Adaptive Compression and Joint Detection for Fronthaul Uplinks in Cloud Radio Access Networks

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

***Abstract*—Cloud radio access network (C-RAN) has recently attracted much attention as a promising architecture for future mobile networks to sustain the exponential growth of data rate. In C-RAN, one data processing center or baseband unit (BBU) communicates with users via distributed remote radio heads (RRHs), which are connected to the BBU via high capacity, low latency fronthaul links. In this paper, we study the compression on fronthaul uplinks and propose a joint decompression algorithm at the BBU. The central premise behind the proposed algorithm is to exploit the correlation between RRHs. Our contribution is threefold. First, we propose a joint decompression and detection (JDD) algorithm which jointly performs decompressing and detecting. The JDD algorithm takes into consideration both the fading and compression effect in a single decoding step. Second, block error rate (BLER) of the proposed algorithm is analyzed in closed-form by using pair-wise error probability analysis. Third, based on the analyzed BLER, we propose adaptive compression schemes subject to quality of service (QoS) constraints to minimize the fronthaul transmission rate while satisfying the pre-defined target QoS. As a dual problem, we also propose a scheme to minimize the signal distortion subject to fronthaul rate constraint. Numerical results demonstrate that the proposed adaptive compression schemes can achieve a compression ratio of 300% in experimental setups.**

***Index Terms*—Cloud radio access network, fronthaul link, joint decompression and detection, optimization.**

I. Introduction

CLOUD radio access network (C-RAN) has been widely accepted as a new architecture for future mobile networks to sustain the ever increasing demand in data rate [1]. In C-RAN, one centralized processor or Baseband Unit (BBU) communicates with users distributed in a graphical area via a number of remote radio heads (RRHs), which act as “soft” relaying nodes and are connected to the BBU via high capacity and low latency fronthaul links. By moving all baseband processing functions from RRHs to a centralized processor, C-RAN enables adaptive load balancing via virtual base station pool [2] and effective network-wide inter-cell interference management thanks to multi-cell processing [3], [4]. The promise of C-RAN over traditional mobile networks includes system throughput improvement, high power efficiency, and dynamic resource management, which eventually result in the cost-saving on CAPital EXpenditure (CAPEX) and OPerating EXpenditure (OPEX) [1], [5]. Because the baseband processing functions are executed at the BBU, the In-phase/Quadrature-phase (I/Q) samples which represent the physical signal obtained through the sampling of the complex baseband signals are exchanged between the RRHs and the BBU, resulting in enormous transmission rate on the fronthaul links. Reducing this rate is extremely important in the implementation of C-RAN since the fronthaul links’ capacity is limited in practice.

Numerous research efforts have recently investigated the compression of C-RAN, mostly from the information-theoretic perspective, which design and optimize the quantization noise to maximize the achievable sum rate [6]–[12]. This problem can be seen as a network multiple-input multiple-output (MIMO) problem with limited backhaul capacity [13]–[15]. The compression process is implemented via a test channel and the quantization noise is modelled as an independent Gaussian random variable, whose variance is linked to the capacity of the test channel. It is shown, in general, that the joint design of the precoding and quantization noise matrix can significantly improve the system sum rate over separate design [7]. Such improvement results from the correlation among the RRHs when distributed source coding is applied [3]. The quality of the received signal at one RRH can be enhanced by exploiting the signal at other RRHs as side information. In [7], a robust distributed compression for uplink baseband signal is proposed based on the Karhunen-Loeve transform. In that work, the correlated data at base stations are assumed to be imperfect and modelled as deterministic additive errors on the bound of eigenvalue of the error matrix. Further performance gain can be achieved by optimizing the test channel [16]. The authors in [8] proposed a hybrid compression and message-sharing strategy that allows a BS to perform a mix of compression and data-sharing on the downlinks. It is shown that the hybrid solution achieves a better rate region than the pure method of compression or data-sharing. In [10], an optimum compression method is derived for sensor networks to compress noisy sensor measurements by minimizing the trace or determinant of the error covariance matrix. Further review on C-RAN is presented in [17] and [18].

From the practical system point of view, various compression techniques have been studied in both time- and frequency-domains (sub-carrier compression) [1]. The key idea in those

Manuscript received January 21, 2015; revised May 29, 2015 and July 21, 2015; accepted August 25, 2015. Date of publication September 1, 2015; date of current version November 13, 2015. Parts of this work were submitted to IEEE GLOBECOM’15. This research was supported in part by the A$*$STAR SERC under Grant 1224104048 and the MOE ARF Tier 2 grant MOE2014-T2-2-002. The associate editor coordinating the review of this paper and approving it for publication was M. Tao.

T. X. Vu and T. Q. S. Quek are with Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: xuanthang_vu@sutd.edu.sg; tonyquek@sutd.edu.sg).

H. D. Nguyen is with the Institute for Inforcomm Research, A-STAR, Singapore 138632 (e-mail: nguyendh@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2015.2475430

techniques is to minimize redundancy of control information in common public radio interface (CPRI) package structure [19]. Lossless compression is proposed to achieve a good compression ratio due to two added nodes at the ends of the fronthaul links that optimize redundancy in both time and frequency domains [20], [21]. Statistical multiplexing gain is achieved from: i) only information data of active users are transmitted via the fronthaul links, ii) a minimum information needed for the reconstruction of the control information since a large amount of control information, which is completely or semi static, is locally generated, and iii) a reduced set of the precoding matrix is transferred. A similar time-domain compression technique is proposed by [22]. Note that most of the compression techniques proposed in time domain are for single base station, which cannot exploit residential gain from the correlation among the RRHs.

In this paper, we study the compression on C-RAN uplinks and propose a near-optimal receiver at the BBU. In practical systems, the received signal at a RRH is first uniformly quantized into bits sequence by a analogue-to-digital converter (ADC).[1] These bits are then transmitted to the BBU via ideal fronthaul links. Compression ratio can be managed by changing the resolution of the ADC. The proposed compression method is not limited to time domain and can be applied to frequency domain with little modification. From the observation that treating the decompression and demodulation separately leads to a very suboptimal solution [9], we propose a joint decompression and demodulation (JDD) algorithm that jointly performs decompressing and detecting in a single step and effectively exploits the correlation among the RRHs, which achieves significant improvement in the information-theoretic sense [3], [16]. Our first goal is to minimize the transmission rate on the fronthaul links with an acceptable distortion of the decompressed signal so that the BBU can support a maximum number of RRHs. This design criterion is different from that in [3] and [16], which aims to fully occupy the fronhaul link capacity. Our objective comes from practical situations where most applications can tolerate an acceptable non-zero block error rate (BLER). A second goal is to minimize the signal distortion given a fronthaul rate constraint. Analytical closed-form expression for the BLER is derived using pair-wise error probability (PEP) analysis, which is shown as a function of the channel fadings, thermal noise, and quantization noise. Based on the analyzed BLER expression, two adaptive compression schemes with a quality of service (QoS) constraint are proposed to maximize the compression ratio while satisfying a given BLER target. We also present a JDD scheme which aims to minimize the distortion given a predetermined fronthaul rate.

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

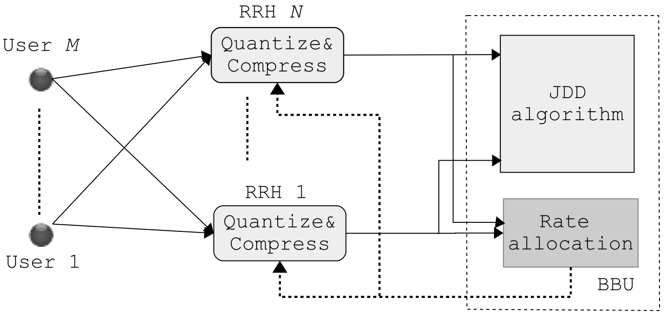


Fig. 1. Block diagram of uplinks in C-RAN with joint decompression and detection algorithm at the BBU. The adaptive compression scheme employs rate allocation block to feedback optimal sampling rate to the RRHs.

II. System Model

We consider a C-RAN system consisting of M users denoted by $\mathrm{U}_1, \ldots, \mathrm{U}_M$, N RRHs denoted by $\mathrm{R}_1, \ldots, \mathrm{R}_N$, and one BBU, as shown in Fig. 1. The users communicate with the RRHs via wireless medium, while the RRHs connect to the BBU by high-speed, low-latency optical fibre (or wireless) links, which are known as fronthaul links [1]. A distinguished feature of the RRHs compared with classical base station (BS) is that the RRH's function is much simpler than that of traditional BS because all baseband processing functions are immigrated to the BBU. Therefore, a RRH can be seen as a "soft" relaying node that forwards I/Q signal to the BBU. The users and RRHs are equipped with a single antenna. In practical system, a multiple-antenna RRH can be seen as a band of single-antenna RRHs[2] because all baseband processing functions are performed at the BBU. Due to limited capacity on the fronthaul links, I/Q signal needs to be compressed before being sent to the processing center [3]. The BBU decompresses the received signal from the RRHs and then performs further processing. In the following, we focus on the compression and decompression on fronthaul uplinks.

We assume that all nodes are synchronous and all wireless channels are block Rayleigh fading. The BBU is assumed to know all the channel state information (CSI) in the network. Denote by c_m a modulated symbol emitted by user U_m. The modulated symbol c_m, $1 \le m \le M$, belongs to the source codebook $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$, where $|.|$ denotes the cardinality of a set. The source codebook satisfies unit power constraint, e.g., $\mathbb{E}_{s \in \mathcal{S}} |s|^2 = 1$. Denote $\mathbf{c} = [c_1, \ldots, c_M]^T$ as a codeword transmitted by the users, where $(.)^T$ represents the vector/matrix transpose. The received signal at R_n is given by

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

where $\mathbf{P} = \mathrm{diag}([\sqrt{P_1}, \ldots, \sqrt{P_M}])$, P_m is the average transmit power of user U_m, h_{nm} is the channel fading coefficient between U_m and R_n, including the path loss, which is a complex Gaussian random variable with zero mean and variance $\sigma^2_{h_{nm}}$, $\mathbf{h}_n = [h_{n1}, \ldots, h_{nM}]$ is the channel vector from all users to R_n, and z_n is independent and identically distributed (i.i.d.) Gaussian noise with zero mean and variance σ^2.

[1] Other non-linear quantization methods can also be applied.

[2] These RRHs are subject to a sum rate constraint.

Upon receiving analogue signals from the users, the RRHs quantize and compress them into digital bits and then forward these bits to the BBU.

A. Uniform Compression Scheme

To reduce the transmission rate on fronthaul links, the received signal at each RRH is compressed before being sent to the BBU. In this study, we consider uniform quantization as the compression method because of its low-complexity and practical implementation [23]. This compression method can be realized by flexibly tuning the ADC's resolution. Therefore, a target compression ratio can be achieved by changing the resolution of the ADC. In the case where the resolution of the ADC is fixed due to some hardware constraints, this compression method can be performed by truncating some least important bits in the ADC's output. The compression is executed 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

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

where η_n is a scaling factor that restricts $\bar{y}_n^R$ and $\bar{y}_n^I$ within $[-1, 1]$ with high probability. The value of η_n can be calculated for a given codebook $\mathcal{S}$ and the channel fading coefficients $\mathbf{h}_n$. In this work, we use the "three-sigma" rule [25] 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, the BBU computes $\eta_n = 3\sqrt{\|\mathbf{h}_n\mathbf{P}\|^2 + \sigma^2}$, which is assumed to be known at the n-th RRH because its overhead is negligible compared with data.

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:

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

where "a" represents either "R" or "I"; and the function round(x) denotes the closest integer of x. The quantization error at R$_n$ is given as $q_n = y_n - \tilde{y}_n = q_n^R + iq_n^I$. When the absolute value of y_n is large compared to quantization step, q_n^R and q_n^I can be well modelled as uniform random variables with the support $[-\delta_n, \delta_n]$, where $\delta_n = \eta_n 2^{-Q_n-1}$. We observe via intensive simulations that with the three-sigma rule, such assumption is still feasible even with a small number of quantization bits. After 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 Demodulation Algorithm

In this section, we propose a JDD algorithm that performs decompressing and detecting for the source codeword simultaneously by exploiting the structure of the quantizer and the codebook. The BBU is assumed to know the 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 sequences, the BBU optimally estimates the source codeword by using the maximum a posteriori (MAP) receiver as follows:

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

where (a) $\Pr\{\tilde{y}_1, \ldots, \tilde{y}_N\}$ is constant for any codeword, and (b) 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 $y_n = \mathbf{h}_n\mathbf{P}\mathbf{c} + z_n$. It is worth to mention that for real signal, the linear quantizer outputs y if the distance between the input and y 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 quantization is performed independently for the real and the imaginary parts, we have

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

To derive the above probability, we remind that for the given the codeword and the fading channels, y_n^R and y_n^I are Gaussian distributed with the same variance $\sigma^2/2$ and mean $\mathcal{R}(\mathbf{h}_n\mathbf{P}\mathbf{c})$ and $\mathcal{I}(\mathbf{h}_n\mathbf{P}\mathbf{c})$, respectively, where $\mathcal{R}(x)$ and $\mathcal{I}(x)$ are the real and imaginary parts of x. Therefore, the conditional probability density function (PDF) of y_n^R and y_n^I are, respectively, given by

$$f\left(y_n^M|\mathbf{c}\right) = \frac{1}{\sqrt{\pi}\sigma}\exp\left(-\frac{\left|y_n^R - \mathcal{R}(\mathbf{h}_n\mathbf{P}\mathbf{c})\right|^2}{\sigma^2}\right),$$

$$f\left(y_n^I|\mathbf{c}\right) = \frac{1}{\sqrt{\pi}\sigma}\exp\left(-\frac{\left|y_n^I - \mathcal{I}(\mathbf{h}_n\mathbf{P}\mathbf{c})\right|^2}{\sigma^2}\right).$$

By substituting the above PDFs into $\Pr\{\tilde{y}_n|\mathbf{c}\}$ we obtain

$$\begin{aligned}\Pr\{\tilde{y}_n|\mathbf{c}\} &= \int_{\tilde{y}_n^R-\delta_n}^{\tilde{y}_n^R+\delta_n} f\left(y_n^R|\mathbf{c}\right) dy_n^R \times \int_{\tilde{y}_n^I-\delta_n}^{\tilde{y}_n^I+\delta_n} f(y_n^I|\mathbf{c}) dy_n^I\\ &= \frac{1}{4}\\ &\times\left[\text{erfc}\left(\frac{\tilde{y}_n^R - \mathcal{R}(\mathbf{h}_n\mathbf{P}\mathbf{c}) - \delta_n}{\sigma}\right) - \text{erfc}\left(\frac{\tilde{y}_n^R - \mathcal{R}(\mathbf{h}_n\mathbf{P}\mathbf{c}) + \delta_n}{\sigma}\right)\right]\\ &\times\left[\text{erfc}\left(\frac{\tilde{y}_n^I - \mathcal{I}(\mathbf{h}_n\mathbf{P}\mathbf{c}) - \delta_n}{\sigma}\right) - \text{erfc}\left(\frac{\tilde{y}_n^I - \mathcal{I}(\mathbf{h}_n\mathbf{P}\mathbf{c}) + \delta_n}{\sigma}\right)\right],\end{aligned} \tag{3}$$

where erfc(.) denotes the complementary error function. Substituting (3) into (2), we then obtain a decoding rule for codeword $\hat{\mathbf{c}}$.

IV. Performance Analysis

This section analyzes the BLER of the proposed JDD algorithm. 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 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) as follows:

$$\text{BLER} \leq \text{APEP} = \frac{1}{|\mathcal{S}|^M} \sum_{\mathbf{c} \in \mathcal{S}^M} \sum_{\mathbf{c} \neq \tilde{\mathbf{c}} \in \mathcal{S}^M} \Pr\{\mathbf{c} \rightarrow \tilde{\mathbf{c}}\}. \tag{4}$$

where $\Pr\{\mathbf{c} \rightarrow \tilde{\mathbf{c}}\}$ 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 can be well justified when the absolute value of the input is much larger than the quantization step. Under such assumption, the compressed signal from the n-th RRH is modeled as

$$\tilde{y}_n = \mathbf{h}_n \mathbf{P} \mathbf{c} + z_n + q_n, \tag{5}$$

where $q_n = q_n^R + i q_n^I$ being quantization noise at R_n. Since both q_n^R and q_n^I are uniformly distributed in $[-\delta_n, \delta_n]$ (see Section II for more details), it is straightforward to verify that q_n has zero mean and variance which is computed as

$$\begin{aligned}\sigma_{q_n}^2 &= \operatorname{Var}(q_n) = \operatorname{Var}\left(q_n^R\right) + \operatorname{Var}\left(q_n^I\right) \\ &= \frac{1}{2\delta_n} \int_{-\delta_n}^{\delta_n} \left|q_n^R\right|^2 dq_n^R + \frac{1}{2\delta_n} \int_{-\delta_n}^{\delta_n} \left|q_n^I\right|^2 dq_n^I = \frac{2\delta_n^2}{3}.\end{aligned}$$

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 computed in (3). A pair-wise error occurs if the metric of the transmitted codeword is smaller than that of another candidate:

$$\Pr\{\mathbf{c} \rightarrow \tilde{\mathbf{c}}\} = \Pr\left\{\mathbb{M}(\mathbf{c}) < \mathbb{M}(\tilde{\mathbf{c}})\right\}. \tag{6}$$

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 feasible point. Applying to the function erfc(.) in (3) with $x_0 = (\tilde{y}_n^R - \mathcal{R}(\mathbf{h}_n \mathbf{P} \mathbf{c}))/(\sigma)$ for the real part and $x_0 = (\tilde{y}_n^I - \mathcal{I}(\mathbf{h}_n \mathbf{P} \mathbf{c}))/(\sigma)$ for the imaginary part, $\Pr\{\tilde{y}_n | \mathbf{c}\}$ can be written as

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

Remark 1: The derivation of $\Pr\{\tilde{y}_n | \mathbf{c}\}$ in (3) is exact and (3) can be used as the (exact) decoding metric. 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, an approximation using first-order Taylor's series (7) can be used instead to avoid such problems.

Substituting (7) into $\mathbb{M}(\mathbf{c})$, we obtain $\mathbb{M}(\mathbf{c}) = K \exp(-\mathbb{D}(\mathbf{c}))$, where $K = \prod_{n=1}^N \delta_n^2 / (\pi \sigma^2)^N$ 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:

$$\Pr\{\mathbf{c} \rightarrow \tilde{\mathbf{c}}\} = \Pr\left\{\underbrace{\mathbb{D}(\mathbf{c}) - \mathbb{D}(\tilde{\mathbf{c}})}_{\mathcal{I}(\mathbf{c}, \tilde{\mathbf{c}})} > 0\right\}, \tag{8}$$

where

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

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

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

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

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

On the other hand, because q_n is uniformly distributed, 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., $\bar{Z}_2 \sim \mathcal{N}(\mu_{Z_2}, \sigma_{\bar{Z}_2}^2)$, where $\mu_{\bar{Z}_2} = \mathbb{E}\{Z_2\} = 0$ and $\sigma_{\bar{Z}_2}^2 = \mathbb{E}\{|Z_2|^2\} = 2 \sum_{n=1}^N \sigma_{q_n}^2 |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2 = \frac{4}{3} \sum_{n=1}^N \delta_n^2 |\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})|^2$. 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:

$$\begin{aligned}\Pr\{\mathbf{c} \rightarrow \tilde{\mathbf{c}}\} &= \Pr\{Z > \psi\} = \frac{1}{2} \\ &\times \operatorname{erfc}\left(\frac{\sum_{n=1}^N \left|\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})\right|^2}{\sqrt{4\sigma^2 \sum_{n=1}^N \left|\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})\right|^2 + \frac{8}{3} \sum_{n=1}^N \delta_n^2 \left|\mathbf{h}_n \mathbf{P}(\tilde{\mathbf{c}} - \mathbf{c})\right|^2}}\right).\end{aligned} \tag{9}$$

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 QoS Constraint

In practical systems, various applications might require different QoSs 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 maximize the compression efficiency. In this section, we first propose two adaptive compression schemes to maximize the compression efficiency under a certain target BLER so that a front-haul link can support a maximal number of antennas. Such schemes are desirable for systems which support large front-haul feedback and/or require stringent BLER QoS. Furthermore, we also consider an adaptive compression design which minimizes the BLER, specifically the PEP as a proxy of BLER, given a compression efficiency. Compared to the previous two counterparts, this design focuses on systems with stricter constraint on the front-haul bandwidth.

A. Minimization of the Number of Bits Given the BLER

In this subsection, we consider systems which require a certain BLER QoS while tolerating a possible large front-haul bandwidth. In particular, we would like to minimize the number of bits for quantization under the QoS constraint as follows:

$$\begin{aligned} \underset{\{Q_n \geq 1\}_{n=1}^N}{\texttt{minimize}} \quad & \sum_{n=1}^{N} Q_n \qquad (10) \\ \texttt{s.t.} \quad & \frac{1}{|\mathcal{S}|^M} \sum_{\tilde{\mathbf{c}} \neq \mathbf{c} \in \mathcal{S}^M} \sum \Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\} \leq BLER_0, \end{aligned}$$

where $BLER_0$ is the predefined BLER target, and $\Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\}$ is given in (9).

1) PEP-Based Algorithm: 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:

$$\begin{aligned} \underset{\{Q_n \geq 1\}_{n=1}^N}{\texttt{minimize}} \quad & \sum_{n=1}^{N} Q_n \qquad (11) \\ \texttt{s.t.} \quad & \frac{1}{2}\text{erfc}\left(\sqrt{\Phi_{\tilde{\mathbf{c}},\mathbf{c}}}\right) \leq \frac{BLER_0}{|\mathcal{S}|^M - 1}, \forall\, \tilde{\mathbf{c}} \neq \mathbf{c}, \end{aligned}$$

where

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

and the 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 $\text{Pe}(\mathbf{c})$ be error probability when $\mathbf{c}$ was transmitted and $\hat{\mathbf{c}} \neq \mathbf{c}$ is received, i.e., $\text{Pe}(\mathbf{c}) = \Pr\{\hat{\mathbf{c}} \in \mathcal{S}^M \setminus \mathbf{c} | \mathbf{c}\}$, where $\mathcal{S}^M \setminus \mathbf{c}$ denotes the set of codewords except $\mathbf{c}$. Obviously, $\Pr\{\hat{\mathbf{c}} \in \mathcal{S}^M \setminus \mathbf{c}|\mathbf{c}\} \leq \sum_{\hat{\mathbf{c}} \neq \mathbf{c}} \text{PEP}\{\mathbf{c} \to \hat{\mathbf{c}}\}$. This confirms that the optimal objective value of (11) is an upper-bound for that of (10). Note that (11) is an integer programming problem, which is difficult to solve. We therefore resort to a convex formulation of (11) by relaxing the integer constraint of Q_n.

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

$$\begin{aligned} \underset{\left\{\mu_n \leq \frac{1}{4}\right\}_{n=1}^N}{\texttt{minimize}} \quad & \sum_{n=1}^{N} -\frac{1}{2}\log_2(\mu_n) \qquad (12) \\ \texttt{s.t.} \quad & (13), \forall\, \tilde{\mathbf{c}} \neq \mathbf{c}, \end{aligned}$$

where

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

In (13), α is an auxiliary variable satisfying $\frac{1}{2}\text{erfc}(\sqrt{\alpha}) = \frac{BLER_0}{|\mathcal{S}|^M - 1}$. Note that we can consider α as the maximum PEP that the $BLER_0$ constraint can still be satisfied. The problem (10) with a BLER constraint has been effectively transformed to a PEP-based counterpart [24]. We denote the scheme that solved (12) as MinBits-PEP.

It can be proved that (12) is a convex optimization problem and thus can be solved efficiently by using, e.g., the primal-dual interior point method [26]. Furthermore, (11) is substantially simpler than (10) and is more preferable under systems requiring low complexity. The integer quantization bit Q_n can be obtained from μ_n simply by choosing the smallest following integer of $\hat{Q}_n = 1 - \frac{1}{2}\log_2 \mu_n$, i.e., $\lfloor \hat{Q}_n \rfloor$. In general, there is no bound for the optimality loss of such approximation. However, as the constraint threshold BLER tends to 0, the loss also tends to 0. The reasoning is that each Q_n becomes large in such case, which leads to a small $\frac{\hat{Q}_n - \lfloor \hat{Q}_n \rfloor}{\hat{Q}_n}$.

2) SDR-Based Algorithm: In this subsection, we propose an approximated solution for problem (10) using semidefinite programming relaxation (SDR). The problem (10) is first rewritten as

$$\begin{aligned} \underset{\{Q_n \geq 1\}_{n=1}^N}{\texttt{minimize}} \quad & \sum_{n=1}^{N} Q_n \qquad (14) \\ \texttt{s.t.} \quad & \frac{1}{|\mathcal{S}|^M} \sum_{\tilde{\mathbf{c}} \neq \mathbf{c} \in \mathcal{S}^M} \sum \frac{1}{2}\text{erfc})\sqrt{\alpha_{\tilde{\mathbf{c}},\mathbf{c}}}) \leq BLER_0, \\ & \alpha_{\tilde{\mathbf{c}},\mathbf{c}} \leq \Phi_{\tilde{\mathbf{c}},\mathbf{c}}, \forall\, \tilde{\mathbf{c}} \neq \mathbf{c} \end{aligned}$$

Let us introduce μ_n as in Section V-A1, and $\mathbf{x} \in \mathbb{R}^{(L+N+1)\times 1}$

$$\mathbf{x} = [\alpha_1 \ldots \alpha_L \mu_1 \ldots \mu_N 1]^T,$$

and $\mathbf{A}_{\tilde{\mathbf{c}},\mathbf{c}} \in \mathbb{R}^{(L+N+1)\times(L+N+1)}$ which is given in (15), shown at the bottom of the page, where $L = |\mathcal{S}|^M \times (|\mathcal{S}|^M - 1)$. Problem (14) can be expressed as

$$\begin{aligned} \underset{\left\{\mu_n \le \frac{1}{4}\right\}_{n=1}^{N}, x}{\texttt{minimize}} \quad & \sum_{n=1}^{N} -\frac{1}{2}\log_2(\mu_n) \\ \texttt{s.t.} \quad & \frac{1}{|\mathcal{S}|^M} \sum\sum_{\tilde{\mathbf{c}}\neq\mathbf{c}\in\mathcal{S}^M} \frac{1}{2}\text{erfc}(\sqrt{\alpha_{\tilde{\mathbf{c}},\mathbf{c}}}) \le BLER_0, \\ & trace(\mathbf{A}_{\tilde{\mathbf{c}},\mathbf{c}}\mathbf{x}\mathbf{x}^T) \le \left(\sum_{n=1}^{N} \left|\mathbf{h}_n\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})\right|^2\right)^2, \end{aligned} \tag{16}$$

By defining $\mathbf{X} = \mathbf{x}\mathbf{x}^T$, problem (16) is equivalent to the following:

$$\begin{aligned} \underset{\mathbf{X}}{\texttt{minimize}} \quad & \sum_{n=L+1}^{L+N} -\frac{1}{4}\log_2\left([\mathbf{X}]_{n,n}\right) \\ \texttt{s.t.} \quad & \frac{1}{|\mathcal{S}|^M} \sum_{l=1}^{L} \frac{1}{2}\text{erfc}\left([\mathbf{X}]_{l,l}^{1/4}\right) \le BLER_0, \\ & trace(\mathbf{A}_{\tilde{\mathbf{c}},\mathbf{c}}\mathbf{X}) \le \left(\sum_{n=1}^{N} \left|\mathbf{h}_n\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})\right|^2\right)^2, \\ & [\mathbf{X}]_{n,n} \le \frac{1}{16}, n = L+1, \ldots, L+N, \\ & rank(\mathbf{X}) = 1. \end{aligned} \tag{17}$$

The SDR of (17) is obtained by ignoring the rank constraint. It can be shown that the SDR of (17) is a convex optimization problem and is solvable by using, e.g., the primal-dual interior point method [26]. We denote the scheme that solve (17) without the rank constraint as MinBits-SDR.

Compared with the PEP-based minimization, the BLER-based minimization is expected to achieve higher compression ratio with the trade-off of higher computing complexity. This is because the PEP-based solution guarantees all PEP satisfying the target QoS, which can result in a smaller BLER than necessary. Consequently, the PEP-based solution requires more fronthaul rate to achieve a better BLER.

Remark 2: In general, a SDR solution of (17) might violate the rank-one constraint, which is, in fact, a generic problem of SDR. To obtain an approximated (vector) solution $\mathbf{x}^*$ for (17) from a SDR counterpart $\mathbf{X}^*$, we implement the Gaussian randomization procedure [28]. Please refer to [28] for more details of such procedure and its approximation accuracies under several setups.

Remark 3: To facilitate the computation of the first constraint in (17), a tight approximation of the erfc$(x) \simeq \frac{1}{6}e^{-x^2} + \frac{1}{2}e^{-4x^2/3}$ can also be employed [27, eq. (14)]. The resulting problem (18) is still convex and solvable as follows:

$$\begin{aligned} \underset{\mathbf{X}}{\texttt{minimize}} \; & \sum_{n=L+1}^{L+N} -\frac{1}{4}\log_2\left([\mathbf{X}]_{n,n}\right) \\ \texttt{s.t.} \; & \frac{1}{|\mathcal{S}|^M}\sum_{n=1}^{L} \frac{1}{12}\exp\left(-[\mathbf{X}]_{n,n}^{\frac{1}{2}}\right) + \frac{1}{4}\exp\left(-\frac{4}{3}[\mathbf{X}]_{n,n}^{\frac{1}{2}}\right) \le BLER_0, \\ & trace(\mathbf{A}_{\tilde{\mathbf{c}},\mathbf{c}}\mathbf{X}) \le \left(\sum_{n=1}^{N} |\mathbf{h}_n\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})|^2\right)^2, \forall\, \tilde{\mathbf{c}} \neq \mathbf{c} \\ & [\mathbf{X}]_{n,n} \le \frac{1}{16}, n = L+1, \ldots, L+N. \end{aligned} \tag{18}$$

B. Minimization of the Maximum PEP Given the Number of Bits Q_{sum}

In this section, we investigate the dual problem of (10) which is solved in Section V-A1 and Section V-A2 by using PEP- and SDR-based algorithms. Particularly, we want to minimize the BLER given that $\sum_{n=1}^{N} Q_n \le Q_{sum}$. This problem arises under systems with limited front-haul bandwidth but less stringent BLER constraint. The problem, however, is difficult to solve. Therefore, similar to Section V-A1, we will consider an alternative problem based on PEP which gives an upper-bound solution for the original optimization. The alternative problem is mathematically expressed as follows:

$$\begin{aligned} \underset{\{Q_n \ge 1\}_{n=1}^{N}}{\texttt{minimize}} \quad & \max_{\tilde{\mathbf{c}}\neq\mathbf{c}} \Pr\{\mathbf{c} \to \tilde{\mathbf{c}}\} = \frac{1}{2}\text{erfc}(\sqrt{\Phi_{\tilde{\mathbf{c}},\mathbf{c}}}) \\ \texttt{s.t.} \quad & \sum_{n=1}^{N} Q_n \le Q_{sum} \end{aligned} \tag{19}$$

Because erfc(.) is a monotonic function, by introducing an auxiliary varable α, we reformulate (19) as:

$$\begin{aligned} \underset{\{Q_n \ge 1\}_{n=1}^{N}}{\texttt{maximize}} \quad & \alpha \\ \texttt{s.t.} \quad & \alpha \le \Phi_{\tilde{\mathbf{c}},\mathbf{c}}, \forall\, \tilde{\mathbf{c}} \neq \mathbf{c}, \\ & \sum_{n=1}^{N} Q_n \le Q_{sum}, \end{aligned} \tag{20}$$

$$\mathbf{A}_{\tilde{\mathbf{c}},\mathbf{c}} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ 0 \ldots 0 \frac{8}{3}\eta_1^2 \left|\mathbf{h}_1\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})\right|^2 \ldots \frac{8}{3}\eta_N^2 \left|\mathbf{h}_N\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})\right|^2 4\sigma^2 \sum_{n=1}^{N} \left|\mathbf{h}_n\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})\right|^2 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \tag{15}$$

TABLE I
MinBLER-PEP Algorithm

```
1. Initialize $\alpha_H$, $\alpha_L$ and $\epsilon$.
2. $\alpha_M = (\alpha_H + \alpha_L)/2$.
3. If (23) is feasible, then $\alpha_L := \alpha_M$.
   Otherwise $\alpha_H := \alpha_M$.
4. Repeat step 2 and 3 until $|\alpha_H - \alpha_L| \le \epsilon$.
```

where $\Phi_{\tilde{\mathbf{c}},\mathbf{c}}$ has been defined in Section V-A. In another form, (20) is identical to the following problem:

$$\begin{aligned} \underset{\{Q_n\ge 1\}_{n=1}^N}{\texttt{maximize}} \quad & \alpha \\ \texttt{s.t.} \quad & (22), \forall \tilde{\mathbf{c}} \ne \mathbf{c}, \\ & \sum_{n=1}^{N} Q_n \le Q_{sum}, \end{aligned} \tag{21}$$

where the first constraint in problem (21) is given as follows:

$$\begin{aligned} &\frac{8}{3}\sum_{n=1}^{N}\eta_n^2\left|\mathbf{h}_n\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})\right|^2 2^{-2(Q_n+1)} \\ &\le \frac{\left(\sum_{n=1}^{N}\left|\mathbf{h}_n\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})\right|^2\right)^2}{\alpha} - 4\sigma^2\sum_{n=1}^{N}\left|\mathbf{h}_n\mathbf{P}(\tilde{\mathbf{c}}-\mathbf{c})\right|^2. \end{aligned} \tag{22}$$

Similar to Section V-A1, we have used the maximum PEP as a proxy for the BLER minimization. The auxiliary variable α effectively represents the minimum SNR across all users which is achievable under the Q_{sum} constraint. Note that given α, (21) is a convex optimization problem and thus solvable using standard methods. We therefore resort to a bisection technique to solve (21). The resulting optimization problem, whose steps are given in Algorithm 1, is given as follows:

$$\begin{aligned} \underset{\{Q_n\ge 1\}_{n=1}^N}{\texttt{maximize}} \quad & \alpha_M \\ \texttt{s.t.} \quad & \Gamma(\alpha_M), \forall \tilde{\mathbf{c}} \ne \mathbf{c} \\ & \sum_{n=1}^{N} Q_n \le Q_{sum}, \end{aligned} \tag{23}$$

where $\Gamma(\alpha_M)$ is obtained by replacing α in (22) by α_M. We denote this scheme as MinBLER-PEP (Table I).

Remark 4: The decoder and optimization derived in this paper are based on the ML receiver and thus have a complexity that exponentially increases with the number of users. In order to reduce the computing complexity, one can consider sphere decoder, which significantly reduces the computing complexity by reducing the search circle, but can achieve a close performance as the ML decoder.

VI. Simulation Results

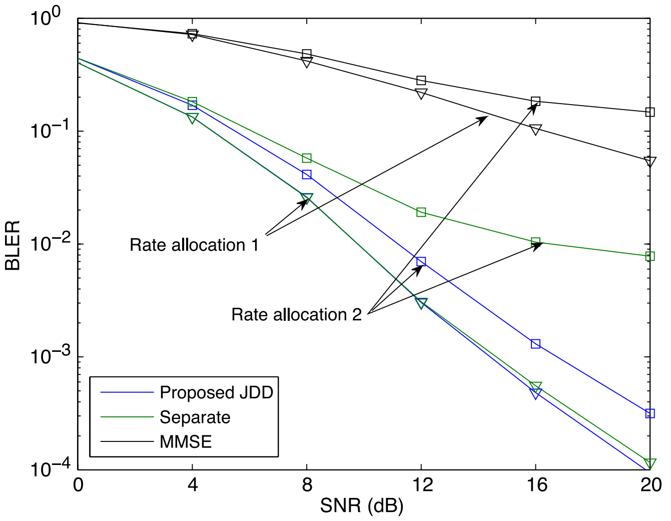


Fig. 2. BLER of the proposed JDD algorithm compared with separate decompression and detection, and with MMSE receivers. Rate allocation 1: $[Q_1, Q_2, Q_3] = [4, 4, 4]$. Rate allocation 2: $[Q_1, Q_2, Q_3] = [2, 4, 6]$.

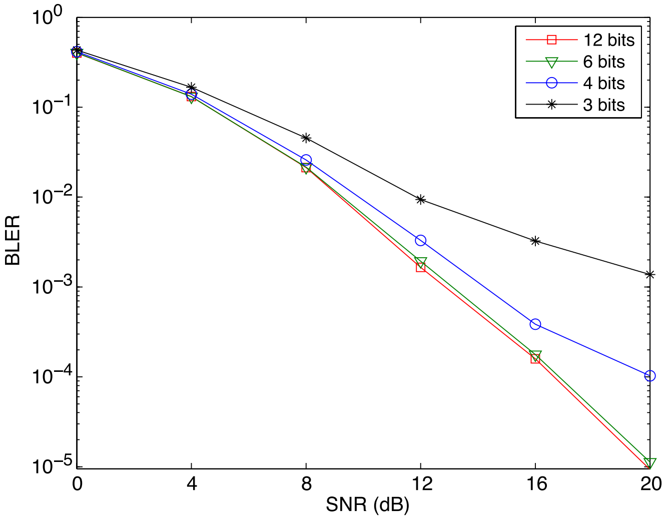
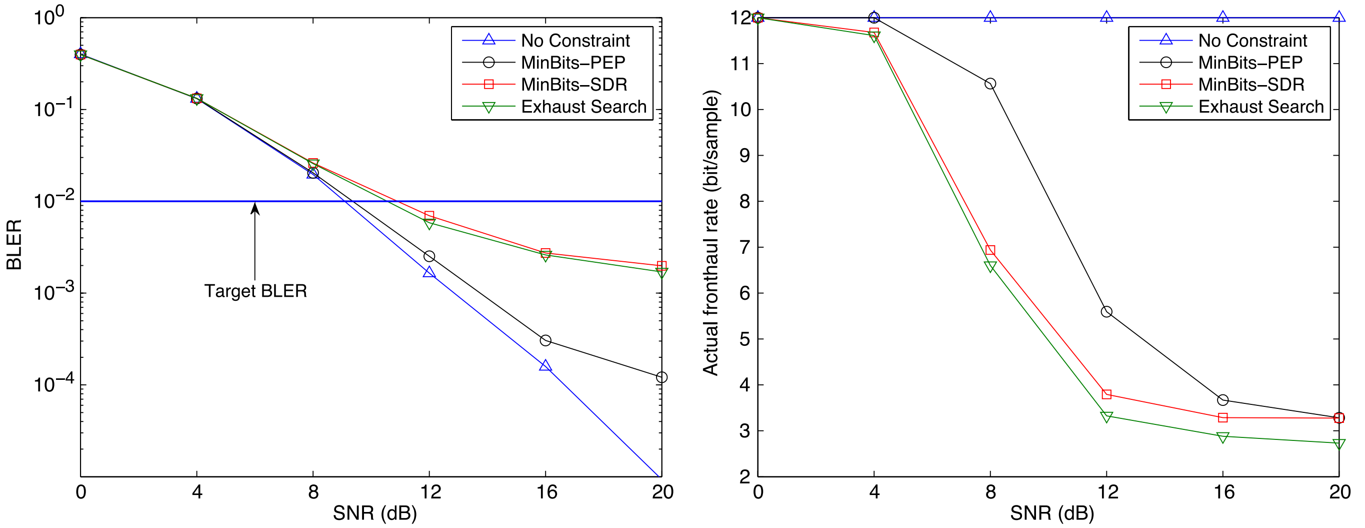


Fig. 3. BLER of the joint decompression and detection algorithm with different fronthaul sampling resolution $Q = 3, 4, 6, 12$ bits.

To demonstrate the effectiveness of the proposed algorithms, simulation is carried out on a network that consists of $M = 3$ users and $N = 3$ RRHs. We assume block Rayleigh fading channel and symmetric network with equal user's transmit energy, e.g., $P_1 = \ldots = P_M = P$. The average SNR is defined as P/σ^2. In addition, $\sigma_{h_{nn}}^2 = 1, \forall n$, and $\sigma_{h_{nm}}^2 = 0.5$ for $n \ne m$. During each channel realization, the users emit a message comprising of $K = 1000$ data symbols that belong to a QPSK codebook, i.e., $\mathcal{S} = \{-1-1i, -1+1i, 1-1i, 1+1i\}$. All the RRHs apply uniform quantization. The BBU is assumed to know CSI of the entire network.

Fig. 2 compares the proposed JDD algorithm with two references: Separate decompression and detection (SDD) and minimum mean square error (MMSE) [29] receivers. The SDD receiver detects the source symbols merely based on the Hamming distance between the received signal and the trial codeword. The MMSE receiver separates the received signal vector into M orthogonal sub-streams by multiplying it with a matrix $\mathbf{W}$: $\hat{\mathbf{x}} = \mathbf{W}^H\tilde{\mathbf{y}}$, where $\mathbf{H} = [\mathbf{h}_1{}^T, \ldots, \mathbf{h}_N{}^T]^T$, $\tilde{\mathbf{y}} = [\tilde{y}_1, \ldots, \tilde{y}_N]^T$, $\mathbf{W} = (\mathbf{H}\mathbf{H}^H + N\mathbf{I})^{-1}\mathbf{H}$, and $(.)^H$ and $(.)^{-1}$ denote the Hermitian transpose and inversion operator, respectively. We note that SDD and MMSE receivers ignore the quantization noise and use the quantized output $\tilde{\mathbf{y}}$ as the correct output. It is shown that the proposed algorithm, which takes into consideration effect of quantization noise, achieves the best performance compared with the references. Such expected gain results from



(a)

(b)

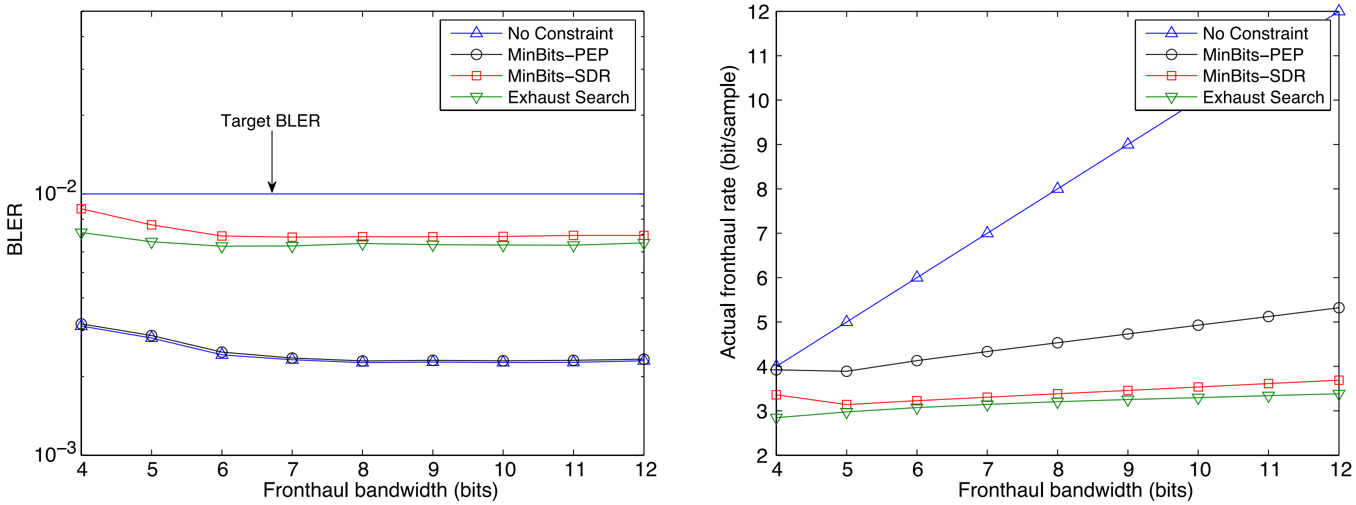
Fig. 4. BLER performance (a) and compression efficiency (b) of the adaptive compression schemes with target $BLER_0 = 1\text{e-}2$ for different SNR. The fronthaul bandwidth $Q = 12$ bits.

the searching over all combinations of the source symbols of the proposed algorithm. In addition, the proposed JDD algorithm yields larger gain over SDD receiver in non-uniform fronthaul bandwidths (Rate allocation 2). This is because the quantization noise in this case, which is ignored in SDD receiver, has more impact on the overall system performance.

Fig. 3 presents the BLER performance of the proposed receiver versus SNR for different fronthaul sampling rates. The BLER is measured as the probability that at least one of M source symbols is decoded with error. Here we assume that the RRHs have uniform fronthaul bandwidth allocation. It is observed that the fronthaul bandwidth Q has effects on both BLER and diversity order. In general, larger Q results in better BLER, and the BLER will saturate as Q decreases. At low SNRs, the contribution of quantization noise is small. For example, under SNRs between 0 dB and 8 dB, sampling at 4 bits per sample or 12 bits per sample achieves almost similar BLER. This is due to the fact that under small/medium SNR regime, thermal noise is large, and therefore, is dominant compared with the quantization noise. In contrast, under the high SNR regime, the thermal noise is comparable to or even smaller than the quantization noise. Decreasing Q in this case can result in severe loss in BLER. It is also observed that a 6-bit quantizer achieves almost similar performance as a 12-bit counterpart under the considered setting and the observing SNR range. This interesting observation suggests an adaptive compression scheme to minimize the fronthaul rate while maintaining the BLER under a given QoS.

Fig. 4 presents the BLER of adaptive compression versus SNR for given QoS constraint. The premise is that different applications require various QoS, e.g., BLER levels. For a given BLER, we want to maximize the compression efficiency, or equivalently to minimize the fronthaul transmission rate. Two adaptive compression schemes based on PEP constraint (Section V-A1, named MinBits-PEP) and BLER constraint (Section V-A2, named MinBits-SDR) are presented. In addition, the scheme without QoS constraint which fully occupies the fronthaul bandwidth (named “No Constraint”), and the “Exhaustive Search” optimization are also plotted. For the Exhaustive Search, we check every combination of the rate allocation $[Q_1, Q_2, \dots, Q_N]$ and find the one which gives the minimum fronthaul rate while satisfying the BLER target. Such scheme yields the optimal rate allocation, but hindered by its NP-complete complexity. 300 channel realizations are conducted in the simulation. The threshold $BLER_0$ is equal to 1e-2. The BLER performance is shown in Fig. 4(a) and the actual fronthaul rate is presented in Fig. 4(b). At very low SNR, the BLER does not satisfy the target QoS because the channel is too poor. Even using all 12 bits for quantization does not satisfy the target BLER. One good thing is that although not satisfying the target BLER, the proposed adaptive compressions achieve smaller fronthaul rate (higher compression ratio). More specifically, at 8 dB the MinBits-PEP algorithm saves 1.4 bits and the MinBits-SDR algorithm saves 5 bits. Moving to higher SNR regime (from 12 dB in the figure), both adaptive schemes meet the target QoS while significantly improve the compression ratio. Because the No Constraint scheme always uses 12 bits for quantization, its fronthaul rate is 12 bits per sample in the whole SNR range. On the other hand, a compression ratio of 350% is observed by both adaptive schemes, which only require 3.4 bits per sample to achieve a BLER less than or equal 1e-2. It is also shown in the figure that the MinBits-SDR achieves close performance to the Exhaustive Search, which confirms the effectiveness of the SDR optimization. Furthermore, the MinBits-SDR obtains a better compression efficiency than the MinBits-PEP at low and medium SNR. This result is obvious since the MinBits-PEP minimizes the worst PEP while the MinBits-SDR targets the BLER. When the SNR is large, it prefers to employ MinBits-PEP because it yields smaller BLER.

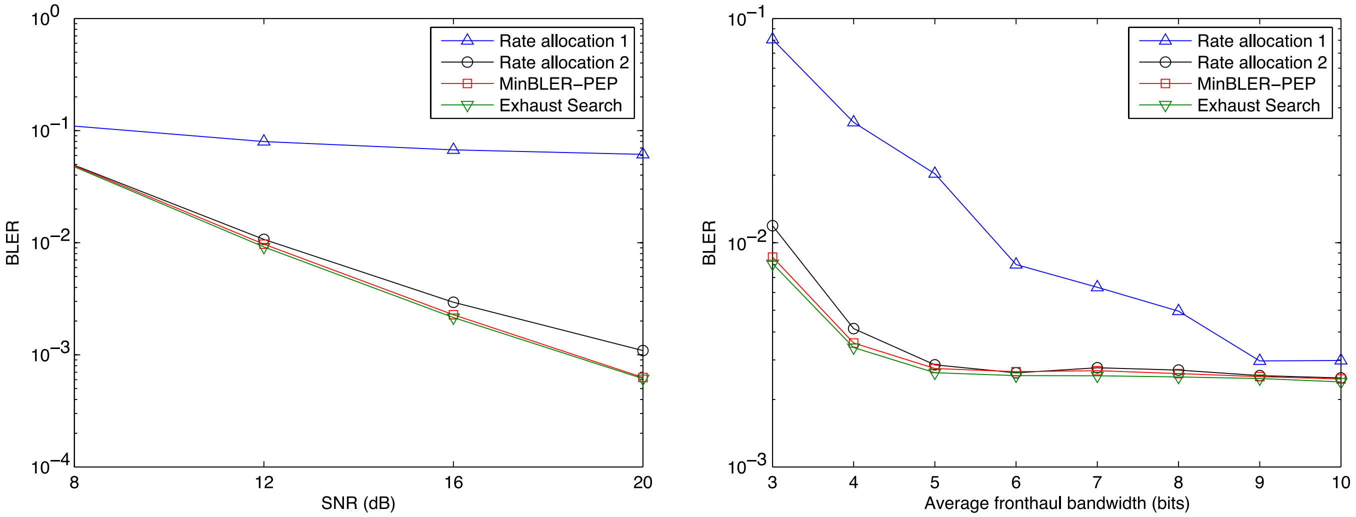
Fig. 5 shows BLER and compression efficiency of adaptive compressions for different fronthaul bandwidth Q at SNR = 12 dB. Similar to the previous simulation, the target $BLER_0$ is set at 1e-2. The results show that all schemes satisfy the BLER target while significantly reduce the fronthaul rate. The



(a)

(b)

Fig. 5. BLER performance (a) and compression efficiency (b) of adaptive compression with target $BLER_0 = 1\text{e-}2$ for different fronthaul bandwidths, SNR $= 12$ dB.



(a)

(b)

Fig. 6. BLER comparison of the optimal rate allocation scheme for different Q_{sum}. Rate allocation 1: $[Q_1, Q_2, Q_3] = [n, n, n]$, Rate allocation 2: $[Q_1, Q_2, Q_3] = [n-2, n, n+2]$, where n is the average fronthaul bandwidth. In Fig. a, $Q_{sum} = 9$. In Fig. b, SNR $= 12$ dB. (a) BLER v.s. SNR. (b) BLER v.s. fronthaul bandwidth.

MinBits-SDR does exactly what we require: it satisfies the BLER of 1e-2, but no more. In that sense, it is the best since it satisfies the constraint but with less number of bits. The MinBits-PEP acheives similar BLER as that of the No Constraint scheme while the MinBits-SDR obtains slightly worse BLER. However, they are all satisfied the target BLER of 1e-2. Specifically, the compression efficiency obtained by the MinBits-PEP increases from 120% at $Q = 5$ to 240% at $Q = 12$; while the MinBits-SDR achieves better compression efficiency from 160% at $Q = 5$ to about 330% at $Q = 12$. This observation is consistent with the result in Fig. 3, which shows that at 12 dB the 4-bit sampling and 12-bit sampling yield approximately similar performance.

The above results aim at minimizing the actual fronthaul rate subject to the BLER constraint. Based on the analysed BLER in Section IV, a reciprocal problem is how to allocate the fronthaul bandwidth $\{Q_n\}_{n=1}^{N}$ to minimize the BLER for a given sum of $Q_{sum} = \sum_{n=1}^{N} Q_n$. Because the n-th RRH uses all Q_n bit for quantization, we refer Q_n as fronthaul rate in this paragraph for simplicity. The optimization problem is described in Section V-B. Fig. 6(a) shows the BLER of the MinBLER-PEP and two other allocation schemes: uniform rate allocation (Rate allocation 2), e.g., $Q_1 = Q_2 = Q_3$, and non-uniform rate allocation (Rate allocation 1), e.g., $[Q_1, Q_2, Q_3] = [Q-2, Q, Q+2]$. The sum rate $Q_{sum} = 9$. In addition, the performance of Exhaustive Search is also presented. For the Exhaustive Search scheme in Fig. 6, we check every combination of the rate allocation $[Q_1, Q_2, Q_3]$ and find the one which gives the minimum BLER based on (4) while satisfying the Q_{sum} constraint. As expected, our MinBLER-PEP algorithm achieves almost the same BLER as the Exhaustive Search, which yields the best BLER, but is limited by its NP-complete complexity. It is observed that the non-uniform rate allocation scheme obtains the worse BLER because there is always one RRH using fewer quantization bits than needed. At low SNR, both uniform and optimal rate allocations get similar performance because at this SNR the thermal noise is dominant. As SNR increases, performance gain provided by the optimal rate allocation is larger.

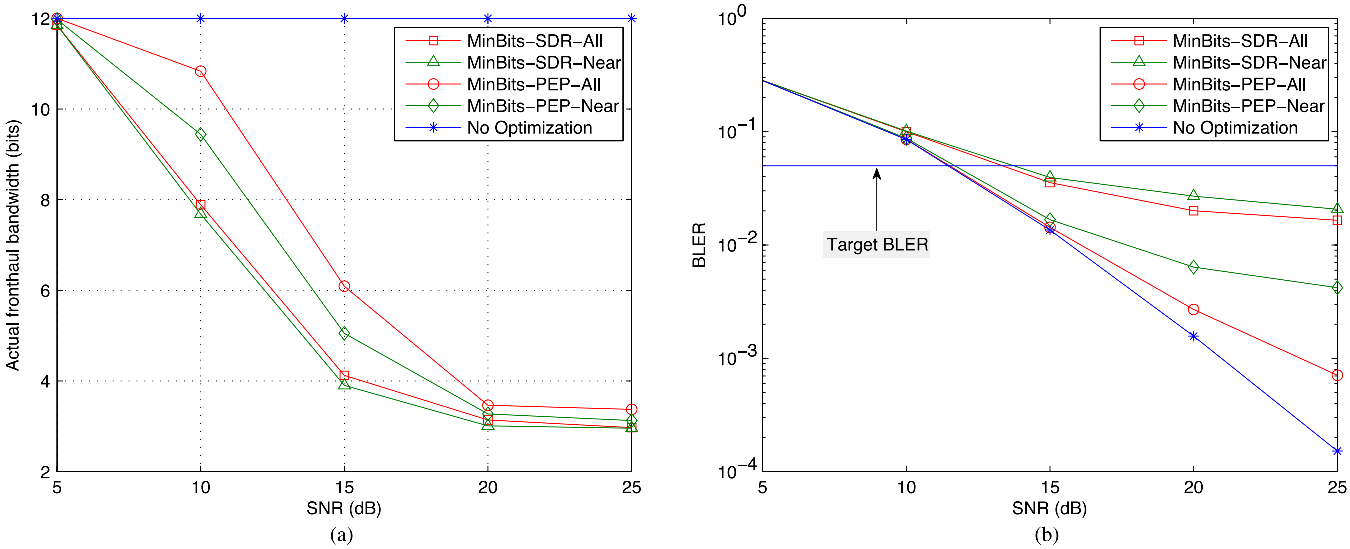


Fig. 7. Performance comparison between all-symbol scheme and nearest-symbol scheme. 4-PAM modulation, $N = M = 2$. Average fronthaul bandwidth is equal to 12 bits. (a) Compression efficiency. (b) BLER performance.

Fig. 6(b) shows the BLER for different number of Q_{sum} at 12 dB. The performance of non-uniform allocation is the worst when Q_{sum} is small but it approaches the optimal BLER when Q_{sum} increases. The optimal allocation outperforms uniform rate allocation with small Q_{sum} and achieves similar BLER as Q_{sum} increases. Other observation is that the optimal BLER remains constant for a large range of Q_{sum}. This is explained from Fig. 3 that at 12 dB, 4-bit quantization nearly achieves the same performance as 12-bit quantization. Spending more bits in this case only bring a small improvement in BLER. In conclusion, the MinBLER-PEP based rate allocation is more effective in high SNR and when the fronthaul bandwidth is small. Otherwise, uniform rate allocation is less complex but still can achieve the best performance.

Remark 5: The schemes MinBits-SDR and MinBits-PEP derived in the present paper are based on the bound (4), which consider all PEPs. When the number of users or the modulation order is large, this bound can be too relax. In this case, considering the nearest symbols only might be practically preferred due to its lower complexity (less PEP constraints) and better compression gain. Such possible gain is demonstrated in Fig. 7.

VII. Conclusion and Discussions

We have proposed a near-optimal receiver structure for cloud radio access networks, which takes into consideration quantization effect of capacity-limited fronthaul links and exploits the correlation among the remote radio heads. Analytical result has been derived for the pair-wise error probability. Based on the analysed PEP, two adaptive compression schemes have been proposed to improve the compression efficiency while satisfying the QoS constraint. The proposed optimization problem comes from practical situation in which most applications tolerate a target QoS, e.g., BLER. A compression efficiency of 350% can be achieved by the proposed optimization schemes. In addition, an optimal rate allocation, which minimizes the BLER given the compression efficiency, has also been proposed based on the theoretical PEP.

For future extensions of this work, several directions are promising. The first problem is to develop low-complexity decoder for the proposed architecture, e.g, sphere decoding. The second problem is to consider imperfect CSI scenario. In such practical systems, due to a large number of nodes in CRAN, it is usually very difficult, if not impossible, to acquire accurate CSI of every link. In this case, understanding the impact of CSI error on the system performance is crucial. Third, the current work assumes the network is fully connected and the BBU has knowledge of all connections. However, in reality, the BBU might only know the instantaneous value of a subset of the channels because some links cannot be estimated. Therefore, a robust receiver design is required. Finally, we can also consider adaptive modulation. In practical systems with time-varying channels, it is necessary to choose the suitable modulation schemes as well as to optimize the usage of fronthauls bandwidth. This technique, however, requires another layer of optimization.

References

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

[2] Z. Zhu *et al.*, "Virtual base station pool: Towards a wireless network cloud for radio access networks," in *Proc. ACM Int. Conf. Comput. Frontiers*, New York, NY, USA, Mar. 2011, pp. 1–10.

[3] 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.

[4] 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, Jun. 2011.

[5] "Mobile Traffic Forecasts 2010–2020," Univ. Mobile Telecommun. Syst. (UMTS), London, U.K., Tech. Rep. 44, 2011.

[6] Y. Zhou, W. Yu, and D. Toumpakaris, "Uplink multi-cell processing: Approximate sum capacity under a sum backhaul constraint," in *Proc. IEEE Inf. Theory Workshop*, Sevilla, Spain, Sep. 2013, pp. 1–5.

[7] 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, Aug. 2013.

[8] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," in *Proc. IEEE Inf. Theory Appl. Workshop*, San Diego, CA, USA, Feb. 2014, pp. 1–6.

[9] 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 Commun.*, vol. 13, no. 3, pp. 1555–1567, Mar. 2014.

[10] 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.

[11] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5068–5081, Sep. 2015.

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

[13] D. Gesbert *et al.*, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[14] J. Zhao, T. Q. S. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer constraints," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.

[15] Z. Jian, T. Q. S. Quek, and Z. Lei, "Wireless backhaul in heterogeneous cellular networks: Fast admission control and large system analysis," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2128–2143, Oct. 2015.

[16] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint decompression and decoding for cloud radio access networks," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 503–506, Mar. 2013.

[17] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun. Mag.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[18] T. Q. S. Quek, M. Peng, O. Simeone, and W. Yu, *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, to be published.

[19] "Common Public Radio Interface," Specification V6.0, 2013.

[20] J. Lorca and L. Cucala, "Lossless compression technique for the fronthaul of LTE/LTE-advanced cloud-RAN architectures," in *Proc. IEEE Int. Symp. Workshops World Wireless, Mobile Multimedia Netw.*, Madrid, Spain, Jun. 2013, pp. 1–9.

[21] A. Nanba and S. Agata, "A new IQ data compression scheme for fronthaul link in centralized RAN," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun.*, London, U.K., Sep. 2013, pp. 210–214.

[22] K. F. Nieman and B. L. Evans, "Time-domain compression of complex-baseband LTE signals for cloud radio access networks," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Austin, TX, USA, Dec. 2013, pp. 1198–1201.

[23] "Feasibility study for further advancements for E-UTRA (LTE-advanced)," Eur. Telecommun. Standards Inst. (ETSI), Sophia-Antipolis Cedex, France, Tech. Rep. ETSI TR 136 912, 2010.

[24] T. X. Vu, H. D. Nguyen, and T. Q. S. Quek, "Joint decoding and adaptive compression with QoS constraint for uplinks in cloud radio access networks," in *Proc. IEEE Global Commun. Conf.*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[25] R. Gray and D. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[27] M. Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Trans. Commun.* vol. 2, no. 4, pp. 840–845, Jul. 2003.

[28] Z. Q. Luo, W. K. Ma, A. M. C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, Mar. 2010.

[29] 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.

Thang X. Vu (S'11–M'15) was born in Hai Duong, Vietnam. He received the B.S. and the M.Sc. degrees both in electronics and telecommunications engineering, from the VNU University of Engineering and Technology, Vietnam, in June 2007 and September 2009, respectively, and the Ph.D. degree in electrical engineering from the University Paris-Sud, France, in January 2014.

From 2007 to 2009, he was with the Department of Electronics and Telecommunications, VNU University of Engineering and Technology as a Research Assistant. In 2010, he received the Allocation de Recherche fellowship to study Ph.D. in France. From September 2010 to May 2014, he was with the Laboratory of Signals and Systems (LSS), a joint laboratory of CNRS, CentraleSupelec and University Paris-Sud XI, France. Since July 2014, he has been a Postdoctoral Research Fellow at the Information Systems Technology and Design (ISTD) pillar, Singapore University of Technology and Design (SUTD), Singapore. His research interests are in the field of wireless communications, with particular interests of cloud radio access networks (C-RAN), cooperative diversity, channel and network decoding, and iterative decoding.

Hieu D. Nguyen (S'10–M'14) received the B.S. (First-Class Hons.) degree from the Vietnam National University, Hanoi, Vietnam, in 2009, and the Ph.D. degree from National University of Singpaore, Singpaore, in 2013, all in electrical engineering. Since October 2013, he has been with the Advanced Communication Technology Department, Institute for Infocomm Research (I2R), A-STAR, Singapore, where he is now a Research Scientist. His current research interests include wireless communications and information theory, focusing on distributed antenna systems, cloud radio access networks, large-system analysis, stochastic geometry, and interference channels.

Tony Q. S. Quek (S'98–M'08–SM'12) received the B.E. and M.E. degrees in electrical and electronics engineering from Tokyo Institute of Technology, Tokyo, Japan, respectively. He received the Ph.D. degree in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, MA, USA. Currently, he is an Assistant Professor with the Information Systems Technology and Design Pillar at Singapore University of Technology and Design (SUTD). He is also a Scientist with the Institute for Infocomm Research. His main research interests are the application of mathematical, optimization, and statistical theories to communication, networking, signal processing, and resource allocation problems. Specific current research topics include heterogeneous networks, green communications, smart grid, wireless security, Internet-of-Things, big data processing, and cognitive radio.

He has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is serving as the technical chair for the PHY & Fundamentals Track for IEEE WCNC in 2015, the Communication Theory Symposium for IEEE ICC in 2015, the PHY & Fundamentals Track for IEEE EuCNC in 2015, and the Communication and Control Theory Symposium for IEEE ICCC in 2015. He is currently an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE WIRELESS COMMUNICATIONS LETTERS, and an Executive Editorial Committee Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was Guest Editor for the IEEE SIGNAL PROCESSING MAGAZINE (Special Issue on Signal Processing for the 5G Revolution) in 2014, and the IEEE WIRELESS COMMUNICATIONS MAGAZINE (Special Issue on Heterogeneous Cloud Radio Access Networks) in 2015.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the IEEE Globecom 2010 Best Paper Award, the CAS Fellowship for Young International Scientists in 2011, the 2012 IEEE William R. Bennett Prize, the IEEE SPAWC 2013 Best Student Paper Award, and the IEEE WCSP 2014 Best Paper Award.