# Distributed Signal Compressive Quantization and Parallel Interference Cancellation for Cloud Radio Access Network

Hang Ma, *Member, IEEE*, Beibei Wang, *Senior Member, IEEE*, and K. J. Ray Liu, *Fellow, IEEE*

***Abstract*— The explosive growth of wireless traffic requires revolutionary wireless communication techniques. The cloud radio access network (C-RAN) is a solution to leverage the spatial multiplexing gain by utilizing a number of antennas distributed in a certain area. However, in most of the current literature, due to the limited front-haul capacity, each remote radio head (RRH) can use only a single antenna, and thus the total number of antennas can be utilized are limited. In this work, we propose a distributed baseband signal compressive quantization scheme for the uplink of a multi-antenna C-RAN where each RRH uses multiple antennas. At each RRH, the baseband signals of multiple time instants are embedded into a vector with low dimension using delay-and-add so that more bits can be allocated to each value, and the quantization noise power caused by the front-haul link capacity deficit is reduced. The delay-and-add operation is low complexity that can be realized using basic buffering and adding, and it does not require channel information in the RRHs. Therefore, the low deployment cost feature of C-RAN is preserved. As a result, a large number of antennas can be utilized by deploying a lot of multi-antenna RRHs, which provide rich spatial diversity that assists the detection which happens in the baseband units. In the symbol detection phase, the corresponding weight vectors are designed to detect the symbols from the compressive quantized baseband signal. A parallel interference cancellation algorithm is proposed to further improve the accuracy of the symbol detection. Numerical results show that the proposed scheme is efficient in tackling the front-haul capacity challenge. We also apply the proposed scheme to orthogonal frequency-division multiplexing-based multi-antenna C-RAN, where we find that the system can utilize larger bandwidth with limited front-haul capacity. It facilitates the deployment of C-RAN based on both 4G and 5G wireless communications.**

***Index Terms*— Interference suppression, broadband communication, least mean square methods, antenna arrays, MIMO, data compression, quantization.**

## I. Introduction

With the proliferation of new mobile devices and applications, the demand for ubiquitous wireless services has increased dramatically in recent years. It has been projected that by the year 2020, the volume of the wireless traffic will rise to about 1000 times that of the year 2010 [1]. This explosion of wireless traffic will be a new challenge to wireless networks. In fact, people have started to feel the impact in some places, such as at the airport, conference and stadiums where it is difficult to access the wireless network with hundreds of other devices around.

To address the aforementioned challenge, many techniques have been developed. In recent years, massive multiple-input and multiple-output (massive MIMO) has drawn the attention of researchers. Following the same idea of leveraging the spatial multiplexing gain as the traditional MIMO, the massive MIMO utilizes a much larger number of antennas, and some interesting features have been discovered, such as the improved spectral efficiency and energy efficiency which can be achieved by maximum ratio combining (MRC) in the uplink or maximum ratio transmission (MRT) in the downlink [2], [3]. However, despite the advantages, the massive MIMO is limited by a few bottlenecks. For example, due to the limited separation between antennas, the channels are always correlated, which undermines the advantages made possible with massive antennas [4], [5].

With the same initiative of utilizing extra antennas, another promising approach is the cloud radio access network (C-RAN) [6]–[9] where multiple remote radio heads (RRHs) are connected to a central baseband units (BBUs) pool powered by high performance computing. In the uplink transmission, instead of decoding the received symbols, each RRH just performs the basic receptions and forwards the baseband signal to the BBU pool for centralized processing. In the C-RAN, the BBUs are able to utilize the multiple antennas distributed in a certain area and work like a virtual MIMO. Due to the centralized processing, the cooperation between RRHs becomes much easier, and therefore the cell edge coverage is improved. Moreover, the cost of deployment is reduced since each RRH only needs the basic transmission and reception functions. However, a major bottleneck is the limited capacity of the front-haul link which connects the RRH and the BBU pool [10]. Many works have been done to alleviate the impact of the bottleneck in the C-RAN. Several baseband signal compression methods are proposed in [11]–[13] where the baseband signal is compressed before transmitted through the front-haul links. Although baseband signal compression can alleviate the traffic in the front-haul

Manuscript received September 14, 2017; revised January 29, 2018 and April 3, 2018; accepted April 3, 2018. Date of publication April 13, 2018; date of current version September 14, 2018. The associate editor coordinating the review of this paper and approving it for publication was D. Niyato. *(Corresponding author: Hang Ma.)*

H. Ma was with the Department of Electrical and Computer Engineering, University of Maryland at College Park, College Park, MD 20742 USA. He is now with Google Inc., Mountain View, CA 94043 USA (e-mail: mahang2010@gmail.com).

B. Wang and K. J. R. Liu are with the Department of Electrical and Computer Engineering, University of Maryland at College Park, College Park, MD 20742 USA, and also with Origin Wireless Inc., College Park, MD 20742 USA (e-mail: bebewang@umd.edu; kjrliu@umd.edu).

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.2018.2826545

under certain cases, it introduces extra computation complexity at the RRH side, which makes this approach less cost-effective. An alternative solution is the sparse beamforming [14]–[16] where each terminal device (TD) is associated with a cluster of RRHs. However, the data rate in the front-haul link is related to the cluster size, and a larger cluster requires a higher front-haul link capacity [14]. As a result, the limited front-haul link capacity makes it impossible to take full advantage of the available spatial diversity. Ma *et al.* [17] explore to utilize the time-reversal communication [18], [19] as the air interface in order to alleviate the traffic load in the front-haul links without incurring compression or decompression in the RRH. Despite all the above efforts, the C-RAN is mostly utilizing single antenna RRHs due to the limited front-hual capacity, which limits the total number of available antennas.

We notice that if the total number of antennas is large, the C-RAN works like a virtual massive MIMO, and many nice properties of massive MIMO will apply. For example, the massive antennas help focus the energy into smaller region [20], such that more TDs can be accommodated efficiently. The nice properties of massive MIMO are essential to the next generation of wireless communication systems. However, in the current literature where each RRH always uses a single antenna, the total number of antennas that can be utilized is limited. Recently, motivated by the advantages of massive antennas, the multi-antenna C-RAN is investigated [21]. Since the amount of baseband signal data in the front-haul link is proportional to the number of antennas used by the RRH, a spatial filtering method is designed in [22] to compress the baseband signal in the uplink. However, the design of the spatial filter is performed at the BBUs due to the limited computation capability of RRHs, which brings a lot of overhead in the already-tight front-haul links. Moreover, the compression of the baseband signal using the spatial filter involves massive multiplication operations, which brings up the cost of each RRH.

Recall that in the massive MIMO, the MRT precoding [20] and MRC receiver [23] are widely used in downlink and uplink, respectively. In the uplink, the performance of the low cost MRC receiver approximates the the matched filter bound (MFB) detector [23]. It is due to the rich spatial diversity provided by the massive antennas. In the multi-antenna C-RAN, the rich spatial diversity is also available due to the fact that the total number of antennas is large. In this work, we aim to leverage this rich spatial diversity to design low cost but effective baseband signal processing schemes which can facilitate the C-RAN tackle the front-haul capacity challenge caused by using multi-antenna RRHs. Instead of elaborately designed compressor, quantizer and decoder, we propose a novel baseband signal compressive quantization at the RRH side coupled with the weight vectors designed at the BBU side. The compressive quantization consists of two phases: the delay-and-add and the quantization. The delay-and-add simply buffers the baseband signals of multiple time instants and sums the delayed versions of them, which is then quantized and transmitted to the BBUs through the front-haul link. By the delay-and-add, the raw baseband signals from multiple time instants are embedded in a vector with lower dimension compared with the direct concatenation, and more bits can be allocated to each value in the quantization. As a result, the quantization noise is mitigated. On the other hand, due to the overlapping of baseband signals from multiple time instants, extra interference is created. Unlike the quantization noise, the interference has its own structure, and is easy to mitigate and cancel out with the rich spatial diversity provided by the massive antennas. In other words, different from general baseband signal compression algorithms, the compressive quantization is tailored for the multi-antenna C-RAN where massive antennas are available. In the BBU side, the linear minimum mean square error (LMMSE) and MRC detectors are designed to detect the symbols from the signal processed by the RRHs using compressive quantization. To fully utilize the computation power of the BBUs, a parallel interference cancellation algorithm is proposed to further reduce the influence of interference. Moreover, we extend the proposed scheme to an orthogonal frequency-division multiplexing (OFDM) based multi-antenna C-RAN and find that the proposed scheme helps the OFDM based multi-antenna C-RAN utilize larger bandwidth with limited front-haul capacity. This extension is applicable for C-RAN based on both OFDMA in 4G wireless communications, as well as non-orthogonal multiple access (NOMA) in 5G wireless communications. Numerical results in various settings illustrate the effectiveness of the proposed schemes.

The rest of the paper is organized as follows: in section II, the system model is introduced. We propose and analyze the compressive quantization scheme as well as the corresponding symbol detection schemes in section III. In section IV, the proposed scheme is applied to the OFDM based multi-antenna C-RAN. Numerical results are shown in section V and section VI concludes the paper.

## II. System Model

We consider the uplink communication of the cloud radio access network (C-RAN). As shown in Fig. 1, a group of $N$ terminal devices (TDs) are distributed in an area and they work in the same band. Each TD transmits the signal, which is received by $M$ remote radio heads (RRHs) distributed in the same area. All the RRHs periodically report the received baseband signal to the baseband unit (BBU) pool through front-haul links with limited capacity. Without loss of generality, we assume the $i$-th RRH is equipped with $Q_i$ antennas where $Q_i > 1$, and each TD is equipped with single antenna.

In this work, we assume that the instant channel state information (CSI) is available at the BBUs through channel estimation prior to the uplink transmission. The channel estimation is periodically performed for each TD to update the out-of-date CSIs. The channel impulse response between the $i$-th RRH and the $j$-th TD is $\mathbf{h}_{i,j} = [h_{i,j}^{(1)}, h_{i,j}^{(2)}, \cdots, h_{i,j}^{(Q_i)}]^T$ where $h_{i,j}^{(m)}$ is the channel impulse response between the $m$-th antenna of RRH $i$ and TD $j$.

In the uplink transmission, all the RRHs periodically report the received baseband signal to BBUs and the BBUs are responsible for detecting the symbols transmitted by the TDs using the received signal. The signal received by the $i$th RRH

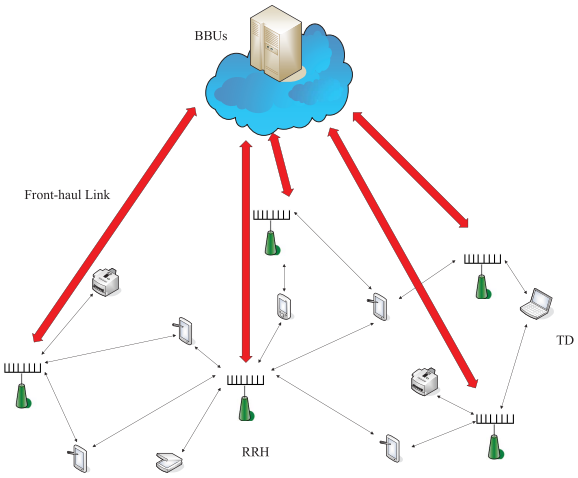



Fig. 1. The System Model.

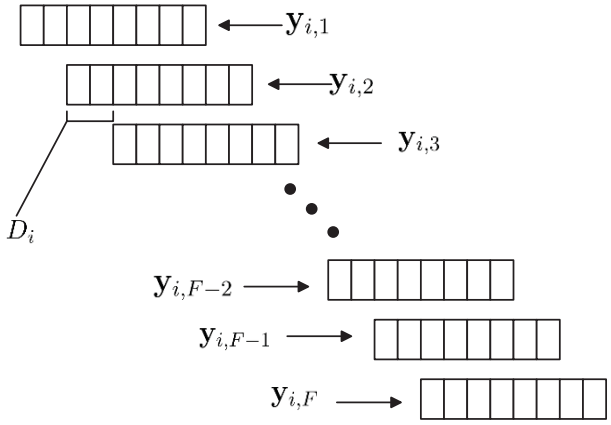



Fig. 2. An example of the delay-and-add phase with $Q_i = 8$, $D_i = 2$.

at time instant $t$ can be represented as

$$\mathbf{y}_{i,t} = \mathbf{H}_i \mathbf{x}_t + \mathbf{n}_{i,t} \tag{1}$$

where $\mathbf{x}_t \in \mathbb{C}^{N\times 1}$ is the vector containing the intended symbols of all the TDs at time $t$, $\mathbf{n}_{i,t} \in \mathbb{C}^{Q_i\times 1}$ is the additive white gaussian noise (AWGN) received by RRH $i$ at time $t$. The matrix $\mathbf{H}_i$ of size $Q_i \times N$ is

$$\mathbf{H}_i = [\mathbf{h}_{i,1}, \mathbf{h}_{i,2}, \cdots, \mathbf{h}_{i,N}]. \tag{2}$$

As shown in (1), the dimension of the baseband signal $\mathbf{y}_{i,t}$ increases as more antennas are used at each RRH. The baseband signals transmitted directly through the front-haul links will suffer from severe quantization noise. Let the capacity of the front-hual link be $C_i$ bits per second, and the wireless bandwidth be $J$ Hz. We analyze the quantization noise under uniform quantization. The average number of bits allocated to an individual value in $\mathbf{y}_{i,t}$ is

$$B_i = \frac{C_i}{J \cdot Q_i}. \tag{3}$$

Typically, the distribution $\mathbf{y}_{i,t}$ is close to complex Gaussian distribution since each component is a summation of signal transmitted from multiple TDs. Thus, the quantization noise $\mathbf{q}_i[k]$ can be approximated as a complex random variable whose real and imaginary parts are uniformly distributed in the range $(-\frac{L_i}{2}, \frac{L_i}{2})$ where $L_i = \frac{2K_i}{2^{b_i}-1}$ is the quantization level [24] of the baseband signal of $i$-th RRH, $b_i = B_i/2$ is the number of bits used to represent the real/imaginary part of $\mathbf{y}_{i,t}[k]$, and $[-K_i, K_i]$ is the dynamic range of the real/imaginary part of $\mathbf{y}_{i,t}[k]$.

By [24], the power of the quantization noise can be written as

$$E[\|q_i[k]\|^2] = \frac{(L_i)^2}{12} + \frac{(L_i)^2}{12} = \frac{(L_i)^2}{6}. \tag{4}$$

It can be seen from (3) and (4) that with limited front-haul capacity, the number of bits allocated to each individual value of $\mathbf{y}_{i,t}[k]$ decreases as $Q_i$ increases, and in consequence the quantization noise power increases. In other words, under direct quantization, the limited front-haul capacity is the main obstacle in transmitting the $\mathbf{y}_{i,t}$'s to the BBUs. In the next section, we will introduce the compressive quantization scheme which is able to quantize $\mathbf{y}_{i,t}$'s more efficiently so that the signal fits the front-haul capacity.

## III. Compressive Quantization and Symbol Detection

To tackle the aforementioned quantization noise problem, in this section, we propose the distributed baseband signal compressive quantization as well as the corresponding symbol detection schemes. We will first introduce the compressive quantization which happens in the RRH, and then the corresponding weight vectors that can be used to detect the symbols from the compressive quantized baseband signal. Finally, we propose a parallel interference cancellation algorithm which further improves the symbol detection accuracy.

### A. Compressive Quantization

The frame based compressive quantization scheme consists of the delay-and-add phase and the quantization phase. Let $F$ denote the frame length. In the delay-and-add phase, the $\mathbf{y}_{i,t}$'s obtained in $t = 1, 2, \cdots, F$ are buffered. As shown in Fig. 2, instead of quantizing and transmitting $\mathbf{y}_{i,t}$ to the BBUs at each time instant, each $\mathbf{y}_{i,t}$ is delayed for $D_i$ time instants from the previous one. The vector $\mathbf{z}_i$ is the summation of the delayed versions of $\mathbf{y}_{i,t}$'s. More specifically, $\mathbf{z}_i$ can be written as

$$\mathbf{z}_i[k] = \sum_{t=1}^{F} \tilde{\mathbf{y}}_{i,t}[k] \tag{5}$$

where

$$\tilde{\mathbf{y}}_{i,t}[k] = \begin{cases} \mathbf{y}_{i,t}[k-(t-1)\cdot D_i], & \text{if } (t-1)\cdot D_i + 1 \\ & \leq k \leq (t-1)\cdot D_i + Q_i \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

To model the proposed compressive quantization scheme, we introduce the equivalent channel $\mathbf{H}_i^{(eq)}$ as

$$\mathbf{H}_i^{(eq)} = \left[\tilde{\mathbf{H}}_i^{(1)}, \tilde{\mathbf{H}}_i^{(2)}, \cdots, \tilde{\mathbf{H}}_i^{(F)}\right] \tag{7}$$

where

$$\tilde{\mathbf{H}}_i^{(r)} = \begin{pmatrix} \mathbf{0}_{(r-1)D_i \times N} \\ \mathbf{H}_i \\ \mathbf{0}_{(F-r)D_i \times N} \end{pmatrix}. \tag{8}$$

The overlapped baseband signal $\mathbf{z}_i$ can be written as

$$\mathbf{z}_i = \mathbf{H}_i^{(eq)} \mathbf{x}^{(eq)} + \mathbf{n}_i^{(eq)} \tag{9}$$

where $\mathbf{x}^{(eq)} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_F^T]^T$ and $\mathbf{n}_i^{(eq)} = \mathbf{A} \cdot [\mathbf{n}_{i,1}^T, \mathbf{n}_{i,2}^T, \cdots, \mathbf{n}_{i,F}^T]^T$. The noise transformation matrix $\mathbf{A} \in \mathbb{R}^{(Q_i+(F-1)D_i)\times(Q_iF)}$ can be written as

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_F] \tag{10}$$

and

$$\mathbf{a}_k = \begin{pmatrix} \mathbf{0}_{(k-1)D_i \times Q_i} \\ \mathbf{I}_{Q_i \times Q_i} \\ \mathbf{0}_{(F-k)D_i \times Q_i} \end{pmatrix}. \tag{11}$$

The quantization phase follows the delay-and-add phase. Uniform quantization is applied to each element in $\mathbf{z}_i$. More specifically, the quantized signal $\tilde{\mathbf{z}}_i$ can be represented as

$$\begin{aligned} \tilde{\mathbf{z}}_i &= \mathbf{z}_i + \mathbf{q}_i^{(CQ)} \\ &= \mathbf{H}_i^{(eq)} \mathbf{x}^{(eq)} + \mathbf{n}_i^{(eq)} + \mathbf{q}_i^{(CQ)} \end{aligned} \tag{12}$$

where $\mathbf{q}_i^{(CQ)}$ is the quantization noise caused by compressive quantization. In the following, we will show the quantization noise is reduced by the proposed method.

The total length of vector $\mathbf{z}_i$ containing the information of $\mathbf{y}_{i,t}$'s is $Q_i + (F-1) \cdot D_i$. Compared with the aggregation of $\mathbf{y}_{i,t}$'s of the length $Q_i \cdot F$, the dimension of the vector is much reduced. For instance, as in the example shown in Fig. 2 where $Q_i = 8$ and $D_i = 2$, the number of components is reduced to nearly $\frac{1}{4}$ of the original value. As a result, with limited front-haul capacity, more bits can be allocated to each individual component. More specifically, similar to (3), we have

$$B_i^{(CQ)} = \frac{F \cdot C_i}{J \cdot (Q_i + (F-1) \cdot D_i)}. \tag{13}$$

By (5), the distribution of $\mathbf{z}_i$ is close to complex Gaussian distribution as the elements in $\mathbf{y}_{i,t}$ are approximated as complex Gaussian random variables. Similar to (4), the quantization noise power under the compressive quantization can be written as

$$E[\|\mathbf{q}_i^{(CQ)}[k]\|^2] = \frac{(L_i^{(CQ)})^2}{12} + \frac{(L_i^{(CQ)})^2}{12} = \frac{(L_i^{(CQ)})^2}{6} \tag{14}$$

where

$$L_i^{(CQ)} = \frac{2K_i^{(CQ)}}{2^{B_i^{(CQ)}}} \tag{15}$$

is the quantization level under the compressive quantization and $K_i^{(CQ)}$ is the dynamic range of $\mathbf{z}_i[k]$.

In fact, $\mathbf{z}_i[k]$ is formed by the summation of $\frac{Q_i}{D_i}$ independent components. While $K_i^{(CQ)}$ increases linearly with $\frac{1}{D_i}$, $B_i^{(CQ)}$ also increases linearly with $\frac{1}{D_i}$ by (13). In other words, the numerator of (15) increases linearly while the denominator increases exponentially. As a result, the $L_i^{(CQ)}$ decreases with $\frac{1}{D_i}$ and the quantization noise power is reduced as well by (14).

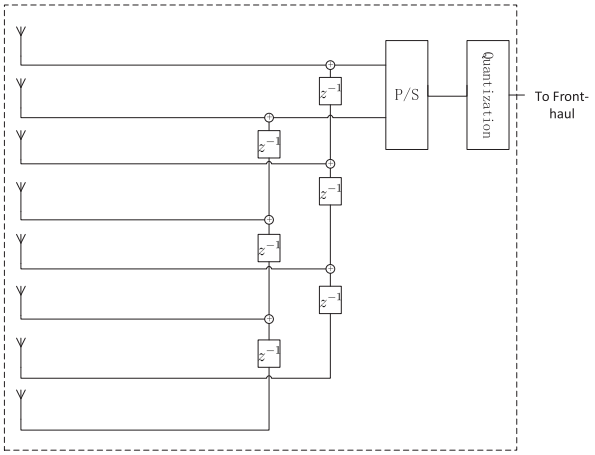



Fig. 3. An example of the DSP circuit for compressive quantization with $Q_i = 8$, $D_i = 2$.

In summary, in the proposed compressive quantization scheme, the baseband signals from multiple time instants are first reduced into a lower dimension vector using the delay-and-add and then quantized. Because of the reduction in dimension, more bits can be allocated to each individual value in quantization, and the quantization noise power is reduced. On the other hand, due to the reduction of dimension, the independent baseband signals of multiple time instants are overlapped together, which creates some additional interference and makes the detection of the symbols more difficult. The parameter $D_i$ in the proposed scheme can be tuned to balance the quantization noise power and the interference power. Since the length of $\mathbf{z}_i$ is $Q_i + (F-1) \cdot D_i$, with smaller $D_i$, the raw baseband signals are densely overlapped together, which creates much interference while significantly reduces the quantization noise power. With larger $D_i$, the reduction in quantization noise power is not so aggressive, while the extra interference created by the compressive quantization is also mild. In next subsection, we will propose weight vector designs to detect the symbols from the compressive quantized baseband signals.

Before we introduce the weight vectors that are used in the BBUs to detect the symbols, it is worth analyzing the complexity of the compressive quantization in the RRHs. By (5), the delay-and-add does not require the channel information, which reduces the communication overhead between the BBUs and the RRHs since the channel information is obtained in the BBUs. This is a desirable feature in the C-RAN since the front-haul link capacity is usually the bottleneck limiting the performance of the system. Moreover, as an example shown in Fig. 3, the processing in the RRHs only involves buffering and adding without multiplying, which can be easily implemented using low cost hardware. It enables the RRH to compress the baseband signal from multiple antennas without complicated processing, which preserves the advantages of the C-RAN system that the RRHs are cheap and can be deployed with low cost.

### B. The Weight Vector Design

As introduced in last subsection, each RRH processes the baseband signal using compressive quantization and transmits them to the BBUs via the front-haul links with limited capacity. The BBUs collect the baseband signals transmitted by each RRH and detect the intended symbol of each TD. In this subsection, two weight vector designs are proposed to detect the intended symbols from the compressive quantized baseband signals.

The aggregate of the baseband signal from all the RRHs collected at the BBUs can be written as

$$\begin{aligned}\tilde{\mathbf{z}} &= [\tilde{\mathbf{z}}_1^T, \tilde{\mathbf{z}}_2^T, \cdots, \tilde{\mathbf{z}}_M^T]^T \\ &= \mathbf{H}^{(eq)}\mathbf{x}^{(eq)} + \mathbf{n}^{(eq)} + \mathbf{q}^{(CQ)}\end{aligned} \tag{16}$$

where $\mathbf{H}^{(eq)} = [(\mathbf{H}_1^{(eq)})^T, (\mathbf{H}_2^{(eq)})^T, \cdots, (\mathbf{H}_M^{(eq)})^T]^T$, $\mathbf{n}^{(eq)} = [(\mathbf{n}_1^{(eq)})^T, (\mathbf{n}_2^{(eq)})^T, \cdots, (\mathbf{n}_M^{(eq)})^T]^T$, $\mathbf{q}^{(CQ)} = [(\mathbf{q}_1^{(CQ)})^T, (\mathbf{q}_2^{(CQ)})^T, \cdots, (\mathbf{q}_M^{(CQ)})^T]^T$.

Without loss of generality, in the rest of the paper, we assume that $D_i = D$ for $i \in \{1, 2, \cdots, N\}$, while the analysis can be easily extended to the case that the RRHs use different $D_i$'s. The BBUs use weight vectors to combine the elements of $\mathbf{z}$ so as to optimally detect the transmitted symbols of all the TDs. In this work, we first determine the optimal weight vectors $\mathbf{W}_{LMMSE}$ to minimize the mean square error (MSE) of the detection. The problem can be formulated as

$$\min_{\mathbf{W}} \sum_{k=1}^{N} E[\|\hat{\mathbf{x}}^{(eq)} - \mathbf{x}^{(eq)}\|^2] \tag{17}$$

where $\hat{\mathbf{x}}^{(eq)} = \mathbf{W}\tilde{\mathbf{z}}$. The solution to (17) can be written as [25]

$$\begin{aligned}&\mathbf{W}^{(LMMSE)} \\ &= [\mathbf{\Sigma}_x - \mathbf{\Sigma}_x(\mathbf{H}^{(eq)})'\mathbf{\Sigma}_e^{-1}\mathbf{H}^{(eq)}(\mathbf{\Sigma}_x^{-1} + (\mathbf{H}^{(eq)})'\mathbf{\Sigma}_e^{-1}\mathbf{H}^{(eq)})^{-1}] \\ &\quad \cdot (\mathbf{H}^{(eq)})'\mathbf{\Sigma}_e^{-1}\end{aligned} \tag{18}$$

where $\mathbf{\Sigma}_x = E[\mathbf{x}^{(eq)} \cdot (\mathbf{x}^{(eq)})']$, $\mathbf{\Sigma}_e = E[(\mathbf{n}^{(eq)} + \mathbf{q}^{(CQ)}) \cdot (\mathbf{n}^{(eq)} + \mathbf{q}^{(CQ)})']$. Since $\mathbf{n}^{(eq)}$ and $\mathbf{q}^{(CQ)}$ are independent, $\mathbf{\Sigma}_e = E[(\mathbf{n}^{(eq)}) \cdot (\mathbf{n}^{(eq)})'] + E[\mathbf{q}^{(CQ)} \cdot (\mathbf{q}^{(CQ)})']$.

The error covariance matrix for estimating $\mathbf{x}^{(eq)}$ using $\mathbf{W}^{(LMMSE)}$ can be written as

$$\mathbf{\Sigma}_{\mathbf{d}} = \mathbf{\Sigma}_x - \mathbf{\Sigma}_x(\mathbf{H}^{(eq)})'(\mathbf{H}^{(eq)}\mathbf{\Sigma}_{\mathbf{x}}(\mathbf{H}^{(eq)})' + \mathbf{\Sigma}_e)^{-1}\mathbf{H}^{(eq)}\mathbf{\Sigma}_x. \tag{19}$$

Recall that $\mathbf{z}_i[k]$ is a summation of $\frac{Q_i}{D_i}$ independent random variables from $\mathbf{y}_{i,t}$. Each component in $\mathbf{y}_{i,t}$ approximates Gaussian distribution due to it's the summation of signals transmitted from $N$ TDs through independent channels. If both $\frac{Q_i}{D_i}$ and $N$ are not big enough, the quantization noise power obtained in (14) which relies on the Gaussian assumption is less accurate. The $\mathbf{W}^{(LMMSE)}$ obtained in (18) also loses part of the optimality.

By (18), the calculation of $\mathbf{W}^{(LMMSE)}$ becomes difficult as the number of antennas increases. Under this condition, similar to that in the massive MIMO, some simple weight vector design such as maximum-radio combining (MRC) is effective. The MRC weight vector design is

$$\mathbf{W}^{(MRC)} = \mathbf{\Lambda} \cdot (\mathbf{H}^{(eq)})' \tag{20}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{\Lambda}_{i,i} = \frac{1}{(\mathbf{h}_i^{(eq)})' \cdot \mathbf{h}_i^{(eq)}}$ and $\mathbf{h}_i^{(eq)}$ is the $i$-th column of $\mathbf{H}^{(eq)}$.

To gain a better understanding of the performance of the symbol detection, let $\mathbf{x}[s]$ denote the symbol transmitted by the TD $i$ at time $j$, i.e., $s = (j-1) * F + i$. The $\hat{\mathbf{x}}^{(eq)}[s]$ can be written as

$$\begin{aligned}\hat{\mathbf{x}}^{(eq)}[s] &= \mathbf{w}_s^H\tilde{\mathbf{z}} \\ &= \mathbf{w}_s^H\left(\mathbf{H}^{(eq)}\mathbf{x}^{(eq)} + \mathbf{n}^{(eq)} + \mathbf{q}\right) \\ &= \mathbf{w}_s^H\mathbf{h}_s^{(eq)}\mathbf{x}^{(eq)}[s] + \mathbf{w}_s^H\sum_{\substack{k=1 \\ k\neq s}}^{N*F}\mathbf{h}_k^{(eq)}\mathbf{x}^{(eq)}[k] \\ &\quad + \mathbf{w}_s^H\left(\mathbf{n}^{(eq)} + \mathbf{q}\right) \\ &= \mathbf{w}_s^H\mathbf{h}_s^{(eq)}\mathbf{x}^{(eq)}[s] \\ &\quad + \mathbf{w}_s^H\left(\sum_{k=1}^{(i-1)*F}\mathbf{h}_k^{(eq)}\mathbf{x}^{(eq)}[k]\right. \\ &\quad \left. + \sum_{k=i*F+1}^{N*F}\mathbf{h}_k^{(eq)}\mathbf{x}^{(eq)}[k]\right) \\ &\quad + \mathbf{w}_s^H\sum_{\substack{k=(i-1)F+1 \\ k\neq s}}^{i*F}\mathbf{h}_k^{(eq)}\mathbf{x}^{(eq)}[k] + \mathbf{w}_s^H\left(\mathbf{n}^{(eq)} + \mathbf{q}\right)\end{aligned} \tag{21}$$

where $\mathbf{w}_s^H$ is the $s$-th row of $\mathbf{W}$, which can be either $\mathbf{W}^{(LMMSE)}$ or $\mathbf{W}^{(MRC)}$. The first term in the last equality of (21) is the intended signal, the second term is inter-transmission interference (ITI) that caused by the transmission of other time instants, the third term is the inter-user interference (IUI) that caused by the symbols transmitted by other TDs at the same time instant, and the last term is noise which is the sum of the received AWGN and the quantization noise.

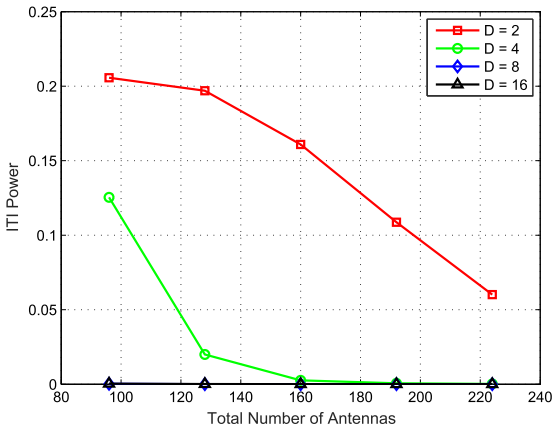
The signal power can be written as

$$\begin{aligned}P_{sig} &= E_X[\|\mathbf{w}_s^H\mathbf{h}_s^{(eq)}\mathbf{x}^{(eq)}[s]\|^2] \\ &= \alpha_s\|\mathbf{w}_s^H\mathbf{h}_s^{(eq)}\|^2\end{aligned} \tag{22}$$

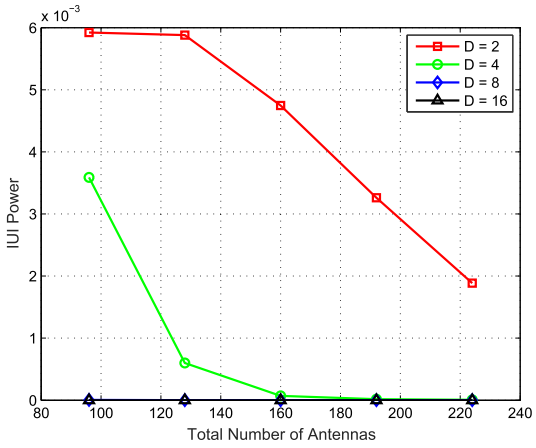
where $\alpha_s = E[\|\mathbf{x}^{(eq)}[s]\|^2]$ is the power of $\mathbf{x}^{(eq)}[s]$. In this work, we assume that the transmitting power of all the symbols $\alpha_s$'s are predetermined before the transmission and the information is available in the BBUs.

Accordingly, the power of the ITI and the IUI can be written as

$$\begin{aligned}&P_{iti} \\ &= E_X[\left\|\mathbf{w}_s^H\left(\sum_{k=1}^{(i-1)*F}\mathbf{h}_k^{(eq)}\mathbf{x}^{(eq)}[k] + \sum_{k=i*F+1}^{N*F}\mathbf{h}_k^{(eq)}\mathbf{x}^{(eq)}[k]\right)\right\|^2] \\ &= \sum_{k=1}^{(i-1)*F}\alpha_k\|\mathbf{w}_s^H\mathbf{h}_k^{(eq)}\|^2 + \sum_{k=i*F+1}^{N*F}\alpha_k\|\mathbf{w}_s^H\mathbf{h}_k^{(eq)}\|^2\end{aligned} \tag{23}$$




(a) The ITI power




(b) The IUI power

Fig. 4. The ITI and IUI power decrease as more antennas are available.

and

$$P_{iui} = E_X[\left\| \mathbf{w}_s^H \sum_{\substack{k=(i-1)F+1 \\ k\neq s}}^{i*F} \mathbf{h}_k^{(eq)} \mathbf{x}^{(eq)}[k] \right\|^2] = \sum_{\substack{k=(i-1)F+1 \\ k\neq s}}^{i*F} \alpha_k \|\mathbf{w}_s^H \mathbf{h}_k^{(eq)}\|^2. \quad (24)$$

The power of the noise can be written as

$$P_{noise} = E[\|\mathbf{w}_s^H(\mathbf{n}^{(eq)} + \mathbf{q})\|^2] = \mathbf{w}_s^H \sigma_e \mathbf{w}_s. \quad (25)$$

By combining (22) to (25), the SINR of the symbol $\mathbf{x}^{(eq)}[s]$ can be written as

$$SINR = \frac{P_{sig}}{P_{iti} + P_{iui} + P_{noise}}. \quad (26)$$

From the analysis above, compared with the signal detection using $\mathbf{y}_i$'s, we can see extra interference ITI in (23) is created by overlapping the baseband signal of multiple time instants, which makes the detection of the symbol $\mathbf{x}^{(eq)}[s]$ more difficult than using the original baseband signal. On the other hand, by applying the compressive quantization, each individual value in $\mathbf{z}_i$ is allocated more bits, and in consequence the quantization noise power in (4) is reduced. Therefore, the compressive quantization scheme reduces the quantization noise power at the cost of introducing extra interference power.

In this work, since all the RRHs are equipped with multiple antennas, the total number of antennas can be quite large, and the C-RAN system is in essence a virtual massive MIMO. In a massive MIMO system, it had been shown that when more antennas are available, the channel information vectors of multiple users are near orthogonal to each other, and the interference power diminishes [2], [3]. This phenomenon is also observed in this work. As shown in Fig. 4a and Fig. 4b where $Q_i = 16$ for $i = 1, 2, \cdots, M$, we increase the total number of antennas by deploying more RRHs. Both the ITI power and IUI power diminishes for different $D$'s using the LMMSE weight vectors. Moreover, since $Q_i$ and $B_i^{(CQ)}$ does not change for each RRH, the quantization noise power does not increase according to (14) and (15). As a result, the SINR in (26) is improved. In other words, the proposed compressive quantization scheme transforms the quantization power into interference power, which can be tackled in the BBUs by the corresponding weight vectors, especially when there are massive antennas available.

### C. The Parallel Interference Cancellation

In last subsection, we introduce the weight vectors to detect the symbols transmitted by the TDs from the compressive quantized baseband signals. The weight vectors are constrained in the linear domain. In this subsection, we go one step beyond by using the parallel interference cancellation method to further mitigate the interference, which is in the non-linear domain.

Unlike the AWGN and the quantization noise, the interference terms in (23) and (24) have their own structures, which can be further explored to improve the accuracy of the detection. Since the channel information $\mathbf{H}^{(eq)}$ is available in the BBUs, the ITI and IUI can be reconstructed and isolated from the intended signal if the relevant transmitted symbols are available. More specifically, the symbols initially detected by using the weight vectors are used to generate the approximate interference terms in (23) and (24), which are then subtracted from the original baseband signal in (21). After that, the updated estimator is employed to detect each symbol. In the following, the steps for the parallel interference cancellation scheme will be introduced. Without loss of generality, we assume that transmitted symbols are modulated using QPSK, while the proposed algorithm can be easily extended to other modulations.

The initially detected symbol $\bar{\mathbf{x}}^{(eq)}[s]$ can be written as

$$\bar{\mathbf{x}}^{(eq)}[s] = \mathtt{sign}(Re(\hat{\mathbf{x}}^{(eq)}[s])) + \mathtt{sign}(Im(\hat{\mathbf{x}}^{(eq)}[s])) * j \quad (27)$$

where $\hat{\mathbf{x}}^{(eq)}$ is obtained using either LMMSE or MRC weight vectors.

Next, we generate the ITI and IUI cancellation vectors for symbol $\mathbf{x}[s]$.

$$ITI[s] = \left( \sum_{k=1}^{(i-1)*F} \mathbf{h}_k^{(eq)} \bar{\mathbf{x}}^{(eq)}[k] + \sum_{k=i*F+1}^{N*F} \mathbf{h}_k^{(eq)} \bar{\mathbf{x}}^{(eq)}[k] \right) \tag{28}$$

$$IUI[s] = \sum_{\substack{k=(i-1)F+1 \\ k \neq s}}^{i*F} \mathbf{h}_k^{(eq)} \bar{\mathbf{x}}^{(eq)}[k] \tag{29}$$

Since the ITI and IUI cancellation vectors only depend on the initial detected symbols $\bar{\mathbf{x}}^{(eq)}$, the calculation can be done in parallel for each symbol $\mathbf{x}[s]$. When the ITI and IUI cancellation vectors are available, they are subtracted from the orignal baseband signal, which can be written as

$$\tilde{\mathbf{z}}^{(IC)}[s] = \tilde{\mathbf{z}}[s] - ITI[s] - IUI[s]. \tag{30}$$

When $\tilde{\mathbf{z}}^{(IC)}$ is obtained, the ITI and IUI are eliminated from the original baseband signal[1] $\tilde{\mathbf{z}}$, and we can use the simple MRC weight vector [26] to update the corresponding transmitted symbols. The updated estimated symbol $\hat{\mathbf{x}}^{(eq,IC)}[s]$ is calculated as

$$\hat{\mathbf{x}}^{(eq,IC)}[s] = \mathbf{m}_s' \tilde{\mathbf{z}}^{(IC)} \tag{31}$$

where $\mathbf{m}_s = \mathbf{h}_s^{(eq)}$ is the weight vector for symbol $\mathbf{x}^{(eq)}[s]$. The detected symbols are updated using $\hat{\mathbf{x}}^{(eq,IC)}$. More specifically,

$$\begin{aligned} &\bar{\mathbf{x}}^{(eq,IC)}[s] \\ &\quad = \mathtt{sign}(Re(\hat{\mathbf{x}}^{(eq,IC)}[s])) + \mathtt{sign}(Im(\hat{\mathbf{x}}^{(eq,IC)}[s])) * j. \end{aligned} \tag{32}$$

The steps for the parallel interference cancellation based symbol updating are summarized in Algorithm 1.

**Algorithm 1** Parallel Interference Cancellation Based Symbol Updating

1 Make the initial decision using (27)
2 Generate the approximate IUI and ISI terms using (28) and (29)
3 Subtract the IUI and ISI terms from the original baseband signal using (30)
4 Update the symbols using (31) and (32)

With this algorithm, we can further leverage the advantages of the C-RAN architecture to improve the detection accuracy of each symbol: the BBUs are empowered by high performance computing, and holds the necessary information that enables us to perform the algorithm.

Looking at the overall compressive quantization and detection process, we can see it is tailored for the C-RAN architecture. As illustrated in Fig. 3, the compressive quantization happens at the RRH side, which is as simple as possible so that the advantage of low deployment cost of the C-RAN architecture is preserved. The receiver design and the parallel interference happen in the BBUs where high performance computing power and the necessary information are readily available. The time complexity of the computing is dominated by the design of LMMSE receiver, which is $O(N^3F^3)$. The space complexity is dominated by the memory required to store the $\mathbf{H}^{(eq)}$, which is $O(N * F * \sum_{i=1}^{M} Q_i)$.

When compared with the spatial filter method proposed in [22], the compressive quantization has advantage in terms of complexity. The RRH side uses the overlap and sum method, which is $O(n)$ complexity compared with the the spatial filter where it's $O(n^2)$ due to matrix multiplication. Moreover, since the filter design is usually done in the BBU side, the spatial filter causes additional overhead in the front-haul links. In the BBUs, the complexity of both schemes are $O(n^3)$ due to the matrix inversion in calculating the MMSE or LMMSE estimators.

## IV. Application to the OFDM Based C-RAN

In section III, we have proposed the compressive quantization algorithm which applies to the single tap channel. Due to the rapid growth of demand for wireless communications, larger bandwidth are being utilized. Typically, with larger bandwidth, the traffic load in the front-haul link will increase due to higher sampling rate. In this section, we apply the proposed algorithm to the uplink of OFDM based C-RAN, which significantly reduces the traffic load in the front-haul links caused by increasing the bandwidth.

We consider the uplink of OFDM based multi-antenna C-RAN where the TDs use OFDMA [27] to access the wireless medium. Assume that all the TDs are equipped with a single antenna while RRH $i$ is equipped with $Q_i$ antennas where $Q_i > 1$. All the RRHs work in the same spectrum. Each TD maps the symbols to the allocated subcarriers before adding the cyclic prefix and transmitting them. Without loss of generality, we assume that the transmitted signal is received by all the RRHs.

Note that the OFDM based RRHs setting does not make any assumption on the subcarrier allocations among TDs. Therefore, it can work for both the orthogonal subcarrier allocations such as OFDMA in 4G wireless communications, as well as the non-orthogonal multiple access (NOMA) [28] in 5G communications.

At the RRH side, for the baseband signal received by each antenna, the cyclic prefix is removed and the fast Fourier transform (FFT) is periodically performed. We define a cycle as the time between two consecutive FFTs. As shown in Fig. 5, after the FFT, the baseband signal of the same subcarrier from all the antennas are collected in $\mathbf{v}_{i,c,t} = [\mathbf{v}_{i,c,t}[1], \mathbf{v}_{i,c,t}[2], \cdots, \mathbf{v}_{i,c,t}[Q_i]]^T$ where $\mathbf{v}_{i,c,t}[k]$ denotes the component in the subcarrier $c$ of the $k$-th antenna of RRH $i$ at the $t$-th cycle.

By the special property of OFDM in transforming multipath channel into a number of equivalent interference-free

[1]There are residue IUI and ISI power due to the errors in determining $\bar{\mathbf{x}}^{(eq)}$, which can be ignored if the original BER is not too high.

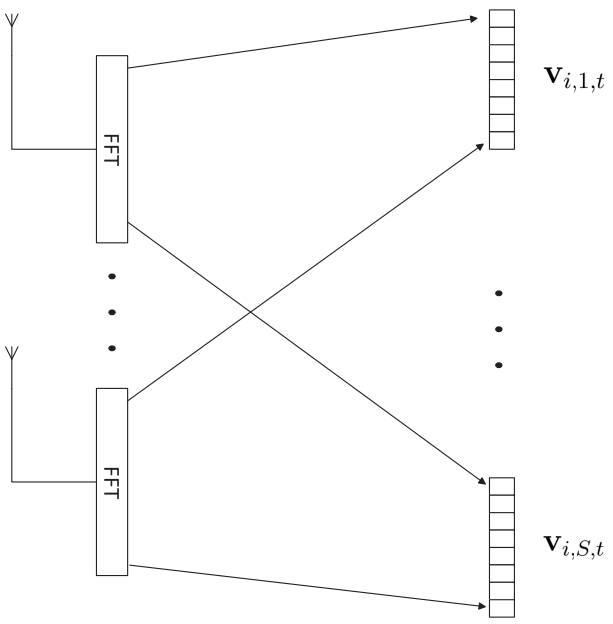



Fig. 5. An example of re-organizing the baseband signal with $Q_i = 8$.

subchannels [29], $\mathbf{v}_{i,c,t}[k]$ can be written as

$$\mathbf{v}_{i,c,t} = \begin{bmatrix} \mathbf{g}_{i,c,t}[1] \\ \mathbf{g}_{i,c,t}[2] \\ \vdots \\ \mathbf{g}_{i,c,t}[Q_i] \end{bmatrix} x_c + \mathbf{n}_{i,c,t} \triangleq \mathbf{g}_{i,c,t} x_c + \mathbf{n}_{i,c,t} \qquad (33)$$

where $x_c$ is the value transmitted in the subcarrier $c$, $\mathbf{g}_{i,c}[k]$ is the equivalent channel impulse response on subcarrier $c$ between the $k$-th antenna of RRH $i$ and the TD which the subcarrier is allocated to at cycle $t$. $\mathbf{n}_{i,c,t}$ is the noise at RRH $i$ in subcarrier $c$ at cycle $t$. It is shown that an equivalent channel is established in the subcarrier $c$.

As the system utilizes larger bandwidth, there are more subcarriers and thus more $\mathbf{v}_{i,c,t}$'s are obtained at each cycle. Directly transmitting them through the limited capacity front-haul to the BBUs becomes difficult. To tackle this problem, we apply the proposed compressive quantization and detection scheme to the system. Instead of transmitting the baseband signal $\mathbf{v}_{i,c,t}$'s separately, we partition them into multiple groups and apply the compressive quantization to the baseband signals within each group. More specifically, let $\beta_r$ denote the set of subcarriers in group $r$ and the cardinality of $\beta_r$ is $N_r$, then we can calculate the baseband signal $\mathbf{u}_{i,t}^r$ for the group $r$ of the $i$-th RRH at cycle $t$, which can be written as

$$\begin{aligned} \mathbf{u}_{i,t}^r &= \sum_{c\in\beta_r} \mathbf{v}_{i,c,t} \\ &= \sum_{c\in\beta_r} \mathbf{g}_{i,c,t} x_c + \sum_{c\in\beta_r} \mathbf{n}_{i,c,t} \\ &= [\mathbf{g}_{i,1,t}\mathbf{g}_{i,2,t}\cdots\mathbf{g}_{i,N_r,t}] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N_r} \end{bmatrix} + \sum_{c\in\beta_r} \mathbf{n}_{i,c,t} \\ &\triangleq \mathbf{G}_{i,t}\mathbf{x}_t + \mathbf{n}_{i,t}. \end{aligned} \qquad (34)$$

Since (34) is equivalent to (1), the proposed compressive quantization scheme can be applied to each group, and all the subcarrier components can be recovered at the BBUs by performing the weight vectors and the parallel interference cancellation schemes for each group.

By applying the compressive quantization, the baseband signals are reduced in both frequency and time domain. Compared with directly transmitting the baseband signals through the front-haul links, the baseband signals corresponding to different subcarriers do not need to be transmitted separately. As a result, the traffic load in the front-haul links caused by adopting larger bandwidth is less severe.

To illustrate this effect, we analyze the front-haul traffic load for the proposed method. For RRH $i$, since the baseband signal for each group is transmitted separately, we just analyze the baseband signal for one group. Assume the bandwidth utilized in the system is $J$ Hz, which is corresponding to $S$ subcarriers. Then the rate of each baseband component $\mathbf{v}_{i,c,t}^r$ is $\frac{J}{S}$ Hz, i.e., the separation between $\mathbf{u}_{i,t}^r$ and $\mathbf{u}_{i,t+1}^r$ is $\frac{S}{J}$ seconds. Assume $U_i$ bits are used for each individual value of the signal reduced using delay-and-add, the group $r$ of RRH $i$ requires $(Q_i + (F-1)D_i)U_i$ bits in $\frac{SF}{J}$ seconds. Therefore, the traffic load in the front-link is

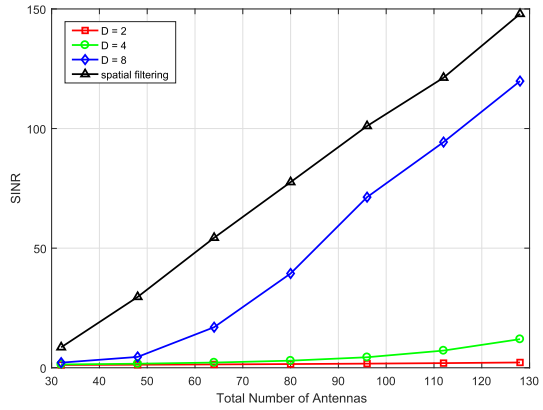
$$\begin{aligned} C_i &= \frac{(Q_i + (F-1)D_i)U_i}{\frac{SF}{J}} \\ &= \frac{(Q_i + (F-1)D_i)U_i}{F} \cdot \frac{J}{S} \\ &\approx D_i U_i \cdot \frac{J}{S} \end{aligned} \qquad (35)$$

where the approximation is due to the fact that the frame length $F$ is always much larger than $Q_i$.

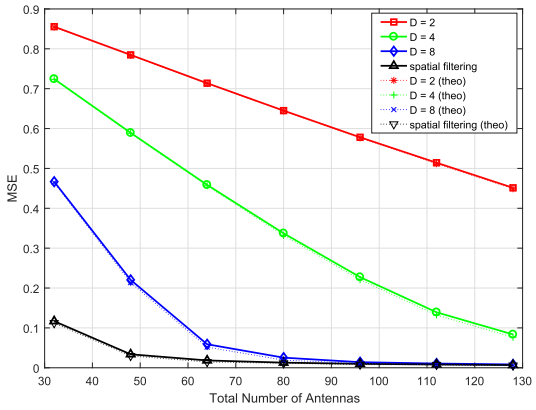
Typically, the number of subcarriers $S$ increases with the utilized bandwidth $J$ and $\frac{J}{S}$ keeps constant. Therefore, if $D_i$ keeps constant, the traffic load in the front-haul link only depends on $U_i$. If $U_i$ also keeps constant, the traffic load in the front-haul link keeps constant independent of the bandwidth $J$ adopted. In fact, when larger bandwidth is used, there are more subcarriers in each group. As a result, more raw baseband signal components are added together and the dynamic range increases. Larger $U_i$ is required for each value and the traffic load in the front-haul link grows accordingly. However, as will be shown in section V, $U_i$ only increases slightly as $J$ increase to maintain good performance. Therefore, the proposed compressive quantization scheme facilitates the usage of larger bandwidth in the OFDM based multi-antenna C-RAN system.

## V. Numerical Results

In this section, we use numerical results to illustrate the effectiveness of the proposed scheme. We first evaluate the performance of the proposed compressive quantization scheme under the single tap channel model. It is shown that the proposed scheme is better than existing schemes in tackling the front-haul link capacity deficit. The performance gap between LMMSE and MRC weight vectors are also investigated. It is shown that the gap decreases with more antennas available or by performing the proposed parallel interference cancellation scheme. We further evaluate the performance




(a) Average SINR




(b) Average MSE

Fig. 6. The Average SINR and MSE of C = 1.2 Gb/s.

of the proposed scheme to the OFDM based C-RAN under multipath channel. It is shown that the proposed scheme can facilitate the C-RAN utilize larger bandwidth with limited front-haul capacity.

### *A. Performance of The Proposed Scheme Under Single-Tap Channel*

Suppose there are $N$ TDs and $M$ RRHs distributed in an area where each TD is equipped with a single antenna, while each RRH is equipped with multiple antennas and has limited front-haul capacity. All the TDs and RRHs work in the same spectrum. All the TDs simultaneously transmit the signal through the wireless medium to all the RRHs. Each RRH receives a mixture of the signal transmitted by all the TDs. Assume there is no large scale fading. The baseband signal is processed at each RRH using the compressive quantization scheme proposed in this paper before it is transmitted through the front-haul links to the BBUs where the LMMSE weight vector proposed in (18) is applied to detect the symbols transmitted by all the TDs. In this experiment, all the antennas at all RRHs are actively receiving signals from all TDs at all time instants. There is no assumption on sparsity of the received baseband signals.

In the first experiment, the bandwidth is 10 MHz, and the signal to noise ratio (SNR) is 10 db,[2] the number of TDs is $N = 30$. We fix $Q_i = 16$ and vary the number of RRHs so as to change the total number of antennas $T$. We run the experiment over multiple channel realizations where the $h_{i,j}^{(k)}$'s are drawn from independent complex Gaussian distributions and calculate the average SINR and MSE. In Fig. 6 through 8, we show the average SINR and MSE of each single TD using LMMSE weight vectors under various per front-haul link capacities.

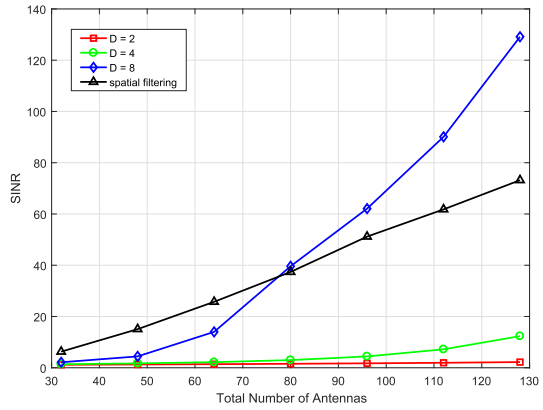
We use $D_i = D, \forall i$ in this experiment and change the parameter $D$ so as to adjust the interference power and the quantization noise. It is shown that for each of the $D$, the SINR of each TD improves as more antennas are available where the interference power is mitigated. Moreover, the theoretical values of MSE in (19) match well with the experimental values.

We continue to look at Fig. 9 for a special experiment. In this experiment, the SNR is 15 db while all the other parameters are the same as the experiments shown in Fig. 6 through 8. We can observe from Fig. 9 that the curve $D = 4$ starts to outperform the $D = 8$ curve when the front-haul link capacity is $0.4\ Gb/s$, while the $D = 8$ setting works better when the front-haul link capacity is relatively higher. It is because that the $D = 4$ setting reduces the baseband signal into less values than $D = 8$, and thus each individual value can use more bits. On the other hand, the interference becomes more severe at $D = 4$. Therefore, when the front-haul link capacity is low, smaller $D$ is beneficial in suppressing the quantization noise power. When the front-haul link capacity is relatively high, larger $D$ is beneficial in creating less interference power. In other words, the proposed scheme provides the flexibility in the tradeoff between the quantization noise and interference. The spatial filtering method proposed in [22] is an example where no such flexibility is provided since the baseband signal of multiple time instants are transmitted separately. It is shown in Fig. 6 through 9 that it does not perform well when the front-haul link capacity is low.

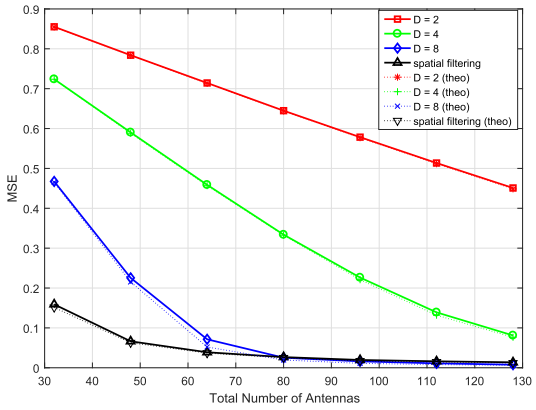
We further apply the interference cancellation method proposed in section III-C to the multi-RRH uplink system. All the 30 TDs use QPSK as the modulation. As shown in Fig. 10a through 10c, the interference cancellation scheme improves the BER when the original BER is not too high. Moreover, it is observed that the improvement of BER is more significant in the compressive quantization scheme than in the spatial filtering scheme in [22]. The reason is that the compressive quantization scheme transforms the quantization noise power into interference power. Unlike the quantization noise, the interference has its own structure and can be further suppressed by the interference cancellation scheme.

It can be seen from (18) that as the total number of antennas increases, the calculation of $\mathbf{W}^{(LMMSE)}$ becomes difficult.

[2]Typically, the AWGN is composed of the environmental noise, the device thermal noise and the quantization noise caused by the ADC. In this work, even with 8-bit commodity ADC, the quantization noise power is approximately $\frac{1}{10000}$ of the signal power. Therefore, the 10 db SNR is compatible with commodity ADC.

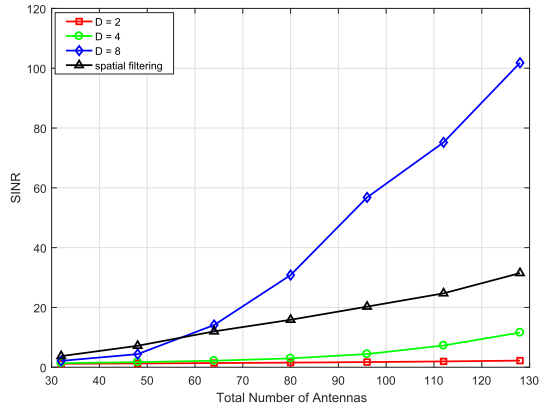



(a) Average SINR

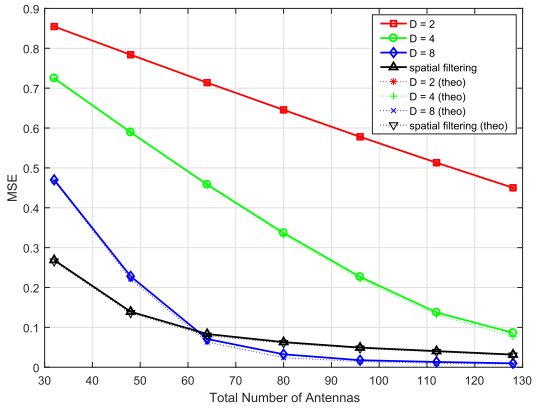



(b) Average MSE

Fig. 7. The Average SINR and MSE of C = 1.0 Gb/s.

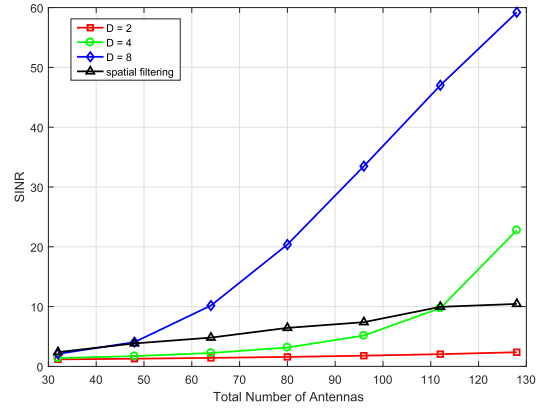



(a) Average SINR

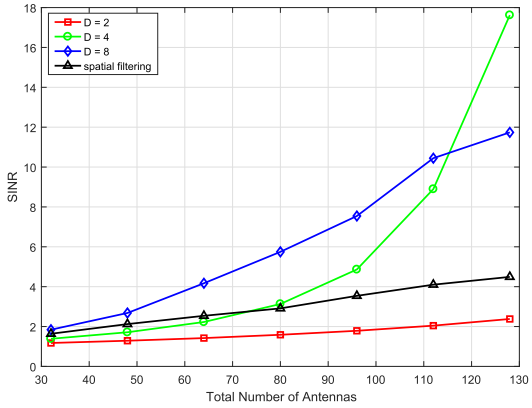



(b) Average MSE

Fig. 8. The Average SINR and MSE of C = 800 Mb/s.




(a) Average SINR for C = 600 Mb/s

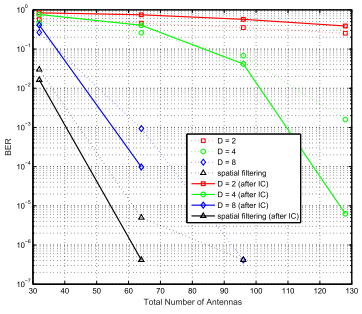



(b) Average SINR for C = 400 Mb/s

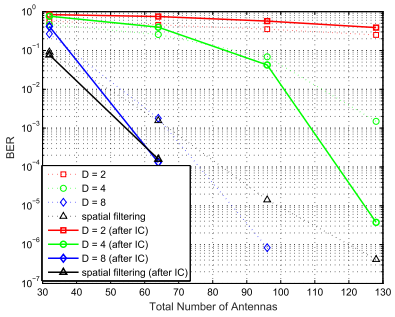
Fig. 9. Comparing the choice of $D$.

On the other hand, when there are significantly more antennas than users, the entire system becomes a virtual massive MIMO and the basic MRC weight vectors become effective. As the performance gap between the LMMSE and MRC weight vectors have been observed in [30] and [23], we compare the performance of the LMMSE and MRC weight vectors in the proposed system under the condition $Q_i = 16$ an $N = 10$. The front-haul capacity $C = 0.6\ Gb/s$ is used. The symbols are modulated using QPSK.

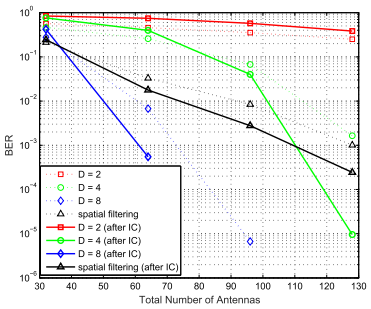
As shown in Fig. 11a and Fig. 11b, the performances of the LMMSE weight vectors and the MRC weight vectors get close as more antennas are available. In other words, with massive




(a) The BER performance of C = 1.0 Gb/s

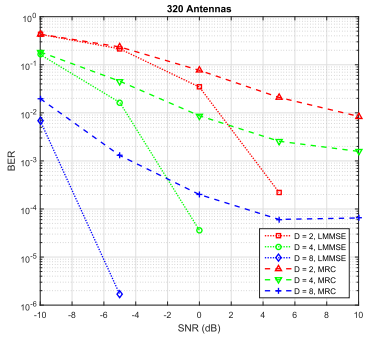



(b) The BER performance of C = 800 Mb/s

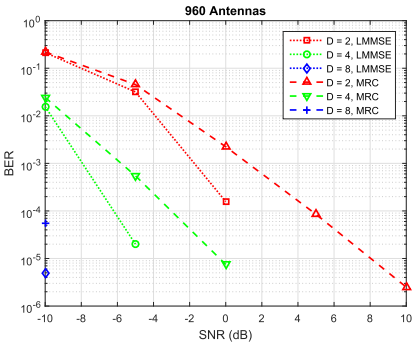



(c) The BER performance of C = 600 Mb/s

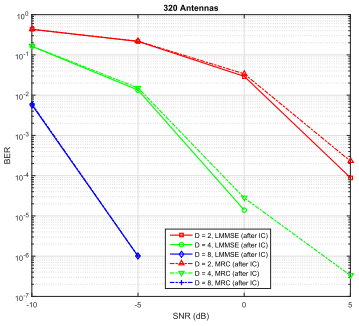
Fig. 10. Comparing BER performance with different front-haul capacity.




(a) The BER performance of 320 antennas without PIC




(b) The BER performance of 960 antennas without PIC




(c) The BER performance of 320 antennas with PIC

Fig. 11. Comparing BER performance with different number of antennas.

antennas, while calculating $\mathbf{W}^{(LMMSE)}$ becomes difficult, it can be replaced by $\mathbf{W}^{(MRC)}$ that is much easier to obtain.

Another way to bypass the difficulty in calculating $\mathbf{W}^{(LMMSE)}$ is to use the proposed parallel interference cancellation (PIC) scheme. By comparing Fig. 11a and Fig. 11c, the system performance is improved by performing PIC for both LMMSE and MRC weight vectors without changing the number of antennas. Moreover, the performance gap is reduced after the PIC. The reason is that the PIC cancels much of the interference which the MRC weight vectors are not able to deal with as well as the LMMSE weight vectors do.

### *B. Application to The OFDM Based Multi-Antenna C-RAN Under Multipath Channel*

Next, we apply the proposed scheme to the OFDM based broadband uplink system. In this experiment, each RRH is equipped with 8 antennas. The bandwidth of the system is 20 MHz, which is corresponding to 1200 subcarriers. We group the 1200 subcarriers into 40 groups with 30 subcarriers in each group. The bits are allocated equally among all groups. Within each group, the baseband signals within each group are processed using the compressive quantization proposed in section IV and the LMMSE weight vectors are used in the BBUs. In Table I, the EVM [31] performance under different modulations and front-haul capacities are presented. It is shown that even with the 40 Mb/s front-haul link capacity which is relatively low, the EVM performance is satisfactory. In contrast, typically, at 20 MHz bandwidth, even the single-antenna RRH has to use fiber optics with multiple Gb/s capacity to have satisfactory performance [32]. It is due to the fact that the baseband signals from multiple subcarriers and multiple time instants are combined together and transmitted in the front-haul links, which alleviates a lot of traffic load in the front-haul links. This phenomenon is similar to the "time-reversal tunneling effect" observed in [17] where the unique spatial and temporal focusing effects of time-reversal communication make it possible to detect the symbols transmitted by multiple TDs from the blended baseband signals. In this work, the spatial diversity provided

TABLE I

THE EVM OF THE COMPRESSIVE QUANTIZATION SCHEME APPLIED TO THE OFDM BASED C-RAN

| | $T = 80$ | $T = 112$ | $T = 128$ |
|---|---|---|---|
| QPSK, C = 0.02 | 19.64 | 12.31 | 10.75 |
| QPSK, C = 0.03 | 9.99 | 5.47 | 4.63 |
| QPSK, C = 0.04 | 7.71 | 5.84 | 3.49 |
| 16QAM, C = 0.02 | 19.72 | 12.40 | 10.80 |
| 16QAM, C = 0.03 | 9.14 | 6.89 | 4.38 |
| 16QAM, C = 0.04 | 7.65 | 5.61 | 4.22 |
| 64QAM, C = 0.02 | 19.72 | 12.26 | 10.80 |
| 64QAM, C = 0.03 | 9.29 | 6.63 | 4.65 |
| 64QAM, C = 0.04 | 8.08 | 4.65 | 4.54 |

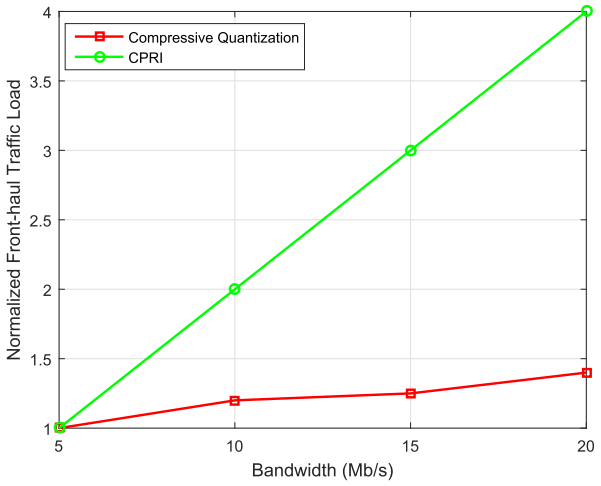



Fig. 12. The comparison of traffic load in front-haul link between CPRI and compressive quantization.

by the large number of available antennas help the BBUs effectively detect the symbols transmitted by the TDs from the compressive quantized baseband signal.

To further illustrate the advantage of the compressive quantization scheme, we compare the growth of the front-haul traffic load in the front-haul links. In this experiment, instead of using fixed bandwidth, we gradually change the bandwitdh from 5 MHz to 20 MHz. Let $\phi^{(CQ)}(J)$ and $\phi^{(CPRI)}(J)$ denote the traffic load in each front-haul link for bandwidth $J$ using compressive quantization and the CPRI compression [33] under the condition that the EVM is below 5% with QPSK modulation. We plot the normalized traffic load for each compression scheme where $\epsilon^{(CQ)}(J) = \frac{\phi^{(CQ)}(J)}{\phi^{(CQ)}(J)|_{J=5MHz}}$ and $\epsilon^{(CPRI)}(J) = \frac{\phi^{(CPRI)}(J)}{\phi^{(CPRI)}(J)|_{J=5MHz}}$. As shown in Fig. 12, the traffic load for the system using compressive quantization grows slower than that using the CPRI compression. The reason is that the compressive quantization scheme allows the baseband signals in multiple subcarriers to be combined together and transmitted.

## VI. Conclusion

In this work, we propose the compressive quantization and symbol detection scheme for the uplink of multi-antenna cloud radio access network (C-RAN) to tackle the limited front-haul capacity challenge. The proposed scheme is tailored to the C-RAN architecture in that the complexity of the compressive quantization which happens in the remote radio head (RRH) is low and can be implemented using basic buffering and adding operations, due to which the low deployment cost feature of the C-RAN is preserved and massive antennas can be utilized by deploying multiple RRHs. With the spatial diversity provided by the massive antennas, the baseband units (BBUs) are able to efficiently detect the symbols transmitted by the terminal devices (TDs) from the compressive quantized baseband signal. Numerical results show that the proposed scheme is efficient for the multi-antenna C-RAN in tackling the front-haul capacity deficit challenge. We further extend the proposed scheme to the OFDM based multi-antenna C-RAN. It is also shown that the proposed scheme can facilitate the OFDM based C-RAN utilize larger bandwidth with limited front-haul capacity.

## References

[1] A. Osseiran *et al.*, "The foundation of the mobile and wireless communications system for 2020 and beyond: Challenges, enablers and technology solutions," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1–5.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[4] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[5] A. L. Moustakas, S. H. Simon, and A. M. Sengupta, "MIMO capacity through correlated channels in the presence of correlated interferers and noise: A (not so) large N analysis," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2545–2561, Oct. 2003.

[6] "C-RAN: The road towards green RAN," ChinaMobile, Beijing, China, White Paper, Oct. 2011. [Online]. Available: https://ieeexplore.ieee.org/document/8039203/

[7] M. Webb, Z. Li, P. Bucknell, T. Moulsley, and S. Vadgama, "Future evolution in wireless network architectures: Towards a 'cloud of antennas,'" in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2012, pp. 1–5.

[8] C.-L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent progress on C-RAN centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[9] Y. D. Beyene, R. Jäntti, and K. Ruttik, "Cloud-RAN architecture for indoor DAS," *IEEE Access*, vol. 2, pp. 1205–1212, Oct. 2014.

[10] R. Wang, H. Hu, and X. Yang, "Potentials and challenges of C-RAN supporting multi-RATs toward 5G mobile networks," *IEEE Access*, vol. 2, pp. 1187–1195, 2014.

[11] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.

[12] X. Rao and V. K. N. Lau, "Distributed fronthaul compression and joint signal recovery in cloud-RAN," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1056–1065, Feb. 2015.

[13] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Inter-cluster design of precoding and fronthaul compression for cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 369–372, Apr. 2014.

[14] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.

[15] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Downlink multicell processing with limited-backhaul capacity," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 3:1–3:10, Feb. 2009. [Online]. Available: https://dl.acm.org/citation.cfm?id=1598770

[16] R. Zakhour and D. Gesbert, "Optimized data sharing in multicell MIMO with finite backhaul capacity," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6102–6111, Dec. 2011.

[17] H. Ma, B. Wang, Y. Chen, and K. J. R. Liu, "Time-reversal tunneling effects for cloud radio access network," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 3030–3043, Apr. 2016.

[18] B. Wang, Y. Wu, F. Han, Y. H. Yang, and K. J. R. Liu, "Green wireless communications: A time-reversal paradigm," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1698–1710, Sep. 2011.

[19] Y. Chen *et al.*, "Time-reversal wireless paradigm for green Internet of Things: An overview," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 81–98, Feb. 2014.

[20] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[21] S. Park, C.-B. Chae, and S. Bahk, "Before/after precoding massive MIMO systems for cloud radio access networks," *J. Commun. Netw.*, vol. 15, no. 4, pp. 398–406, Aug. 2013.

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

[23] F. A. de Figueiredo, J. P. Miranda, F. L. Figueiredo, and F. A. Cardoso. (2015). "Uplink performance evaluation of massive MU-MIMO systems." [Online]. Available: https://arxiv.org/abs/1503.02192
[24] B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2008. [Online]. Available: https://books.google.com/books?id=8q-xcGeEJDwC
[25] B. Hajek, *Random Processes for Engineers*. Cambridge, U.K.: Cambridge Univ. Press, 2015. [Online]. Available: https://books.google.com/books?id=76uwBgAAQBAJ
[26] K. J. R. Liu, A. K. Sadek, W. Su, and A. Kwasinski, *Cooperative Communications and Networking*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[27] M. Morelli, "Timing and frequency synchronization for the uplink of an OFDMA system," *IEEE Trans. Commun.*, vol. 52, no. 2, pp. 296–306, Feb. 2004.
[28] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst.*, Aug. 2014, pp. 781–785.
[29] O. Edfors, M. Sandell, J.-J. Van De Beek, S. K. Wilson, and P. O. Borjesson, "OFDM channel estimation by singular value decomposition," *IEEE Trans. Commun.*, vol. 46, no. 7, pp. 931–939, Jul. 1998.
[30] K. T. Truong and R. W. Heath, Jr., "The viability of distributed antennas for massive MIMO systems," in *Proc. 47th Asilomar Conf. Signals, Syst. Comput.*, Monterey, CA, USA, Nov. 2013, pp. 1318–1323.
[31] R. A. Shafik, M. S. Rahman, and A. H. M. R. Islam, "On the extended relationships among EVM, BER and SNR as performance metrics," in *Proc. Int. Conf. Elect. Comput. Eng. (ICECE)*, Dec. 2006, pp. 408–411.
[32] J. Lorca and L. Cucala, "Lossless compression technique for the fronthaul of LTE/LTE-advanced cloud-RAN architectures," in *Proc. IEEE 14th Int. Symp. Workshops World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2013, pp. 1–9.
[33] B. Guo, W. Cao, A. Tao, and D. Samardzija, "LTE/LTE-A signal compression on the CPRI interface," *Bell Labs Tech. J.*, vol. 18, no. 2, pp. 117–133, Sep. 2013.

**Hang Ma** (M'17) received the B.S. degree in information engineering from Northwestern Polytechnical University, Xi'an, China, in 2010, and the Ph.D. degree in electrical engineering from the University of Maryland at College Park, College Park, MD, USA, in 2016. His research interests include wireless communication and signal processing. He received the honor of Outstanding Graduate of Northwestern Polytechnical University in 2010.

**Beibei Wang** (SM'15) received the B.S. degree in electrical engineering (Hons.) from the University of Science and Technology of China, Hefei, in 2004, and the Ph.D. degree in electrical engineering from the University of Maryland at College Park, College Park, MD, USA, in 2009. She was with the University of Maryland at College Park as a Research Associate from 2009 to 2010. She was with Qualcomm Research and Development from 2010 to 2014. Since 2015, she has been with Origin Wireless Inc., where she is currently a Chief Scientist. Her research interests include wireless communications and signal processing. She received the Graduate School Fellowship, the Future Faculty Fellowship, and the Deans Doctoral Research Award from the University of Maryland at College Park, and the Overview Paper Award from the IEEE Signal Processing Society in 2015. She is a co-author of the *Cognitive Radio Networking and Security: A Game-Theoretic View* (Cambridge University Press, 2010).

**K. J. Ray Liu** (F'03) was named a Distinguished Scholar-Teacher with the University of Maryland at College Park in 2007, where he is a Christine Kim Eminent Professor of Information Technology. He leads the Maryland Signals and Information Group conducting research encompassing broad areas of information and communications technology with recent focus on wireless AI.

Dr. Liu is a fellow of the AAAS. He was recognized by Web of Science as a Highly Cited Researcher. He was a recipient of the 2016 IEEE Leon K. Kirchmayer Award on graduate teaching and mentoring, the IEEE Signal Processing Society 2014 Society Award, the IEEE Signal Processing Society 2009 Technical Achievement Award, and over a dozen of best paper awards. His invention of the Time-Reversal Machine by Origin Wireless Inc., received the 2017 CEATEC Grand Prix Award.

Dr. Liu is the IEEE Vice President, Technical Activities-Elect. He was the President of the IEEE Signal Processing Society, where he has served as the Vice President-Publications and Board of Governors, and as a member of the IEEE Board of Directors as a Division IX Director. He has also served as the Editor-in-Chief of the *IEEE Signal Processing Magazine*.

Dr. Liu has received teaching and research recognitions from the University of Maryland, including the university-level Invention of the Year Award; and the college-level Poole and Kent Senior Faculty Teaching Award, the Outstanding Faculty Research Award, and the Outstanding Faculty Service Award, all from A. James Clark School of Engineering.