Scheduling for Backhaul Load Reduction in CoMP

Tilak Rajesh Lakshmana, Jingya Li, Carmen Botella†, Agisilaos Papadogiannis, and Tommy Svensson
Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden
†Institute of Robotics and Information & Communication Technologies (IRTIC), Universitat de València, València, Spain
{tilak, jingya.li, agisilaos.papadogiannis, tommy.svensson}@chalmers.se, carmen.botella@uv.es

***Abstract*—Coordinated multi-point (CoMP) transmission has received a lot of attention, as a way to improve the system throughput in an interference limited cellular system. For joint processing in CoMP, the user equipments (UEs) need to feed back the channel state information (CSI), typically to their serving base stations (BSs). The BS forwards the CSI to a central coordination node (CCN) for precoding. These precoding weights need to be forwarded from the CCN to the corresponding BSs to serve the UEs. In this work, a feedback load reduction technique is employed via partial joint processing to alleviate the CSI feedback overhead. Similarly, to achieve backhaul load reduction due to the precoding weights, scheduling approaches are proposed. The state of the art block diagonalization solution is compared with our proposed constrained and unconstrained scheduling. Our main contribution is the method of choosing the best subset of the BSs and UEs at the CCN that yields the best sum rate under the constraint of efficient backhaul use. In particular, with constrained scheduling, the choice of a smaller subset proportionally reduces the backhaul load. Simulation results based on a frequency selective WINNER II channel model, show that our proposed constrained scheduling outperforms the block diagonalization approach in terms of the average sum rate per backhaul use.**

***Index Terms*—Backhaul Load Reduction, Scheduling, CoMP, Partial Joint Processing, Zero Forcing**

I. Introduction

In future cellular communication systems, coordinated multi-point (CoMP) transmission is a promising technique proposed to improve the throughput of the user equipments (UEs) at the cell edge, being limited by interference [1]-[2]. To realize these gains, the UEs need to feed back the channel state information (CSI) typically to their serving base station (BS). The CSI is forwarded to the central coordination node (CCN) to form the aggregated channel matrix that is used to create the precoding matrix to jointly mitigate interference. To reduce the overhead of feeding back the CSI, clusters of BSs are formed [2]. In particular, partial joint processing (PJP) was proposed in [3] for feedback load reduction, in which dynamically overlapping clusters of BSs are formed.

PJP can be seen as a framework that attempts to categorize the trade off between how much load can be reduced for a given perceivable loss in the system performance. In this regard, the CSI feedback load reduction can be achieved by limiting the quantity of feedback by the UEs. To this end, a relative thresholding is proposed in [3], where the UE only feeds back the CSI for a set of BSs links that fall within a threshold relative to its strongest BS. The CSI of the BSs that fall outside this threshold are modeled as *zeros* in the aggregated channel matrix. Likewise, in this context, the signaling in the backhaul (BSs-CCN) is primarily due to the distribution of the precoding weights from the CCN to the cooperating BSs. Limiting the feedback causes the aggregated channel matrix to be sparse and poses problems in the case of precoders such as zero forcing (ZF). Under these circumstances, achieving an efficient use of the backhaul is difficult. Hence, the structure in the aggregated channel matrix formed at the CCN needs to be exploited, such that the zeros are correspondingly preserved in the precoder matrix for reducing the backhaul load. To achieve this, backhaul load reduction can be carried out at the medium access control (MAC) layer or physical (PHY) layer, as proposed in [4].

The MAC layer approach is a scheduling based scheme, where disjoint BS subgroups are formed, such that the aggregated channel matrix is block diagonal. The main benefit of this approach is that the inverse of a block diagonal matrix is still block diagonal and that the zeros are preserved. The PHY layer approach is a ZF precoding approach, where the aggregated channel matrix is repeated such that a block diagonal structure is created, and the precoding matrix is created with zeros where needed. The limitations of the PHY layer approach are discussed in [5].

In this paper, we propose a constrained scheduling (CS) and an unconstrained scheduling (US) for backhaul load reduction. In the CS approach, an exhaustive search is carried out to find the best subset of the aggregated channel matrix, such that zeros are avoided in the aggregated channel matrix. This approach directly aims at reducing the backhaul load as the zeros are disallowed. The US approach is similar to the CS approach except that the zeros are allowed to be present in the aggregated channel matrix, and the backhaul load reduction is achieved by explicit nulling of the precoding weights corresponding to the zeros in the aggregated channel matrix. We compare our techniques with the MAC layer scheduling based block diagonalization (BD) technique proposed in [4]. The BD approach achieves the backhaul load reduction by forming a block diagonal structure of the aggregated channel matrix. All the above techniques are evaluated with the PJP based CSI feedback load reduction as proposed in [3]. To summarize our contribution, our proposed CS and US algorithms (i) reduce the backhaul load, (ii) significantly increase the performance, as a larger feasible subset is considered compared to the BD approach, which poses a stricter constraint of being block diagonal, and (iii) the best subset of BSs and UEs are clustered.

The paper is organized as follows, in Section II the system

This work has been supported by the Swedish Agency for Innovation Systems (VINNOVA), within the P36604-1 MAGIC project and the Swedish Research Council VR under the project 621-2009-4555 Dynamic Multipoint Wireless Transmission. The computations were performed on C^{3}SE computing resources.

model is introduced with the focus on how the feedback and backhaul load reduction are achieved in a frequency selective channel. Discussions on the scheduling strategies for backhaul load reduction are presented in Section III. The performance of these scheduling strategies are discussed in Section IV and finally the main results are concluded in Section V. The notation used in this paper is summarized in the footnote.

II. System Model

Consider the cluster layout as shown in Fig. 1, where $K = |\mathcal{K}|$ single antenna BSs need to serve $M = |\mathcal{M}|$ single antenna UEs. $\mathbf{h}_m = [h_{m,1}, h_{m,2}, ..., h_{m,K}]$ is the CSI of the links from the K BSs to the mth UE. In this work, we study block-fading channels where the CSI available at the CCN is considered to be error free, i.e., the quantization loss and the backhaul delays are assumed to be negligible [6]. In a wideband system, consisting of a number of subcarriers, each UE feeds back the CSI for a given frequency resource. The CSI being fed back can be applied to a group of subcarriers. The CSI feedback process is performed under the PJP framework proposed in [3], using a relative active set thresholding as summarized in [5, Algo.1]. The CSI feedback from the mth UE based on the channel from K BSs can be represented as

$$\widetilde{\mathbf{h}}_m = \mathbf{h}_m \odot \mathbf{t}_{x,m}, \tag{1}$$

where $\mathbf{t}_{x,m}(k) \in \{0,1\}, \forall k = 1,\ldots,K$. The operation is independently performed over a collection of subcarriers (frequency adaptive thresholding) [7], where all the M UEs report the CSI for all the subcarriers. When the mth UE feeds back the CSI of kth BS, it is denoted "1" while a "0" denotes that the CSI was not fed back. The feedback load reduction can be seen as a masking operation by a binary threshold vector $\mathbf{t}_{x,m}$ via element wise multiplication with the CSI measured by the mth UE. The subscript x denotes the threshold value in dB. When $x = 0$ dB, it represents a scenario where only the strongest BS is serving the UE, while the threshold of $x = \infty$ dB represents that all the links are fed back. If $K = 3$ then the mth UE will feed back in one of the following ways: $\mathbf{t}_{x,m} = \{\{1,0,0\}, \{0,1,0\}, \{0,0,1\}, \{1,1,0\}, \{0,1,1\}, \{1,0,1\}$, and $\{1,1,1\}\}$. However, with relative thresholding [3], $\mathbf{t}_{x,m} = \{0,0,0\}$ will never occur, as it enables the UE to feedback at least its strongest BS.

For backhaul load reduction, consider a subset of the cluster formed at the CCN, with $N = |\mathcal{N}|$ UEs and $L = |\mathcal{L}|$ BSs, where $\mathcal{N} \subseteq \mathcal{M}$ and $\mathcal{L} \subseteq \mathcal{K}$. The maximum number of BSs that can be chosen is $L_{\max} = K$ and the maximum number of UEs that can be scheduled in a given subcarrier group/resource

Boldface upper-case letters represent matrices, $\mathbf{X}$, boldface lower-case letters represent vectors, $\mathbf{x}$, and italics represent scalars, x. The $\mathbb{C}^{m\times n}$ is a complex valued matrix of size $m \times n$. The $(\cdot)^T$ and $(\cdot)^H$ is the transpose and conjugate transpose, respectively. $\mathbf{E}_x\{\cdot\}$ is the expectation with respect to x. The $||\cdot||_F$ is the Frobenius norm. $\mathbf{X}(i,j)$ is the (i,j)th element of matrix $\mathbf{X}$ and $\mathbf{x}(i)$ is the ith element of the vector $\mathbf{x}$. The ith row of a matrix $\mathbf{X}$ is $\mathbf{X}(i,:)$ and the jth column of a matrix $\mathbf{X}$ is $\mathbf{X}(:,j)$. The sets are indicated in calligraphic letters and $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$. The operator $\odot$ is the element wise multiplication.

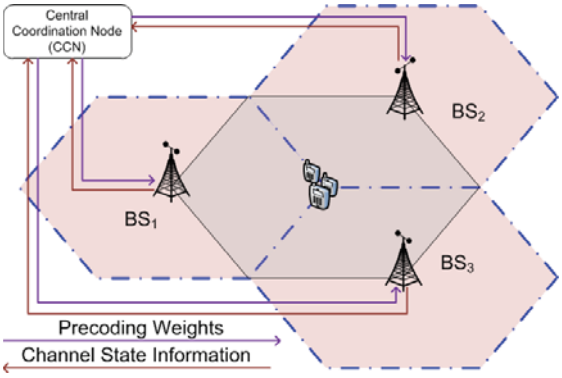

Fig. 1. The cluster layout

is $N_{\max} = L_{\max}$. The discrete time signal received at the N selected UEs, $\mathbf{y} \in \mathbb{C}^{N\times 1}$ is

$$\mathbf{y} = \mathbf{H}\overline{\widetilde{\mathbf{W}}}\mathbf{x} + \mathbf{n}, \tag{2}$$

where $\mathbf{H} \in \mathbb{C}^{N\times L}$ is the channel matrix for the subset of the cluster. $\widetilde{\mathbf{W}} \in \mathbb{C}^{L\times N}$ is the precoding matrix and $\mathbf{n}$ is the receiver noise at the UEs, which are spatially and temporally white with variance σ^2.

A linear ZF precoding is considered in this work. The precoding matrix is firstly calculated as the Moore-Penrose pseudoinverse of the aggregated channel matrix $\widetilde{\mathbf{H}}$

$$\widetilde{\mathbf{W}} = \widetilde{\mathbf{H}}^H(\widetilde{\mathbf{H}}\widetilde{\mathbf{H}}^H)^{-1}, \tag{3}$$

where $\widetilde{\mathbf{H}} = \mathbf{H} \odot \mathbf{T}_x$ and $\mathbf{T}_x = \left[\mathbf{t}_{x,1}^T, \mathbf{t}_{x,2}^T, \ldots, \mathbf{t}_{x,N}^T\right]^T$. Then, the columns of $\widetilde{\mathbf{W}}$ are normalized to have a unit norm [4]. Finally, based on equal user rate power allocation [8], the precoding matrix can be obtained as

$$\overline{\widetilde{\mathbf{W}}} = \sqrt{\frac{P_{\max}}{\left(\max\limits_{l=1,\ldots L} ||\widetilde{\mathbf{W}}(l,:)||_F^2\right)}} \cdot \widetilde{\mathbf{W}}, \tag{4}$$

where $P_{\max}$ is the maximum power at which a BS can transmit on a given resource, i.e., we are not considering optimal power allocation over the parallel resources. The signal to interference plus noise ratio (SINR) for the nth UE is given as

$$\text{SINR}_n = \frac{||\mathbf{h}_n\overline{\widetilde{\mathbf{W}}}(:,n)||^2}{\sum\limits_{j\in\mathcal{N},j\neq n}||\mathbf{h}_n\overline{\widetilde{\mathbf{W}}}(:,j)||^2 + \sigma^2}. \tag{5}$$

The sum rate in bps/Hz for scheduling the N different UEs on the same frequency/time resource is

$$R_{\text{tot}} = \sum_{n\in\mathcal{N}} \log_2(1 + \text{SINR}_n). \tag{6}$$

Due to feedback load reduction, the channel matrix $\widetilde{\mathbf{H}}$ might have zero elements depending on the threshold x dB. Hence, a sparse channel matrix is used to obtain the precoding matrix, $\overline{\widetilde{\mathbf{W}}}$. The zeros in $\widetilde{\mathbf{H}}$ pose problems for the ZF precoder for backhaul load reduction. For example, if the nth UE does not feed back the CSI for the lth BS, then $\widetilde{\mathbf{H}}(n,l) = 0$. Applying

the pseudoinverse in (3), the sparse aggregated channel matrix $\widetilde{\mathbf{H}}$ of size $N \times L$, will create $\widetilde{\mathbf{W}}$ of size $L \times N$, however, it could lead to $\widetilde{\mathbf{W}}(l,n) \neq 0$. This will lead to unnecessary backhaul, given that the UE has not fed back the CSI while the ZF solution still tries to serve the nth UE from the lth BS. Also, consider the situation where the nth UE has fed back the CSI for the $(l+1)$th BS such that $\widetilde{\mathbf{H}}(n,l+1) \neq 0$, however (3) might lead to a situation where $\widetilde{\mathbf{W}}(l+1,n) = 0$. This is poor backhauling as the uplink resources are already being spent for the nth UE to feed back the CSI.

Hence, suitable scheduling strategies need to be developed in achieving an efficient use of the backhaul. These are discussed in the subsequent section.

III. Scheduling

Let a set of $\mathcal{N} \subseteq \mathcal{M}$ UEs be chosen to be served from a set of $\mathcal{L} \subseteq \mathcal{K}$ BSs. To maintain orthogonality with a linear ZF precoder, the number of UEs chosen are $N = |\mathcal{N}|$ and the BSs chosen are $L = |\mathcal{L}|$, such that $N \leq L$. The particular choice of the set of UEs and BSs are driven by the combination that maximizes the sum rate as

$$\{\mathcal{L}^*, \mathcal{N}^*\} = \arg \max_{\{\mathcal{L},\mathcal{N}:|\mathcal{N}|\leq|\mathcal{L}|\}} \sum_{n\in\mathcal{N}} \log_2\left(1+\widehat{\text{SINR}}_n\right) \quad (7)$$

$$\widehat{\text{SINR}}_n = \frac{\sum_{l\in\mathcal{L}} \widetilde{\mathbf{H}}(n,l)\,\mathbf{W}(l,n)}{\sum_{\substack{j\neq n \\ j\in\mathcal{N}}} \sum_{l\in\mathcal{L}} \widetilde{\mathbf{H}}(n,l)\,\mathbf{W}(l,j) + \sigma^2}, \quad (8)$$

where $\widetilde{\mathbf{H}}$ is the channel sub-matrix of size $N \times L$ related to the set $\{\mathcal{N},\mathcal{L}\}$ of UEs and BSs. Applying (3) and (4) to this $\widetilde{\mathbf{H}}$ results in the precoding matrix $\mathbf{W}$. In the following subsections, we evaluate the scheduling strategies considered in this work.

A. Block Diagonalization (BD)

In [4], a MAC-layer approach is proposed where disjoint subgroups of BSs are formed to preserve the block diagonal structure of the aggregated channel matrix. An important property of a block diagonal structure is that it is preserved even under matrix inversion. This property is key to backhaul load reduction which conserves an equivalent feedback load reduction for the scheduled UEs belonging to the set $\mathcal{N}$, i.e., if $\widetilde{\mathbf{H}}(n,l) = 0$ then $\widetilde{\mathbf{W}}(l,n) = 0$. However, it should be noted that this BD approach based on [4] always requires $N_{\max}$ UEs to be scheduled. Therefore, with feedback load reduction, M UEs feeding back the CSI results in $N_{\max} = L_{\max} = K$ UEs being scheduled. The choice of $N_{\max}$ corresponds to an exhaustive search for the best combination of UEs that maximizes the sum rate given that the aggregated channel matrix is block diagonal. The BD approach can be summarized as choosing the combination of the UEs as in Algorithm 1. The block diagonal channel matrix $\widetilde{\mathbf{H}}_{\text{BD}}$ is extracted based on $\mathbf{T}_x$ and correspondingly $\widetilde{\mathbf{W}}_{\text{BD}}$ is obtained from (3).

Due to the BD structure, the positions of zeros in $\widetilde{\mathbf{H}}_{\text{BD}}$ and $\widetilde{\mathbf{W}}_{\text{BD}}$ are identical, and the aggregated channel matrix needs to be a square matrix such that $N_{\max} = L_{\max}$. This gives rise to some ill-effects. Consider $N < N_{\max}$ and $L < L_{\max}$ then $N < L$ is not considered which could potentially produce a better sum rate translating to a better system performance. This is due to the stringent constraint of the BD approach that $N_{\max} = L_{\max} = K$, where the feasible set is considerably reduced. Also, the feedback can be significantly reduced at the cluster center, via small relative thresholds [7].

Algorithm 1 BD approach: Note that the BD algorithm [4] always considers $L = L_{\max} = K$.

```
M UEs feed back the CSI as defined in Section II
for every N from M such that N_max = |N| = L_max do
    Form T_x(n,l) ∈ {0,1}, ∀n = 1,...,N_max; ∀l = 1,...,L_max
    if permuted T_x is block diagonal then
        Found T_x to have a block diagonal structure, evaluate (7)
        Save T_x based on the best {L*, N*} achieved so far
    else
        Failed to find a block diagonal structure
    end if
end for
return Schedule the subset formed with {L*, N*} using T_x
```

For example, a threshold of 5 dB creates a sparse aggregated channel matrix which is difficult to block diagonalize. One of the ways to overcome this limitation of the BD approach is to increase the threshold towards infinity. However, this increases the feedback load. On the contrary, a full aggregated channel matrix with few zeros also renders the BD approach difficult to realize. As a generalization, the BD approach proposed in [4] can be extended to consider the cases when $N = L < K$. However, in this work, we confine our study to the original algorithm proposed in [4].

B. Unconstrained Scheduling (US)

With feedback load reduction, the aggregated channel matrix is sparse depending on the threshold. The choice of a feasible subset of BSs and UEs, $\{\mathcal{L}^*, \mathcal{N}^*\}$, that produces the best sum rate is summarized in Algorithm 2.

Algorithm 2 US approach

```
M UEs feed back the CSI as defined in Section II
Assign L = K
while L ≥ 1 do
    for L : L ⊆ K; |L| = L do
        for every N from M such that N = |N| ≤ L do
            Form T_x(n,l) ∈ {0,1}, ∀n = 1,..,N; ∀l = 1,..,L
            Evaluate (7)
            Save T_x based on the best {L*, N*} achieved so far
        end for
    end for
    L = L − 1
end while
return Schedule the subset formed with {L*, N*} using T_x
```

The scheduled BSs and UEs in matrix form can be written as $\mathbf{T}_x(n,l) \in \{0,1\}, \forall n = 1,\ldots,N$ and $\forall l = 1,\ldots,L$. Hence, the sparse aggregated channel matrix can be written as $\widetilde{\mathbf{H}}_{\text{US}} = \mathbf{H} \odot \mathbf{T}_x$. Compared to the BD approach, the US approach has a flexibility in considering $N \leq L$, and $\widetilde{\mathbf{W}}$ is obtained from $\widetilde{\mathbf{H}}_{\text{US}}$ by applying (3) and (4). This is followed by explicit

nulling of the precoded weights as $\widetilde{\widetilde{\mathbf{W}}}_{\mathrm{US}} = \widetilde{\widetilde{\mathbf{W}}} \odot (\mathbf{T}_x)^T$, to achieve backhaul load reduction based on the nulls due to feedback load reduction. However, explicit nulling gives rise to multi-user interference to remain in the system. It should be noted that the explicit nulling is automatically taken care of in the BD approach in Section III-A. Explicit nulling seems like an intuitive approach but the ZF precoder has its own limitations when there are zeros in the aggregated channel matrix (see Section II).

C. Constrained Scheduling (CS)

The CS approach is similar to the US approach with an important constraint that the aggregated channel matrix $\widetilde{\mathbf{H}}_{\mathrm{CS}}$ is full due to the proper selection of UEs and BSs, i.e., $\widetilde{\mathbf{H}}_{\mathrm{CS}} = \mathbf{H} \odot \mathbf{T}_x$, where $\widetilde{\mathbf{H}}_{\mathrm{CS}} \in \mathbb{C}^{N \times L}$ and $\widetilde{\mathbf{H}}_{\mathrm{CS}}(i,j) \neq 0, \forall i,j$ as $\mathbf{T}_x(n,l) \in \{1\}, \forall n = 1,\ldots,N$ and $\forall l = 1,\ldots,L$. This simplifies the ZF in (3). The CS approach is summarized in Algorithm 3. The main advantage of this approach is that the backhaul load reduction is automatically achieved by this constrained scheduling approach as smaller subset of a matrix, $\widetilde{\mathbf{H}}_{\mathrm{CS}}$, is formed from $\mathbf{H}$. Also, multi-user interference is removed from the system.

Algorithm 3 CS approach

```
M UEs feed back the CSI as defined in Section II
Assign L = K
while L ≥ 1 do
  for L : L ⊆ K; |L| = L do
    for every N from M such that N = |N| ≤ L do
      Form T_x(n,l) ∈ {1}, ∀n = 1,...,N; ∀l = 1,...,L
      Evaluate (7)
      Save T_x based on the best {L*, N*} achieved so far
    end for
  end for
  L = L − 1
end while
return Schedule the subset formed with {L*, N*} using T_x
```

Illustrative Example: To illustrate the above algorithms with an example, consider $\mathbf{T}_x = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. This subset is feasible with the BD approach, while the CS approach requires the zeros to be removed. Hence, a feasible subset $\mathbf{T}_x$ after removing the zeros can be any of these $\left\{ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \end{bmatrix}, [1] \right\} \Rightarrow \mathcal{S}_{\mathrm{CS}} = \{2\times 2, 1\times 2, 1\times 1\}$ while $\{3\times 3, 2\times 3, 1\times 3\} \notin \mathcal{S}_{\mathrm{CS}}$ for this particular case of $\mathbf{T}_x$. As for the US approach, all possible combinations are feasible. From our proposed algorithms, what clearly falls out is that they offer an inherent seamless mode switching capability between CoMP and single cell 1×1 scenario. When N users are selected to be served from L BSs, they are expressed as $N \times L$. Table I summarizes the possible combinations of the various user scheduling strategies described above. In all the scheduling strategies, it should be noted that the UEs that are not currently being served can be expected to be served in another resource, thereby achieving user fairness.

Table I
SUMMARY OF THE SCHEDULING APPROACHES WITH $K = 3$

	BD	US	CS
Feasible Set†, $\mathcal{S}_{\mathrm{algo}}$	$\{3\times 3\}$	$\{3\times 3, 2\times 3, 2\times 2, 1\times 3, 1\times 2, 1\times 1\}$	$\{3\times 3, 2\times 3, 2\times 2, 1\times 3, 1\times 2, 1\times 1\}$
Search	Exhaustive	Exhaustive	Exhaustive
Cardinality	$\lvert\mathcal{S}_{\mathrm{BD}}\rvert < \lvert\mathcal{S}_{\mathrm{CS}}\rvert$	$\lvert\mathcal{S}_{\mathrm{US}}\rvert$	$\lvert\mathcal{S}_{\mathrm{CS}}\rvert \leq \lvert\mathcal{S}_{\mathrm{US}}\rvert$
Zeros‡	Allowed	Allowed	Not Allowed
Interference	Removed	Partially	Removed

† The subscript "algo" refers to BD or US or CS.
‡ The zeros in the aggregated channel matrix, $\widetilde{\mathbf{H}}$, formed at the CCN.

IV. PERFORMANCE EVALUATION

Consider the cluster center where M single antenna UEs moving at 3 kmph are dropped as shown in Fig. 1. The radius of the cell is $R = 500$ m. These UEs are uniformly dropped in an ellipsoid in $\mathbb{R}^2$, whose center is the cluster center. The major and minor axis of the ellipsoid are $(2\Delta x, 2\Delta y)$ where $0 \leq \Delta x \leq \frac{R}{16}$, $0 \leq \Delta y \leq \frac{h/2}{16}$ and h is the height of the hexagon or cluster. $K = 3$ single antenna BSs are positioned as shown in Fig. 1. A realistic WINNER II channel model [9] corresponding to scenario B1: urban micro-cell, non-line of sight with pathloss and shadow fading is considered with 500 independent channel realizations at 2 GHz center frequency. The signal to noise ratio at the cell-edge (reference value for one user located at the cell-edge) is fixed at 15 dB. For the B1 scenario, the channel provided by the WINNER II model is converted to the frequency domain with a 256-fast Fourier transform, where 32 consecutive subcarriers correspond to one resource for simplicity. The feedback load reduction is performed for one such resource, $\mathbf{T}_x$, where x takes the values $0, 5$ and 40 dB. The results presented are averaged over the Monte Carlo simulations over all the resources.

Figure 2 shows the average sum rate of the various scheduling algorithms considered in this work. As expected for lower thresholds, the unconstrained scheduler, US, performs better than the constrained scheduler, CS. However, this is achieved at the expense of the backhaul. This is due to the smaller sets of $\mathcal{L}$ and $\mathcal{N}$ being formed with CS unlike US. For CS and US, the sum rate increases with the increase in the feedback threshold. However in the case of BD, for lower number of UEs, the 0 dB threshold outperforms the 5 dB. This is related to Fig. 3 where the original BD algorithm is unable to find a block diagonal structure. The BD 0 dB case has a better chance of finding an identity matrix that results in block diagonal struture than the BD 5 dB, as the 0 dB thresholding maps to the UEs feeding back atleast the strongest BSs, while the 5 dB allows the UE to feed back more BSs, thereby making it hard to find a block diagonal structure. However, the situation improves when the number of UEs increase and the scheduler is able to find a block diagonal structure. Theoretically, an ∞ dB threshold is the case when CS, US and the generalized BD scheduling algorithm converge to the same solution.

The BD algorithm performs reasonably well in terms of sum rate, however, it is not feasible when the number of UEs are small with lower threshold. Figure 3 captures this in terms of the probability of failure to find a block diagonal

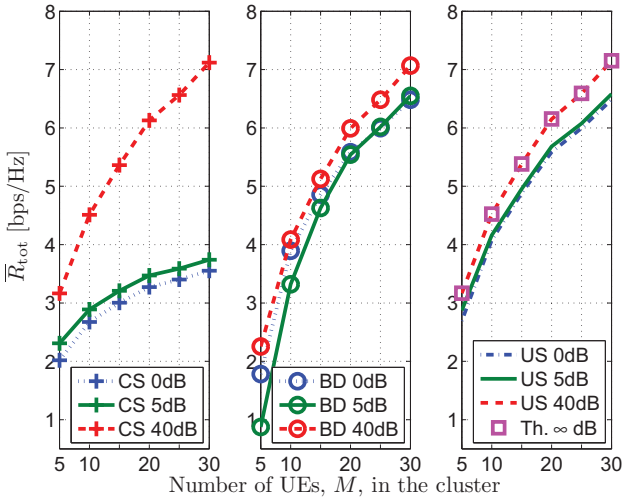

Fig. 2. Average sum rate versus the increase in the number of UEs.

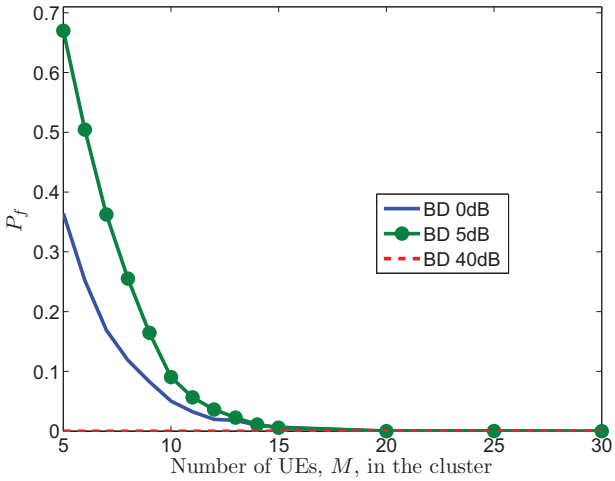

Fig. 3. Probability of failure to find a block diagonal subset, P_f

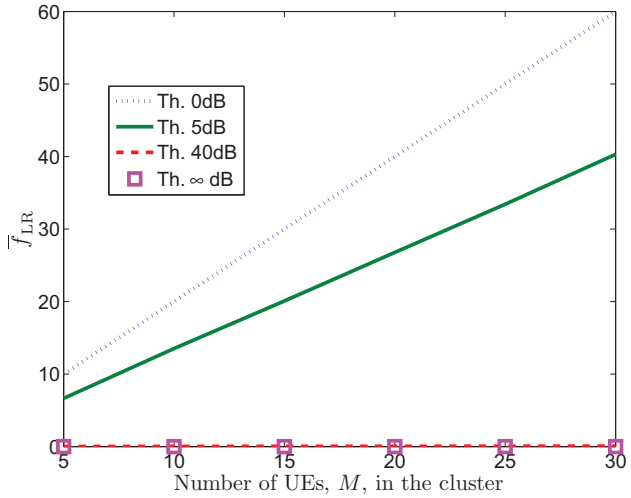

Fig. 4. Average feedback load reduction, $\overline{f}_{\text{LR}}$, achieved via PJP

subset of UEs, P_f, such that $N_{\max} = L_{\max} = K = 3$, for a given threshold for feedback load reduction. This failure maps to Algorithm 1, step 8. P_f goes to zero when the number of UEs exceeds 25 for all the thresholds considered in this work. With small number of UEs, the failure is due to the relative thresholding. The ratio of the number of unsuccessful attempts to the total number of attempts to find a block diagonal structure when performing the exhaustive search is 77.8%, 94.9%, and 0.6% for 0 dB, 5 dB, and 40 dB, respectively. These values do not change with the increase in the number of UEs. Let us consider the number of UEs to be 30. This translates to the total number of attempts being $\binom{30}{3} = \frac{30!}{3!27!} = 4060$. With a feedback load reduction threshold of 5 dB, corresponding to the cluster center [10, Fig. 4.19], the BD approach cannot be evaluated for potentially 94.9% of the time. With a threshold of 40 dB, the failure to find a block diagonal structure is as low as 0.6%, this is due to the fact that a bigger threshold allows the UE to feed back the CSI from more BSs, causing $\widetilde{\mathbf{H}}$ to be a full matrix more often than not. Hence, the BD procedure can easily be applied to a 40 dB threshold. However, it should be noted that this failure can be avoided if the BD approach in [4] is generalized, such that the subsets $\{2\times 2, 1\times 1\}$ are also included. This is treated as part of our future work.

We define the average feedback load reduction, $\overline{f}_{\text{LR}}$ as the average of the number of zeros in a sparse aggregated channel matrix $\widetilde{\mathbf{H}} \in \mathbb{C}^{M\times K}$ i.e., the cardinality of set $\mathcal{S}_{\text{FB}} = \left\{\widetilde{\mathbf{H}}\,(i,j) = 0, \forall i,j \in \mathbb{N}^{+}, i \leq M, j \leq K\right\}$. The average feedback load reduction is calculated as

$$\overline{f}_{\text{LR}} = \mathbf{E}_{\widetilde{\mathbf{H}}}\left\{|\mathcal{S}_{\text{FB}}|\right\}. \tag{9}$$

Figure 4 shows the $\overline{f}_{\text{LR}}$ due to various thresholds that were applied to all the scheduling algorithms considered in this work. As more number of UEs feed back the CSI, the same needs to be available at the CCN. The savings in the feedback load is linear whose slope decreases with increasing threshold. As expected, the feedback load reduction with threshold of 40 dB and ∞ dB has poor savings.

Now we discuss the impact on the backhaul due to the precoding weights. We define the normalized average backhaul load reduction as the relative difference in the total number of UEs and BSs to the cardinality of the set, $\mathcal{S}_{\text{BH}}$ consisting of non-zeros in the precoded matrix, $\widetilde{\widetilde{\mathbf{W}}}_{\text{algo}} \in \mathbb{C}^{L\times N}$, i.e., $\mathcal{S}_{\text{BH}} = \left\{\widetilde{\widetilde{\mathbf{W}}}_{\text{algo}}\,(j,i) \neq 0, \forall i,j \in \mathbb{N}^{+}, i \leq N, j \leq L\right\}$. The normalized average backhaul load reduction is calculated as

$$\overline{b}_{\text{LR}} = \frac{(L_{\max}N_{\max}) - \mathbf{E}_{\widetilde{\mathbf{H}}}\left\{|\mathcal{S}_{\text{BH}}|\right\}}{L_{\max}N_{\max}}, \tag{10}$$

where $L_{\max} = K = 3$ and $N_{\max} = L_{\max}$ as the maximum number of UEs served is limited by the maximum number of BSs selected. This is captured in Fig. 5 for the various scheduling algorithms considered in this work. The CS approach has nearly 90% backhaul savings with the smallest feedback load reduction threshold, and the savings diminish as the threshold increases. As the number of UEs grows, it is interesting to note that with CS 40 dB and US 40 dB, the savings are nearly similar, with both undergoing an exponential decay. An ∞ dB threshold also shows this decay, resulting in savings in the backhaul. This is due to the fact that the scheduler is capable of

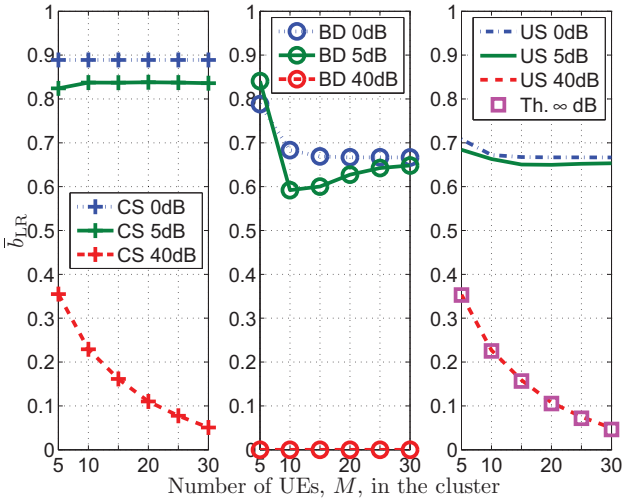

Fig. 5. Average normalized backhaul load reduction, $\bar{b}_{\mathrm{LR}}$

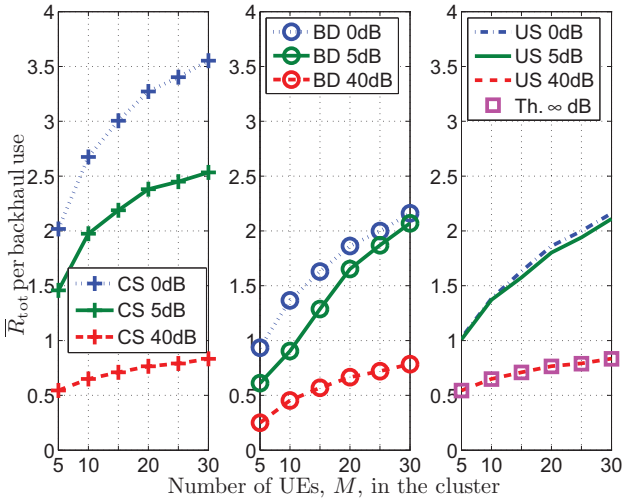

Fig. 6. The average sum rate per backhaul use

finding a smaller set of BSs and UEs that can achieve a better sum rate. With smaller thresholds, the CS and US both tend to have higher savings in the backhaul. There is no backhaul savings when the feedback threshold is 40 dB in the case of BD, as the aggregated channel matrix is full. The BD 5 dB has better savings in the backhaul compared to BD 0 dB when M is small. This is due to the failure to find a block diagonal structure that results in the savings in the backhaul as observed in Fig. 3.

The metric average sum rate per backhaul use is considered, as the user data will be routed at the CCN based on the non-zero precoding weight. This will dominate the backhaul compared to the CSI feedback [5, Fig. 1]. Hence, the average sum rate per backhaul use is calculated as

$$\underset{\text{per backhaul use}}{\overline{R}_{\mathrm{tot}}} = \frac{\overline{R}_{\mathrm{tot}}}{\left(1-\bar{b}_{\mathrm{LR}}\right) L_{\max} N_{\max}} = \frac{\overline{R}_{\mathrm{tot}}}{\mathbf{E}_{\widetilde{\mathbf{H}}}\left\{|\mathcal{S}_{\mathrm{BH}}|\right\}}, \quad (11)$$

and Fig. 6 captures this metric. It can be observed that our proposed CS algorithm performs the best compared to all the other algorithms, providing the best sum rate per backhaul use.

The limitation of the proposed approaches is that they need to perform an exhaustive search to find the best possible set of BSs and UEs that gives the best sum rate. However, a greedy based user selection can be easily implemented based on the proposed algorithm in order to reduce the complexity [11].

V. Conclusion

In this work, we explore scheduling techniques that can efficiently use the backhaul for distributing the precoding weights (from CCN to corresponding BSs) under feedback load reduction achieved via partial joint processing for coordinated multipoint transmission. We proposed the constrained and unconstrained scheduling schemes, comparing them to the state of the art MAC layer block diagonalization technique for backhaul load reduction. The constrained scheduling achieves the best tradeoff in terms of the sum rate per backhaul use. The block diagonalization technique performs well in terms of the sum rate when the number of users is large, however, they fail to find a block diagonal structure when the number of users are small.

As part of our future work, combining the constrained scheduling and the block diagonalization technique can harness the gains of both these approaches and overcome their limitations simultaneously. This combined technique can achieve a better tradeoff between the sum rate and backhaul use. Also, generalizing the block diagonalization technique, such that $N = L \leq K$ can improve these preliminary results.

References

[1] L. Daewon, S. Hanbyul, et al., "Coordinated Multipoint Transmission and Reception in LTE-Advanced: Deployment Scenarios and Operational Challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148-155, Feb. 2012.

[2] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and Wei Yu, "Multi-Cell MIMO Cooperative Networks: A New Look at Interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[3] C. Botella, T. Svensson, X. Xu, and H. Zhang, "On the performance of joint processing schemes over the cluster area," in *Proc. IEEE Veh. Technol. Conf.*, pp. 1-5, May 2010.

[4] A. Papadogiannis, H.J. Bang, D. Gesbert, and E. Hardouin, "Efficient Selective Feedback Design for Multicell Cooperative Networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 1, pp. 196-205, Jan. 2011.

[5] T.R. Lakshmana, C. Botella, and T. Svensson, "Partial Joint Processing with Efficient Backhauling using Particle Swarm Optimization," *EURASIP J. Wireless Commun. and Netw.*, vol. 2012, 2012.

[6] B. Makki and T. Eriksson, "On Hybrid ARQ and Quantized CSI Feedback Schemes in Quasi-Static Fading Channels," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 986-997, Apr. 2012.

[7] T.R. Lakshmana, C. Botella, T. Svensson, X. Xu, J. Li and X. Chen, "Partial Joint Processing for Frequency Selective Channels," in *Proc. IEEE VTC Fall*, Sept. 2010.

[8] H. Zhang and H. Dai, "Cochannel Interference Mitigation and Cooperative Processing in Downlink Multicell Multiuser MIMO Networks," *EURASIP J. Wireless Commun. and Netw.*, vol. 2004, no. 2, pp. 222-235, 2004.

[9] P. Kyösti, J. Meinilä, et al., "D1.1.2 WINNER II channel models: Part I channel models," IST-4-027756 WINNER II, September 2007.

[10] ARTIST4G D1.2, Innovative advanced signal processing algorithms for interference avoidance, *ARTIST4G technical deliverable*, https://ict-artist4g.eu/projet/work-packages/wp1/documents/d1.2/d1.2.pdf, pp. 84. Accessed 17 Aug. 2012.

[11] J. Li, T. Svensson, C. Botella, T. Eriksson, X. Xu, and X. Chen, "Joint Scheduling and Power Control in Coordinated Multi-Point Clusters", in *Proc. IEEE VTC Fall*, Sept 2011.