PERFORMANCE COMPARISON OF DATA-SHARING AND COMPRESSION STRATEGIES FOR CLOUD RADIO ACCESS NETWORKS

Pratik Patil, Binbin Dai, and Wei Yu

Department of Electrical and Computer Engineering
University of Toronto, Toronto, Ontario M5S 3G4, Canada
Emails: {ppatil, bdai, weiyu}@comm.utoronto.ca

ABSTRACT

This paper provides a system-level performance comparison of two fundamentally different transmission strategies for the downlink of a cloud radio access network. The two strategies, namely the data-sharing strategy and the compression-based strategy, differ in the way the limited backhaul is utilized. While the data-sharing strategy uses the backhaul to carry raw user data, the compression strategy uses the backhaul to carry compressed beamformed signals. Although these strategies have been individually studied in the literature, a fair comparison of the two schemes under practical network settings is challenging because of the complexity in jointly optimizing user scheduling, beamforming, and power control for system-level performance evaluation, along with the need to optimize cooperation clusters for the data-sharing strategy and quantization noise levels for the compression strategy. This paper presents an optimization framework for both the data-sharing and compression strategies, while taking into account losses due to practical modulation in terms of gap to capacity and practical quantization in terms of gap to rate-distortion limit. The main conclusion of this paper is that the compression-based strategy, even with a simple fixed-rate uniform quantizer, outperforms the data-sharing strategy under medium to high capacity backhauls. However, the data-sharing strategy outperforms the compression strategy under low capacity backhauls primarily because of the large quantization loss at low backhaul capacity with compression.

1. INTRODUCTION

The ultra-dense cell deployment in the next generation (5G) wireless networks calls for efficient management of inter-cell interference. Cloud radio access network (C-RAN) has emerged as a promising cellular architecture that allows joint signal processing across base-stations (BSs) for interference mitigation purposes whereby the BSs are connected to a centralized cloud-computing based processor. This paper compares the performance of two fundamentally different transmission strategies for the downlink C-RAN, where the BSs essentially act as relays in transmitting data from the central processor to the remote users.

In the *data-sharing* strategy, the central processor shares the data of each user to a cluster of BSs which then compute the beamformed signals to be transmitted. In the *compression* strategy, the central processor itself computes the beamformed signals to be transmitted by each BS, which are then quantized and sent to the BSs through capacity-limited backhaul links. Individually, both the data-sharing and compression strategies have been studied in the context of C-RAN. However, a fair system-level comparison between the two strategies under practical network settings is still not available in the literature due to the challenges in solving the corresponding network optimization problems involving user scheduling, beamforming, power control, along with the optimization of clusters for the data-sharing strategy and the optimization of quantization noise levels for the compression strategy. This paper tackles such a system-level performance evaluation and tries to answer the question of under what condition one strategy performs better than the other.

One contribution of this paper is that we model and take into account loss due to practical modulation schemes in terms of gap to capacity for both strategies. In addition, for the compression strategy, we introduce a similar notion of gap to rate-distortion limit to account for quantization losses due to non-ideal quantizers used in practice. Further, we propose a novel algorithm for the joint optimization of the beamformers and quantization noise levels for the compression strategy based on an equivalence between weighted sum rate (WSR) maximization and weighted minimum mean square error (WMMSE) optimization.

We show through simulations on a heterogeneous cellular topology that whether one strategy is superior to the other largely depends on the backhaul capacity constraint in the system. If the available backhaul capacity is medium to high, the compression strategy outperforms the data-sharing strategy, even with a simple fixed-rate uniform scalar quantizer. However, if the available backhaul capacity is low, the data-sharing strategy outperforms the compression strategy. Intuitively, under low backhaul capacity the quantization noises introduced in the compression strategy dominate the interference, in which case it is better to just share the data directly with a limited set of BSs rather than to compress.

We note that in our previous work [1], a comparison between the data-sharing strategy and the compression strategy is made. But the system considered in [1] is limited to only a sum backhaul constraint, instead of the per-BS backhaul constraints considered here. Moreover, in [1], the data-sharing strategy does not select an optimized cluster of BSs for each user; the compression strategy does not consider the joint optimization of the beamformers and the quantization noise levels; further only a fixed user scheduling is assumed.

This paper restricts attention to linear precoding strategies and does not consider nonlinear precoding based on dirty-paper coding [2]. A hybrid between the data-sharing and compression strategies is also possible and is discussed in [1]. For more references on the data-sharing strategy, we refer the readers to [3] and for the compression strategy to [4].

2. SYSTEM MODEL

Consider a downlink C-RAN consisting of L single-antenna BSs serving K single-antenna remote users. All L BSs are connected to a central processor with capacity-limited backhaul links. (We use the term backhaul, because the links carry digital data. These links are sometimes referred to as fronthaul links in the C-RAN literature, especially when they carry compressed analog signals.) The capacity of the backhaul link connecting lth BS to the central processor is denoted by C_l, $l = 1, \ldots, L$. We assume one data stream per user, and that the central processor has access to the data and perfect CSI for all K users in the network.

Let x_l denote the complex signal transmitted by BS l and $\mathbf{x} \in \mathbb{C}^{L\times 1} = [x_1, \ldots, x_L]^T$ be the aggregate signal from all the BSs. The received signal at user k can be written as

$$y_k = \mathbf{h}_k^H \mathbf{x} + z_k, \quad k = 1, 2, \ldots, K \tag{1}$$

where $\mathbf{h}_k \in \mathbb{C}^{L\times 1} = [h_{1,k}, \ldots, h_{L,k}]^T$ is the channel to the user k from all the BSs, and z_k is the additive complex Gaussian noise with zero-mean and variance σ^2. Each BS l has a transmit power budget denoted by P_l. Let s_k denote the data of kth user distributed as complex Gaussian with zero-mean and unit variance, which is available at the central processor.

3. DATA-SHARING STRATEGY

In the data-sharing strategy, a cluster of BSs locally form beamformers to cooperatively serve each user. The data for that user is replicated at all the participating BSs in the cluster via the backhaul links. A crucial decision is to select an appropriate cluster of BSs for each user for interference mitigation, while staying under the limited backhaul capacity.

Let the beamforming vector for user k from all the BSs be $\mathbf{w}_k \in \mathbb{C}^{L\times 1} = [w_{1,k}, w_{2,k}, \ldots, w_{L,k}]^T$, where $w_{l,k}$ denotes the component of the beamformer from BS l. If BS l does not participate in cooperatively serving user k, then $w_{l,k} = 0$. The beamformed signals transmitted from all the BSs can then be written as

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{w}_k s_k. \tag{2}$$

At user k, the signal-to-interference-plus-noise ratio (SINR) can be expressed as

$$\text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j\neq k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2}. \tag{3}$$

The information theoretical achievable rate for user k is related to SINR as $R_k = \log(1+\text{SINR}_k)$. However, this rate expression assumes Gaussian signaling, while in practice QAM constellations are typically used for the Gaussian channel in the moderate and high SINR regime. With moderate coding, to achieve a given data rate we still need an SINR higher than what is suggested above. This extra amount of power is usually captured by a so-called SNR gap. Denoting the gap by Γ_m, we can rewrite the achievable rate for user k as

$$R_k = \log\left(1 + \frac{\text{SINR}_k}{\Gamma_m}\right). \tag{4}$$

The optimization problem of finding the optimal set of BS clusters and beamformers for the data-sharing scheme can now be formulated as a WSR maximization problem under per-BS power constraints and per-BS backhaul constraints:

$$\underset{w_{l,k}}{\text{maximize}} \quad \sum_{k=1}^{K} \alpha_k R_k \tag{5a}$$

$$\text{subject to} \quad \sum_{k=1}^{K} |w_{l,k}|^2 \leq P_l, \quad \forall l \tag{5b}$$

$$\sum_{k=1}^{K} \left\{|w_{l,k}|^2\right\} R_k \leq C_l, \quad \forall l \tag{5c}$$

where α_k denotes the priority weight associated with user k and the indicator function $\left\{|w_{l,k}|^2\right\}$ denotes if BS l participates in beamforming to user k, and if so, the user rate R_k is included in the backhaul constraint C_l. The beamforming coefficients are computed at the central processor, and are assumed to be transmitted to the BSs without any error. We neglect the backhaul consumption for transmitting the beamformers. This formulation considers joint design of BS clustering and beamforming. It also implicitly does power control and user scheduling. This optimization problem is solved repeatedly and the BS clusters are dynamically optimized in each time slot as the priority weights are updated.

The presence of the backhaul constraint (5c) makes the optimization problem challenging. In this paper, we follow the approximation suggested in [3] to first write the indicator function as a l_0 norm which is then approximated as a weighted l_1 norm as

$$\left\{|w_{l,k}|^2\right\} = \left\||w_{l,k}|^2\right\|_0 \approx \beta_{l,k}|w_{l,k}|^2, \tag{6}$$

where $\beta_{l,k}$ is a constant weight associated with BS l and user k and is updated iteratively according to

$$\beta_{l,k} = \frac{1}{|w_{l,k}|^2 + \tau}, \quad \forall\, k, l \tag{7}$$

for some regularization constant $\tau > 0$ and $|w_{l,k}|^2$ from the previous iteration. This simplifies the constraint (5c) to

$$\sum_{k=1}^{K} \beta_{l,k} |w_{l,k}|^2 R_k \leq C_l, \quad \forall l \tag{8}$$

which is equivalent to a generalized power constraint, if R_k is assumed fixed and heuristically chosen from the previous iteration. The resulting optimization problem can then be solved using an equivalence between the WSR maximization and the WMMSE problem.

The only difference between the formulation (5) and that in [3] is the gap factor Γ_m. We can easily verify that the equivalence between WSR optimization and WMMSE extends even with the gap Γ_m. Below we summarize the overall algorithm for the optimization of the data-sharing strategy. Though we do not have theoretical guarantee of its convergence in general, it is observed to converge in simulations.

Algorithm 1 WSR maximization for data-sharing strategy

Initialization: $\{\beta_{l,k}\}, \{\mathbf{w}_k\}, \{R_k\}$;

Repeat:

1. For fixed $\{\mathbf{w}_k\}$, compute the MMSE receivers $\{u_k\}$ and the corresponding MSE $\{e_k\}$ according to (11) and (9);
2. Update the MSE weights $\{\rho_k\}$ according to (10);
3. For fixed $\{u_k\}$, $\{\rho_k\}$, and $\{R_k\}$ in (12c), find the optimal transmit beamformer $\{w_{l,k}\}$ by solving (12);
4. Update $\{\beta_{l,k}\}$ as in (7). Compute the achievable rate $\{R_k\}$ according to (4);

Until convergence

The quantities used in the WMMSE approach in the above algorithm are as follows. The mean square error (MSE) for user k is defined as

$$e_k = |u_k|^2 \left(\Gamma_m \left(\sum_{j \neq k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 \right) + |\mathbf{h}_k^H \mathbf{w}_k|^2 \right) - 2\,\mathrm{Re}\{u_k^H \mathbf{h}_k^H \mathbf{w}_k\} + 1. \tag{9}$$

The optimal MSE weight ρ_k under fixed $\{\mathbf{w}_k\}$ and $\{u_k\}$ is given by

$$\rho_k = e_k^{-1}. \tag{10}$$

The optimal receive beamformer u_k under fixed $\{\mathbf{w}_k\}$ and $\{\rho_k\}$ is given by

$$u_k = \left(\Gamma_m \left(\sum_{j \neq k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 \right) + |\mathbf{h}_k^H \mathbf{w}_k|^2 \right)^{-1} \mathbf{h}_k^H \mathbf{w}_k. \tag{11}$$

The optimization of transmit beamformers $\{\mathbf{w}_k\}$ under fixed $\{u_k\}$, $\{\rho_k\}$ and fixed $\{R_k\}$ is the following quadratically constrained quadratic programming (QCQP) problem:

$$\underset{w_{l,k}}{\text{minimize}} \quad \sum_{k=1}^{K} \mathbf{w}_k^H \mathbf{A}_k \mathbf{w}_k - \mathrm{Re}\{\mathbf{b}_k^H \mathbf{w}_k\} \tag{12a}$$

$$\text{subject to} \quad \sum_{k=1}^{K} |w_{l,k}|^2 \leq P_l, \quad \forall l \tag{12b}$$

$$\sum_{k=1}^{K} R_k \beta_{l,k} |w_{l,k}|^2 \leq C_l, \quad \forall l \tag{12c}$$

where $\mathbf{A}_k \in \mathbb{C}^{L \times L}$ and $\mathbf{b}_k \in \mathbb{C}^{L \times 1}$ are defined to be

$$\mathbf{A}_k = \sum_{j \neq k} \alpha_j \rho_j |u_j|^2 \Gamma_m \mathbf{h}_j \mathbf{h}_j^H + \alpha_k \rho_k |u_k|^2 \mathbf{h}_k \mathbf{h}_k^H \tag{13}$$

$$\mathbf{b}_k = 2 \alpha_k \rho_k u_k \mathbf{h}_k \tag{14}$$

4. COMPRESSION STRATEGY

In the compression strategy, the central processor computes the beamformed analog signals to be transmitted by the BSs. These signals have to be compressed before they can be forwarded to the corresponding BSs through the finite-capacity backhaul links. The process of compression introduces quantization noises; the quantization noise levels depend on backhaul capacities.

In the data-sharing strategy, the beamformed signal $\mathbf{x}$ as given by (2) is computed at the BSs. In the compression strategy, $\mathbf{x}$ is computed at the central processor, then compressed, sent over the backhaul links, and reproduced by the BSs. We model the quantization process for $\mathbf{x}$ as

$$\hat{\mathbf{x}} = \mathbf{x} + \mathbf{e}, \tag{15}$$

where $\mathbf{e}$ is the quantization noise with covariance matrix $\mathbf{Q} \in \mathbb{C}^{L \times L}$ modelled as complex Gaussian and assumed to be independent of $\hat{\mathbf{x}}$. The received SINR for user k can then be written as

$$\mathrm{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j \neq k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma^2 + |\mathbf{h}_k^H \mathbf{Q} \mathbf{h}_k|}. \tag{16}$$

This paper considers independent quantization at each BS, in which case $\mathbf{Q}$ is a diagonal matrix with diagonal entries q_l. (Multivariate compression is also possible and has been studied in [4].) Assuming an ideal vector quantizer, the quantization noise level q_l and the backhaul capacity C_l (from rate-distortion theory) are related as

$$\log \left(1 + \frac{\sum_{k=1}^{K} |w_{l,k}|^2}{q_l} \right) \leq C_l. \tag{17}$$

However, the quantizers used in practice for compression can be far from ideal. In order to capture these losses, we introduce a notion of gap to rate-distortion limit. Following [5],

we note that operational distortion achieved by virtually all practical quantizers at high resolution follow the relation

$$\delta(R) = \Gamma_q \mathrm{var}(X) 2^{-R}. \tag{18}$$

where $\mathrm{var}(X)$ is the variance of the signal being quantized, R is the rate of quantizer, and Γ_q is a constant that depends on the particular choice of quantizer. For example, for a fixed-rate (uncoded) uniform scalar quantizer, $\Gamma_q = \frac{\sqrt{3}\pi}{2}$, which is around 2.72. For a uniform scalar quantizer followed by variable-rate entropy coding we get $\Gamma_q = \frac{\pi e}{6}$ which is around 1.42. Note that $\Gamma_q = 1$ corresponds to the distortion achievable by the best possible vector quantization scheme. Accounting for this, we can rewrite the relation above as

$$\log\left(1 + \frac{\Gamma_q \sum_{k=1}^{K} |w_{l,k}|^2}{q_l}\right) \leq C_l. \tag{19}$$

Note that $\sum_{k=1}^{K} |w_{l,k}|^2$ is the power of the signal that is quantized for BS l. The achievable rate for user k, R_k, is again as given by (4).

The design of the compression strategy can now be stated as a WSR maximization problem over the transmit beamformers and the quantization noise levels:

$$\underset{w_{l,k},q_l}{\text{maximize}} \quad \sum_{k=1}^{K} \alpha_k R_k \tag{20a}$$

$$\text{subject to} \quad \sum_{k=1}^{K} |w_{l,k}|^2 - \frac{2^{C_l}-1}{\Gamma_q}\, q_l \leq 0, \quad \forall l \tag{20b}$$

$$\sum_{k=1}^{K} |w_{l,k}|^2 + q_l \leq P_l, \quad \forall l \tag{20c}$$

The constraint (20b) is just a reformulation of (19) while the constraint (20c) is the power constraint on the compressed signal x_l. Finding the globally optimal solution to the above problem is challenging. An iterative approach based on the majorize-minimization (MM) algorithm has been suggested in [4]. The algorithm in [4] transforms $\mathbf{w}_k\mathbf{w}_k^H$ into a non-negative definite matrix variable $\mathbf{R}_k$ and ignores the rank constraint on $\mathbf{R}_k$ in the optimization. In this paper, we propose a novel way to solve (20) by reformulating it as an equivalent WMMSE problem and then using the block coordinate descent method between the transmit beamformers $\{\mathbf{w}_k\}$ and the quantization noise levels $\{q_l\}$, the receive beamformers $\{u_k\}$, and the MSE weights $\{\rho_k\}$. The algorithm can be shown to reach a stationary point of (20). The explicit equivalence is not stated here for brevity. The numerical procedure is presented as Algorithm 2.

In Algorithm 2, the optimization of the transmit beamformers $\{\mathbf{w}_k\}$ and the quantization noise levels $\{q_l\}$ under fixed $\{u_k\}$ and $\{\rho_k\}$ is the following convex program:

Algorithm 2 WSR maximization for compression strategy

Initialization: $\{\mathbf{w}_k\}, \{q_l\}$;

Repeat:

1. For fixed $\{\mathbf{w}_k\}, \{q_l\}$, compute the MMSE receivers $\{u_k\}$ and the corresponding MSE $\{e_k\}$ according to (11) and (9) with σ^2 replaced by $\sigma^2 + |\mathbf{h}_k^H \mathbf{Q} \mathbf{h}_k|$ in both equations;
2. Update the MSE weights $\{\rho_k\}$ according to (10);
3. For fixed $\{u_k\}$ and $\{\rho_k\}$, find the optimal transmit beamformers $\{\mathbf{w}_k\}$ and quantization noise levels $\{q_l\}$ by solving the convex optimization problem (21);

Until convergence

$$\min_{w_{l,k},q_l} \sum_{k=1}^{K} \mathbf{w}_k^H \mathbf{A}_k \mathbf{w}_k - \mathrm{Re}\{\mathbf{b}_k^H \mathbf{w}_k\} + \alpha_k \rho_k |u_k|^2 \Gamma_m |\mathbf{h}_k^H \mathbf{Q} \mathbf{h}_k| \tag{21a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} |w_{l,k}|^2 - \frac{2^{C_l}-1}{\Gamma_q}\, q_l \leq 0, \quad \forall l \tag{21b}$$

$$\sum_{k=1}^{K} |w_{l,k}|^2 + q_l \leq P_l, \quad \forall l \tag{21c}$$

where $\mathbf{A}_k$ and $\mathbf{b_k}$ are as defined in (13) and (14).

We further observe that the convex optimization problem (21) has a particular structure that can be exploited. Observe that the two constraints (21b) and (21c) provide a lower and an upper bound on $\{q_l\}$, respectively. Since the objective (21a) is monotonically decreasing in $\{q_l\}$, we can replace the inequality with equality in the constraint (21b) and substitute $\{q_l\}$ from (21b) into the objective (21a) and the constraint (21c). This results in a QCQP problem in only a single set of variables $\{\mathbf{w}_k\}$, which can be efficiently solved by standard solvers.

5. PERFORMANCE EVALUATION

We consider a 7-cell wrapped-around two-tier heterogeneous network with simulation parameters as listed in Table 1. All the macro-BSs and pico-BSs are connected to a centralized processor by capacity-limited backhaul links. We compare the performance of the two strategies under varying backhaul capacities. The combined background noise and interference caused by two tiers of cells outside the 7-cells is estimated to be -150 dBm/Hz. We assume an SNR gap of $\Gamma_m = 9$ dB (corresponding to uncoded QAM transmission) and a gap to rate-distortion limit of $\Gamma_q = 4.3$ dB (corresponding to uncoded fixed-rate uniform scalar quantizer). At each time slot, we solve the respective network optimization problems and update the weights in WSR maximization according to the proportional fair criterion.

Channel bandwidth	10 MHz
Distance between cells	0.8 km
Number of users/cell	30
Number of macro-BSs/cell	1
Number of pico-BSs/cell	3
Max. Tx power at macro-BS	43 dBm
Max. Tx Power at pico-BS	30 dBm
Antenna gain	15 dBi
Background noise	-169 dBm/Hz
Path loss from macro-BS to user	$128.1 + 37.6 \log_{10}(d)$
Path loss from pico-BS to user	$140.7 + 36.7 \log_{10}(d)$
Log-normal shadowing	8 dB
Rayleigh small scale fading	0 dB
SNR gap (Γ_m)	9 dB
Rate-distortion gap (Γ_q)	4.3 dB

Table 1. Simulation Parameters

Fig. 1 shows the cumulative distribution of user rates under varying backhaul capacities for both strategies. For reference, we also include the full cooperation case with infinite backhaul capacity and the baseline scheme of no cooperation with each user connected to the strongest BS. When the backhaul capacity is low at 40 Mbps/macro-BS and 20 Mbps/pico-BS, the data-sharing strategy outperforms the compression strategy. The 50-percentile rate for the data-sharing strategy is about 3 times that of the compression strategy. If we double the backhaul capacity to 80 Mbps/macro-BS and 40 Mbps/pico-BS, the compression strategy becomes comparable to the data-sharing strategy and both have about the same 50-percentile user rates. At this operating point, the sum backhaul capacity is about 6 times that of the average sum rate per cell. We also observe that the compression strategy favours low rate users while the data-sharing strategy favours high rate users. A reason for this is that the compression strategy under low backhaul capacity is limited by the quantization noises which are about the same for all the BS signals resulting in more uniform user rates.

We observe that with moderate-to-high backhaul capacity of 160 Mbps/macro-BS and 80 Mbps/pico-BS, the compression strategy outperforms the data-sharing strategy with the 50-percentile rate for the compression strategy more than 2.5 times than that of data-sharing. Increasing the backhaul in this regime improves the compression strategy drastically, while the data-sharing strategy sees only a moderate increase. This is because, at low backhaul capacity, the performance of the compression strategy is limited by the quantization noises. An increase in backhaul capacity reduces the quantization noise levels exponentially, while a similar increase in the backhaul capacity does not buy as much for the data-sharing strategy. Finally with a backhaul of 240 Mbps/macro-BS and 120 Mbps/pico-BS, the compression strategy performs close to the full cooperation limit, while for the data-sharing strategy, backhaul capacities of 1200 Mbps/macro-BS and 600 Mbps/pico-BS are needed to get as close.

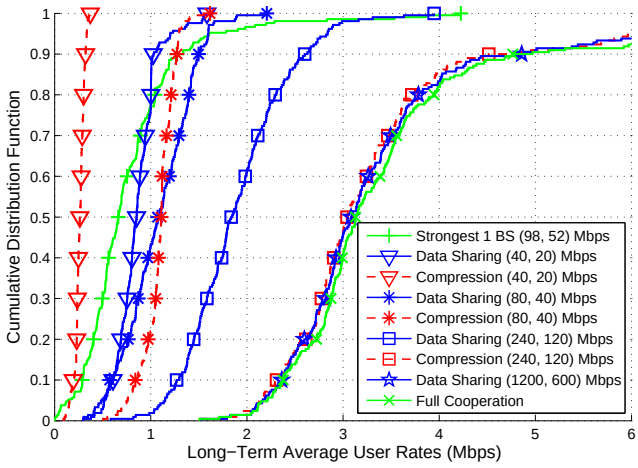

Fig. 1. Comparison of cumulative distribution of user rates for the data-sharing and compression strategies.

6. CONCLUSIONS

This paper compares two fundamentally different strategies, the data-sharing and the compression strategy, for the downlink C-RAN under realistic network settings considering various practical aspects. Our main conclusion is that the backhaul capacity constraint is crucial in deciding which strategy to adopt. The compression strategy offers better user rates for moderate-to-high backhaul capacity, due to its ability to have full cooperation before quantization. But it suffers from high quantization loss at low backhaul capacity in which case it is better to do data-sharing with limited cooperation cluster.

REFERENCES

[1] P. Patil and W. Yu, "Hybrid compression and message-sharing strategy for the downlink cloud radio-access network," in *Proc. Inf. Theory Applicat. Workshop*, Feb. 2014.

[2] O. Simeone, O. Somekh, H. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," *EURASIP J. Advances Singal Process.*, Feb. 2009.

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

[4] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

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