Joint Decoding and Adaptive Compression with QoS Constraint for Uplinks in Cloud Radio Access Networks

Thang X. Vu[†], *Member, IEEE*, Tony Q.S. Quek[†], *Senior Member, IEEE*, and Hieu D. Nguyen*, *Member, IEEE*

[†]Singapore University of Technology and Design, Singapore 487372, Singapore
Email: {xuanthang_vu, tonyquek}@sutd.edu.sg.
* Institute for Inforcomm Research, A*STAR, Singapore 138632, Singapore
Email: nguyendh@i2r.a-star.edu.sg

Abstract—Cloud Radio Access Network (C-RAN) is a promising candidate for future mobile networks to sustain the exponentially increasing demand for data rate. The centralized architecture enables C-RAN to exploit multi-cell cooperation and interference management effectively. In C-RAN, one baseband unit (BBU) communicates with users through distributed Remote Radio Heads (RRHs) which are connected to the BBU via high capacity, low latency fronthaul links and perform "soft" relaying. However, the architecture of C-RAN imposes a shortage of fronthaul bandwidth because raw In-phase/Quadrature-phase (I/Q) samples are exchanged between the RRHs and the BBU. In this paper, we leverage on advanced signal processing to improve the compression efficiency in fronthaul uplinks. Specifically, we propose a joint decoding algorithm at the BBU that exploits the correlation among the RRHs and jointly performs decompressing and decoding. An upper bound of the Block Error Rate (BLER) of the proposed algorithm is derived using pair-wise error probability analysis. Based on the BLER upper bound, we propose an adaptive compression scheme which minimizes the fronthaul transmission rate while satisfying a target quality of service constrain on the BLER. Our proposed adaptive compressor originates from practical scenarios in which most applications tolerate certain non-zero BLER thresholds.

Index Terms—Cloud radio access network, joint decompression and decoding, adaptive compression, transmission error probability.

I. INTRODUCTION

Cloud Radio Access Network (C-RAN) is a novel architecture for future mobile networks which can sustain the ever increasing demand in data rate [1]. C-RAN usually consists of one centralized Baseband Unit (BBU) and a number of distributed Remote Radio Heads (RRHs) which serve users in a geographical area. The RRHs communicate with users through wireless medium and are connected to the BBU via high capacity, low latency fronthaul links. In C-RAN, RRHs act as soft-relaying nodes and do not perform baseband processing functions. With centralized processing at the BBU, C-RAN enables adaptive load balancing via virtual base station pool [2], and multi-cell processing to mitigate network-wide inter-cell interference [3]. C-RAN is expected to achieve system throughput improvement, high power efficiency, and

dynamic resource management, which in turn results in cost-saving on CAPital EXpenditure (CAPEX) and OPerating EXpenditure (OPEX) [1], [4]. However, the architecture of C-RAN requires raw In-phase/Quadrature-phase (I/Q) samples to be exchanged between the RRHs and the BBU since all baseband processing are performed the data center. Because the I/O samples represent the physical signal obtained through sampling of complex baseband signals [1], it imposes a burden of bandwidth on C-RAN fronthaul links. Therefore, reducing the transmission rate on fronthaul links is extremely important C-RAN practical implementation since the fronthaul link's capacity is limited in practice.

Various compression techniques have been proposed to reduce the fronthaul transmission rate in both time and frequency domains via, e.g., sub-carrier compression [1], [5–7]. Good compression ratio is achieved thanks to the reduction of control information redundancy in Common Public Radio Interface (CPRI) structure - introduced in [8]. In [5], [6], lossless compression is proposed which optimizes the redundancy in both time and frequency domains and in turn achieves good compression gain. Such improvement is the result from: i) only information data of active users are transmitted on fronthaul links, ii) a large amount of control information is locally generated, and iii) a reduced set of precoding matrix is transferred. A similar time-domain compression technique is presented in [7]. The above mentioned compression techniques, however, are proposed for single base station scenarios, and thus might not exploit correlation among the RRHs. Recently, joint compression techniques have been proposed to optimize quantization noise power [9–11]. The compression process is implemented via a test channel in which the quantization noise is modeled as an independent Gaussian random variable. It is shown that in general the joint design of precoding and quantization noise matrix can significantly improve the system sum rate, compared to the separate design [9]. This improvement comes from the correlation between RRHs to which distributed source coding can be applied [12]. A hybrid compression and message-sharing is proposed in [13] that allows the BSs to alternatively chose to perform a mix of compression and data-sharing on downlinks. We would emphasize that these joint compression techniques are investigated under
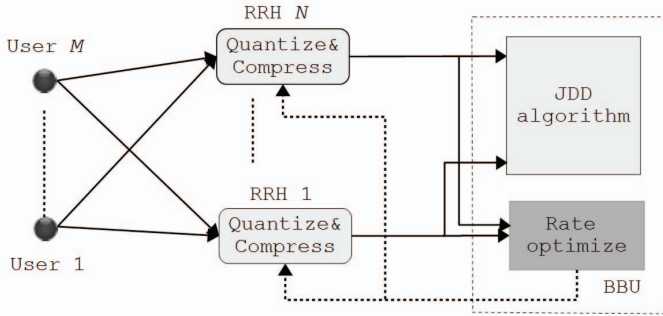
Fig. 1: Block diagram of uplinks in C-RAN. The adaptive compression scheme optimizes sampling rate needed and then feedbacks them to the RRHs.

information theoretic aspect. As a consequence, the fronthaul links' capacity is fully occupied with high probability under such schemes.

In this paper, we study the C-RAN uplink compression from practical deployment consideration. In practical systems, most applications tolerate certain Quality of Service (QoS), e.g., Block Error Rate (BLER) can be non-zero but less than a threshold, which depends on specific applications. For example, a text message application usually support higher BLER threshold than a video call. Specifically, we aim to minimize the transmission rate in fronthaul links with an acceptable distortion of the compressed signal so that the BBU can support more RRHs, which is fundamentally different from [9–11]. First, we propose a Joint Decompression and Decoding (JDD) algorithm that effectively exploits the correlation among the RRHs, which has been utilized for significant throughput improvement as shown under the theoretical aspect in [12], [14]. An upper bound of the BLER of the JDD algorithm is analysed by using Pair-wise Error Probability (PEP) analysis and a union bound. It is shown that the PEP is a function of the channel gain, thermal noise, and quantization noise. Based on the derived BLER, an adaptive compression scheme with QoS constraint is proposed to maximize the compression ratio while satisfying a predefined target BLER threshold. Numerical results show that a compression ratio of 400% can be achieved by the proposed adaptive compressor under the considered scenario.

The rest of the paper is organized as follows. Section II introduces the system model and compression scheme used throughout this paper. Section III presents the proposed JDD algorithm. The performance of JDD algorithm is analysed in Section IV. Section V presents the adaptive compression scheme. Section VI shows numerical results. Finally, conclusions and discussions are given in Section VII.

II. SYSTEM MODEL

We consider a C-RAN system consisting of one BBU, N RRHs denoted by $R_1, ..., R_N$, and $M \leq N$ users denoted by $U_1, ..., U_M$ as shown in Fig. 1. The users communicate with the RRHs via wireless medium. RRHs are connected to the BBU by high-speed, low-latency optical cables, which are

known as fronthaul links [1]. A distinguished feature of C-RAN is that RRH's function is much simpler than that of traditional Base Station (BS) because baseband processing is immigrated to the BBU. Therefore, a RRH can be seen as a "soft" relaying node that forwards I/Q signal to the BBU. In this paper, we assume that each user or RRH is equipped with a single antenna. It is worth to note that in practical system, a multiple-antenna RRH can be seen as a band of single-antenna RRHs because all baseband processing is performed at the BBU[1]. To increase the transmission efficiency of the fronthaul links, the I/Q signal should be compressed before being sent to the BBU [12]. In this paper, we focus on the joint decompression and decoding for uplinks at the BBU.

Assume that all nodes are perfectly time synchronized and denote c_m as a modulated symbol emitted by source U_m. The modulated symbol c_m, $1 \leq m \leq M$, thus belongs to the source codebook $\mathcal{S} = \{s_1, ..., s_{|\mathcal{S}|}\}$, where $|.|$ denotes the cardinality of a set. Without loss of optimality, the source codebook is normalized to satisfy a unit power constraint, e.g., $\mathbb{E}_{s \in \mathcal{S}}|s|^2 = 1$. Denote by $\mathbf{c} = [c_1, ..., c_M]^T$ a codeword transmitted by the sources, where $(.)^T$ indicates the vector/matrix transpose. The signal received at R_n is given as follows:

$$y_n = \sum_{m=1}^{M} h_{nm}\sqrt{P_m}c_m + z_n = \mathbf{h}_n\mathbf{P}\mathbf{c} + z_n, \qquad (1)$$

where $\mathbf{P} = \text{diag}([\sqrt{P_1}, ..., \sqrt{P_M}])$, P_m is the average received power at R_n from U_m, including the path loss; h_{mn} is the channel fading coefficient from U_m to R_n, $\mathbf{h}_n = [h_{n,1}, ..., h_{n,M}]$ is the channel vector from all sources to R_n; and z_n is a complex i.i.d. Gaussian noise with zero mean and variance σ^2.

Upon receiving the analogue signal from all users, the RRH quantizes the aggregated signal into digital bits. These bits are then compressed and propagated to the BBU via error-free fronthaul links.

A. Uniform compression scheme

To reduce the transmission rate on fronthaul links, the received signal at every RRH is compressed before being sent to the BBU. In this study, we consider uniform quantization [15] as the compression method because of its low-complexity and practical implementation. The compression is performed on the real and imaginary parts separately [1]. Let y_n^R and y_n^I be the real and imaginary parts of y_n, respectively. The received signal at the n-th RRH is first normalized as

$$\bar{y}_n = \frac{y_n^R}{\eta_n} + i\frac{y_n^I}{\eta_n} = \bar{y}_n^R + i\bar{y}_n^M,$$

where η_n is a scaling factor that restricts $\bar{y}_n^R$ and $\bar{y}_n^M$ within $[-1, 1]$. The value of η_n can be calculated for a given codebook $\mathcal{S}$ and the channel fading coefficients $\mathbf{h}_n$. In this paper, we use the "three-sigma" rule [16] in which η_n is equal to three times the square root of the power of y_n. For a given $\mathbf{h}_n$, it is straightforward to compute the power of y_n as $|\mathbf{h}_n\mathbf{P}|^2 + \sigma^2$. Apply the "three-sigma" rule, we have $\eta_n = 3\sqrt{|\mathbf{h}_n\mathbf{P}|^2 + \sigma^2}$.

---

[1]Those single-antenna RRHs are subjected to a sum rate constraint.

In the next step, the normalized signal $\bar{y}_n$ is quantized into $\tilde{y}_n = \tilde{y}_n^R + i\tilde{y}_n^I$ by an uniform quantizer whose resolution equals to Q_n bits. The compressed signal can be calculated from the normalized signal as follows:

$$\tilde{y}_n^a = \eta_n \frac{\text{round}(\bar{y}_n^a \times 2^{Q_n})}{2^{Q_n}},$$

where "a" represents either "R" or "I"; and the function $\text{round}(x)$ denotes the closest integer of x. The quantization error at R_n is given as $q_n = y_n - \tilde{y}_n = q_n^I + iq_n^M$. When the absolute value of y_n is large compared to quantization step, q_n^I and q_n^M can be well modeled as uniform random variables with the support $[-\delta_n, \delta_n]$, where $\delta_n = \eta_n 2^{-Q_n-1}$. After the compression, $\tilde{y}_n$ is converted into a bit sequence which is later sent to the BBU via error-free fronthaul links.

III. Joint Decompression and Decoding algorithm

In this section, we propose a JDD algorithm that performs decompression and decoding for the source codeword simultaneously, by exploiting the structure of the quantizer and the codebook. The BBU is assumed to know the Channel State Information (CSI) of all wireless links. The CSI can be obtained via, e.g., channel estimation with pilot transmission in training period. Given the compressed bit sequence, the BBU optimally estimates the source codeword using the maximum a posteriori (MAP) decoding rule as follows:

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} \Pr\{\mathbf{c}|\tilde{y}_1, \ldots, \tilde{y}_N\}$$
$$\overset{(a)}{=} \arg\max_{\mathbf{c}} \Pr\{\mathbf{c}, \tilde{y}_1, \ldots, \tilde{y}_N\}$$
$$\overset{(b)}{=} \arg\max_{\mathbf{c}} \Pr\{\mathbf{c}\} \prod_{n=1}^{N} \Pr\{\tilde{y}_n|\mathbf{c}\}, \quad (2)$$

where (a) is because $\Pr\{\tilde{y}_1, \ldots, \tilde{y}_N\}$ is constant for any codeword, and (b) is because the noise z_n's and compressed signals are independent given the source codeword.

In (2), $\Pr\{\tilde{y}_n|\mathbf{c}\}$ is the probability that the quantizer outputs $\tilde{y}_n$ from the observation $y_n = \mathbf{h}_n\mathbf{Pc} + z_n$. It is worthy mentioning that with real signal, the linear quantizer produces output y from the input x if $|y - x|$ is less than or equal to the quantization error. For the complex signal y_n, the quantizer outputs $\tilde{y}_n$ if both $|y_n^R - \tilde{y}_n^R|$ and $|y_n^I - \tilde{y}_n^I|$ are less than the quantization error. Because the quantizing step is performed independently for the real and the imaginary parts, we have

$\Pr\{\tilde{y}_n|\mathbf{c}\}$
$= \Pr\{y_n^R \in [\tilde{y}_n^R - \delta_n, \tilde{y}_n^R + \delta_n] \cap y_n^I \in [\tilde{y}_n^I - \delta_n, \tilde{y}_n^I + \delta_n]\}$
$= \Pr\{y_n^R \in [\tilde{y}_n^R - \delta_n, \tilde{y}_n^R + \delta_n]\}\Pr\{y_n^I \in [\tilde{y}_n^I - \delta_n, \tilde{y}_n^I + \delta_n]\}.$

To derive the above probability, we remind that given the codeword and the fading channels, y_n^R and y_n^I are Gaussian distributed with the same variance $\sigma^2/2$ but difference means, i.e., $\mathcal{R}(\mathbf{h}_n\mathbf{Pc})$ and $\mathcal{I}(\mathbf{h}_n\mathbf{Pc})$, respectively. Thus, we obtain:

$$\Pr\{\tilde{y}_n|\mathbf{c}\} = \frac{1}{4} \times \quad (3)$$
$$\left[\text{erfc}\left(\frac{\tilde{y}_n^R - \mathcal{R}(\mathbf{h}_n\mathbf{Pc}) - \delta_n}{\sqrt{2}\sigma}\right) - \text{erfc}\left(\frac{\tilde{y}_n^R - \mathcal{R}(\mathbf{h}_n\mathbf{Pc}) + \delta_n}{\sqrt{2}\sigma}\right)\right] \times$$

$$\left[\text{erfc}\left(\frac{\tilde{y}_n^I - \mathcal{I}(\mathbf{h}_n\mathbf{Pc}) - \delta_n}{\sqrt{2}\sigma}\right) - \text{erfc}\left(\frac{\tilde{y}_n^I - \mathcal{I}(\mathbf{h}_n\mathbf{Pc}) + \delta_n}{\sqrt{2}\sigma}\right)\right],$$

where erfc(.) denotes the complementary error function, and $\mathcal{R}(x)$ and $\mathcal{I}(x)$ are the real and imaginary parts of x. Substituting (3) into (2) we obtain a decoding rule for codeword $\hat{\mathbf{c}}$.

Remark 1: The derivation of $\Pr\{\tilde{y}_n|\mathbf{c}\}$ in (3) is exact. However, under high SNR regime and fading channel, the argument of erfc(.) function in (3) can be very large, resulting in over buffer and wrongly decoding. For a practical implementation of our scheme, approximation using first-order Taylor's series can be used instead to avoid such problems.

IV. Performance Analysis

In this paper, the BLER is defined as the probability of receiving codeword $\hat{\mathbf{c}}$ when a codeword $\mathbf{c} \neq \hat{\mathbf{c}}$ was transmitted. A block error event occurs when at least one out of M symbols c_m, $1 \leq m \leq M$, is decoded with error. Since the BLER is difficult to investigate, we instead resort to the union bound on the BLER and consider the Average Pairwise Error Probability (APEP) given as follows:

$$\text{BLER} \leq \text{APEP} = \frac{1}{|\mathcal{S}|^M} \sum_{\mathbf{c} \in \mathcal{S}^M} \sum_{\tilde{\mathbf{c}} \in \mathcal{S}^M, \tilde{\mathbf{c}} \neq \mathbf{c}} \Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\}, \quad (4)$$

where $\Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\}$ is the instantaneous PEP of receiving $\tilde{\mathbf{c}}$ when $\mathbf{c}$ was transmitted, which depends on the channel fading coefficients, and $\tilde{\mathbf{c}}$ is the only candidate.

To evaluate the PEP, we model the quantization effect by an uniformly distributed random variable that is independent from the input. This assumption is well justified when the absolute value of the input is much larger than the quantization resolution. Under such assumption, the compressed signal from the n-th RRH is modeled as follows:

$$\tilde{y}_n = \mathbf{h}_n\mathbf{Pc} + z_n + q_n, \quad (5)$$

where q_n is the quantization noise at R_n and is uniformly distributed in $[-\delta_n, \delta_n]$. It is straightforward to verify that q_n has zero mean and variance $\sigma_{q_n}^2 = \delta_n^2/3$. Denote $\mathbb{M}(\mathbf{c}) = \prod_{n=1}^{N} \Pr\{\tilde{y}_n|\mathbf{c}\}$ as the detection metric of codeword $\mathbf{c}$, where $\Pr\{\tilde{y}_n|\mathbf{c}\}$ is given in (3). A pair-wise error occurs if the metric of the transmitted codeword is smaller than that of another candidate:

$$\Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\} = \Pr\{\mathbb{M}(\mathbf{c}) < \mathbb{M}(\tilde{\mathbf{c}})\}. \quad (6)$$

The computation of (6) based on the exact expression in (3) is very complicated due to the multi-fold product of erfc(.) functions. As an alternative, we use the first order Taylor approximation $f(x) \simeq f(x_0) + f'(x_0)(x - x_0)$, with x_0 is any accessible point. Applying to the function erfc(.) in (3) with $x_0 = (\tilde{y}_n^R - \mathcal{R}(\mathbf{h}_n\mathbf{Pc}))/(\sqrt{2}\sigma)$ for the real part and $x_0 = (\tilde{y}_n^I - \mathcal{I}(\mathbf{h}_n\mathbf{Pc}))/(\sqrt{2}\sigma)$ for the imaginary part, the probability $\Pr\{\tilde{y}_n|\mathbf{c}\}$ can be written in a simplified form as follows:

$$\Pr\{\tilde{y}_n|\mathbf{c}\} \simeq \frac{\delta_n}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|\tilde{y}_n - \mathbf{h}_n\mathbf{Pc}|^2}{2\sigma^2}\right). \quad (7)$$

Substituting (7) into $\mathbb{M}(\mathbf{c})$ we obtain $\mathbb{M}(\mathbf{c}) = K \exp(-\mathbb{D}(\mathbf{c}))$, where $K = \prod_{n=1}^{N} \delta_n/(\sqrt{2\pi}\sigma)^N$ is a constant and $\mathbb{D}(\mathbf{c}) = \sum_{n=1}^{N} |\tilde{y}_n - \mathbf{h}_n \mathbf{P} \mathbf{c}|^2$. Then the PEP is derived as:

$$\Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\} = \Pr\{\underbrace{\mathbb{D}(\mathbf{c}) - \mathbb{D}(\tilde{\mathbf{c}})}_{\mathcal{I}(\mathbf{c},\tilde{\mathbf{c}})} > 0\}, \qquad (8)$$

where

$$\mathcal{I}(\mathbf{c},\tilde{\mathbf{c}}) = \sum_{n=1}^{N} \left[\tilde{y}_n^T \mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c}) + (\tilde{\mathbf{c}} - \mathbf{c})^T \mathbf{P} \mathbf{h}_n^T \tilde{y}_n \right.$$
$$\left. + |\mathbf{h}_n \mathbf{P} \mathbf{c}|^2 - |\mathbf{h}_n \mathbf{P} \tilde{\mathbf{c}}|^2 \right].$$

Substituting (5) into $\mathcal{I}(\mathbf{c},\tilde{\mathbf{c}})$ we have:

$$\mathcal{I}(\mathbf{c},\tilde{\mathbf{c}}) = \sum_{n=1}^{N} [z_n^T \mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c}) + (\tilde{\mathbf{c}} - \mathbf{c})^T \mathbf{P} \mathbf{h}_n^T z_n]$$
$$+ \sum_{n=1}^{N} [q_n^T \mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c}) + (\tilde{\mathbf{c}} - \mathbf{c})^T \mathbf{P} \mathbf{h}_n^T q_n] - \psi,$$

where $\psi = \sum_{n=1}^{N} |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2$.

Let us define $Z_1 = \sum_{n=1}^{N} [z_n^T \mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c}) + (\tilde{\mathbf{c}} - \mathbf{c})^T \mathbf{P} \mathbf{h}_n^T z_n]$ and $Z_2 = \sum_{n=1}^{N} [q_n^T \mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c}) + (\tilde{\mathbf{c}} - \mathbf{c})^T \mathbf{P} \mathbf{h}_n^T q_n]$. Because each z_n is a complex Gaussian random variable with zero mean and variance σ^2, and z_n's are mutually independent, Z_1 is also a Gaussian random variable with zero mean and variance

$$\sigma_{Z_1}^2 = 2\sigma^2 \sum_{n=1}^{N} |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2.$$

On the other hand, because q_n is uniformly distributed in $[-\delta_n, \delta_n]$, it is complicated to compute the exact joint PDF of Z_2. For ease of analysis, we model Z_2 by a Gaussian variable $\bar{Z}_2$ that has similar mean and variance as Z_2, i.e., $\bar{Z}_2 \sim \mathcal{N}(\mu_{Z_2}, \sigma_{Z_2}^2)$, where $\mu_{\bar{Z}_2} = \mathbb{E}\{Z_2\} = 0$ and

$$\sigma_{\bar{Z}_2}^2 = \mathbb{E}\{|Z_2|^2\} = \frac{4}{3} \sum_{n=1}^{N} \delta_n^2 |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2.$$

Then the sum $Z = Z_1 + Z_2$ is also a Gaussian random variable with zero mean and variance $\sigma_Z^2 = \sigma_{Z_1}^2 + \sigma_{\bar{Z}_2}^2$. Therefore we can compute the PEP as follows:

$$\Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\} = \Pr\{Z > \psi\} = \frac{1}{2} \times \qquad (9)$$

$$\mathrm{erfc}\left(\frac{\sum_{n=1}^{N} |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2}{\sqrt{4\sigma^2 \sum_{n=1}^{N} |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2 + \frac{8}{3} \sum_{n=1}^{N} \delta_n^2 |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2}} \right).$$

It is observed from (9) that the PEP depends on the relative distance between $\mathbf{c}$ and $\tilde{\mathbf{c}}$ distorted by the fading channels, thermal noise power σ^2, and the compression noise δ_n. Substituting (9) into (4) we obtain the upper bound for the BLER.

V. ADAPTIVE COMPRESSION UNDER BLER CONSTRAINT

In practical systems, various applications might require different QoS depending on specific contexts. For example,

a voice message usually requires a lower QoS compared to a video call. A flexible compression scheme should be capable to adapt the compression ratio to satisfy a predefined QoS and to maximize the compression efficiency. In this section, we propose an adaptive compression scheme to maximize the compression efficiency under a certain BLER target so that a fronthaul link can support a maximal number of antennas. More specifically, we would like to minimize the number of bits for quantization under the BLER constraint as follows:

$$\begin{array}{ll} \underset{\{Q_n\}_{n=1}^{N}}{\text{minimize}} & \sum_{n=1}^{N} Q_n \qquad (10) \\ \text{s.t.} & \frac{1}{|\mathcal{S}|^M} \sum_{\tilde{\mathbf{c}} \neq \mathbf{c} \in \mathcal{S}^M} \Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\} \leq \text{BLER}, \\ & Q_n \geq 1, \forall n, \end{array}$$

where $\Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\}$ is given in (9). The problem in (10) is difficult to solve due to its non-convexity. We instead propose an alternative approach which gives us an upper-bound of (10) as follows:

$$\begin{array}{ll} \underset{\{Q_n\}_{n=1}^{N}}{\text{minimize}} & \sum_{n=1}^{N} Q_n \qquad (11) \\ \text{s.t.} & \frac{1}{2} \mathrm{erfc}\left(\sqrt{\Phi_{\tilde{\mathbf{c}},\mathbf{c}}}\right) \leq \text{BLER}/(|\mathcal{S}|^M - 1), \forall \tilde{\mathbf{c}} \neq \mathbf{c}, \\ & Q_n \geq 1, \forall n, \end{array}$$

where

$$\Phi_{\tilde{\mathbf{c}},\mathbf{c}} = \frac{\left(\sum_{n=1}^{N} |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2\right)^2}{4\sigma^2 \sum_{n=1}^{N} |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2 + \frac{8}{3} \sum_{n=1}^{N} \delta_n^2 |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2},$$

and the first constraint in (11) is obtained by using $\Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\}$ in (9). We note that the optimal solution of (11) always satisfies (10), i.e., the optimal objective value of (11) is an upper-bound for that of (10). The proof is as follows. Let $\mathrm{Pe}(\mathbf{c})$ be error probability when $\mathbf{c}$ was transmitted and $\hat{\mathbf{c}} \neq \mathbf{c}$ is received, i.e. $\mathrm{Pe}(\mathbf{c}) = \Pr\{\hat{\mathbf{c}} \in \mathcal{S}^M \setminus \mathbf{c}|\mathbf{c}\}$, where $\mathcal{S}^M \setminus \mathbf{c}$ denotes the set of codewords except $\mathbf{c}$. Obviously, $\Pr\{\hat{\mathbf{c}} \in \mathcal{S}^M \setminus \mathbf{c}|\mathbf{c}\} \leq \sum_{\hat{\mathbf{c}} \neq \mathbf{c}} \mathrm{PEP}\{\mathbf{c} \to \hat{\mathbf{c}}\}$.

By introducing $\mu_n = 2^{-2(Q_n+1)}$, we can reformulate (11) as

$$\begin{array}{ll} \underset{\{\mu_n\}_{n=1}^{N}}{\text{miniimize}} & \sum_{n=1}^{N} -\frac{1}{2} \log_2(\mu_n) \qquad (12) \\ \text{s.t.} & (13), \forall \tilde{\mathbf{c}} \neq \mathbf{c} \\ & \mu_n \leq \frac{1}{16}, \forall n, \end{array}$$

where

$$4\sigma^2 \sum_{n=1}^{N} |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2 + \frac{8}{3} \sum_{n=1}^{N} \eta_n^2 |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2 \mu_n$$
$$\leq \frac{\left(\sum_{n=1}^{N} |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2\right)^2}{\alpha}. \qquad (13)$$

with α in (13) is defined as $\frac{1}{2} \mathrm{erfc}(\sqrt{\alpha}) = \frac{\text{BLER}}{|\mathcal{S}|^M - 1}$. It can be proved that (12) is a convex optimization problem and thus

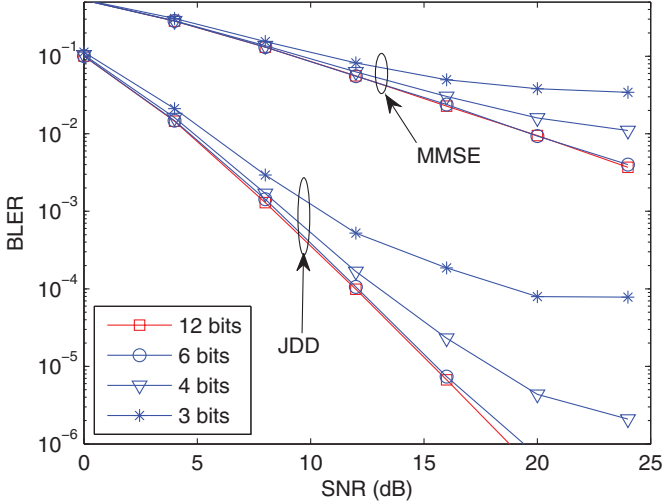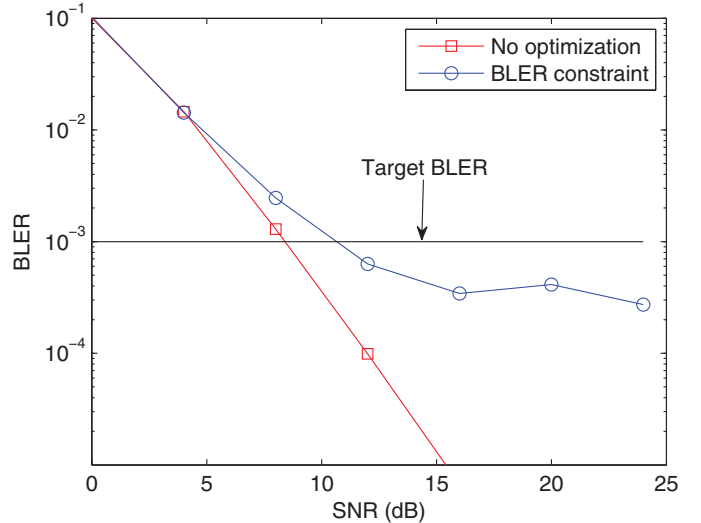Fig. 2: BLER of the JDD algorithm with different fronthaul sampling resolution $Q = 3, 4, 6, 12$ bits.

(a) BLER performance.

(b) Compression ratio.

Fig. 3: Performance comparison between the adaptive compression scheme with a BLER target equals 10^{-3} and the fixed compression scheme. Fronthaul bandwidth $Q = 12$ bits. The fixed compression scheme fully occupies the fronthaul bandwidth.

can be solved efficiently by using, e.g., the primal-dual interior point method [17]. Furthermore, (11) is substantially simpler than (10) and is more preferable for its low-complexity.
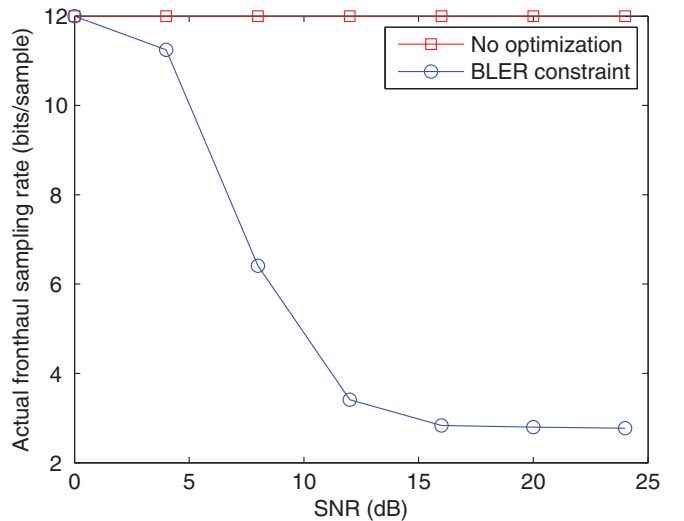
VI. SIMULATION RESULTS

The proposed algorithm is evaluated for a C-RAN system consisting of $M = 3$ users and $N = 3$ RRHs under block Rayleigh fading channel, i.e., h_{mn}'s are independent identically distributed (i.i.d.) random variables, each distributed as $\mathcal{CN}(0, 1)$. Symmetric network is assumed with equal source's transmit power, e.g., $P_1 = ... = P_M = P$. We consider a QPSK modulation with the codebook $\mathcal{S} = \{-1 - 1i, -1 + 1i, 1 - 1i, 1 + 1i\}/\sqrt{2}$. The average SNR is defined as P/σ^2. The BBU is assumed to know CSI of all the wireless channels.

Fig. 2 presents the BLER performance of the proposed JDD receiver with respect to the SNR for different fronthaul sampling rates. As for reference, the BLER of Minimum Mean Square Error (MMSE) receiver [18] is also illustrated in Fig. 2. We note that the MMSE performs decompression and decoding separately. It first implements the decompression by adding zeros to the removed bits and then applies standard MMSE decoder to decode the source data. It is observed that the proposed algorithm achieves a huge gain over the MMSE-based scheme. This gain results from the searching over all combinations of the source symbols of the proposed algorithm. The effect of the quantization resolution Q (fronthaul sampling rate in bits per sample (bps)) has the same trend on both receivers: in general larger Q leads to better BLER and the BLER will saturate as Q decreases. Under low SNR regime, the contributions of quantization resolution is small. For example, when the SNR is between 0dB and 12dB, sampling rate at 4 bps and at 10 bps yield almost similar BLER. This is due to the fact that for a small SNR value, the thermal noise is large, and therefore, is dominant compared with quantization noise. In contrast, with higher SNR, thermal noise is comparable to or even smaller than the quantization

noise. Reducing the sampling rate Q in this case can result in severe loss in BLER. One interesting observation is that a 6-bit quantizer achieves almost similar BLER as a 12-bit quantizer in the observing SNR range. This result suggests an adaptive compression scheme to minimize fronthaul rate.

Fig. 3 presents the performance of the proposed adaptive compression versus SNR for the target BLER = 10^{-3}. The adaptive compressor will minimize the actual fronthaul sampling rate while attempting to satisfy the BLER target. The fronthaul bandwidth constraint is 12 bits. For reference, we present a fixed compressor which fully occupies 12 bits and is denoted as *No optimization* in the figure. Simulation is conducted under 1000 channel realizations. Fig. 3a shows
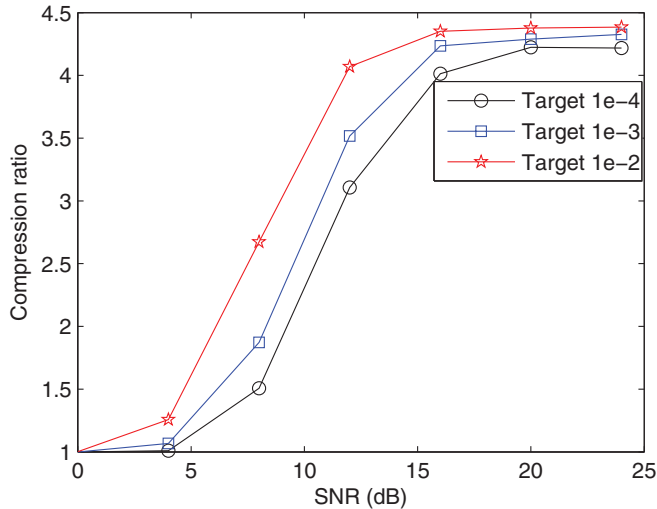
Fig. 4: Compression ratio performance of the proposed adaptive compressor for different target BLERs. The fronthaul bandwidth equals 12 bits.

the BLER performance of the proposed adaptive compressor with bit minimization and the fixed compression scheme. The actual fronthaul rate is presented in Fig. 3b. In low SNR region, both the adaptive compressor and the fixed compressor do not satisfy the target BLER because the channel is too poor. However, the adaptive scheme achieves better effective fronthaul rate as it uses less than 12 bits. More specifically, at 8dB the adaptive scheme saves 5.5 bits. When SNR increases, the adaptive compressor meets the target BLER while significantly improving the compression efficiency. Because the non-optimization scheme always uses 12 bits for quantization, its actual fronthaul rate is always 12 bits per sample. On the other hand, a compression ratio of 400% is achieved by the proposed adaptive compressor, which only requires approximately 3 bits per sample to achieve a BLER less than or equal 10^{-3}.

Fig. 4 presents the compression ratio of the proposed adaptive compressor for different target BLERs. The compression ratio is defined as the ratio of the fronthaul bandwidth divided by the actual fronthaul rate after optimization. Numerical results show that a smaller target BLER achieves a lower compression ratio. This observation is logical since a compressor with a better BLER target requires more sampling bits. In particular, the compressor with target $\text{BLER} = 10^{-4}$ gains a compression ratio equal to 3.1 while the compression with target $\text{BLER} = 10^{-2}$ achieves a compression ratio of 4.1. When SNR increases, the compression efficiency saturates, which suggests that increasing transmitting power under such regime does not improve the compression ratio.

VII. CONCLUSIONS AND DISCUSSIONS

We have proposed a joint decompression and decoding receiver for uplinks in cloud radio access networks. The proposed receiver takes into consideration the quantization effect of capacity-limited fronthaul links, and exploits the correlation

among the Remote Radio Heads. An upper bound of the BLER has been derived in closed-form expression by using pair-wise error probability analysis. Based on the analysed result, an adaptive compression schemes has been proposed to increase the fronthaul compression ratio while satisfying the predetermined BLER constraint. Numerical results show that compression ratio of 400% can be achieved by the proposed optimization scheme, i.e., the proposed adaptive compression scheme can support four times more number of RRHs than the fixed compression scheme. A possibly higher compression ratio is also achievable by combining the proposed scheme with control-signal compression techniques [5–7]. Such extensions will be considered in subsequent studies.

REFERENCES

[1] ChinaMobile, "C-RAN the road towards green RAN," 2011, white paper.
[2] Z. Zhu, P. Gupta, Q. Wang, S. Kalyanaraman, Y. Lin, H. Franke, and S. Sarangi, "Virtual base station pool: Towards a wireless network cloud for radio access networks," in *Proceedings of the 8th ACM International Conference on Computing Frontiers*, New York, NY, USA, 2011, pp. 34:1–34:10.
[3] P. Marsch and G. Fettweis, "Uplink CoMP under a constrained backhaul and imperfect channel knowledge," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1730–1742, Oct. 2011.
[4] UMTS, "Mobile traffic forecasts 2010-2020," Tech. Rep. 44, 2011.
[5] J. Lorca and L. Cucala, "Lossless compression technique for the fronthaul of LTE/LTE-advanced cloud-ran architectures," in *IEEE 14th International Symposium and Workshops on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'2013)*, 2013, pp. 1–9.
[6] A. Nanba, S. nd Agata, "A new IQ data compression scheme for fronthaul link in centralized RAN," in *IEEE 24th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'2013 Workshops)*, Sep. 2013, pp. 210–214.
[7] K. F. Nieman and B. L. Evans, "Time-domain compression of complex-baseband LTE signals for cloud radio access networks," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, 2013, pp. 1198–1201.
[8] –, "Common public radio intreface (CPRI); interface specification," 2013, CPRI Specification V6.0.
[9] 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.
[10] 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. Commun.*, vol. 13, no. 3, pp. 1555–1567, Mar. 2014.
[11] Y. Wang, H. Wang, and L. Scharf, "Optimum compression of a noisy measurement for transmission over a noisy channel," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1279–1289, Mar. 2014.
[12] A. Sanderovich, O. Somekh, H. V. Poor, and S. Shamai, "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3457–3478, Aug. 2009.
[13] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," in *IEEE Information Theory and Application workshop*, 2014.
[14] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint base station selection and distributed compression for cloud radio access networks," in *IEEE Globecom Workshops (GC'2012 Workshops)*, pp. 1134–1138, 2012.
[15] "Feasibility study for further advancements for E-UTRA (LTE-advanced)," Tech. Rep. ETSI TR 136 912, 2010.
[16] R. Gray and D. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.
[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.
[18] P. Li, D. Paul, R. Narasimhan, and J. Cioffi, "On the distribution of sinr for the mmse mimo receiver and performance analysis," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 271–286, Jan. 2006.