise $\mathbf{N}_d = (\mathbf{Q}_p + \mathbf{E})\,\mathbf{X}_d + \mathbf{Z}_d + \mathbf{Q}_d$, which has i.i.d. zero-mean entries with power

$$\sigma_n^2 = P_d\left(\sigma_p^2 + \sigma_e^2\right)1 + \sigma_d^2. \tag{14}$$

We observe that, as in Eq. (7), $\mathbf{N}_d$ is not Gaussian distributed and is not independent of $\mathbf{X}_d$.

3) Ergodic Rate: Let C_p and C_d denote respectively the fronthaul rates allocated for the transmission of the quantized channel estimates in Eq. (11) and of the quantized received signals in Eq. (12) on the fronthaul link from the RRH to the BBU. An achievable ergodic rate is given as [17]:

$$R = \frac{T_d}{T} E\left[\log_2 \det\left(\mathbf{I}_{N_r} + \rho_{\text{eff}}\widehat{\mathbf{H}}\widehat{\mathbf{H}}^{\dagger}\right)\right], \tag{15}$$

with the effective SNR

$$\rho_{\text{eff}} = \frac{P_d}{N_t\sigma_n^2} = \frac{P_d}{N_t\left(1 + \sigma_d^2 + P_d\left(\sigma_p^2 + \sigma_e^2\right)\right)}; \tag{16}$$

$\widehat{\mathbf{H}}$ being distributed as in Eq. (11); and with σ_e^2 in Eq. (10). Moreover, if the conditions

$$C_p = \frac{N_r N_t}{T}\log_2\left(\frac{\alpha - \sigma_e^2}{\sigma_p^2}\right) \tag{17a}$$

$$\text{and}\quad C_d = \frac{N_r T_d}{T}\log_2\left(1 + \left(\frac{\alpha P_d + 1}{\sigma_d^2}\right)\right), \tag{17b}$$

are satisfied, then a quantization scheme exists that guarantees the desired quantization error (σ_p^2, σ_d^2) [17]. The ergodic achievable rate in Eq. (15) can now be optimized over the fronthaul allocation (C_p, C_d) under the fronthaul constraint $\bar{C} = C_p + C_d$, with C_p and C_d in Eq. (17), by maximizing the effective SNR ρ_{eff} in Eq. (16) using a line search [22] in a bounded interval.

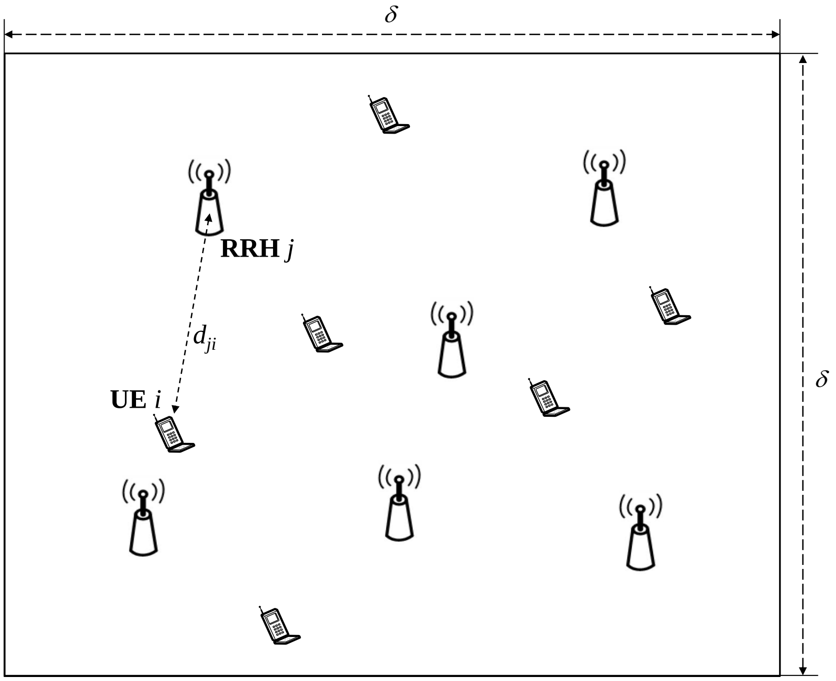

Fig. 2. Set-up under consideration for the numerical results, where RRHs and UEs are located in a square with side δ. All RRHs are connected to the same BBU.

4) Adaptive quantization: The alternative functional split studied here enables the RRHs to performs adaptive quantization of the data as a function of the estimated CSI in each coherence block. Specifically, rather than performing separate quantization of CSI and data, the data is quantized in each coherence period with a different accuracy depending on the corresponding CSI: a better channel quality calls for a more accurate quantization of the data field, and vice versa for worse CSI. We note that this is not possible in the conventional approach in which CSI is not estimated at the RRHs. Further details can be found elsewhere [17].

D. Numerical Results

In this section, we evaluate the performance of the discussed conventional and alternative strategies for the uplink. For the latter, we consider both the basic and adaptive implementations mentioned in the previous section. To this end, we consider a system with $N_R = N_U = 2$ RRHs and UEs with $N_t = N_r = 4$

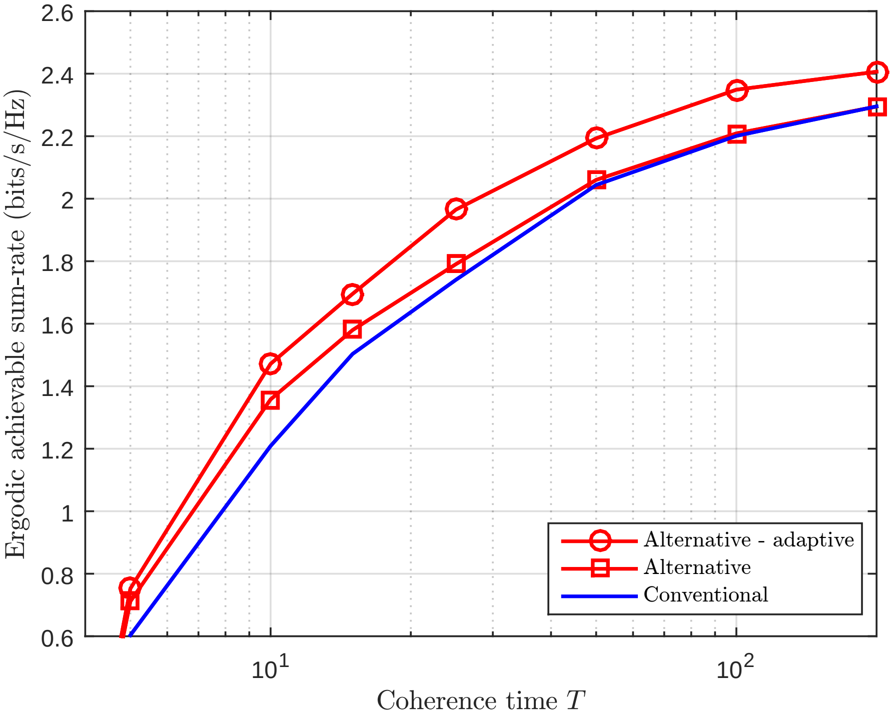

Fig. 3. Ergodic achievable sum-rate vs. coherence time ($N_R = N_U = 2$, $N_t = N_r = 4$, $\bar{C}$ = 6 bits/s/Hz, and $\bar{P} = 10dB$).

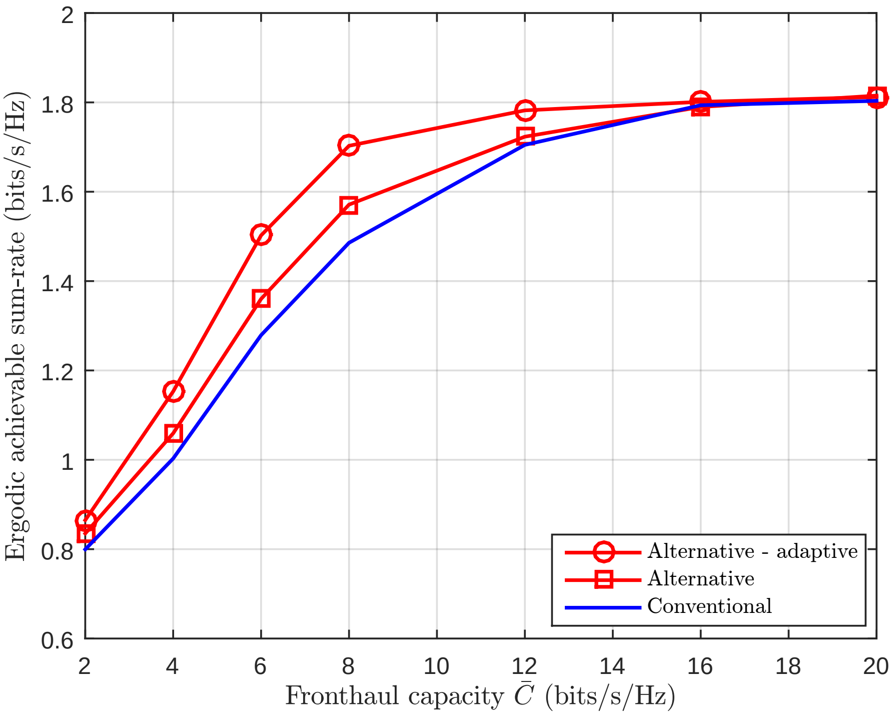

Fig. 4. Ergodic achievable sum-rate vs. fronthaul capacity ($N_R = N_U = 2$, $N_t = N_r = 4$, $\bar{P} = 10dB$, and T = 10).

antennas. The positions of the RRHs and the UEs are fixed[1] in the area with side $\delta = 500m$ as in Fig. 2. In the path loss formula Eq. (4), we set the reference distance to $d_0 = 50m$ and the path loss exponent to $\eta = 3$. Throughout, we assume that each RRH has the same fronthaul capacity $\bar{C}$, that is $\bar{C}_j = \bar{C}$ for $j \in \mathcal{N}_R$. We optimize over the power allocation (P_p, P_d) and we set $T_p = N_t$, which was shown to be optimal in [18] for a point-to-point link with no fronthaul limitation.

The effect of an increase of the coherence time on the ergodic achievable sum-rate is investigated in Fig. 3 with fronthaul capacity $\bar{C}$ = 6 bits/s/Hz, and power $\bar{P}$ = 10dB. As expected from information-theoretic considerations, Fig. 3 demonstrates that the alternative approach is advantageous, although most of the gains are accrued by means of adaptive quantization. Moreover, it is observed that the performance of the conventional approach without adaptive quantization approaches that of the alternative approach as the coherence time T increases. This is because, for large coherence time T, the fraction of fronthaul capacity devoted to training becomes negligible and hence accurate CSI can be obtained at the BBU.

In Fig. 4, we set the power as $\bar{P} = 10dB$ and the coherence time as $T = 10$, and we plot the ergodic achievable sum-rate versus the fronthaul capacity $\bar{C}$. The main conclusions are consistent with those discussed above for Fig. 3. Moreover, it is seen that the performance gain of the alternative functional split is relevant as long as $\bar{C}$ is not too large, in which case the performance is limited by the uplink SNR and not by the limited fronthaul capacity.

IV. Downlink: Where to Perform Channel Encoding and Precoding?

In this section, we turn to the downlink and address the issue of whether it is more advantageous to implement channel encoding and precoding at the RRHs rather than at the BBU as in the conventional implementation. Specifically, we compare the following two approaches:

- the conventional approach, in which the BBU performs channel coding and precoding and then quantizes and forwards the resulting baseband signals on the fronthaul links to the RRHs;

[1] The positions of RRHs are set as $\boldsymbol{p}_{R,1}$ = [307.50 233.18]^T and $\boldsymbol{p}_{R,2}$ = [430.3 192.64]^T, where $\boldsymbol{p}_{R,i}$ is the position of i-th RRH with coordinate origin at the lower left corner, and the positions of UEs as $\boldsymbol{p}_{U,1}$ = [363.7 316.66]^T and $\boldsymbol{p}_{U,2}$ = [438.17 107.09]^T, where $\boldsymbol{p}_{U,j}$ is the position of j-th UE.

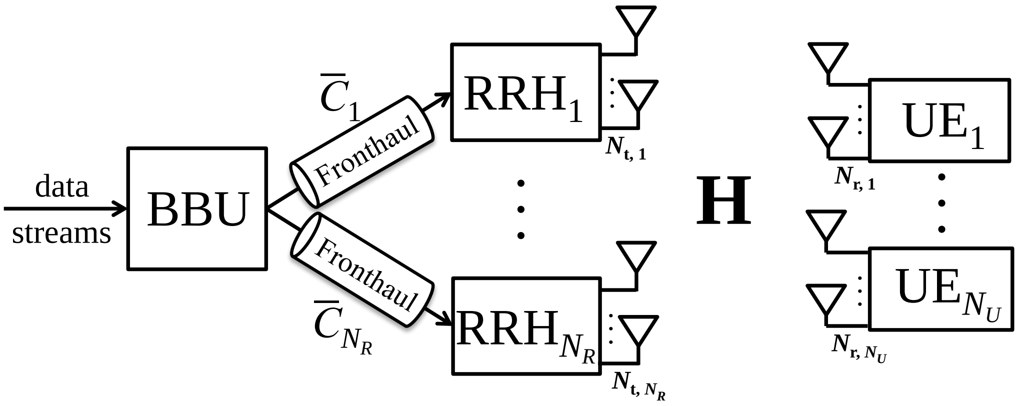

Fig. 5. Downlink of a C-RAN system consisting of N_R RRHs and N_U UEs. The BBU is connected to each i-th RRH with a fronthaul link of capacity $\bar{C}_i$.

- channel encoding and precoding at the RRHs in which the BBU does not perform precoding but rather forwards separately the information messages of a subset of UEs, along with the quantized precoding matrices to the all RRHs, which then perform channel encoding and precoding.

The conventional approach has been studied under a simplified quasi-static, rather than ergodic, channel model [23], [24], while the alternative functional split was investigated by Park *et al.* [25]. This section is adapted from our earlier paper [26], to which we refer for further details and proofs. We also note that we focus here on linear precoding, or beamforming, and separate quantization for each RRH, and that related discussion on non-linear precoding and joint fronthaul quantization can be found in the literature [24].

We start by detailing the system model in Sec. IV-A. In Sec. IV-B, we study the conventional approach, while the alternative functional split mentioned above is studied in IV-C. In Sec. IV-D, numerical results are presented.

A. System Model

We consider the counterpart downlink C-RAN model of the uplink set-up studied in Sec III, in which a cluster of N_R RRHs provides wireless service to N_U UEs as illustrated in Fig. 5. Most of the baseband processing for all the RRHs in the cluster is carried out at a BBU that is connected to each i-th RRH via a fronthaul link of finite capacity $\bar{C}_i$. Each i-th RRH has $N_{t,i}$ transmit antennas and each j-th UE has $N_{r,j}$ receive antennas. We denote the set of all RRHs as $\mathcal{N}_R = \{1, \ldots, N_R\}$ and the set of all UEs as $\mathcal{N}_U = \{1, \ldots, N_U\}$, and we define the number of total transmit antennas as $N_t = \sum_{i=1}^{N_R} N_{t,i}$ and of total receive antennas as $N_r = \sum_{j=1}^{N_U} N_{r,j}$. Moreover, we adopt a block-ergodic channel model in which the fading channels are constant within a coherence period but vary in an ergodic fashion across a large number of coherence periods.

Within each channel coherence period of duration T channel uses, the baseband signal transmitted by the i-th RRH is given by a $N_{t,i} \times T$ complex matrix $\mathbf{X}_i$, where each column corresponds to the signal transmitted from the $N_{t,i}$ antennas in a channel use. The $N_{r,j} \times T$ signal $\mathbf{Y}_j$ received by the j-th UE in a given channel coherence period, where each column corresponds to the signal received by the $N_{r,j}$ antennas in a channel use, is given by

$$\mathbf{Y}_j = \mathbf{H}_j \mathbf{X} + \mathbf{Z}_j, \tag{18}$$

where $\mathbf{Z}_j$ is the $N_{r,j} \times T$ noise matrix, which consist of i.i.d. $\mathcal{CN}(0, 1)$ entries; $\mathbf{H}_j = [\mathbf{H}_{j1}, \ldots, \mathbf{H}_{jN_R}]$ denotes the $N_{r,j} \times N_t$ channel matrix for j-th UE, where $\mathbf{H}_{ji}$ is the $N_{r,j} \times N_{t,i}$ channel matrix from the i-th RRH to the j-th UE; and $\mathbf{X}$ is the collection of the signals transmitted by all the RRHs, i.e., $\mathbf{X} = [\mathbf{X}_1^T, \ldots, \mathbf{X}_{N_R}^T]^T$.

We consider the scenario in which the BBU has instantaneous information about the channel matrix $\mathbf{H}$ as well as the case in which the BBU is only aware of the distribution of the channel matrix $\mathbf{H}$, i.e., it has *stochastic CSI*. Instead, the UEs always have full CSI about their corresponding channel matrices, as we will state more precisely in the next sections. The transmit signal $\mathbf{X}_i$ has a power constraint given as $T^{-1} E[\|\mathbf{X}_i\|^2] \leq \bar{P}_i$.

While the analysis applies more generally, in order to elaborate on the CSI requirements of the BBU, we consider as a specific channel model of interest the standard Kronecker model, in which the channel matrix $\mathbf{H}_{ji}$ is written as

$$\mathbf{H}_{ji} = \mathbf{\Sigma}_{R,ji}^{1/2} \widetilde{\mathbf{H}}_{ji} \mathbf{\Sigma}_{T,ji}^{1/2}, \tag{19}$$

where the $N_{t,i} \times N_{t,i}$ matrix $\mathbf{\Sigma}_{T,ji}$ and the $N_{r,j} \times N_{r,j}$ matrix $\mathbf{\Sigma}_{R,ji}$ are the transmit-side and receiver-

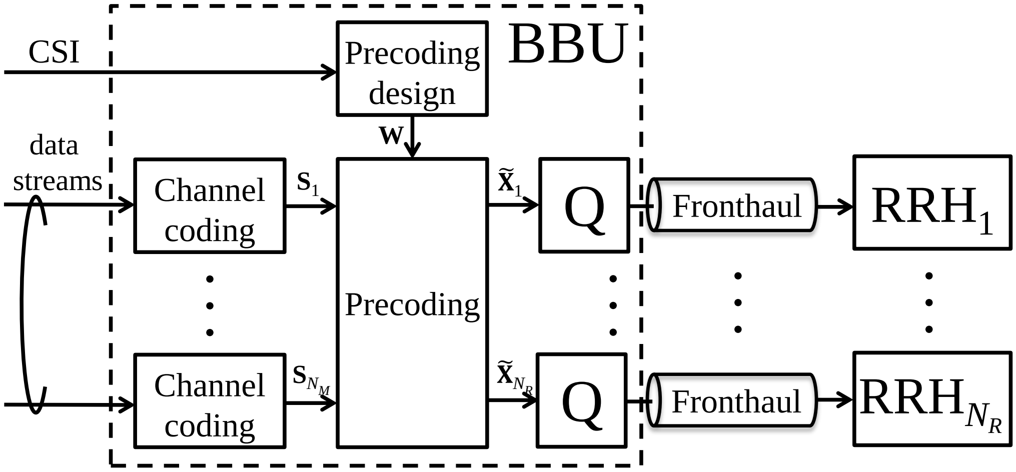

Fig. 6. Downlink: Conventional approach (“Q” represents fronthaul compression).

side spatial correlation matrices, respectively, and the $N_{r,j} \times N_{t,i}$ random matrix $\widetilde{\mathbf{H}}_{ji}$ has i.i.d. $\mathcal{CN}(0,1)$ variables and accounts for the small-scale multipath fading [27]. With this model, stochastic CSI entails that the BBU is only aware of the correlation matrices $\mathbf{\Sigma}_{T,ji}$ and $\mathbf{\Sigma}_{R,ji}$. Moreover, in case that the RRHs are placed in a higher location than the UEs, one can assume that the receive-side fading is uncorrelated, i.e., $\mathbf{\Sigma}_{R,ji} = \mathbf{I}_{N_{r,j}}$, while the transmit-side covariance matrix $\mathbf{\Sigma}_{T,ji}$ is determined by the one-ring scattering model (see [27] and references therein). In particular, if the RRHs are equipped with $\lambda/2$-spaced uniform linear arrays, we have $\mathbf{\Sigma}_{T,ji} = \mathbf{\Sigma}_T(\theta_{ji}, \Delta_{ji})$ for the j-th UE and the i-th RRH located at a relative angle of arrival θ_{ji} and having angular spread Δ_{ji}, where the element (m,n) of matrix $\mathbf{\Sigma}_T(\theta_{ji}, \Delta_{ji})$ is given by

$$[\mathbf{\Sigma}_T(\theta_{ji}, \Delta_{ji})]_{m,n} = \frac{\alpha_{ji}}{2\Delta_{ji}} \int_{\theta_{ji}-\Delta_{ji}}^{\theta_{ji}+\Delta_{ji}} \exp^{-j\pi(m-n)\sin(\phi)} \, d\phi, \tag{20}$$

with the path loss coefficient α_{ji} between the j-th UE and the i-th RRH being given as Eq. (4).

B. Conventional Approach

We first describe the conventional approach in Sec. IV-B1. Then, we discuss the joint optimization of fronthaul quantization and precoding with perfect instantaneous channel knowledge at the BBU in Sec. IV-B2 and under the assumption of stochastic CSI at the BBU in Sec. IV-B3.

1) Problem Formulation: With the conventional scheme as illustrated in Fig. 6, the BBU performs channel coding and precoding, and then quantizes the resulting baseband signals so that they can be forwarded on the fronthaul links to the corresponding RRHs. Specifically, channel coding is performed separately for the information stream intended for each UE. This step produces the data signal $\mathbf{S} = [\mathbf{S}_1^\dagger, \dots, \mathbf{S}_{N_U}^\dagger]^\dagger$ for each coherence block, where $\mathbf{S}_j$ is the $M_j \times T$ matrix containing, as rows, the $M_j \leq N_{r,j}$ encoded data streams for the j-th UE. We define the number of total data streams as $M = \sum_{j=1}^{N_U} M_j$ and assume the condition $M \leq N_t$. Following standard random coding arguments, we take all the entries of matrix $\mathbf{S}$ to be i.i.d. as $\mathcal{CN}(0, 1)$. The encoded data $\mathbf{S}$ is further processed to obtain the transmitted signals $\mathbf{X}$ as detailed below.

The precoded data signal computed by the BBU for any given coherence time can be written as $\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{S}$, where $\mathbf{W}$ is the $N_t \times M$ precoding matrix. With instantaneous CSI, a different precoding matrix $\mathbf{W}$ is used for different coherence times in the coding block, while, with stochastic CSI, the same precoding matrix $\mathbf{W}$ is used for all coherence times.

In both cases, the precoded data signal $\widetilde{\mathbf{X}}$ can be divided into the $N_{t,i} \times T$ signals $\widetilde{\mathbf{X}}_i$ corresponding to i-th RRH for all $i \in \mathcal{N}_R$ as $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}_1^T, \dots, \widetilde{\mathbf{X}}_{N_R}^T]^T$, with $\widetilde{\mathbf{X}}_i = \mathbf{W}_i^r\mathbf{S}$, where $\mathbf{W}_i^r$ is the $N_{t,i} \times N_r$ precoding matrix for the i-th RRH, which is obtained by properly selecting the rows of matrix $\mathbf{W}$ (as indicated by the superscript "r" for "rows"): the matrix $\mathbf{W}_i^r$ is given as $\mathbf{W}_i^r = \mathbf{D}_i^{rT}\mathbf{W}$, with the $N_t \times N_{t,i}$ matrix $\mathbf{D}_i^r$ having all zero elements except for the rows from $\sum_{k=1}^{i-1} N_{t,k} + 1$ to $\sum_{k=1}^{i} N_{t,k}$, that contain an $N_{t,i} \times N_{t,i}$ identity matrix.

The BBU quantizes each sequence of baseband signal $\widetilde{\mathbf{X}}_i$ for transmission on the i-th fronthaul link to the i-th RRH independently. We write the compressed signals $\mathbf{X}_i$ for the i-th RRH as

$$\mathbf{X}_i = \widetilde{\mathbf{X}}_i + \mathbf{Q}_{x,i}, \tag{21}$$

where the quantization noise matrix $\mathbf{Q}_{x,i}$ is assumed to have i.i.d. $\mathcal{CN}(0, \sigma_{x,i}^2)$ entries. Note that the advantages of joint quantization across multiple RRHs are explored in [24] for static channels. Based on Eq. (21), the design of the fronthaul compression reduces to the optimization of the quantization noise variances $\sigma^2_{x,1}, \ldots, \sigma^2_{x,N_R}$. The power transmitted by i-th RRH is computed as

$$P_i\left(\mathbf{W}, \sigma^2_{x,i}\right) = \frac{1}{T} E[||\mathbf{X}_i||^2] = \text{tr}\left(\mathbf{D}_i^{rT}\mathbf{W}\mathbf{W}^\dagger\mathbf{D}_i^r + \sigma^2_{x,i}\mathbf{I}\right), \tag{22}$$

where we have emphasized the dependence of the power $P_i(\mathbf{W}, \sigma^2_{x,i})$ on the precoding matrix $\mathbf{W}$ and quantization noise variances $\sigma^2_{x,i}$. Moreover, using standard rate-distortion arguments, the rate required on the fronthaul between the BBU and i-th RRH in a given coherence interval can be quantified by $I(\widetilde{\mathbf{X}}_i; \mathbf{X}_i)/T$ (see, e.g., [19, Ch. 3]), yielding [26]

$$C_i\left(\mathbf{W}, \sigma^2_{x,i}\right) = \log\det\left(\mathbf{D}_i^{rT}\mathbf{W}\mathbf{W}^\dagger\mathbf{D}_i^r + \sigma^2_{x,i}\mathbf{I}\right) - N_{t,i}\log\left(\sigma^2_{x,i}\right), \tag{23}$$

so that the fronthaul capacity constraint is $C_i(\mathbf{W}, \sigma^2_{x,i}) \leq \bar{C}_i$.

We assume that each j-th UE is aware of the effective receive channel matrices $\widetilde{\mathbf{H}}_{jk} = \mathbf{H}_j\mathbf{W}_k^c$ for all $k \in \mathcal{N}_U$ at all coherence times, where $\mathbf{W}_k^c$ is the $N_t \times N_{r,j}$ precoding matrix corresponding to k-th UE, which is obtained from the precoding matrix $\mathbf{W}$ by properly selecting the columns as $\mathbf{W} = [\mathbf{W}_1^c, \ldots, \mathbf{W}_{N_U}^c]$. We collect the effective channels in the matrix $\widetilde{\mathbf{H}}_j = [\widetilde{\mathbf{H}}_{j1}, \ldots, \widetilde{\mathbf{H}}_{jN_U}] = \mathbf{H}_j\mathbf{W}$. The effective channel $\widetilde{\mathbf{H}}_j$ can be estimated at the UEs via downlink training.

Under these assumptions, the ergodic achievable rate for the j-th UE is computed as $E[R_j^{conv}(\mathbf{H}, \mathbf{W}, \boldsymbol{\sigma}_x^2)]$, with $R_j^{conv}(\mathbf{H}, \mathbf{W}, \boldsymbol{\sigma}_x^2) = I_{\mathbf{H}}(\mathbf{S}_j; \mathbf{Y}_j)/T$, where $I_{\mathbf{H}}(\widetilde{\mathbf{S}}_j; \mathbf{Y}_j)$ represents the mutual information for a fixed realization of the channel matrix $\mathbf{H}$, the expectation is taken with respect to $\mathbf{H}$ and

$$\begin{aligned} R_j^{conv}(\mathbf{H}, \mathbf{W}, \boldsymbol{\sigma}_x^2) = {} & \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\mathbf{W}\mathbf{W}^\dagger + \boldsymbol{\Omega}_x\right)\mathbf{H}_j^\dagger\right) \\ & - \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\sum_{k\in\mathcal{N}_U\setminus j}\mathbf{W}_k^c\mathbf{W}_k^{c\dagger} + \boldsymbol{\Omega}_x\right)\mathbf{H}_j^\dagger\right). \end{aligned} \tag{24}$$

In Eq. (24), the covariance matrix $\boldsymbol{\Omega}_x$ is a diagonal with diagonal blocks given as diag$([\sigma^2_{x,1}\mathbf{I}, \ldots, \sigma^2_{x,N_R}\mathbf{I}])$ and $\boldsymbol{\sigma}_x^2 = [\sigma^2_{x,1}, \ldots, \sigma^2_{x,N_R}]^T$.

The ergodic achievable weighted sum-rate can be optimized over the precoding matrix $\mathbf{W}$ and the compression noise variances $\boldsymbol{\sigma}_x^2$ under fronthaul capacity and power constraints. In the next subsections, we consider separately the cases with instantaneous and stochastic CSI.

2) Instantaneous CSI: In the case of instantaneous channel knowledge at the BBU, the design of the precoding matrix $\mathbf{W}$ and the compression noise variances $\boldsymbol{\sigma}_x^2$, is adapted to the channel realization $\mathbf{H}$ for each coherence block. To emphasize this fact, we use the notation $\mathbf{W}(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$. The problem of optimizing the ergodic weighted achievable sum-rate with given weights $\mu_j \geq 0$ for $j \in \mathcal{N}_M$ is then formulated as

$$\begin{aligned}
\underset{\mathbf{W}(\mathbf{H}),\boldsymbol{\sigma}_x^2(\mathbf{H})}{\text{maximize}} \quad & \sum_{j\in\mathcal{N}_U} \mu_j E\left[R_j^{conv}\left(\mathbf{H}, \mathbf{W}(\mathbf{H}), \boldsymbol{\sigma}_x^2(\mathbf{H})\right)\right] && \text{(25a)}\\
\text{s.t.} \quad & C_i\left(\mathbf{W}, \sigma_{x,i}^2(\mathbf{H})\right) \leq \bar{C}_i, && \text{(25b)}\\
& P_i\left(\mathbf{W}(\mathbf{H}), \sigma_{x,i}^2(\mathbf{H})\right) \leq \bar{P}_i, && \text{(25c)}
\end{aligned}$$

where Eq. (25b)-(25c) apply for all $i \in \mathcal{N}_R$ and all channel realizations $\mathbf{H}$. Due to the separability of the fronthaul and power constraints across the channel realizations $\mathbf{H}$, the problem in Eq. (25) can be solved for each $\mathbf{H}$ independently. Note that the achievable rate in Eq. (25a) and the fronthaul constraint in Eq. (25b) are non-convex. However, the functions $R_j^{conv}(\mathbf{H}, \mathbf{W}(\mathbf{H}), \boldsymbol{\sigma}_x^2(\mathbf{H}))$ and $C_i(\mathbf{W}(\mathbf{H}), \sigma_{x,i}^2(\mathbf{H}))$ are difference of convex (DC) functions of the covariance matrices $\widetilde{\mathbf{V}}_j(\mathbf{H}) = \widetilde{\mathbf{W}}_j^c(\mathbf{H})\widetilde{\mathbf{W}}_j^{c\dagger}(\mathbf{H})$ for all $j \in \mathcal{N}_U$ and the variance $\boldsymbol{\sigma}_x^2(\mathbf{H})$. The resulting rank-relaxed problem can be tackled via the Majorization-Minimization (MM) algorithm as detailed in [24], from which a feasible solution of problem in Eq. (25) can be obtained. We refer to [24] for details.

3) Stochastic CSI: With only stochastic CSI at the BBU, in contrast to the case with instantaneous CSI, the same precoding matrix $\mathbf{W}$ and compression noise variances $\boldsymbol{\sigma}_x^2$ are used for all the coherence blocks. Accordingly, the problem of optimizing the ergodic weighted achievable sum-rate can be reformulated as

$$\begin{aligned}
\underset{\mathbf{W},\boldsymbol{\sigma}_x^2}{\text{maximize}} \quad & \sum_{j\in\mathcal{N}_U} \mu_j E\left[R_j^{conv}\left(\mathbf{H}, \mathbf{W}, \boldsymbol{\sigma}_x^2\right)\right] && \text{(26a)}\\
\text{s.t.} \quad & C_i\left(\mathbf{W}, \sigma_{x,i}^2\right) \leq \bar{C}_i, && \text{(26b)}\\
& P_i\left(\mathbf{W}, \sigma_{x,i}^2\right) \leq \bar{P}_i, && \text{(26c)}
\end{aligned}$$

TABLE I

DESIGN OF FRONTHAUL COMPRESSION AND PRECODING: CONVENTIONAL APPROACH WITH STOCHASTIC CSI

```
**Initialization**: Initialize the covariance matrices $\mathbf{V}^{(0)}$ and the quantization noise variances $\boldsymbol{\sigma}_x^{2\ (0)}$, and set $n = 0$.
**repeat (outer loop)**
    $n \leftarrow n + 1$
    Generate a channel matrix realization $\mathbf{H}^{(n)}$ using the available stochastic CSI.
    **Initialization**: Initialize $\mathbf{V}^{(n,0)} = \mathbf{V}^{(n-1)}$ and $\boldsymbol{\sigma}_x^{2\ (n,0)} = \boldsymbol{\sigma}_x^{2\ (n-1)}$, and set $r = 0$.
    **repeat (inner loop)**
        $r \leftarrow r + 1$
            $\max_{\mathbf{V},\boldsymbol{\sigma}_x^2} \quad \frac{1}{n}\sum_{l=1}^{n}\sum_{j\in\mathcal{N}_U}\mu_j\widetilde{R}_j^{conv}\left(\mathbf{H}^{(l)},\mathbf{V},\boldsymbol{\sigma}_x^2|\mathbf{V}^{(l-1)},\boldsymbol{\sigma}_x^{2\ (l-1)}\right)$
            $\text{s.t.} \quad \widetilde{C}_i\left(\mathbf{V},\sigma_{x,i}^2|\mathbf{V}^{(n,r-1)},\sigma_{x,i}^{2\ (n,r-1)}\right) \leq \bar{C}_i,$
            $\quad\quad P_i\left(\mathbf{V},\sigma_{x,i}^2\right) \leq \bar{P}_i, \quad \text{for all } i \in \mathcal{N}_R.$
        Update $\mathbf{V}^{(n,r)} \leftarrow \mathbf{V}$ and $\boldsymbol{\sigma}_x^{2\ (n,r)} \leftarrow \boldsymbol{\sigma}_x^2$.
    **until** a convergence criterion is satisfied.
    Update $\mathbf{V}^{(n)} \leftarrow \mathbf{V}^{(n,r)}$ and $\boldsymbol{\sigma}_x^{2\ (n)} \leftarrow \boldsymbol{\sigma}_x^{2\ (n,r)}$.
**until** a convergence criterion is satisfied.
**Solution**: Calculate the precoding matrix $\mathbf{W}$ from the covariance matrices $\mathbf{V}^{(n)}$ via rank reduction as $\mathbf{W}_j = \gamma_j\nu_{\text{max}}^{(M_j)}(\mathbf{V}_j^{(n)})$ for all $j \in \mathcal{N}_U$, where $\gamma_j$ is obtained by imposing $P_i\left(\mathbf{W},\sigma_{x,i}^2\right) = \bar{P}_i$ using Eq. (22).
```

where Eq. (26b)-(26c) apply to all $i \in \mathcal{N}_R$. In order to tackle this problem, we adopt the Stochastic Successive Upper-bound Minimization (SSUM) method [28], whereby, at each step, a stochastic lower bound of the objective function is maximized around the current iterate[2]. To this end, similar to [24], we can recast the optimization over the covariance matrices $\mathbf{V}_j = \mathbf{W}_j^c\mathbf{W}_j^{c\dagger}$ for all $j \in \mathcal{N}_U$, instead of the precoding matrices $\mathbf{W}_j^c$ for all $j \in \mathcal{N}_U$. We observe that, with this choice, the objective function is expressed as the average of DC functions, while the constraint in Eq. (26b) is also a DC function, with respect to the covariance $\mathbf{V} = [\mathbf{V}_1 \ldots \mathbf{V}_{N_U}]$ and the quantization noise variances $\boldsymbol{\sigma}_x^2$. Due to the DC structure, locally tight (stochastic) convex lower bounds can be calculated for objective function in Eq. (26a) and the constraint in Eq. (26b) (see, e.g., [30]).

The algorithm proposed in [26] is based on SSUM [28] and contains two nested loops. At each outer iteration n, a new channel matrix realization $\mathbf{H}^{(n)} = [\mathbf{H}_1^{T\ (n)}, \ldots, \mathbf{H}_{N_U}^{T\ (n)}]$ is drawn based on the availability

[2] We mention here that an alternative method to attack the problem is the strategy introduced in [29].

of stochastic CSI at the BBU. For example, with the model in Eq. (19), the channel matrices are generated based on the knowledge of the spatial correlation matrices. Following the SSUM scheme, the outer loop aims at maximizing a stochastic lower bound on the objective function, given as

$$\frac{1}{n}\sum_{l=1}^{n}\widetilde{R}_j^{conv}\left(\mathbf{H}^{(l)},\mathbf{V},\boldsymbol{\sigma}_x^2|\mathbf{V}^{(l-1)},\boldsymbol{\sigma}_x^{2\,(l-1)}\right), \tag{27}$$

where $\widetilde{R}_j^{conv}(\mathbf{H}^{(l)},\mathbf{V},\boldsymbol{\sigma}_x^2|\mathbf{V}^{(l-1)},\boldsymbol{\sigma}_x^{2\,(l-1)})$ is a locally tight convex lower bound on $R_j^{conv}(\mathbf{H},\,\mathbf{W},\boldsymbol{\sigma}_x^2)$ around solution $\mathbf{V}^{(l-1)}$, $\boldsymbol{\sigma}_x^{2\,(l-1)}$ obtained at the $(l-1)$ the outer iteration when the channel realization is $\mathbf{H}^{(l)}$. This can be calculated as (see, e.g., [28])

$$\begin{aligned}\widetilde{R}_j^{conv}(\mathbf{H}^{(l)},\mathbf{V},\boldsymbol{\sigma}_x^2|\mathbf{V}^{(l-1)},\boldsymbol{\sigma}_x^{2\,(l-1)}) &\triangleq \log\det\left(\mathbf{I}+\mathbf{H}_j^{(l)}\left(\sum_{k=1}^{N_U}\mathbf{V}_k+\boldsymbol{\Omega}_x\right)\mathbf{H}_j^{(l)\,\dagger}\right)\\ &-f\left(\mathbf{I}+\mathbf{H}_j^{(l)}\boldsymbol{\Lambda}_j^{(l-1)}\mathbf{H}_j^{(l)\,\dagger},\mathbf{I}+\mathbf{H}_j^{(l)}\boldsymbol{\Lambda}_j\mathbf{H}_j^{(l)\,\dagger}\right),\end{aligned} \tag{28}$$

where $\boldsymbol{\Lambda}_j=\sum_{k=1,k\neq j}^{N_U}\mathbf{V}_k+\boldsymbol{\Omega}_x$, $\boldsymbol{\Lambda}_j^{(l-1)}=\sum_{k=1,k\neq j}^{N_U}\mathbf{V}_k^{(l-1)}+\boldsymbol{\Omega}_x$, the covariance matrix $\boldsymbol{\Omega}_x^{(l)}$ is a diagonal matrix with diagonal blocks given as $\mathrm{diag}([\sigma_{x,1}^{2\,(l)}\mathbf{I},\ldots,\,\sigma_{x,N_R}^{2\,(l)}\mathbf{I}])$ and the linearized function $f(\mathbf{A},\mathbf{B})$ is obtained from the first-order Taylor expansion of the log det function as

$$f(\mathbf{A},\mathbf{B})\triangleq\log\det\left(\mathbf{A}\right)+\frac{1}{\ln 2}\mathrm{tr}\left(\mathbf{A}^{-1}\left(\mathbf{B}-\mathbf{A}\right)\right). \tag{29}$$

Since the maximization of Eq. (27) is subject to the non-convex DC constraint in Eq. (26b), the inner loop tackles the problem via the MM algorithm i.e., by applying successive locally tight convex lower bounds to the left-hand side of the constraint in Eq. (26b) [31]. Specifically, given the solution $\mathbf{V}^{(n,r-1)}$ and $\boldsymbol{\sigma}_x^{2\,(n,r-1)}$ at $(r-1)$-th inner iteration of the n-th outer iteration, the fronthaul constraint in Eq. (26b) at the r-th inner iteration can be locally approximated as

$$\begin{aligned}&\widetilde{C}_i\left(\mathbf{V},\sigma_{x,i}^2|\mathbf{V}^{(n,r-1)},\sigma_{x,i}^{2\,(n,r-1)}\right)\triangleq \\ &f\left(\sum_{k=1}^{N_U}\mathbf{D}_i^{rT}\mathbf{V}_k^{(n,r-1)}\mathbf{D}_i^r+\sigma_{x,i}^{2\,(n,r-1)}\mathbf{I},\sum_{k=1}^{N_U}\mathbf{D}_i^{rT}\mathbf{V}_k\mathbf{D}_i^r+\sigma_{x,i}^2\mathbf{I}\right)-N_{t,i}\log\left(\sigma_{x,i}^2\right).\end{aligned} \tag{30}$$

The resulting combination of SSUM and MM for the solution of problem in Eq. (26) is summarized in Table Algorithm I. The algorithm is completed by calculating, from the obtained solution $\mathbf{V}^*$ of the relaxed problem, the precoding matrix $\mathbf{W}$ by using the standard rank-reduction approach [32], which is

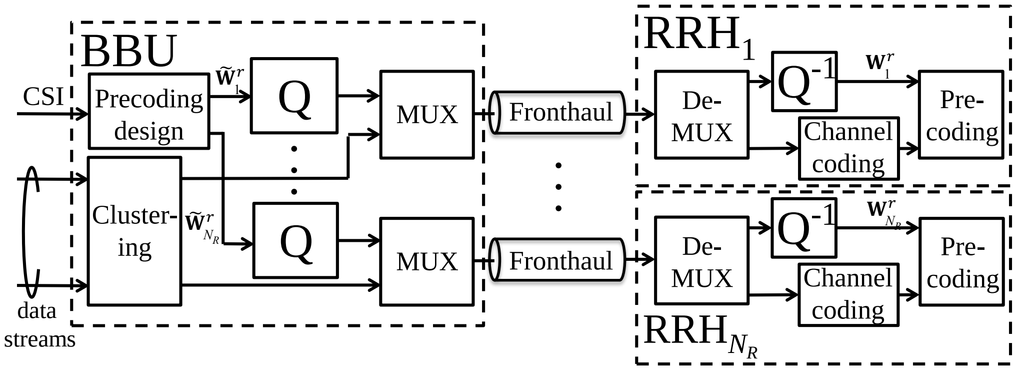

Fig. 7. Downlink: Alternative functional split ("Q" and "Q^{-1}" represents fronthaul compression and decompression, respectively).

given as $\mathbf{W}_j^* = \gamma_j \nu_{\max}^{(M_j)}(\mathbf{V}_j^*)$ with the normalization factor γ_j, selected so as to satisfy the power constraint with equality, namely $P_i(\mathbf{W}, \sigma_{x,i}^2) = \bar{P}_i$.

We finally note that, since the approximated functions in Eq. (28) and Eq. (30) are local lower bounds, the algorithm provides a feasible solution of the relaxed problem at each inner and outer iteration (see, e.g., [28]).

C. Channel Encoding and Precoding at the RRHs

With this alternative functional split, the BBU calculates the precoding matrices, but does not perform precoding. Instead, as illustrated in Fig. 7, it uses the fronthaul links to communicate the information messages of a given subset of UEs to each RRH, along with the corresponding compressed precoding matrices. Each RRH can then encode and precode the messages of the given UEs based on the information received from the fronthaul link. As it will be discussed, with this approach, a preliminary clustering step is generally advantageous whereby each UE is assigned to a subset of RRHs. In the following, we first describe the strategy in Sec. IV-C1. Then we discuss the design problem for fronthaul quantization and precoding under instantaneous CSI in Sec. IV-C2 and with stochastic CSI in Sec. IV-C3.

1) Problem Formulation: As shown in Fig. 7, the precoding matrix $\widetilde{\mathbf{W}}$ and the information streams are separately transmitted from the BBU to the RRHs, and the received information bits are encoded and precoded at each RRH using the received precoding matrix. Note that, with this scheme, the transmission overhead over the fronthaul depends on the number of UEs supported by a RRH, since the RRHs should receive all the corresponding information streams.

Given the above, we allow for a preliminary clustering step at the BBU whereby each RRH is assigned by a subset of the UEs. We denote the set of UEs assigned by i-th RRH as $\mathcal{M}_i \subseteq \mathcal{N}_U$ for all $i \in \mathcal{N}_R$. This implies that i-th RRH only needs the information streams intended for the UEs in the set $\mathcal{M}_i$. We also denote the set of RRHs that serve the j-th UE, as $\mathcal{B}_j = \{i|j \in \mathcal{M}_i\} \subseteq \mathcal{N}_R$ for all $j \in \mathcal{N}_U$. We use the notation $\mathcal{M}_i[k]$ and $\mathcal{B}_j[m]$ to respectively denote the k-th UE and m-th RRH in the sets $\mathcal{M}_i$ and $\mathcal{B}_j$, respectively. We define the number of all transmit antennas for the RRHs, which serve the j-th UE, as $N_{t,\mathcal{B}_j}$. We assume here that the sets of UEs assigned by i-th RRH are given and not subject to optimization (see Sec. IV-D for further details).

The precoding matrix $\widetilde{\mathbf{W}}$ is constrained to have zeros in the positions that correspond to RRH-UE pairs such that the UE is not served by the given RRH. This constraint can be represented as

$$\widetilde{\mathbf{W}} = \left[\mathbf{E}_1^c \widetilde{\mathbf{W}}_1^c, \ldots, \mathbf{E}_{N_U}^c \widetilde{\mathbf{W}}_{N_U}^c\right], \tag{31}$$

where $\widetilde{\mathbf{W}}_j^c$ is the $N_{t,\mathcal{B}_j} \times N_{r,j}$ precoding matrix intended for j-th UE and RRHs in the cluster $\mathcal{B}_j$, and the $N_t \times N_{t,\mathcal{B}_j}$ constant matrix $\mathbf{E}_j^c$ ($\mathbf{E}_j^c$ only has either a 0 or 1 entries) defines the association between the RRHs and the UEs as $\mathbf{E}_j^c = [\mathbf{D}_{\mathcal{B}_j[1]}^c, \ldots, \mathbf{D}_{\mathcal{B}_j[|\mathcal{B}_j|]}^c]$, with the $N_r \times N_{r,j}$ matrix $\mathbf{D}_j^c$ having all zero elements except for the rows from $\sum_{k=1}^{j-1} N_{r,k} + 1$ to $\sum_{k=1}^{j} N_{r,j}$, which contain an $N_{r,j} \times N_{r,j}$ identity matrix.

The sequence of the $N_{t,i} \times N_{r,\mathcal{M}_i}$ precoding matrices $\widetilde{\mathbf{W}}_i^r$ intended for each i-th RRH for all coherence times in the coding block is compressed by the BBU and forwarded over the fronthaul link to the i-th RRH. The compressed precoding matrix $\mathbf{W}_i^r$ for i-th RRH is given by

$$\mathbf{W}_i^r = \widetilde{\mathbf{W}}_i^r + \mathbf{Q}_{w,i}, \tag{32}$$

where the $N_{t,i} \times N_{r,\mathcal{M}_i}$ quantization noise matrix $\mathbf{Q}_{w,i}$ is assumed to have zero-mean i.i.d. $\mathcal{CN}(0, \sigma_{w,i}^2)$ entries and to be independent across the index i. Overall, the $N_t \times N_r$ compressed precoding matrix $\mathbf{W}$ for all RRHs is represented as

$$\mathbf{W} = \widetilde{\mathbf{W}} + \mathbf{Q}_w, \tag{33}$$

where $\mathbf{W} = [\mathbf{E}_1^{r\dagger}\mathbf{W}_{w,1}^{\dagger}, \ldots, \mathbf{E}_{N_R}^{r\dagger}\mathbf{W}_{w,N_R}^{\dagger}]^{\dagger}$, $\widetilde{\mathbf{W}}$ and $\mathbf{Q}_w$ are similarly defined.

Similar to Eq. (24), an ergodic rate achievable for j-th UE can be written as $E[R_j^{alt}(\mathbf{H}, \widetilde{\mathbf{W}}, \boldsymbol{\sigma}_w^2)]$, where

$$R_j^{alt}\Big(\mathbf{H}, \widetilde{\mathbf{W}}, \boldsymbol{\sigma}_w^2\Big) = \frac{1}{T} I_{\mathbf{H}}\left(\mathbf{S}_j; \mathbf{Y}_j\right) = \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^{\dagger} + \boldsymbol{\Omega}_w\right)\mathbf{H}_j^{\dagger}\right)$$
$$- \log\det\left(\mathbf{I} + \mathbf{H}_j\left(\sum_{k\in\mathcal{N}_U\setminus j}\widetilde{\mathbf{W}}_k^c\widetilde{\mathbf{W}}_k^{c\dagger} + \boldsymbol{\Omega}_w\right)\mathbf{H}_j^{\dagger}\right). \tag{34}$$

2) Instantaneous CSI: With perfect CSI at the BBU, as discussed in Sec. IV-B2, one can adapt the precoding matrix $\widetilde{\mathbf{W}}(\mathbf{H})$, the user rates $\{R_j(\mathbf{H})\}$ and the quantization noise variances $\boldsymbol{\sigma}_w^2(\mathbf{H})$ to the current channel realization at each coherence block. The rate required to transmit precoding information on the i-th fronthaul in a given channel realizations $\mathbf{H}$ is given by $C_i(\mathbf{H}, \widetilde{\mathbf{W}}_i^r, \sigma_{w,i}^2)/T$, with

$$\frac{1}{T}C_i\left(\mathbf{H}, \widetilde{\mathbf{W}}_i^r, \sigma_{w,i}^2\right) = \frac{1}{T} I_{\mathbf{H}}(\widetilde{\mathbf{W}}_i^r; \mathbf{W}_i^r) \tag{35}$$
$$= \frac{1}{T}\Big\{\log\det\Big(\mathbf{D}_i^{rT}\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^{\dagger}\mathbf{D}_i^r + \sigma_{w,i}^2\mathbf{I}\Big) - N_{t,i}\log\left(\sigma_{w,i}^2\right)\Big\},$$

where the rate $C_i(\widetilde{\mathbf{W}}_i^r, \sigma_{w,i}^2)$ required on i-fronthaul link is defined in Eq. (23). Note that the normalization by T is needed since only a single precoding matrix is needed for each channel coherence interval. Then, under the fronthaul capacity constraint, the remaining fronthaul capacity that can be used to convey precoding information corresponding to the i-th RRH is $\bar{C}_i - \sum_{j\in\mathcal{M}_i} R_j$. As a result, the optimization problem of interest can be formulated as

$$\underset{\widetilde{\mathbf{W}}(\mathbf{H}), \boldsymbol{\sigma}_{w,i}^2(\mathbf{H}), \{R_j(\mathbf{H})\}}{\text{maximize}} \sum_{j\in\mathcal{N}_U} \mu_j R_j(\mathbf{H}) \tag{36a}$$

$$s.t. \quad R_j(\mathbf{H}) \leq R_j^{alt}\left(\mathbf{H}, \widetilde{\mathbf{W}}(\mathbf{H}), \boldsymbol{\sigma}_w^2(\mathbf{H})\right), \tag{36b}$$

$$\frac{1}{T}C_i\Big(\mathbf{H}, \widetilde{\mathbf{W}}_i^r(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H})\Big) \leq \bar{C}_i - \sum_{j\in\mathcal{M}_i} R_j(\mathbf{H}), \tag{36c}$$

$$P_i\left(\widetilde{\mathbf{W}}_i^r(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H})\right) \leq \bar{P}_i, \tag{36d}$$

where the constraints apply to all channel realization, Eq. (36b) applies to all $j \in \mathcal{N}_U$, Eq. (36c) - (36d) apply to all $i \in \mathcal{N}_R$ and the transmit power $P_i(\widetilde{\mathbf{W}}_i^r(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}))$ at i-th RRH is defined in Eq. (22). Similar to Sec. IV-B2, the problem in Eq. (36) can be solved for each channel realization $\mathbf{H}$ independently. In addition, each subproblem can be tackled by using MM algorithm [24].

3) Stochastic CSI: With stochastic CSI at the BBU, the same precoding matrix is used for all the coherence blocks and hence the rate required to convey the precoding matrix $\widetilde{\mathbf{W}}_i^r$ to each i-th RRH becomes negligible. As a result, we can neglect the effect of the quantization noise and set $\sigma_{w,i}^2 = 0$ for all $i \in \mathcal{N}_R$. Accordingly, the fronthaul capacity can be used to transfer the information stream under the constraint $\sum_{j\in\mathcal{M}_i} R_j \leq \bar{C}_i$, for all $i \in \mathcal{N}_R$. Based on the above considerations, the optimization problem of interest is formulated as

$$\underset{\widetilde{\mathbf{W}},\{R_j\}}{\text{maximize}} \quad \sum_{j\in\mathcal{N}_U} \mu_j R_j \tag{37a}$$

$$s.t. \quad R_j \leq E\left[R_j^{alt}\left(\mathbf{H}, \widetilde{\mathbf{W}}, \mathbf{0}\right)\right], \tag{37b}$$

$$\sum_{j\in\mathcal{M}_i} R_j \leq \bar{C}_i, \tag{37c}$$

$$P_i\left(\widetilde{\mathbf{W}}_i^r, 0\right) \leq \bar{P}_i, \tag{37d}$$

where Eq. (37b) applies to all $j \in \mathcal{N}_U$, Eq. (37c)-(37d) apply to all $i \in \mathcal{N}_R$ and the transmit power $P_i(\widetilde{\mathbf{W}}_i^r, \sigma_{w,i}^2)$ at i-th RRH is defined in Eq. (22). In problem Eq. (37), the constraint in Eq. (37b) is not only non-convex but also stochastic. Similar to Sec. IV-B3, the functions $R_j^{alt}(\mathbf{H}, \widetilde{\mathbf{W}})$ are DC functions of the covariance matrices $\widetilde{\mathbf{V}}_j = \widetilde{\mathbf{W}}_j^c \widetilde{\mathbf{W}}_j^{c\dagger}$ for all $j \in \mathcal{N}_U$, hence opening up the possibility to develop a solution based on SSUM. We refer to [26] for details on the resulting algorithm.

D. Numerical Results

In this section, we compare the performance of the conventional approach and the alternative split. To this end, we consider RRHs and UEs to be randomly located in a square area with side $\delta = 500m$ as in Fig. 2. As in Sec. III-D, in the path loss formula Eq. (4), we set the reference distance to $d_0 = 50m$ and the path loss exponent to $\eta = 3$. We assume the spatial correlation model in Eq. (20) with the angular spread $\Delta_{ji} = \arctan(r_s/d_{ji})$, with the scattering radius $r_s = 10m$ and with d_{ji} being the Euclidean distance between the i-th RRH and the j-th UE. Throughout, we consider that the every RRH is subject to the same power constraint $\bar{P}$ and has the same fronthaul capacity $\bar{C}$; that is $\bar{P}_i = \bar{P}$ and $\bar{C}_i = \bar{C}$ for $i \in \mathcal{N}_R$. Moreover, in the alternative split scheme, the UE-to-RRH assignment is carried out by choosing, for each

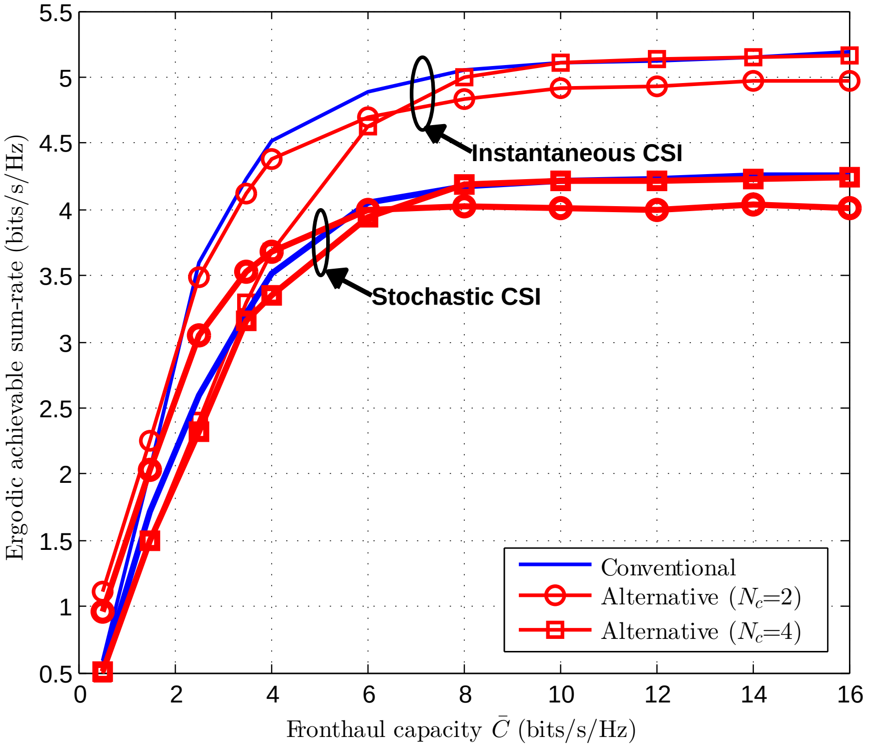

Fig. 8. Ergodic achievable sum-rate vs. the fronthaul capacity $\bar{C}$ ($N_R = N_U = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{P} = 10$ dB, $T = 20$, and $\mu = 1$).

RRH, the N_c UEs that have the largest instantaneous channel norms for instantaneous CSI and the largest average channel matrix norms for stochastic CSI. Note that this assignment is done for each coherence block in the former case, while in the latter the same assignment holds for all coherence blocks. Note also that a given UE is generally assigned to multiple RRHs.

The effect of the fronthaul capacity limitations on the ergodic achievable sum-rate is investigated in Fig. 8, where the number of RRHs and UEs is $N_R = N_U = 4$, the number of transmit antennas is $N_{t,i} = 2$ for all $i \in \mathcal{N}_R$, the number of receive antennas is $N_{r,j} = 1$ for all $j \in \mathcal{N}_U$, the power is $\bar{P} = 10dB$, and the coherence time is $T = 20$. We first observe that, with instantaneous CSI, the conventional approach strategy is uniformly better than the alternative split as long as the fronthaul capacity is sufficiently large (here $\bar{C} > 2$). This is due to the enhanced interference mitigation capabilities of the conventional approach resulting from its ability to coordinate all the RRHs via joint baseband processing without requiring the transmission of all messages on all fronthaul links. Note, in fact, that, with the alternative split, only N_c UEs are served by each RRH, and that making N_c larger entails a significant increase in the fronthaul

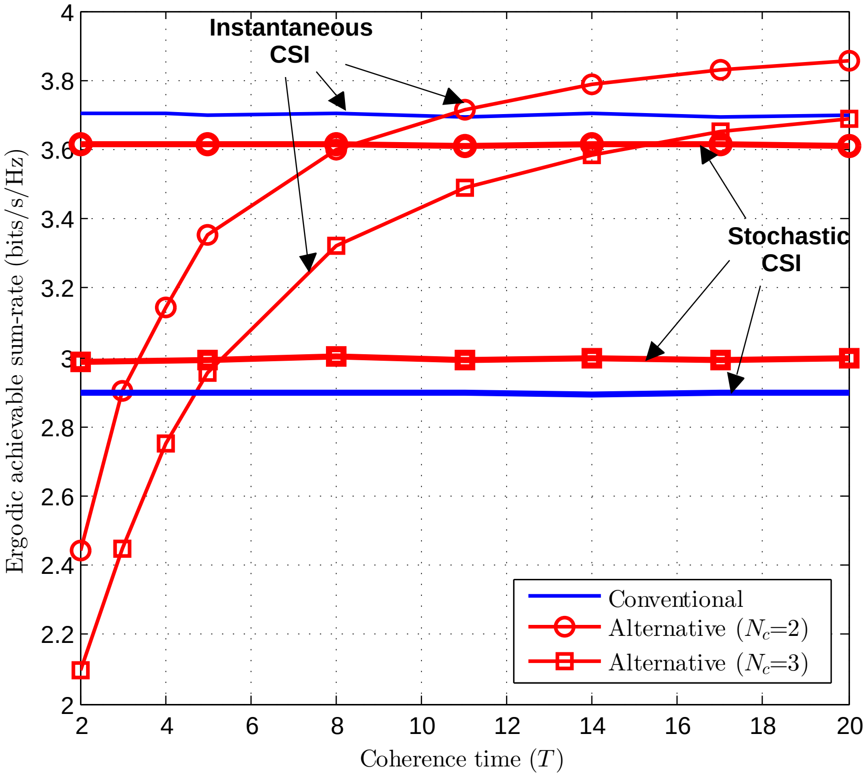

Fig. 9. Ergodic achievable sum-rate vs. the coherence time T ($N_R = N_U = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 2$ bits/s/Hz, $\bar{P} = 20dB$, and $\mu = 1$).

capacity requirements. We will later see that this advantage of the conventional approach is offset by the higher fronthaul efficiency of the alternative split in transmitting precoding information for large coherence periods T (see Fig. 9). Instead, with stochastic CSI, in the low fronthaul capacity regime, here about $\bar{C} < 6$, the alternative split strategy is generally advantageous due to the additional gain that is accrued by amortizing the precoding overhead over the entire coding block. Another observation is that, for small $\bar{C}$, the alternative split schemes with progressively smaller N_c have better performance thanks to the reduced fronthaul overhead. Moreover, for large $\bar{C}$, the performance of the alternative split scheme with $N_c = N_U$, whereby each RRH serves all UEs, approaches that of the conventional scheme.

Fig. 9 shows the ergodic achievable sum-rate as function of the coherence time T, with $N_R = N_U = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 2$ bits/s/Hz, and $\bar{P} = 20$ dB. As anticipated, with instantaneous CSI, the alternative split is seen to benefit from a larger coherence time T, since the fronthaul overhead required to transmit precoding information gets amortized over a larger period. This is in contrast to the conventional approach

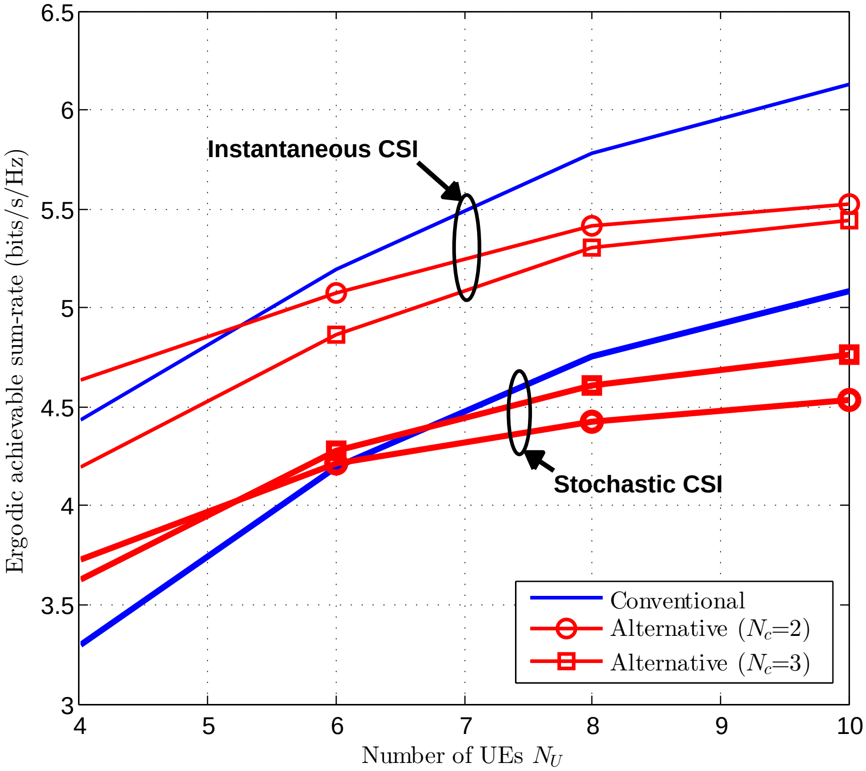

Fig. 10. Ergodic achievable sum-rate vs. the number of UEs N_U ($N_R = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 4$ bits/s/Hz, $\bar{P} = 10$ dB, $T = 10$, and $\mu = 1$).

for which such overhead scales proportionally to the coherence time T and hence the conventional scheme is not affected by the coherence time. As a result, the alternative split can outperform the conventional approach for sufficiently large T in the presence of instantaneous CSI. Instead, with stochastic CSI, the effect is even more pronounced due to the additional advantage that is accrued by amortizing the precoding overhead over the entire coding block.

Finally, in Fig. 10, the ergodic achievable sum-rate is plotted versus the number of UEs N_U for $N_R = 4$, $N_{t,i} = 2$, $N_{r,j} = 1$, $\bar{C} = 4$, $\bar{P} = 10dB$ and $T = 10$. It is observed that the enhanced interference mitigation capabilities of the conventional approach without the overhead associated to the transmission of all messages on the fronthaul links yield performance gains for denser C-RANs, i.e., for larger values of N_U. This remains true for both instantaneous and stochastic CSI cases.

V. Concluding Remarks

In this chapter, we have investigated two important aspects that pertain to the optimal functional split between RRH and BBU at the PHY layer, namely whether uplink channel estimation and downlink encoding/ precoding should be implemented at the RRH or at the BBU. The analysis, based on information-theoretical arguments, and numerical results, built on proposed efficient design algorithms, yields insight into the configurations of network architecture, channel variability and fronthaul capacities in which different functional splits are advantageous. Among the main conclusions, we have argued that the alternative functional split in which uplink channel estimation is performed at the RRH is to be preferred for low or moderate values of the coherence period and fronthaul capacity, and mostly for its capability to enable adaptive quantization based on the channel conditions. Moreover, the alternative functional split in which downlink encoding and precoding are carried out at the RRH is beneficial for lightly loaded networks in the presence of slowly changing channels, particularly under the assumption of stochastic CSI, due to its reduced fronthaul overhead.

We close this chapter with some remark on further related topics and open issues. For the uplink, an aspect that deserves further study is the integration of distributed source coding techniques (or Wyner-Ziv coding) with fronthaul processing for the joint transfer of CSI and data (see [24] for some initial discussion). Analogously, for the downlink, the impact of joint, or multivariate, compression, as proposed in [24], on the optimal functional split in the presence of different degrees of CSI at the BBU is an interesting open problem. Finally, the analysis of alternative RRH-BBU functional splits in conjunction with structured coding, or compute-and-forward, techniques calls for further attention (see [33] and references therein).

References

[1] H. Bo, V. Gopalakrishnan, L. Ji, and S. Lee, “Network function virtualization: Challenges and opportunities for innovations,” *IEEE Comm. Mag.*, vol. 53, no. 2, pp. 90–97, Feb. 2015.

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

[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] 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.

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

[6] Fujitsu, "The benefits of cloud-RAN architecture in mobile network expansion," 2015.

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

[8] B. Guo, W. Cao, A. Tao, and D. Samardzija, "CPRI compression transport for LTE and LTE-A signal in C-RAN," *Proc. Int. ICST Conf. CHINACOM*, pp. 843–839, 2012.

[9] 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.

[10] 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 (WoWMoM)*, pp. 1–9, 2013.

[11] S. Grieger, S. Boob, and G. Fettweis, "Large scale field trial results on frequency domain compression for uplink joint detection," *Proc. IEEE Glob. Comm. Conf.*, pp. 1128–1133, 2012.

[12] A. Vosoughi, M. Wu, and J. R. Cavallaro, "Baseband signal compression in wireless base stations," *Proc. IEEE Glob. Comm. Conf.*, pp. 4505–4511, 2012.

[13] 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.

[14] 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.

[15] H. S. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Info. Th.*, vol. 26, no. 5, pp. 518–521, Sep. 1980.

[16] J. Hoydis, M. Kobayashi, and M. Debbah, "Optimal channel training in uplink network MIMO systems," *IEEE Trans. Sig. Proc.*, vol. 59, no. 6, pp. 2824–2833, Jun. 2011.

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

[18] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Info. Th.*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

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

[20] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Info. Th.*, vol. 42, no. 4, pp. 1152–1159, Jul. 1996.

[21] E. Bjornson and B. E. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Trans. Sig. Proc.*, vol. 58, no. 3, pp. 1807–1820, Mar. 2010.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[23] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," *EURASIP Jour. Adv. Sig. Proc.*, 2009.

[24] 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.

[25] S. Park, C.-B. Chae, and S. Bahk, "Before/after precoded massive MIMO in cloud radio access networks," *Proc. IEEE Int. Conf. on Comm.*, Jun. 2013.

[26] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Fronthaul compression and precoding design for C-RANs over ergodic fading channel," *arXiv:1412.7713*.

[27] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, “Joint spatial division and multiplexing: The large-scale array regime,” *IEEE Trans. Info. Th.*, vol. 59, no. 10, pp. 6441–6463, Oct. 2014.

[28] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, “A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks,” *arXiv:1307.4457*.

[29] Y. Yang, G. Scutari, and D. P. Palomar, “Parallel stochastic decomposition algorithms for multi-agent systems,” *Proc. IEEE Workshop on Sign. Proc. Adv. in Wireless Comm.*, pp. 180–184, Jun. 2013.

[30] D. R. Hunter and K. Lange, “A tutorial on MM algorithms,” *The American Statistician*, vol. 58, no. 1, pp. 30–37, Feb. 2004.

[31] A. Beck and M. Teboulle, “Gradient-based algorithms with applications to signal recovery problems,” *in Convex Optimization in Signal Processing and Communications*, Y. Eldar and D. Palomar, editors, pp. 42-48, Cambridge University Press 2010.

[32] L. Vandenberghe and S. Boyd, “Semidefinite relaxation of quadratic optimization problems,” *SIAM Rev.*, vol. 38, no. 1, pp. 49–95, 1996.

[33] B. Nazer, V. Cadambe, V. Ntranos, and G. Caire, “Expanding the compute-and-forward framework: Unequal powers, signal levels, and multiple linear combinations,” *arXiv:1504.01690*.