Efficient Multi-user Quantize-Forward Relaying in Massive MIMO HetNets

Ahmad Abu Al Haija*, Ben Liang*, Min Dong‡, Gary Boudreau†
* University of Toronto, Toronto, ON, Canada, †Ericsson Canada, Ottawa, ON, Canada
‡ University of Ontario Institute of Technology, Oshawa, ON, Canada

***Abstract*—We utilize the orthogonality and channel hardening properties of massive multiple-input multiple-output (MIMO) systems to propose an efficient uplink transmission scheme for a heterogeneous network (HetNet). Such a network consists of multiple user-equipments (UEs) communicating with a macro-cell base station (MCBS) through a small-cell BS (SCBS) where both BSs have a large number of antennas and deploy zero-forcing (ZF) detection. The SCBS helps relay UEs' information using quantize-forward (QF) relaying with Wyner-Ziv (WZ) binning and multiple-timeslot transmission for the binning indices to the MCBS. The MCBS then deploys separate and sequential decoding for each UE's message. To maximize the rate region, we optimize the quantization levels through geometric programming and further obtain the optimal transmission timeslot durations in terms of the optimal quantization. We show that the proposed scheme has linear codebook size and decoding complexity in the number of UEs, while it achieves the same rate region of other QF schemes that employ joint transmission at the SCBS and/or joint decoding at the MCBS, all of which have exponential complexity. Furthermore, simulation results show that the SCBS should employ finer quantization for UE signals that have strong UE-SCBS links compared with the UE-MCBS links, and the proposed scheme can substantially outperform several existing alternatives under a wide range of parameter settings.**

I. Introduction

As a valuable technology for 5G cellular networks, massive MIMO has received heightened research interest because of its ability to 1) neglect the small scale fading through channel hardening [1], 2) orthogonalize different users' transmission through beamforming and allow concurrent transmission without inter-user interference [2], and 3) achieve close-to-optimal performance with low complexity linear receivers, e.g, the zero forcing (ZF) receiver [2]. Moreover, since massive MIMO arrays can be made rather compact [3] [4], they can be implemented at the MCBS and the SCBS as well. However, the MCBS will still have much more antennas than the SCBS. Given this deployment of massive MIMO, we investigate its impact on the transmission design of heterogeneous networks (HetNets).

Consider the uplink transmission for the HetNet in Fig. 1 where K UEs inside the small cell aim to communicate with the MCBS through the SCBS. For this system, to improve the UEs' transmission rates, several transmission schemes were proposed based on decode-forward (DF) [5]–[7] or quantize-forward (QF) [8], [9] relaying techniques. In LTE-A standard [10], DF relaying is used although QF relaying achieves higher rates in some channel settings, e.g., the relaying links to the SCBS have similar or weaker strength to the direct links to the MCBS [11]. Furthermore, the above works did not consider massive MIMO at the SCBS and the MCBS.

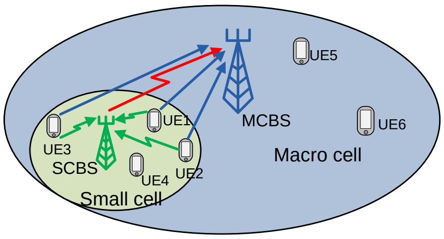

Fig. 1. Uplink Transmission in HetNet.

In QF relaying with multiple antennas, it is challenging to optimize the quantization at the SCBS as optimizing the covariance matrix of the quantization noise vector is generally non-convex. Hence, approximate solutions were obtained via iterative numerical methods [12]–[14]. Furthermore, in [15], we utilize massive MIMO orthogonality and beamforming properties to develop a preliminary QF scheme for the simple case of two UEs. However, for K UEs, it is more challenging to optimize the quantization at the SCBS.

This paper has the following contributions:

- We propose an uplink transmission scheme for K UEs in a massive MIMO HetNet. In this scheme, each UE performs conventional transmission, the SCBS deploys ZF detection, QF relaying with WZ binning and multiple-timeslot transmission to the MCBS. Then, the MCBS deploys ZF detection, and separate and sequential decoding for each UE's message.
- To maximize the rate region of the proposed scheme, we derive the optimal transmission timeslots as functions of the quantization levels. Then, present a method to formulate the optimal quantization problem as a geometric programming (GP) problem, which can be solved by existing tools [16].
- We show that the proposed scheme has linear codebook size and decoding complexity in terms the number of UEs. However, it achieves the same rate region of more complex schemes with joint encoding at the SCBS and/or joint decoding at the MCBS.

The remainder of this paper is organized as follows. Section II presents the uplink channel model of a massive MIMO HetNet with ZF detection at the SCBS and the MCBS. For this channel, Section III describes the proposed transmission scheme and provides its achievable rate region. To maximize this region, Section IV derives the optimal quantization and timeslot durations at the SCBS. Then, Section V shows the efficiency of the proposed scheme. Section VI presents numerical results and Section VII concludes the paper.

II. System Model

We consider the uplink transmission in a HetNet that consists of a macro cell, a small cell, and K UEs ($K > 1$) in the small cell. Each UE has a single antenna while the SCBS (resp. MCBS) has N (resp. M) antennas where we assume $M \gg N \gg K$. Since using massive MIMO techniques at both MCBS and SCBS can reduce the uplink interference from other nodes to a negligible level [2], we ignore transmissions in and from other SCBSs in the same macro cell. The UEs communicate with the MCBS through the SCBS, as shown in Fig. 1. This uplink channel resembles the the multiple-access relay channel (MARC) shown in Fig. 2, where the SCBS resembles the relay ($\mathcal{R}$) and the MCBS resembles the destination ($\mathcal{D}$).

For the MARC in Fig. 2, we assume a block fading channel model where the channel over each link remains constant in each transmission block and changes independently between blocks. Over B transmission blocks where $B \gg 1$, let $\mathbf{h}_{ri,j} = [h_{ri,j}^{(1)}, \cdots, h_{ri,j}^{(N)}]^T$ denote the $N \times 1$ channel vector from UE$_i$ to $\mathcal{R}$ in block j, for $i \in \{1, 2, \ldots, K\}$ and $j \in \{1, \ldots, B\}$, where $h_{ri,j}^{(n)}$ is the channel coefficient from UE$_i$ to the n^{th} antenna of $\mathcal{R}$ in block j. We assume $\mathbf{h}_{ri,j}$ is a complex Gaussian random vector with zero mean and covariance $\sigma_{h,r}^2 \mathbf{I}$. The variance $\sigma_{h,r}^2$ is modeled by the pathloss model as $\sigma_{h,r}^2 = d_{ri}^{\alpha}$, where d_{ri} is the distance between UE$_i$ and $\mathcal{R}$, and α is the pathloss exponent. A similar definition holds for the $M \times 1$ channel vector $\mathbf{h}_{di,j}$ from UE$_i$ to $\mathcal{D}$ and the $M \times N$ channel matrix $\mathbf{H}_{dr}$ from $\mathcal{R}$ to $\mathcal{D}$. We assume all channel coefficients are independent to each other.

At any transmission block $j \in \{1, \ldots, B\}$, given $\mathbf{h}_{ri,j}$, $\mathbf{h}_{di,j}$ and $\mathbf{H}_{dr,j}$, the received signal vectors at $\mathcal{R}$ and $\mathcal{D}$, denoted by $\mathbf{y}_{r,j}$ and $\mathbf{y}_{d,j}$ respectively, are given as follows:

$$\begin{aligned} \mathbf{y}_{r,j} &= \sum_{i=1}^{K} \mathbf{h}_{ri,j} x_{i,j} + \mathbf{z}_{r,j}, \\ \mathbf{y}_{d,j} &= \sum_{i=i}^{K} \mathbf{h}_{di,j} x_{i,j} + \mathbf{H}_{dr,j} \mathbf{x}_{r,j} + \mathbf{z}_{d,j}, \end{aligned} \tag{1}$$

where $x_{i,j}$ is the transmit signal by UE$_i$ for $i \in \{1, 2, \ldots, K\}$ while $\mathbf{x}_{r,j}$ is the $N \times 1$ transmit signal vector from $\mathcal{R}$; $\mathbf{z}_{r,j}$ and $\mathbf{z}_{d,j}$ are $N \times 1$ and $M \times 1$ independent complex AWGN vectors with zero mean and covariance $\mathbf{I}_N$ and $\mathbf{I}_M$, respectively.

We assume that the channel state information (CSI) is known at the respective receivers ($\mathcal{R}, \mathcal{D}$), i.e., $\mathcal{R}$ knows $\mathbf{h}_{ri}$ and $\mathcal{D}$ knows $\mathbf{h}_{di}$ and $\mathbf{H}_{dr}$. Moreover, $\mathcal{R}$ knows (via feedback from $\mathcal{D}$ [17]) the variance of the channel from $\mathcal{R}$ to $\mathcal{D}$, and from each UE to $\mathcal{D}$, through the pathloss information. Such knowledge helps $\mathcal{R}$ optimize its transmission for maximum rate region (see Section IV). Note that the pathloss information over each link is much easier to obtain than the massive MIMO channel itself ($\mathbf{h}_{di}$ and $\mathbf{H}_{dr}$) in each block.

We consider full-duplex relaying at the SCBS. Although full-duplex relaying suffers from self-interference, it can be substantially alleviated by analog and digital cancellation techniques developed in recent literatures and the remaining part appears as an additional additive noise [18]. Hence, we assume perfect cancellation at the SCBSs in our model and focus on its transmission design.

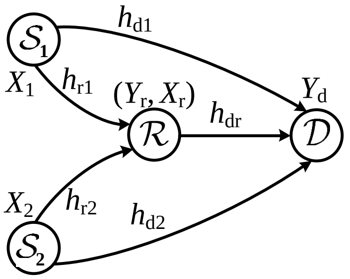

Fig. 2. The channel model of full-duplex MARC, using K=2 as example.

In massive MIMO systems, linear detectors like ZF or maximum ratio combining (MRC) are applied to reduce the inter-user interference [2]. Here, we choose the ZF detector for simplicity, but similar analysis is applicable to other detectors. Receivers at $\mathcal{R}$ and $\mathcal{D}$ apply ZF detection to separate the data streams from different origins. Specifically, define the ZF matrices $\mathbf{A}_{r,j}$ and $\mathbf{A}_{d,j}$ as follows

$$\mathbf{A}_{r,j} \triangleq (\mathbf{G}_{r,j}^H \mathbf{G}_{r,j})^{-1} \mathbf{G}_{r,j}^H, \ \mathbf{A}_{d,j} \triangleq (\mathbf{G}_{d,j}^H \mathbf{G}_{d,j})^{-1} \mathbf{G}_{d,j}^H,$$

where $\mathbf{G}_{r,j} \triangleq [\mathbf{h}_{r1,j} \ \mathbf{h}_{r2,j} \ \ldots \ \mathbf{h}_{rK,j}]$,

$$\mathbf{G}_{d,j} \triangleq [\mathbf{h}_{d1,j} \ \mathbf{h}_{d2,j} \ \ldots \ \mathbf{h}_{dK,j} \ \mathbf{H}_{dr,j}]. \tag{2}$$

Denote $\mathbf{A}_{r,j} = [\mathbf{a}_{r1,j}, \mathbf{a}_{r2,j}, \ldots, \mathbf{a}_{rK,j}]^H$ and $\mathbf{A}_{d,j} = [\mathbf{a}_{d1,j}, \mathbf{a}_{d2,j}, \ldots, \mathbf{a}_{dK,j}, \mathbf{A}_{dr,j}]^H$, where $\mathbf{a}_{ri,j}$ is an $N \times 1$ vector, $\mathbf{a}_{di,j}$ is an $M \times 1$ vector, for $i \in \{1, 2, \ldots, K\}$, and $\mathbf{A}_{dr,j}$ is an $M \times N$ matrix. After applying ZF matrices in (2) to the received signal vectors $\mathbf{y}_{r,j}$ and $\mathbf{y}_{d,j}$ at $\mathcal{R}$ and $\mathcal{D}$) in (1) respectively, we obtain the $K \times 1$ received vector $\tilde{\mathbf{y}}_{r,j}$ at $\mathcal{R}$ and $(K + N) \times 1$ received vector $\tilde{\mathbf{y}}_{d,j}$ at $\mathcal{D}$ as follows

$$\begin{aligned} \tilde{\mathbf{y}}_{r,j} &= [\tilde{y}_{r1,j} \ \tilde{y}_{r2,j} \ \ldots \ \tilde{y}_{rK,j}]^T, \\ \tilde{\mathbf{y}}_{d,j} &= [\tilde{y}_{d1,j} \ \tilde{y}_{d2,j} \ \ldots \ \tilde{y}_{dK,j} \ \tilde{\mathbf{y}}_{dr,j}]^T, \\ \tilde{y}_{ri,j} &= x_{i,j} + \mathbf{a}_{ri,j}^H \mathbf{z}_{r,j}, \ \tilde{y}_{di,j} = x_{i,j} + \mathbf{a}_{di,j}^H \mathbf{z}_{d,j}, \\ \tilde{\mathbf{y}}_{dr,j} &= \mathbf{x}_{r,j} + \mathbf{A}_{dr,j}^H \mathbf{z}_{d,j}, \end{aligned} \tag{3}$$

where $\tilde{y}_{ri,j}$ (resp.$\sim \tilde{y}_{di,j}$) is the signal received at $\mathcal{R}$ (resp.$\sim \mathcal{D}$) from UE$_i$ in block j, and $\tilde{\mathbf{y}}_{dr,j}$ is the $N \times 1$ signal vector received at $\mathcal{D}$ from $\mathcal{R}$ in block j.

III. The QF-WZTD Scheme for Massive MIMO HetNet

The QF-WZTD scheme is based on QF relaying, WZ binning and TD transmission at the SCBS and sliding window decoding with separate and sequential decoding at the MCBS, and we name it as the QF-WZTD scheme. These techniques are explained as follows.

- In QF relaying [11], the SCBS quantizes the received signal and forwards the quantization indices to the MCBS. Since massive MIMO asymptotically orthogonalizes the transmission from different users [2], the SCBS can separately quantize the date steam from each UE as in [11].

- In WZ binning [19], the SCBS partitions the quantization indices of each UE data stream into equal-size bins [11]. That is, after obtaining the quantization indices, the SCBS finds each binning index that includes the quantization index of each data stream.
- In TD, the SCBS transmits different information (codewords) over different time slots (phases). The number of phases is equal to the number of UEs. Hence, each SCBS transmits K binning indices over K phases.
- In separate and sequential decoding, the MCBS sequentially decodes the binning indices, quantization indices and then UEs' messages. Furthermore, the MCBS separately decodes these information triple (binning index, quantization index and the message) for each UE, i.e., the MCBS does not perform simultaneous and joint decoding for UEs' messages.

Next, we describe the transmission scheme in details.

A. Transmission Scheme

In the proposed scheme, the data transmission from the UEs to the MCBS is carried over B transmission blocks as each UE aims to send $B-1$ messages through these blocks.[1] In block $j \in \{1, 2, \ldots, B\}$, the transmission/reception at each node is given as follows.

1) Each UE: transmits a new message to the SCBS and the MCBS, i.e., UE$_k$ sends its new message $w_{k,j}$ as follows.

$$x_{k,j} = \sqrt{P_k} U_i(w_{k,j}),\ k \in \{1, 2, \ldots, K\}, \tag{4}$$

where P_k is the UE$_k$ transmit power and U_k is a Gaussian signal with zero mean and unit variance that conveys $w_{k,j}$ codeword.

2) The SCBS: deploys ZF detection as in (3) and then separately quantizes each UE message, i.e., for $k \in \{1, 2, \ldots, K\}$, the SCBS utilizes $\tilde{y}_{rk,j}$ to estimate the quantization index $\hat{l}_{k,j}$. Such a quantization index exists if its rate R_{qk} satisfies the following constraint [11]:

$$R_{qk} \geq \mathcal{C}\big(\frac{(d_{rk}^{\alpha}/N) + P_k}{Q_k}\big) \triangleq I_{r,k},\ k \in \{1, 2, \ldots, K\}. \tag{5}$$

where $\mathcal{C}(x) = \log(1 + x)$. After obtaining the quantization indices $l_{1,j}, l_{2,j}, \ldots,$ and $l_{K,j}$, the SCBS finds the binning indices $b_{1,j}, b_{2,j}, \ldots,$ and $b_{K,j}$ that include $l_{1,j}, l_{2,j}, \ldots,$ and $l_{K,j}$, respectively. Then, the SCBS transmits these binning indices in block $j + 1$ in K separate phases of durations $\beta_1, \beta_2, \ldots,$ and β_K. Therefore, in block j, the SCBS generates its signals for forwarding as follows:

$$\text{Phase } k\text{: } \mathbf{x}_{rk,j} = \sqrt{\rho_{rk}/(\beta_k N)}\mathbf{U}_{rk}(b_{k,j-1}), \tag{6}$$

where $\mathbf{U}_{rk}(b_{k,j-1})$ is an $N \times 1$ Gaussian random vector with zero mean and covariance $\mathbf{I}_N$ which conveys the codeword of the binning index $b_{k,j-1}$. Moreover, the phase durations $(\beta_1, \beta_2, \ldots, \beta_K)$ and the transmit powers $(\rho_{r1}, \rho_{r2}, \ldots, \rho_{rk})$ at each phase satisfy the following:

$$\sum_{k=1}^{K} \beta_k = 1, \quad \sum_{k=1}^{K} \rho_{rk} = P_r, \quad \beta_k \geq 0, \rho_{rk} \geq 0, \tag{7}$$

where P_r is the transmit power at the SCBS that also deploys power control by transmitting $\mathbf{x}_{rk,j}$ with (ρ_{rk}/β_k) in phase k.

3) The MCBS: performs sliding window decoding over two consecutive blocks (j and $j+1$) to separately and sequentially decode each bin index, quantization index and finally the message of each UE. Specifically, after ZF detection in (3), the received signals from the SCBS over K phases in block $j + 1$ are given by

$$\text{Phase } k\text{: } \tilde{\mathbf{y}}_{drk,j+1} = \mathbf{x}_{rk,j+1} + \mathbf{A}_{dr,j+1}^{H} \mathbf{z}_{dk,j+1}, \tag{8}$$

where $k \in \{1, 2, \ldots, K\}$. The Decoding for each UE is done in a similar approach. For UE$_1$, the MCBS uses $\tilde{\mathbf{y}}_{dr1,j+1}$ in (8) and $\tilde{y}_{d1,j}$ in (3) to sequentially decode

1) the bin index $\hat{b}_{1,j}$ using $\tilde{\mathbf{y}}_{dr1,j+1}$. Reliable decoding is ensured if the bin index rate R_{b1} satisfies [2]

$$R_{b1} \leq \beta_1 N \mathcal{C}\left(\frac{\rho_{r1}(M-N)}{\beta_1 N d_{dr}^{\alpha}}\right). \tag{9}$$

2) the quantization index $\hat{l}_{1,j}$ using $\tilde{y}_{d1,j}$ given that $\hat{l}_{1,j} \in \hat{b}_{1,j}$. Reliable decoding is ensured if

$$R_{q1} - R_{b1} \leq \log(\eta) - \log(\eta - P_1^2)$$
$$\text{where } \eta = \big(\frac{N d_{d1}^{\alpha}}{M-N} + P_1\big)\big(\frac{d_{r1}^{\alpha}}{N} + Q_1 + P_1\big). \tag{10}$$

The constraint is on $R_{q1} - R_{b1}$ instead of R_{q1} as the MCBS only looks for $\hat{l}_{1,j}$ such that $\hat{l}_{1,j} \in \hat{b}_{1,j}$.

3) UE$_1$ message $\hat{w}_{1,j}$ using $\tilde{y}_{d1,j}$ and $\hat{y}_{r1,j}(\hat{l}_{1,j})$. Reliable decoding is ensured if the message transmission rate R_1 satisfies

$$R_1 \leq \mathcal{C}\left(\frac{P_k(M-N)}{d_{dk}^{\alpha}} + \frac{P_k}{(d_{rk}^{\alpha}/N) + Q_k}\right) \triangleq I_1, \tag{11}$$

B. Achievable Rate Region

Let R_k denote the transmission rate for UE$_k$ for $k \in \{1, 2, \ldots, K\}$. The achievable rate region is determined by the rate constraints that ensure reliable decoding at the MCBS. These constraints are derived from the error analysis of the decoding rule at the MCSB as follows.

Theorem 1. *For K-UE massive MIMO HetNet, the achievable rate region of the QF-WZTD scheme consists of all K-tuples rate vectors $(R_1, R_2, \ldots, R_K)$ satisfying*

$$R_k \leq I_k(Q_k), \text{ s.t. } L_k(Q_k) \leq H_k(\beta_k, \rho_{rk}), \tag{12}$$

for all $k \in \{1, 2, \ldots, K\}$ and for all β_k and ρ_{rk} satisfying (7) where

$$I_k(Q_k) = \mathcal{C}\left(\frac{P_k(M-N)}{d_{dk}^{\alpha}} + \frac{P_k}{(d_{rk}^{\alpha}/N) + Q_k}\right),$$
$$L_k(Q_k) = \mathcal{C}\left(\frac{1}{Q_k}\Big[\frac{d_{rk}^{\alpha}}{N} + \frac{P_k}{1 + (P_k(M-N)/d_{dk}^{\alpha})}\Big]\right),$$
$$H_k(\beta_k, \rho_{rk}) = \beta_k N \mathcal{C}\left(\frac{\rho_{rk}(M-N)}{\beta_k N d_{dr}^{\alpha}}\right). \tag{13}$$

[1] This may reduce the average rate region by a factor of (1/B), but this factor becomes negligible as $B \to \infty$ [11].

Proof: $I_k(Q_k)$ ensures reliable decoding for UE$_k$ message as in (11) for $k=1$ while the constraints $L_k(Q_k) \leq H_k(\beta_k, \rho_{rk})$ is obtained from combining the quantization and binning indices constraints in (5), (9) and (10). In fact, $L_k(Q_k) \leq H_k(\beta_k, \rho_{rk})$ is a constraint on the quantization noise variance Q_k, such that the transmission rate of the binning index $b_{k,j-1}$ is bounded by the link from the SCBS to the MCBS. ■

Next, we derive the optimal Q_k^*, β_k^* and ρ_{rk}^* that maximize the rate region.

IV. Optimal Quantization at the SCBS

For practical implementation, it is important to specify the optimal quantization at the SCBS for each UE data stream. As the quantization levels increase, the quantizer becomes finer with smaller noise variances. However, finer quantization requires higher transmission rate for the quantization indices, which may not be sustained by the SCBS-MCBS link. Therefore, we derive the optimal quantization parameters $(Q_1^*, Q_2^*, \ldots, Q_K^*)$ that maximize the rate region in (12).

In Theorem 1, any boundary point of the rate region can be represented by the weighted sum rate $\sum_{i=1}^{K} \mu_i R_i$, where $\mu_i \in [0,1]$ is some priority weighting factor of UE$_i$ rate, while $\sum_{i=1}^{K} \mu_i = 1$. Thus, the rate region boundary is achieved by maximizing the weighted sum rate for some given $\mu_1, \mu_2, \ldots,$ and μ_K over $Q_1, Q_2, \ldots,$ and Q_K. Hence, the optimization problem is formulated as

$$\max_{Q_k, k\in\{1,2,\ldots,K\}} \sum_{k=1}^{K} \mu_k I_k(Q_k), \qquad (14)$$
$$\text{s.t. } L_k(Q_k) \leq H_k(\beta_k, \rho_{rk}),\ Q_k \geq 0.$$

The solution of problem (14) is given as follows.

Theorem 2. *The optimal Q_k^*, ρ_{rk}^* and β_k^* for $k \in \{1,2,\ldots,K\}$ for problem (14) are given as follows*

$$Q_k^* = \frac{(d_{rk}^{\alpha}/N)\big(1 + \frac{P_k(M-N)}{d_{dk}^{\alpha}}\big) + P_k}{\big(1 + \frac{P_k(M-N)}{d_{dk}^{\alpha}}\big)(\lambda_k^* - 1)}, \qquad (15)$$
$$\rho_{rk}^* = \beta_k^* P_r,\ \beta_k^* = \log(\lambda_k^*)/\log(\lambda_s),$$

where λ_k^* is obtained from the solution of the following GP problem:

$$\min_{\lambda_k, k\in\{1,2,\ldots,K\}} \prod_{k=1}^{K} \sum_{i_k=0}^{\mu_k L} \binom{\mu_k L}{i_k} \left(b_k a_k^{-1} \lambda_k^{-1}\right)^{i_k)}$$
$$\text{s.t. } \lambda_s^{-1} \prod_{k=1}^{K} \lambda_k = 1,\ \lambda_k^{-1} \leq 1. \qquad (16)$$

where L is the least common denominator (LCD) between the weighting factors $(\mu_1, \ldots, \mu_K)$ as each factor represents a fraction. Moreover,

$$a_k = 1 + P_k(M-N)/d_{dk}^{\alpha},\ k \in \{1,2,\ldots,K\}$$
$$b_k = P_k N/d_{rk}^{\alpha},\ \lambda_s = \left(1 + \frac{P_r(M-N)}{N d_{dr}^{\alpha}}\right)^N. \qquad (17)$$

Proof: The proof is obtained by three main steps. We first consider a QF-WZ scheme (no TD), which is an upper bound of QF-WZTD as it deploys joint decoding instead of separate decoding for the binning indices [11]. Then, we find the optimal phase durations and power allocations for the QF-WZTD scheme, and show that at optimality QF-WZTD coincides with the QF-WZ scheme. Finally, we transform the problem to a GP problem that can be solved by existing tools in [16].

1) We start the QF-WZ scheme where no TD is deployed and the SCBS generates a common codeword for all tuple of the k binning indices and transmits it (during the whole transmission block) to the MCBS. The MCBS jointly decodes all binning indices by using the received signal from the SCBS and then decodes the quantization indices and UEs' messages as in the proposed QF-WZTD scheme. Such a scheme achieves a rate region similar to (12) in Theorem 1, except the following constraint on $L_k(Q_k)$:

$$\sum_{k=1}^{K} L_k(Q_k) \leq N\mathcal{C}\left(\frac{P_r(M-N)}{N d_{dr}^{\alpha}}\right). \qquad (18)$$

Note that $I_k(Q_k)$ in (13) is maximized by minimizing Q_k. However, we can decrease Q_k as long as the constraint in (18) holds. Hence, the weighted sum rate is maximized when this constraint holds with equality.

2) Considering (18) with equality, let

$$N\mathcal{C}\left(\frac{P_r(M-N)}{N d_{dr}^{\alpha}}\right) = \log(\lambda_s) = \sum_{k=1}^{K} \log(\lambda_k),$$

where λ_s is given in (17) while $\prod_{k=1}^{K} \lambda_k = \lambda_s$. Then, from (18), we obtain Q_k as in (15). To ensure that $Q_k \geq 0$, we have the constraint $\lambda_k \geq 1$.

3) Considering $H_k(\beta_k, \rho_{rk})$ in (13), the proposed QF-WZTD scheme achieves the same performance of the QF-WZ binning when $H_k(\beta_k, \rho_{rk}) = \log \lambda_k$, which occurs with ρ_{rk}^* and β_k^* in (15).

4) By substituting Q_k in (15) into (14), the optimization problem in (14) becomes as follows:

$$\max_{\lambda_k, k\in\{1,2,\ldots,K\}} \sum_{k=1}^{K} \mu_k I_k(Q_k),\ \text{s.t.} \prod_{k=1}^{K} \lambda_k = \lambda_s,\ \lambda_k \geq 1, \qquad (19)$$

where $I_k(Q_k) = \log(a_k(a_k + b_k)) - \log\left(a_k + b_k \lambda_k^{-1}\right),$

5) As λ_k only appears in the negative part of $I_k(Q_k)$, the optimization problem in (19) can be reexpressed as follows

$$\min_{\lambda_k, k\in\{1,2,\ldots,K\}} \sum_{k=1}^{K} \mu_k \log\left(a_k + b_k \lambda_k^{-1}\right)$$
$$\text{s.t.} \prod_{k=1}^{K} \lambda_k = \lambda_s,\ \lambda_k \geq 1. \qquad (20)$$

In (20), while the objective function is convex, the equality constraint is not affine. Hence, the problem is not necessary convex. However, it can be transformed to a convex problem as shown next.

6) By changing the sum of logarithms to the logarithm of a product, optimizing this product is equivalent to (20). Hence, the optimization problem becomes as follows.

$$\min_{\lambda_k, k\in\{1,2,\dots,K\}} \prod_{k=1}^{K} \left(a_k + b_k\lambda_k^{-1}\right)^{\mu_k}$$

$$\text{s.t. } \lambda_s^{-1}\prod_{k=1}^{K}\lambda_k = 1,\ \lambda_k^{-1} \leq 1. \tag{21}$$

7) Next, in (21), let $a_k + b_k\lambda_k^{-1} = a_k(1 + b_k a_k^{-1}\lambda_k^{-1})$. Then, we can remove $\prod_{k=1}^{K} a_k^{\mu_k}$ form the objective function without affecting the optimization problem. The optimization further will not be affected by taking the objective function to the L^{th} power. Hence, we have

$$\min_{\lambda_k, k\in\{1,2,\dots,K\}} \prod_{k=1}^{K} \left(1 + b_k a_k^{-1}\lambda_k^{-1}\right)^{\mu_k L}$$

$$\text{s.t. } \lambda_s^{-1}\prod_{k=1}^{K}\lambda_k = 1,\ \lambda_k^{-1} \leq 1. \tag{22}$$

8) Last, in (22), each $\mu_k L$ is an integer. Hence, by applying multi-binomial expansion to the objective function, it can be expressed as a posynomial as in (16). Moreover, the inequality constraints are posynomials while the equality constraint is a monomial [16]. Hence, the optimization problem in (22) is GP, which can be solved by existing tools in [16].

■

Next, we discuss several proprieties of the proposed QF-WZTD scheme.

V. Discussion

Sections III and IV describe the transmission scheme and its optimal design. Here, we provide several remarks on the transmission scheme and its achievable rate region.

Remark 1. (**Impact of massive MIMO on the transmission design**) As stated in [3], massive MIMO system simplifies the quantization process at the SCBS as compared with a regular MIMO system. First, the optimal quantization element $Q_1, Q_2, \dots,$ and Q_K depend on the large scale fading and their number is equal to number of UEs (K). Optimizing these elements is much simpler than that of a regular MIMO system which requires optimizing the covariance matrix of the quantization noise vector to obtain the rate region boundaries [12], [13]. Second, because of the orthogonality property of massive MIMO [2] and TD transmission, the transmission and decoding for each UE's message is similar to the basic single UE relay channel in [11]. However, all UEs share the same relaying link from the SCBS to the MCBS. Therefore, the quantization at the SCBS is not simply optimized by considering multiple orthogonal single user relay channels. Further details are given in Section IV.

Remark 2. (**Impact of TD on the transmission design at the SCBS**) The TD transmission has several benefits. First, TD transmission leads to a small codebook size and simple decoding because of the separate codeword generation and separate decoding for each bin index, which is much simpler than joint transmission and joint decoding. Specifically, let n be the length of the transmitted codewords, the number of generated codewords will be $\sum_{k=1}^{K} 2^{nR_{bk}}$ for the TD transmission to represent each binning index separately. However, for joint transmission (without TD), the number of the generated codewords will be $2^{\sum_{k=1}^{K} R_{bk}}$ to represent each K-tuple of binning indices. Similar comparison holds for the decoding complexity at the MCBS. Therefore, TD transmission has a linear complexity with the number of UEs while the joint transmission has an exponential complexity. Note that the transmission from all UEs and the decoding of quantization indices and UEs' messages are the same in both schemes (i.e. with or without TD).

Furthermore, TD transmission is more flexible than joint transmission to handle various latency requirements for different UEs. With joint transmission, all UEs' messages that are sent in block j will be decoded at the end of block $j+1$. However, with TD transmission, the SCBS can send the bin index of the UE's data stream with the lowest latency requirement in phase 1 such that the MCBS decodes that UE's message with one phase of delay instead of the whole block as in joint transmission and decoding.

Although TD transmission requires additional optimization of the phase durations and power allocations, their optimal values can be conveniently obtained from the optimal quantization variables (Q_k^*) as shown in Section IV.

Remark 3. (**Impact of WZ binning on the transmission design**) WZ binning simplifies the decoding by facilitating sequential decoding for the binning index, quantization index and a UE's message [11]. WZ binning also reduces the codebook size: the SCBS only transmits the binning indices but not the quantization indices, and the number of binning indices is less than that of the quantization indices [11]. Moreover, although WZ binning is an extra encoding step at the SCBS, only minor computation is needed for sorting the quantization indices into equal-size groups.

Remark 4. (**Impact of massive MIMO, TD transmission and WZ binning on the rate region**) In general, QF-WZ scheme reduces the transmission rates from the general QF scheme (i.e., without WZ binning but with joint decoding for the messages and quantization indices [11]). However, for massive MIMO HetNet, both schemes achieves the same rate for the two-UE case [15]. It can be shown similarily that this conclusion also holds for K-UE case.

Remark 5. (**Impact of the optimal TD transmission on the achievable rate region**) Item 3 in the proof of Theorem 2 shows that even with TD, the proposed QF-WZTD scheme achieves the same performance as the QF-WZ scheme and hence, the general QF scheme without WZ binning or TD (see Remark 4). Furthermore, in the QF-WZTD scheme, the additional variables of phase durations and power allocation are conveniently optimized as direct functions of λ_k^* and hence incur no extra optimization. Moreover, by substituting ρ_{ri}^* in (15) into (12), it is optimal that the SCBS transmits each bin index with the same power P_r in each phase. Hence, there is no need for different power allocation in each transmission

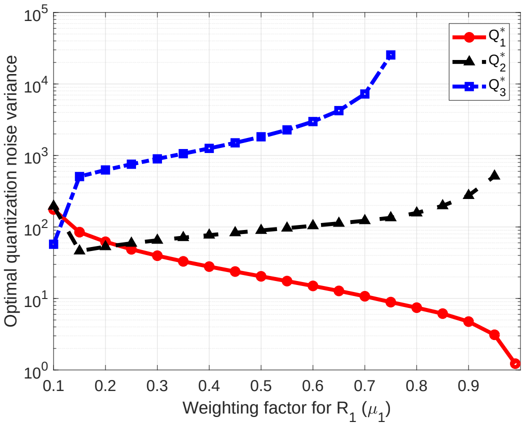

Fig. 3. Optimal quantization noise variances to maximize the weighted sum rate of the QF-WZTD scheme.

phase at the SCBS.

Remark 6. (**Quantization level implication**) In the proof of Theorem 2, item 7 implies that λ_k increases as (b_k/a_k) increases in order to minimize the objective function in (22). Consequently and considering Q_k in (15), item 7 implies that the SCBS deploys finer (resp. coarser) quantization for a UE data stream as its weighting factor increases (resp. decreases) and its link to the SCBS becomes stronger (resp. weaker) compared with that to the MCBS. Furthermore, for a UE with a very low weighting factor or a much weaker link to the SCBS than that to the MCBS, the SCBS does not quantize that UE's data stream and the MCBS decodes its message from only the received signal of that UE.

Remarks 1—6 show the efficiency and effectiveness of our scheme as it achieves the rate performance of more complex schemes.

VI. Numerical Results

We now provide numerical results for the optimal quantization noise variances, optimal phase durations and the weighted sum rate as shown in Theorem 2. In the simulations, we consider three UEs with the same transmission power $P_k = P$ for $k \in \{1, 2, 3\}$ while the SCBS's power is $P_r = 5P_k$. The SCBS (resp. MCBS) has 50 (resp. 500) antennas. The inter-node distances in meters are: $d_{dr} = 100$, $d_{d1} = 105$, $d_{d2} = 110$, $d_{d3} = 120$, $d_{r1} = 30$, $d_{r2} = 40$, and $d_{r3} = 50$. These distances are valid for 5G systems with small cell sizes at the order of 100 m while the path loss exponent $\alpha = 2.7$ is valid for cellular propagations [10]. We define the SNR as the received SNR at the MCBS from UE$_1$ as follows:

$$\text{SNR} = 10 \log_{10} \left(P_1 (M - N) / d_{d1}^{\alpha} \right). \tag{23}$$

In all figures, we set SNR = 1dB and the results are obtained versus μ_1 where $\mu_2 = 0.75(1 - \mu_1)$ and $\mu_3 = 0.25(1 - \mu_1)$.

Fig. 3 shows the optimal Q_1^*, Q_2^*, and Q_3^* that maximize the weighted sum rate. As expected from Remark 6, the SCBS performs finer quantization for UE$_1$ since it is the closest UE to the SCBS, i.e., has the best ratio of its link to the SCBS compared with that to the MCBS. Moreover, as the weighting factor of each UE rate increases, the SCBS performs finer quantization to send a clearer version of that UE's signal to the MCBS. On the other hand, for $\mu_1 \geq 0.8$ where $\mu_3 \leq 0.05$ and since UE$_3$ has the weakest link to the SCBS, the optimal quantization Q_3^* is very high (approaches ∞), i.e., the SCBS does not QF UE$_3$'s signal and the MCBS decodes it through the direct link only. Similarly for the optimal phase durations $(\beta_1^*, \beta_2^*, \beta_3^*)$ in Fig. 4, for UEs with better links to the SCBS and higher weighting factors, the SCBS sends their binning indices over longer durations to increase the total rate. For $\mu_1 \geq 0.8$, $\beta_3^* = 0$ since the SCBS does not QF UE$_3$'s signal.

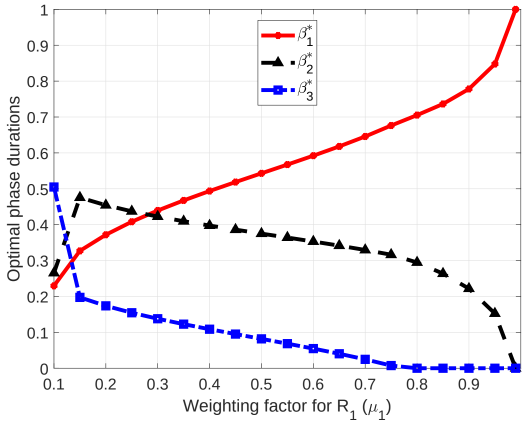

Fig. 4. Optimal phase durations to maximize the weighted sum rate of the QF-WZTD scheme.

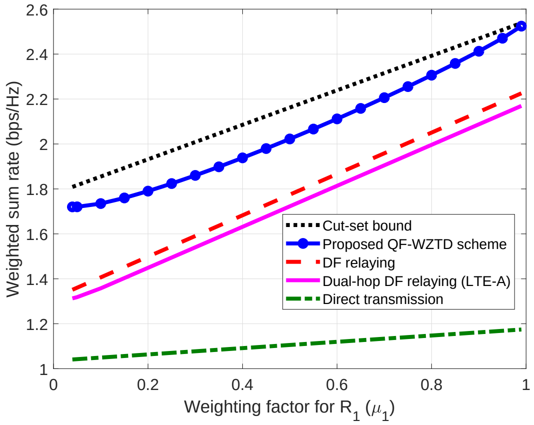

Fig. 5. Weighted sum rate comparison between different schemes with $d_{r1} = 30$, $d_{r2} = 40$ and $d_{r3} = 50$.

Fig. 5 compares between the maximum weighted sum rate for the proposed QF-WZTD scheme, the cut-set outer bound [11], the full-duplex DF scheme [6], the half duplex dual-hop DF scheme (used in LTE-A [10]) and the direct transmission. Results show that our QF-WZTD is close the cut-set bound and outperforms the DF scheme, LTE-A and the direct transmission. Full-duplex DF relaying is always better than half-duplex DF relaying in the LTE-A systems. Note that if the UEs get closer to the SCBS, DF relaying can outperform the QF relaying [11]. However, For the channel setting in Fig. 5, the QF-WZTD is the preferred scheme.

VII. Conclusion

We have utilized massive MIMO features to propose a simple yet efficient uplink transmission scheme for multiple UEs in a HetNet. The proposed scheme is based on QF relaying, Wyner-Ziv binning and multiple-timeslot transmission at the SCBS, and separate and sequential decoding at the MCBS. Such encoding and decoding techniques have linear complexity in the number of UEs. However, they lead to the same rate region of a joint-encoding and joint-decoding scheme with exponential complexity. To maximize the rate region, we have formulated the weighted sum rate problem to obtained the optimal quantization levels and, in turn, the optimal transmission timeslot durations. Our simulation results show that for some channel setting, the proposed scheme performs close to the cut-set outer bound and can substantially outperform several existing alternatives. Furthermore, finer quantization should be used for UEs' data streams as their weighting factors increase and their UE-SCBS links become stronger compared with the UE-MCBS links.

References

[1] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Sept. 2004.

[2] 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.

[3] E. Bjrnson *et al.*, "Massive MIMO: ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[4] E. Bjrnson, L. Sanguinetti, and M. Kountouris, "Deploying dense networks for maximal energy efficiency: small cells meet massive MIMO," in *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, Apr. 2016, pp. 832–847.

[5] Z.-Y. Zhang and W. Lyu, "Interference coordination in full-duplex HetNet with large-scale antenna arrays," *Front. Inform. Technol. Electron. Eng.*, vol. 18, no. 6, pp. 830–840, 2017.

[6] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3037–3063, Sept. 2005.

[7] A. Abu Al Haija and C. Tellambura, "Small-Macro Cell Cooperation for HetNet Uplink Transmission: Spectral Efficiency and Reliability Analyses," in *IEEE J. Sel. Areas Commun.*, vol. 35, no. 1, Jan. 2017.

[8] L. Sankaranarayanan, G. Kramer, and N. B. Mandayam, "Capacity theorems for the multiple-access relay channel," in *IEEE Allerton*, Oct. 2004.

[9] S. H. Lim, Y.-H. Kim, A. El Gamal, and S.-Y. Chung, "Noisy network coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, May 2011.

[10] J. Korhonen, *Introduction to 4G mobile communications*. Artech House, 2014.

[11] R. El Gamal and Y.-H. Kim, *Network Information Theory*, 1st ed. Cambridge University Press, 2011.

[12] S. Simoens, O. Muoz-Medina, J. Vidal, and A. del Coso, "Compress-and-forward cooperative MIMO relaying with full channel state information," *IEEE Trans. Signal Process.*, vol. 58, pp. 781–791, Feb. 2010.

[13] X. Lin, M. Tao, and Y. Xu, "MIMO two-way compress-and-forward relaying with approximate joint eigen-decomposition," *IEEE Commun. Lett.*, vol. 17, no. 9, pp. 1750–1753, Sep. 2013.

[14] Y. Zhou and W. Yu, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4138–4151, Aug. 2016.

[15] A. Abu Al Haija, M. Dong, B. Liang, G. Boudreau, and S. H. Seyedmehdi, "Design and simplification of quantize-forward relaying in massive MIMO HetNets," in *IEEE ICC Workshop: 5G ultra dense network*, May 2018.

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed. Cambridge University Press, 2004.

[17] R. Mudumbai, D. Brown, U. Madhow, and H. Poor, "Distributed transmit beamforming: challenges and recent progress," *IEEE Commun. Mag.*, vol. 47, no. 2, pp. 102–110, Feb. 2009.

[18] D. Bharadia, E. Mcmilin, and S. Katti, "Full duplex radios," *ACM, SIGCOMM*, vol. 43, no. 4, pp. 375–386, Aug. 2013.

[19] M. Aref, "Information flow in relay networks," *Ph.D. dissertation, Stanford Univ.*, Oct. 1980.