

Harmonized Cellular and Distributed Massive MIMO: Load Balancing and Scheduling

Qiaoyang Ye*, Ozgun Y. Bursalioglu[†], Haralabos C. Papadopoulos[†]

*Dept. of ECE, The University of Texas at Austin

[†]Docomo Innovations Inc

Email: qye@utexas.edu, {obursalioglu, hpapadopoulos}@docomoinnovations.com.

Abstract—Multi-tier networks with large-array base stations (BSs) that are able to operate in the “massive MIMO” regime are envisioned to play a key role in meeting the exploding wireless traffic demands. Operated over small cells with reciprocity-based training, massive MIMO promises large spectral efficiencies per unit area with low overheads. Also, near-optimal user-BS association and resource allocation are possible in cellular massive MIMO HetNets using simple admission control mechanisms and rudimentary BS schedulers, since scheduled user rates can be predicted *a priori* with massive MIMO.

Reciprocity-based training naturally enables coordinated multi-point transmission (CoMP), as each uplink pilot inherently trains antenna arrays at all nearby BSs. In this paper we consider a distributed-MIMO form of CoMP, which improves cell-edge performance without requiring channel state information exchanges among cooperating BSs. We present methods for harmonized operation of distributed and cellular massive MIMO in the downlink that optimize resource allocation at a coarser time scale across the network. We also present scheduling policies at the resource block level which target approaching the optimal allocations. Simulations reveal that the proposed methods can significantly outperform the network-optimized cellular-only massive MIMO operation (i.e., operation without CoMP), especially at the cell edge.

I. INTRODUCTION

The exponential growth in wireless traffic is driving the densification of cellular networks. Existing networks of carefully planned conventional macro base stations (BSs) are becoming transformed into dense irregular heterogeneous networks (HetNets), as they are continuously supplemented with various types of BSs, differing in transmit power, physical size, and deployment cost [1].

It has been well recognized that traditional user-BS association schemes are highly suboptimal for HetNets, due to the large disparities in BS transmit power [2]. Moreover, the non-uniform user distribution and irregular deployment of small BSs make load balancing critical. Various approaches have been used to investigate load balancing in HetNets [2], including stochastic geometry approach [3] and techniques from game theory [4]. Some standardization efforts have also been made for load balancing in HetNets, e.g., in the form of *cell range expansion* [5].

Several recent works [6–9] recast load balancing in HetNets as a network utility maximization (NUM) problem.

Paper [6] studies the optimal user-BS association problem in HetNets and shows a great improvement in user rate distribution via systematic load balancing. Papers [7, 8] consider the joint optimization of user association and BS muting – referred to in 3GPP as enhanced intercell interference coordination (eICIC).

In parallel, there is surging interest in equipping BSs with large antenna arrays. With higher-frequency spectrum becoming available, large arrays become feasible even for small cells, as more effective antennas can be packed into a small form factor¹. By exploiting channel reciprocity, massive arrays can be trained in the uplink (whether for uplink or downlink transmission) with low overheads [10]. This enables very large spectral efficiencies per unit area via massive MIMO, i.e., via serving simultaneously many users (although much fewer than antennas), each at a very high rate [10–13]. Attributes of massive MIMO can also be exploited to achieve near-optimal load balancing over massive MIMO HetNets using simple user-BS association methods with cellular transmission (where data for each user is transmitted from a single BS) [9].

In this paper we consider the use of coordinated multi-point transmission (CoMP) as a means for improving network performance, in particular, the cell-edge performance. CoMP is naturally enabled by reciprocity-based training, since a single uplink pilot from a user terminal can train all nearby antennas. In regular cellular layouts with massive MIMO BSs, [11] shows gains to cell-edge users via CoMP.

In this paper we focus on a distributed-MIMO form of CoMP, which does not require channel state information (CSI) exchanges among cooperating BSs and allows us to develop a systematic approach for allocating resources for CoMP and cellular transmission. The methods we develop are based on formulating a NUM problem with respect to user association and resource allocation via extensions of the framework for cellular transmission developed in [9]. We also present scheduling policies at the resource block (RB) level, which target approaching these optimized (coarser time scale) resource allocations. Simulations show that the proposed harmonized CoMP/cellular operation can provide significant gains with respect to cellular-only

¹For example, at 5GHz, 49 antennas (arranged on square grid at half wavelength spacing) can be packed on a 20cm × 20cm antenna patch.

massive MIMO operation [9], especially at the cell edge.

II. SYSTEM MODEL

We consider the downlink of a cellular network comprised of J BSs and K single-antenna users. We use $j \in \mathcal{J} = \{1, 2, \dots, J\}$ and $k \in \mathcal{U} = \{1, 2, \dots, K\}$ to index BSs and users, respectively. We let M_j denote the number of antennas at BS j and assume $M_j \gg 1$. We assume time division duplex (TDD) operation with reciprocity-based CSI acquisition [10, 11]. Hence, every BS antenna in the vicinity of user k can estimate its downlink channel coefficient to user k from the uplink pilot transmitted by user k . This enables the training of large antenna arrays (e.g., $M_j \gg 1$) with pilot overheads proportional to number of simultaneously served users [10]. In contrast to feedback-based CSI acquisition, it also allows a user terminal to train *multiple nearby* BSs, and enables CoMP without additional training overheads.

Transmission resources are split into slots or RBs, with each RB corresponding to a contiguous block of OFDM subcarriers and symbols. In any given RB, we let P_j denote the transmit power at BS j , and assume that this power is equally split among all served user streams. We assume a block-fading channel model where the user-BS channels remain constant within each RB [10–12, 14]. We let $\mathbf{g}_{kj} = \sqrt{\beta_{kj}} \mathbf{h}_{kj}$ denote the $M_j \times 1$ channel vector between BS j and user k on a generic RB, with the slow-fading scalar β_{kj} characterizing the combined effect of distance-based path loss and location-based shadowing, and the vector \mathbf{h}_{kj} capturing fast fading. We assume that the vectors \mathbf{h}_{kj} 's are independent in k and j , and that \mathbf{h}_{kj} has i.i.d. $\mathcal{CN}(0, 1)$ elements (independent Rayleigh fading). We also assume that the thermal noise process at user k is i.i.d. with $\mathcal{CN}(0, \sigma^2)$ samples.

III. MIMO TRANSMISSION

Within each RB, a subset of users are *active*, i.e., are scheduled for transmission. The coded data for any given scheduled user can be transmitted either from a single BS via *cellular transmission*, or from multiple BSs via a CoMP mode referred to as *distributed MIMO transmission*.

A. Prior Art: Cellular Massive MIMO [9]

In setting the stage for the distributed MIMO operation presented in this work, it is worth revisiting load balancing and scheduling for cellular massive MIMO, as considered in [9]. Let S_j denote the maximum number of users served by BS j on any given RB, with $S_j \ll M_j$. Under mild assumptions on fading, the achievable user instantaneous rates on RB t , $r_{kj}(t)$, can be predicted *a priori* in the massive MIMO regime [9]. In particular, there exist deterministic quantities $\{r_{kj}\}$ such that $r_{kj}(t) \xrightarrow{\text{a.s.}} r_{kj}$, for all $k \in \mathcal{U}$ and $j \in \mathcal{J}$, as $M_j, S_j \rightarrow \infty$, with fixed $\nu_j = S_j/M_j \geq 0$ [10–12]. This convergence is very fast with respect to the M_j 's.

Letting $\mathcal{S}_j(t)$ denote the set of users served by BS j on RB t and $x_{kj} = \lim_{T \rightarrow \infty} \frac{|\{t: k \in \mathcal{S}_j(t)\}|}{T}$ denote the

activity fraction of user k on BS j , the long-term averaged throughput of user k can be obtained via [9]

$$R_k = \sum_{j \in \mathcal{J}} x_{kj} r_{kj}, \quad \forall k \in \mathcal{U}. \quad (1)$$

The advantages of the approach in [9] for cellular massive MIMO operation can be summarized as follows:

- (A) The r_{kj} 's are accurate peak-rate proxies, which are independent of scheduled instances and user sets.
- (B) User throughputs depend on activity fractions, via (1).
- (C) The (combinatorial) user-cell association problem is recast as a (convex) NUM problem with respect to the $\{x_{kj}\}$ variables, subject to resource constraints.
- (D) Any $\{x_{kj}\}$ set not violating any resource constraints can be realized by a suitably designed scheduler.

B. CoMP via Distributed MIMO

The distributed MIMO scheme we consider corresponds to a form of CoMP that allows harvesting performance gains at the cell edge, with low operational overheads.

Definition 1. Admissible Distributed MIMO Schemes: An admissible distributed MIMO scheme is a scheme that schedules transmissions for users on a sequence of RBs and, on each RB, satisfies the following:

- (i) All the users served by a given BS j are served in clusters of the same size L for some $L \geq 1$.
- (ii) BS j serves at most $S_j(L)$ users, for some fixed $S_j(L)$, satisfying $S_j \leq S_j(L) \leq LS_j$.
- (iii) The user beams (i.e., the precoding vectors) at BS j are designed as if BS j were engaging in cellular MU-MIMO transmission over all the users it serves.
- (iv) All BSs serving a user transmit the same coded user stream. Each BS transmits the stream on a beam that is (independently) designed for the users at that BS.
- (v) $M_j \gg S_j(L)$, for all L and j considered.

We also assume that, within each RB, the transmit power at each BS is equally split among scheduled users.

Table I provides an example of a scheme complying with Defn. 1, assuming BS clusters of size 1 (cellular transmission) and 2. Four BSs are considered with $P_j = 1$, $S_j(1) = S_j = 2$, and $S_j(2) = 3$. As the table reveals, each BS on RB #1 engages in cellular transmission. On RB #2, BSs pairs jointly serve triplets of users. RBs #3 and #4 provide additional, more interesting, modes. No two users are served by the same BS cluster on RB #3, while on RB #4 BSs 1-2 jointly serve a triplet of users, while BSs 3-4 serve users in cellular transmission. Note also that (at least) 8, 6, 6 and 7 uplink pilot dimensions are needed to enable RBs #1, #2, #3 and #4, respectively. Evidently, the choice of scheduled user sizes, $S_j(L)$, signifies how aggressively pilot dimensions are reused across the network.

It is worth making a few remarks regarding the choice of the distributed MIMO schemes of Defn. 1. First, the schemes of Defn. 1 provide the following CoMP benefits:

- (i) **Performance gains at the cell edge:** The beam-forming (BF) gain provided by the cellular scheme

TABLE I
EXAMPLE OF RBs ENABLED BY DISTRIBUTED MIMO OVER 4 BSs.

RB		BS 1	BS 2	BS 3	BS 4
#1	Cluster Size	1	1	1	1
	User Power	1/2	1/2	1/2	1/2
	Served Users	1,2	3,4	5,6	7,8
#2	Cluster Size	2	2	2	2
	User Power	1/3	1/3	1/3	1/3
	Served Users	1,2,3	1,2,3	4,5,6	4,5,6
#3	Cluster Size	2	2	2	2
	User Power	1/3	1/3	1/3	1/3
	Served Users	1,2,3	1,4,5	2,4,6	3,5,6
#4	Cluster Size	2	2	1	1
	User Power	1/3	1/3	1/2	1/2
	Served Users	1,2,3	1,2,3	4,5	6,7

(that the distributed scheme is based upon) becomes intra-cluster BF gain in the distributed MIMO case, as the same coded data is transmitted from all BSs serving the user. Similarly, the intra-cell interference mitigation capabilities of the cellular scheme are extended across the cluster of BSs from which the user is served. As a result, performance gains can be realized at the cell edge.

- (ii) **Low training overheads:** An uplink pilot from a user terminal trains all nearby BS antennas, whether these are in one or many locations. Thus, CSI acquisition between a user and nearby BSs need not incur additional overheads with respect to cellular transmission.

In addition, the schemes of Defn. 1 possess several important properties that are not in general present in CoMP schemes:

- (a) **Local precoding at each BS:** This is due to item (iii) in Defn. 1. For instance, in the case of linear zero-forcing beamforming (LZFBF), the beam for each user served by BS j is chosen within the null space of the channels of all the other users served by BS j , no matter whether there are additional BSs serving the user on the same RB or not.
- (b) **No need for CSI exchanges among BSs:** Again, due to item (iii) in Defn. 1, BS j only needs CSI between the users it serves and the antennas of BS j in order to generate the user beams at BS j .
- (c) **Flexible scheduling:** The schemes of Defn. 1 enable user-specific BS-cluster transmission, and allow serving users from overlapping but different clusters of BSs on the same RB (see, e.g., RB #3 in Table I).
- (d) **Simple predictors of instantaneous rates:** As shown in [15], the instantaneous user rates can also be predicted *a priori* with CoMP. However, unlike general CoMP settings, where a user's instantaneous rate depends on the other users co-scheduled for transmission on the same RB [15], the schemes of Defn. 1 make a user's instantaneous rate *independent* of the identities of the other users in the scheduling set.

As a result, the cellular-transmission attributes (A)–(C)

exploited in [9] can be appropriately extended to allow resource allocation for the schemes of Defn. 1, in the form of network-optimized activity fractions between users and clusters of BSs. Although, as it turns out, item (D) is not always true with distributed MIMO, i.e., these activity fractions may not necessarily be realizable, as shown in Sec. VI, scheduling policies can be designed that may approximate these fractions in practice sufficiently well.

IV. PEAK RATES AND SCHEDULED THROUGHPUTS

In this section, we develop proxy expressions for the instantaneous user rates and for the scheduled user throughputs that are provided by any given scheduling policies enabling distributed MIMO transmission with either LZFBF or maximum ratio transmission (MRT).

We consider a scheduling policy on RBs $\{1, 2 \dots, T\}$ and assume that all the large-scale coefficients stay fixed within this period. Any such scheduling policy can be described in terms of the scheduling sets $\{\mathcal{S}_C(t); \forall C, \forall t \in \{1, 2 \dots, T\}\}$, where $\mathcal{S}_C(t)$ denotes the set of active users served by cluster C on RB t . Thus, the received signal at an active user $k \in \mathcal{S}_C(t)$ on RB t can be expressed by

$$y_k(t) = \underbrace{\sum_{j \in C} \sqrt{\frac{P_j}{N_j}} \mathbf{g}_{kj}^H \mathbf{f}_{kj} s_k}_{\text{desired}} + \underbrace{\sum_{j \in C} \sum_{\substack{u \in \mathcal{S}_C(t) \\ u \neq k}} \sqrt{\frac{P_j}{N_j}} \mathbf{g}_{kj}^H \mathbf{f}_{uj} s_u}_{\text{intra-cluster interference}} + \underbrace{\sum_{l \notin C} \sum_{u \in \cup_{C': l \in C'} \mathcal{S}_{C'}(t)} \sqrt{\frac{P_l}{N_l}} \mathbf{g}_{kl}^H \mathbf{f}_{ul} s_u}_{\text{inter-cluster interference}} + \underbrace{w_k}_{\text{noise}}, \quad (2)$$

where C denotes the cluster (set) of BSs serving user k on RB t , C' denotes the cluster including BS l , s_u denotes the unit-power stream for user k , and \mathbf{f}_{uj} denotes the unit-norm precoding vector for user u at BS j .

Let r_{kC} denote the peak rate of user k from BS cluster C . It can be shown using the techniques in [15, 16] that, with distributed MIMO based on LZFBF, r_{kC} is given by²

$$r_{kC} = \log_2 \left(1 + \frac{\sum_{j \in C} \sum_{\ell \in C} \sqrt{P_j P_\ell} \beta_{kj} \beta_{k\ell} b_j (|\mathcal{C}|) b_\ell (|\mathcal{C}|)}{\sigma^2 + \sum_{\ell \notin C} P_\ell \beta_{k\ell}} \right), \quad (3)$$

where $b_j(L) = \frac{M_j - S_j(L) + 1}{S_j(L)}$. Similarly, for the case that the distributed MIMO transmission is based on MRT,

$$r_{kC} = \log_2 \left(1 + \frac{\sum_{j \in C} \sum_{\ell \in C} \sqrt{\frac{P_j P_\ell M_j M_\ell \beta_{kj} \beta_{k\ell}}{S_j(|\mathcal{C}|) S_\ell(|\mathcal{C}|)}}}{\sigma^2 + I_{kC} + \sum_{\ell \notin C} P_\ell \beta_{k\ell}} \right), \quad (4)$$

where $I_{kC} = \sum_{j \in C} \frac{S_j(|\mathcal{C}|) - 1}{S_j(|\mathcal{C}|)} P_j \beta_{kj}$ is intra-cluster interference. For completeness, we provide the proof of (3) and (4) in Appendixes A and B, respectively.

²Expression (3) assumes that $\forall j \in C$, BS j serves $S_j(L)$ users, each user at power $P_j/S_j(L)$. In the case that fewer users are served by one of the BSs, the LHS in (3) represents an achievable (lower bound) rate.

Similar to cellular massive MIMO in [9], the long-term user throughput with the admissible distributed MIMO schemes of Defn. 1 can be expressed in terms of the distributed MIMO peak rates and the activity fractions provided by the scheduling policy. In the limit $T \rightarrow \infty$, the throughput of user k can be expressed as³

$$R_k = \sum_{\mathcal{C}} x_{k\mathcal{C}} r_{k\mathcal{C}}, \quad (5)$$

where $x_{k\mathcal{C}} = \lim_{T \rightarrow \infty} \frac{|\{t: 1 \leq t \leq T; k \in \mathcal{S}_{\mathcal{C}}(t)\}|}{T}$ is the activity fraction of user k with respect to cluster \mathcal{C} .

Fig. 1 shows an example of the potential benefits that distributed MIMO transmission can offer, by showing the peak rates of cellular vs pair-BS distributed MIMO transmission as a function of user location. When the user is close the transmitting BS cellular transmission is as good as anything. On the other bend when the user is close to the cell edge between BSs 1 and 2, distributed MIMO transmission from the cluster $\{1, 2\}$ yields about 3 times higher rates than cellular transmission. In the next section we formally consider the problem of allocating resources to users across BSs or clusters of BSs so as to optimize the network-wide system performance.

V. USER-CLUSTER ASSOCIATION AS NUM

In this section, we formulate the user-cluster association problem as a NUM of activity fractions across all users, to optimize a network-wide utility function capturing the operator's notion of (inherently subjective) fairness.

Before formulating the NUM problem, it is worth restricting the domain of scheduling options in order to obtain solutions that are of practical interest. We focus on cluster sizes $L \in \{1, 2, \dots, L_{\max}\}$ for some appropriately chosen maximum⁴ cluster size, L_{\max} . Motivated by the example in Table I, we consider the following architecture.

Definition 2. Uniform Cluster-Size Architecture (UCS): A scheme from Defn. 1 is a UCS architecture, if for each $L \in \{1, 2, \dots, L_{\max}\}$, a $\lambda_L \geq 0$ fraction of the RBs is allocated to serving size- L clusters, and if on any RB from this λ_L fraction the following are satisfied:

- (i) each scheduled user is served by a (user-dependent) cluster of L BSs;
- (ii) for each $j \in \mathcal{J}$, BS j serves no more than $S_j(L)$ users.

In the UCS architecture, users served by different-size clusters are scheduled on distinct RBs. For the example in Table I, such an architecture enables scheduling policies with RBs of types #1, #2, and #3, but not of type #4.

³Convergence to the limiting expressions of interest is very quick [9].

⁴The choice of L_{\max} is a design choice. It depends on the average number of nearby BS arrays that users typically see and the complexity that can be afforded. In our simulations, we set $L_{\max} = 4$.

The NUM subject to the UCS architecture is

$$\max_{\lambda_L, x_{k\mathcal{C}}} \sum_{k \in \mathcal{U}} U \left(\sum_{\mathcal{C}: |\mathcal{C}| \leq L_{\max}} x_{k\mathcal{C}} r_{k\mathcal{C}} \right) \quad (6a)$$

$$\text{s.t.} \sum_{\substack{\mathcal{C}: j \in \mathcal{C} \\ |\mathcal{C}|=L}} \sum_{k \in \mathcal{U}} x_{k\mathcal{C}} \leq \lambda_L S_j(L), \forall j, L \leq L_{\max}, \quad (6b)$$

$$\sum_{\mathcal{C}: |\mathcal{C}|=L} x_{k\mathcal{C}} \leq \lambda_L, \forall k \in \mathcal{U}, L \leq L_{\max}, \quad (6c)$$

$$x_{k\mathcal{C}} \geq 0, \forall k \in \mathcal{U}, \forall \mathcal{C} \text{ with } |\mathcal{C}| \leq L_{\max}, \quad (6d)$$

$$\sum_{L=1}^{L_{\max}} \lambda_L \leq 1, \quad (6e)$$

$$\lambda_L \geq 0, \forall L \leq L_{\max}. \quad (6f)$$

Ineq. (6b) signifies that the total activity fractions of users served by BS j in clusters of size L cannot exceed the product of available RBs and the maximum number of beams that can be spatially multiplexed at BS j in clusters of size L . Ineq. (6c) signifies that the fraction of RBs over which user k is served by clusters of size L cannot exceed the RB fraction allocated to size- L clusters.

It is easy to verify that (6) is a convex optimization problem. Also note that, for $L_{\max} = 1$, (6) specializes to the cellular massive MIMO NUM problem studied in [9].

The second architecture we consider also allows serving users of the type of RB # 4.

Definition 3. Mixed Cluster-Size Architecture (MCS): A scheme from Defn. 1 is a MCS architecture, if a $\lambda_L \geq 0$ fraction of the RBs is allocated $\forall L \in \{2, 3, \dots, L_{\max}\}$, and if within any RB that is part of the λ_L fraction the following are satisfied: (i) each scheduled user is served either in cellular mode, or by a (user-dependent) cluster of L BSs; (ii) $\forall j \in \mathcal{J}$, BS j serves either at most S_j users all in cellular mode, or at most $S_j(L)$ users, all served in clusters of size L .

A convex NUM problem analogous to (6) can be formulated for the MCS architecture.

VI. SCHEDULING POLICIES FOR NUM SOLUTION

In this section, we investigate scheduling policies that yield $\{x_{k\mathcal{C}}\}$ closely matching the solution of (6).

Definition 4. Feasible Schedule: A scheduling policy $\{\mathcal{S}_{\mathcal{C}}(t); \forall \mathcal{C}, \text{ with } |\mathcal{C}| \leq L_{\max}, \forall t \in \{1, 2, \dots, T\}\}$ is feasible with respect to the UCS architecture of Defn. 2 if it satisfies the following:

- (i) For each t , the policy associates with RB t a single cluster size, $L(t)$, for some $L(t) \leq L_{\max}$, i.e., for each \mathcal{C} for which $\mathcal{S}_{\mathcal{C}}(t)$ is non-empty, $|\mathcal{C}| = L(t)$.
- (ii) For each t , each user is served by at most one cluster; that is, $|\sum_{\mathcal{C}} 1\{k \in \mathcal{S}_{\mathcal{C}}(t)\}| \leq 1$ for all $k \in \mathcal{U}$.
- (iii) For each t , and for each $j \in \mathcal{J}$, BS j serves at most $S_j(L(t))$ users; that is, $|\cup_{\mathcal{C}: j \in \mathcal{C}} \mathcal{S}_{\mathcal{C}}(t)| \leq S_j(L(t))$.

It is easy to verify that any feasible schedule yields activity fractions that satisfy (6b)–(6f).

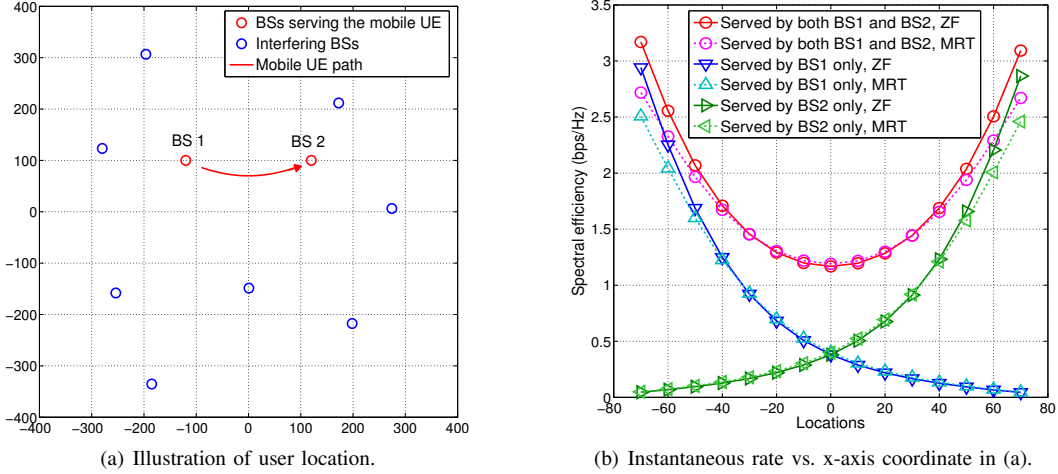


Fig. 1. Illustration of spectral efficiency versus user locations. The location in Fig. 1(a) indicates the x-axis coordinate of the path in Fig. 1(a). The cell edge users (e.g., the areas near the origin) benefit from distributed MIMO.

For instance, in a network of 3 BSs, with $L_{\max} = 2$ and $S_j(2) = 3, \forall j$, no feasible schedule yields $\{x_{kC}\}$ with $\lambda_2 > 0$, for which (6b) is satisfied with equality for all j and $L = 2$. This is because it is impossible to *simultaneously* schedule 3 users at *all* three BSs: at most two BSs can schedule 3 users, while the 3rd would necessarily schedule at most 2 users (i.e., the 3 BSs schedule a total of 4 users, each receiving beams from a BS pair). Clearly, any feasible schedule results in at least one *strict* inequality in (6b). Hence, the coarser time-scale NUM problem (6) does not capture the finer time-scale constraints associated with feasible schedulers. Although, in general, (6) provides an upper bound on the network performance, as we show next, using activity fractions that are the solution to (6), we can design scheduling policies, whose performance is close to the utility provided by the solution to (6).

A. Virtual Queue Based Scheduling Scheme

As in [6, 9], we focus on the proportional fair utility (i.e., $U(x) = \log(x)$ in (6a)) in the rest of this paper. We consider scheduling policies for the UCS architecture comprised of L_{\max} parallel schedulers, one per cluster size $L \in \{1, 2, \dots, L_{\max}\}$. We describe a method for scheduling users over the RBs from the $\lambda_L > 0$ fraction of RBs dedicated to clusters of size L .

We first remark, that as in the cellular settings [6, 9], empirical evidence reveals that, in a “loaded” network, most users are uniquely associated to a single cluster per cluster size, i.e., for most user indices k , there is a single nonzero x_{kC} among all C ’s with the same $|C|$.

Insight regarding this observation can be obtained by examining *Karush-Kuhn-Tucker* (KKT) conditions of (6), which imply

$$\sum_{L'} \sum_{C: |C|=L'} x_{kC} r_{kC} \geq \frac{r_{kC}}{\sum_j \nu_{jL} + \mu_{kL}}, \quad (7)$$

where ν_{jL} and μ_{kL} are the Lagrange multipliers corresponding to (6b) and (6c), respectively.

In a loaded network, where the constraints (6c) are inactive (i.e., $\sum_{C: |C|=L} x_{kC} < \lambda_L \forall k \in \mathcal{U}$), we have the following:

Proposition 1. *If (6c) are inactive $\forall k \in \mathcal{U}$, the number of users that are served by multiple clusters of size L is at most $N_L - 1$, where N_L is the number of size- L clusters.*

Proof: See Appendix C. ■

Given the limited number of fractional users per cluster size L , the scheduler approximates the optimal $\{x_{kC}\}$ by unique association activity fractions, $\{\tilde{x}_{kC}\}$, given by

$$\tilde{x}_{kC} = \begin{cases} x_{kC} & \text{if } C = C^*(k) \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

with $C^*(k) = \arg \max_{C: |C|=L} x_{kC}$.

Letting \mathcal{U}_C denote the users for which $\tilde{x}_{kC} > 0$, we have $\mathcal{U}_C \cap \mathcal{U}_{C'} = \emptyset$ for all $C \neq C'$ with $|C| = |C'|$. We also let $\mathcal{U}^{(L)} = \cup_{C: |C|=L} \mathcal{U}_C$ denote the set of users that receive non-zero activity from clusters of size L .

To assign user k a fraction of RBs close to the desired fraction $\alpha_k = \tilde{x}_{kC}/\lambda_L$, we consider a max-min scheduling policy based on virtual queues (VQ), which assumes user k receives rate $\tilde{R}_k = 1/\alpha_k$ when user k is scheduled for transmission over cluster $C^*(k)$ (i.e., $k \in \mathcal{S}_{C^*(k)}(t)$). The cluster-size L scheduler performs at each t a weighted sum rate maximization (WSRM) of the form [17]:

$$\max_{\tilde{\mathcal{U}} \subseteq \mathcal{U}^{(L)}} \sum_{k \in \tilde{\mathcal{U}}} Q_k(t) \tilde{R}_k, \quad (9a)$$

$$\text{s.t.} \quad \sum_{k \in \tilde{\mathcal{U}}} 1\{j \in C^*(k)\} \leq S_j(L), \quad \forall j \in \mathcal{J}, \quad (9b)$$

where the weight of user k at time t , $Q_k(t)$, is the VQ length of user k at time t . For max-min fairness [17], $Q_k(t)$ is updated as follows:

$$Q_k(t+1) = \max\{0, Q_k(t) - \tilde{R}_k(t)\} + A_k(t), \quad (9c)$$

where

$$\tilde{R}_k(t) = \begin{cases} \tilde{R}_k & \text{if user } k \text{ is scheduled at time } t \\ 0 & \text{otherwise} \end{cases}, \quad (9d)$$

$$A_k(t) = \begin{cases} A_{\max} & \text{if } V > \sum_k Q_k(t) \\ 0 & \text{otherwise} \end{cases}, \quad (9e)$$

with A_{\max} and V chosen sufficiently large [17]. Note that in the absence of constraints (9b), the max-min scheduler (9) schedules user k the desired fraction of RBs, α_k .

Scheduling via (9) is impractical, as it amounts to solving for each RB t an integer linear program of the form (9a)–(9b). A number of heuristic algorithms can be used to provide feasible (though generally suboptimal) solutions to (9). In this paper, we consider a rudimentary greedy algorithm. Letting $K_L = |\mathcal{U}^{(L)}|$ be the total number of users to be served by clusters of size L , the greedy algorithm for size- L clusters at time t operates as follows:

1. Determine a user order $\pi(k)$, where $Q_{\pi(k)}(t)\tilde{R}_{\pi(k)} \geq Q_{\pi(k+1)}(t)\tilde{R}_{\pi(k+1)}$ for all $k \in \mathcal{U}^{(L)}$.
2. Initialization: $k = 1$, and $\tilde{\mathcal{U}} = \emptyset$.
3. If the user set $\tilde{\mathcal{U}} \cup \{\pi(k)\}$ satisfies all the constraints in (9b), set $\tilde{\mathcal{U}} = \tilde{\mathcal{U}} \cup \{\pi(k)\}$.
4. If $k < K_L$, set $k = k + 1$ and go to step 3.
5. Output $\tilde{\mathcal{U}}$ as the scheduling user set at time t .

VII. PERFORMANCE EVALUATION

In this section, we present a brief simulation-based evaluation of the proposed distributed MIMO schemes based on the “wrap-around” checkerboard layout in Fig. 2. There are 4 macros with $M_j = 100$ and $S_j(L) = 10L$, and 32 pico BSs with $M_j = 40$ and $S_j(L) = 4L$. One pico BS is at the center of each white square, while 3 pico BSs are dropped uniformly within each shaded square. Also, 15 and 90 single-antenna users are dropped uniformly in each white and each shaded square, respectively. The macro and pico BS transmit powers are 46dBm and 35dBm, respectively. The path-loss for macro-user links and pico-user links are $128.1 + 37.6 \log_{10} d$ and $140.7 + 36.7 \log_{10} d$, respectively, with the distance d in km.

We consider two distinct macro-pico operation scenarios: (i) macros and picos operate on the same band, with cluster sizes up to $L_{\max} = 4$; (ii) macros are given 20% of the RBs for cellular transmission, and picos are given the remaining 80% for distributed MIMO with $L_{\max} = 4$.

Figs. 3 and 4 compare the proposed distributed MIMO schemes⁵ against network-optimized cellular transmission [9] and max-SINR based association. Fig. 3 shows the user-rate geometric mean for each scheme and each operation scenario considered. As the figure reveals, unique-association (i.e., the $\{\tilde{x}_{kC}\}$) yield almost the same performance as the optimal solution, verifying our conclusion that the number of users served by multiple clusters per architecture is limited. Also, the proposed greedy VQ based scheduler performs within 90% of the NUM optimal value.

⁵The NUM problem for scenario (ii) is a simple extension of (6).

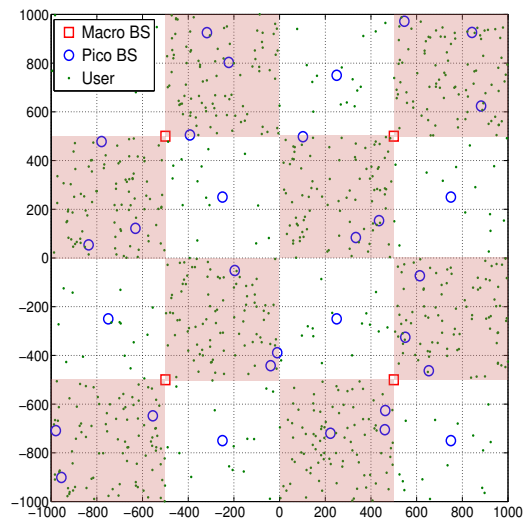


Fig. 2. A 2000m \times 2000m network with 4 macro 32 pico BSs.

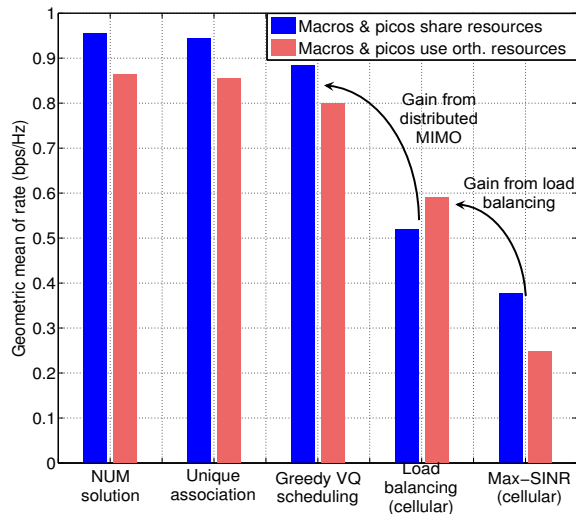


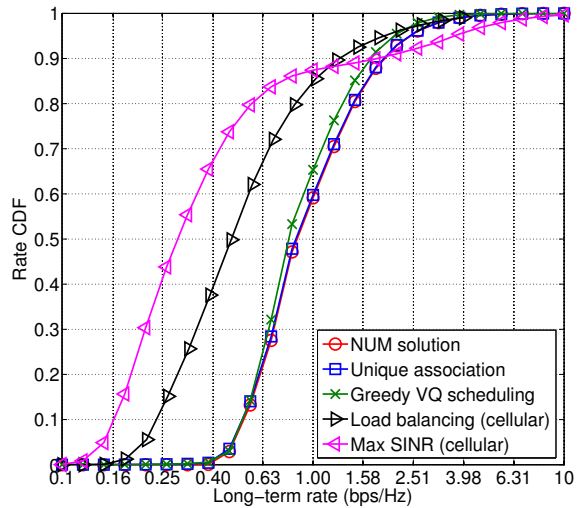
Fig. 3. User-rate geometric means for various schemes under two distinct macro-pico operational scenarios.

More importantly, it significantly outperforms network-optimized cellular operation under both scenarios. Note that though the orthogonal resource allocation with optimal user association in cellular case performs better than the shared resource allocation in our setting, which operation scenario is better highly depends on the system parameters (e.g., channel mode, transmit power and BS density).

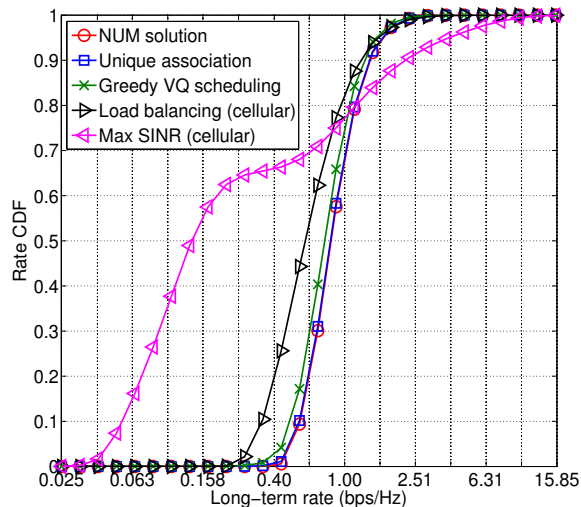
Fig. 4 shows the corresponding user-rate cumulative distribution functions. As the figure shows, the proposed distributed MIMO schemes yield about a $2\times$ gain in 5th percentile rates with respect to the optimal cellular scheme [9], under both macro-pico operation scenarios.

VIII. CONCLUSION

We present techniques for harmonized use of cellular and CoMP transmission over massive MIMO HetNets. The techniques rely on using a class of distributed MIMO transmission schemes, which do not require CSI exchanges



(a) Macro and pico BSs share resources



(b) Macro and pico BSs use disjoint sets of resources

Fig. 4. User-rate CDFs for various schemes.

among BSs, and can enable flexible (user-specific) CoMP transmission. We use properties of the distributed MIMO schemes in the massive MIMO regime to formulate resource allocation as a convex NUM problem, and present scheduling policies whose goal is to approximate the resulting optimized resource allocations. As our simulations show, the proposed operation offers significantly performance gains with respect to the network-optimized cellular-only massive MIMO operation [9], especially at the cell edge. More dynamic settings (e.g., users with high mobility) are left for future work. The investigation of other simulation settings (e.g., different ρ_C) is also of interest.

REFERENCES

- [1] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Comm. Mag.*, vol. 51, pp. 136–144, Mar. 2013.
- [2] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Communications*, vol. 21, pp. 18–25, Apr. 2014.

- [3] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis and design insights," *IEEE Trans. on Wireless Communications*, vol. 12, pp. 2484–2497, May 2013.
- [4] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc., IEEE INFOCOM*, pp. 998–1006, Apr. 2013.
- [5] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, pp. 10–21, June 2011.
- [6] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. on Wireless Communications*, vol. 12, pp. 2706–2716, June 2013.
- [7] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "On/off macrocells and load balancing in heterogeneous cellular networks," *Proc., IEEE Globecom*, Dec. 2013.
- [8] A. Bedekar and R. Agrawal, "Optimal muting and load balancing for eCIC," in *International Symposium on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt)*, pp. 280–287, May 2013.
- [9] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *submitted to IEEE Trans. on Wireless Communications*, 2014. Available at arXiv: <http://arxiv.org/abs/1407.6731>.
- [10] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. on Wireless Communications*, vol. 9, pp. 3590–3600, Nov. 2010.
- [11] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprasad, "Achieving large spectral efficiency with TDD and not-so-many base-station antennas," in *IEEE-APS Topical Conference on Antennas and Propagation in Wireless Communications (APWC)*, pp. 1346–1349, Sep. 2011.
- [12] J. Hoydis, S. Ten Brink, M. Debbah, et al., "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?," *IEEE Journal on Sel. Areas in Communications*, vol. 31, pp. 160–171, Feb. 2013.
- [13] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Comm. Mag.*, vol. 52, pp. 186–195, Feb. 2014.
- [14] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. on Info. Theory*, vol. 56, pp. 2845–2866, June 2010.
- [15] H. Huh, A. M. Tulino, and G. Caire, "Network MIMO with linear zero-forcing beamforming: Large system analysis, impact of channel estimation, and reduced-complexity scheduling," *IEEE Trans. on Info. Theory*, vol. 58, pp. 2911–2934, May 2012.
- [16] Y.-G. Lim, C.-B. Chae, and G. Caire, "Performance analysis of massive MIMO for cell-boundary users," *arXiv preprint arXiv:1309.7817*, Sep. 2013.
- [17] H. Shirani-Mehr, G. Caire, and M. J. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *IEEE Trans. on Communications*, vol. 58, pp. 2055–2066, July 2010.
- [18] K.-K. Wong and Z. Pan, "Array gain and diversity order of multiuser MISO antenna systems," *International Journal of Wireless Information Networks*, vol. 15, pp. 82–89, June 2008.

APPENDIX A

PROOF OF SPECTRAL EFFICIENCY USING ZF PRECODING

In this paper, we assume that each BS has perfect CSI. Techniques in [15, 16] can be applied to derive the results. For completeness, we provide the proof as follows. We use $\mathbb{E} \left[\frac{S}{I+N} \right] \approx \frac{\mathbb{E}[S]}{\mathbb{E}[I+N]}$ to approximate SINR in the calculation of ergodic spectral efficiency in the massive MIMO regime, which is shown to be quite close to the exact asymptotic spectral efficiency [16].

Adopting the ZF precoding, the precoding matrix at BS j is $\mathbf{F}_j = \mathbf{G}_j (\mathbf{G}_j^H \mathbf{G}_j)^{-1} \mathbf{A}_j^{1/2}$, where \mathbf{A}_j is the

normalizing coefficients matrix. In this case, the intra-cell interference is 0. Denoting the k th diagonal element of \mathbf{A}_j by a_{kj} and plugging the precoding matrix $\mathbf{F}_j = \mathbf{G}_j (\mathbf{G}_j^H \mathbf{G}_j)^{-1} \mathbf{A}_j^{1/2}$ into received signal, the SINR at user k from \mathcal{C} is

$$\text{SINR}_{k\mathcal{C}} = \frac{\sum_{j \in \mathcal{C}} \sum_{l \in \mathcal{C}} \sqrt{\frac{P_j P_l a_{kj} a_{kl}}{S_j(|\mathcal{C}|) S_l(|\mathcal{C}|)}}}{\sigma^2 + \left\| \sum_{l \notin \mathcal{C}} \sum_{u \in \cup_{(c':l \in c')} \mathcal{S}_{c'}} \sqrt{\frac{P_l}{S_l(c')}} \mathbf{g}_{kl}^H \mathbf{f}_{ul} s_u \right\|^2}, \quad (10)$$

where \mathcal{C}' denotes the cluster including l that is different from \mathcal{C} . Using similar techniques in the proof of Theorem III-1 in [18], we can show $\frac{a_{kj}}{S_j(|\mathcal{C}|)} \rightarrow \beta_{kj} \frac{(M_j - S_j(|\mathcal{C}|) + 1)}{S_j(|\mathcal{C}|)}$, as $M_j \rightarrow \infty$ with fixed ratio $S_j(|\mathcal{C}|)/M_j \leq 1$. Then we have $\sum_{j \in \mathcal{C}} \sum_{l \in \mathcal{C}} \sqrt{\frac{P_j P_l a_{kj} a_{kl}}{S_j(|\mathcal{C}|) S_l(|\mathcal{C}|)}} \rightarrow \sum_{j \in \mathcal{C}} \sum_{l \in \mathcal{C}} \sqrt{P_j P_l \beta_{kj} \beta_{kl} b_j b_l}$, where $b_j = \frac{M_j - S_j(|\mathcal{C}|) + 1}{S_j(|\mathcal{C}|)}$. As for the interference, we have

$$\begin{aligned} & \left\| \sum_{l \notin \mathcal{C}} \sum_{u \in \cup_{(c':l \in c')} \mathcal{S}_{c'}} \sqrt{\frac{P_l}{S_l(|\mathcal{C}|)}} \mathbf{g}_{kl}^H \mathbf{f}_{ul} s_u \right\|^2 \\ &= \sum_{l \notin \mathcal{C}} \sum_{u \in \cup_{(c':l \in c')} \mathcal{S}_{c'}} \left\| \sqrt{\frac{P_l}{S_l(|\mathcal{C}|)}} \mathbf{g}_{kl}^H \mathbf{f}_{ul} \right\|^2 \\ &\rightarrow \sum_{l \notin \mathcal{C}} P_l \beta_{kl}, \end{aligned}$$

where the last step follows from that channels and precoders of different users are independent. Based on the approximation $\mathbb{E} \left[\frac{S}{I+N} \right] \approx \frac{\mathbb{E}[S]}{\mathbb{E}[I+N]}$, we complete the proof by plugging the above results into (10).

APPENDIX B PROOF OF SPECTRAL EFFICIENCY USING MRT PRECODING

We first give the following properties of MRT in the massive MIMO regime.

1) We have $\|\mathbf{g}_{kj}\|^2 = \mathbf{g}_{kj}^H \mathbf{g}_{kj} = \beta_{kj} \sum_{i=1}^{M_j} h_{kj,i}^* h_{kj,i}$. Recalling that $h_{kj,i}$ are i.i.d. Gaussian, we have $\frac{1}{M_j} \|\mathbf{g}_{kj}\|^2 \rightarrow \beta_{kj} \mathbb{E}[h_{kj,1}^* h_{kj,1}] = \beta_{kj}$, as M_j and $S_j(|\mathcal{C}|)$ become large with a fixed ratio $S_j(|\mathcal{C}|)/M_j \leq 1$.

2) Plugging \mathbf{f}_{kj} , we have $\left| \mathbf{g}_{kj}^H \mathbf{f}_{jn} \right|^2 = \left| \mathbf{g}_{kj}^H \frac{\mathbf{g}_{nj}}{\|\mathbf{g}_{nj}\|} \right|^2 = \left| \frac{\sqrt{\beta_{kj} \beta_{nj}}}{\|\mathbf{g}_{nj}\|} \sum_{i=1}^{M_j} h_{kj,i}^* h_{nj,i} \right|^2$, which converges to $\frac{\beta_{kj} \beta_{nj}}{M_j \|\mathbf{g}_{nj}\|^2} \mathbb{E} \left[|h_{kj,1}^* h_{nj,1}|^2 \right] + M_j (M_j - 1) \mathbb{E} \left[h_{kj,1}^* h_{jn,1} h_{kj,2}^* h_{nj,2} \right] = \beta_{kj}$ as $M_j \rightarrow \infty$, since $h_{kj,i}$ and $h_{nj,i}$ are i.i.d. Gaussian for $n \neq k$.

Using the above two properties and similar techniques

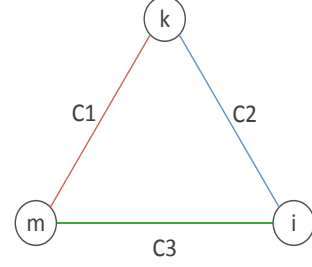


Fig. 5. The graph representation of the associations of three users.

in Appendix A, we have

$$\begin{aligned} \text{SINR}_{k\mathcal{C}} & \approx \frac{\left(\sum_{j \in \mathcal{C}} \sqrt{\frac{P_j}{S_j(|\mathcal{C}|)}} \|\mathbf{g}_{kj}\| \right)^2}{1 + \sum_{j \in \mathcal{C}} (S_j(|\mathcal{C}|) - 1) \frac{P_j}{S_j(|\mathcal{C}|)} \beta_{kj} + \sum_{l \notin \mathcal{C}} P_l \beta_{kl}} \\ & = \frac{\sum_{j \in \mathcal{C}} \sum_{l \in \mathcal{C}} \sqrt{\frac{P_j P_l M_j M_l \beta_{kj} \beta_{kl}}{S_j c_{j\mathcal{C}} S_l c_{l\mathcal{C}}}}}{1 + \sum_{j \in \mathcal{C}} (S_j(|\mathcal{C}|) - 1) \frac{P_j}{S_j(|\mathcal{C}|)} \beta_{kj} + \sum_{l \notin \mathcal{C}} P_l \beta_{kl}}. \end{aligned} \quad (11)$$

Plugging (11) into $\log_2(1 + \text{SINR}_k)$, we complete the proof.

APPENDIX C PROOF OF PROPOSITION 1

We use the techniques similar to the proof of Proposition 3 in [7], where a graph is used to represent the association, and KKT conditions (7) restrict the structure of the graph.

We denote the graph by G_1 , whose nodes represent the users, and the edge between two nodes represents the BS cluster that serves the two users in the considered architecture. Each node has an ID indicating the user index, while each edge has a color that identifies the BS cluster. For example, Fig. 5 shows that user k is served by both clusters \mathcal{C}_1 and \mathcal{C}_2 , and user m is served by both clusters \mathcal{C}_1 and \mathcal{C}_3 .

In a heavily loaded network, where the constraints (6c) are inactive (i.e., $\sum_{\mathcal{C}:|\mathcal{C}|=L} x_{k\mathcal{C}} < \lambda_L$) in the optimal solutions, we have $\mu_{kL} = 0$ for all k . If there are two users k and m being served by size- L clusters \mathcal{C}_1 and \mathcal{C}_2 (i.e., $x_{k\mathcal{C}_1} > 0$, $x_{k\mathcal{C}_2} > 0$, $x_{m\mathcal{C}_1} > 0$, $x_{m\mathcal{C}_2} > 0$), we have $R_k = \frac{r_{k\mathcal{C}_1}}{\sum_{j \in \mathcal{C}_1} \nu_{jL}} = \frac{r_{k\mathcal{C}_2}}{\sum_{j \in \mathcal{C}_2} \nu_{jL}}$ and $R_m = \frac{r_{m\mathcal{C}_1}}{\sum_{j \in \mathcal{C}_1} \nu_{jL}} = \frac{r_{m\mathcal{C}_2}}{\sum_{j \in \mathcal{C}_2} \nu_{jL}}$ from KKT condition (7), where $R_k = \sum_{\mathcal{C}} x_{k\mathcal{C}} r_{k\mathcal{C}}$. Thus, we have

$$\frac{r_{k\mathcal{C}_1}}{r_{k\mathcal{C}_2}} = \frac{r_{m\mathcal{C}_1}}{r_{m\mathcal{C}_2}}, \quad (12)$$

which is true with probability 0. Therefore, it is almost sure that any two users can share at most one same cluster in each architecture. Similarly, we consider an example of three users k, m, i and clusters $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ as illustrated in Fig. 5. We consider the following three cases:

1) If clusters \mathcal{C}_1 , \mathcal{C}_2 and \mathcal{C}_3 are different, we have

$$\begin{aligned} \frac{r_{k\mathcal{C}_1}}{r_{k\mathcal{C}_2}} &= \frac{\sum_{j \in \mathcal{C}_1} \nu_{jL}}{\sum_{j \in \mathcal{C}_3} \nu_{jL}} \frac{\sum_{j \in \mathcal{C}_3} \nu_{jL}}{\sum_{j \in \mathcal{C}_2} \nu_{jL}} \\ &= \frac{r_{m\mathcal{C}_1} r_{i\mathcal{C}_3}}{r_{m\mathcal{C}_3} r_{i\mathcal{C}_2}}, \end{aligned} \quad (13)$$

which is true with probability 0.

2) If $\mathcal{C}_1 = \mathcal{C}_2 \neq \mathcal{C}_3$, we have that users m and i are served both by clusters \mathcal{C}_1 and \mathcal{C}_3 , which is true with probability 0 from (12).

3) If $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}_3$, we have that users k , m and i are served by the same cluster, which is possible. In this case, the graph becomes a *complete graph*.

Therefore, the graph G_1 with three users either contains a loop with the same color edges or no loop. We can get a similar result for graph G_1 with more than three users, where the users served by the same BS cluster constitute a complete graph. Thus, we generate a new graph, denoted by G_2 , where the node represents a cluster. There is an edge between two nodes in G_2 , if these two nodes (i.e., clusters) have a common vertex in G_1 (i.e., there is at least one user served by both these two clusters). Thus the number of users who are served by more than one cluster is limited by the edge of G_2 . Note that there are N_L nodes and no loop in G_2 . Thus, G_2 is a tree, which has the maximal number of edges being one less than the number of nodes (i.e., $N_L - 1$). Therefore, the number of users served by multiple BS clusters equals the number of edges in graph G_2 , which is no more than $N_L - 1$.