Received September 21, 2020; accepted October 12, 2020. Date of publication 21 October 2020; date of current version 25 November 2020.

Digital Object Identifier 10.1109/OJCOMS.2020.3032351

Energy-Efficient Resource Allocation in Single-RF Load-Modulated Massive MIMO HetNets

MAHTAB ATAEESHOJAI[1], ROBERT C. ELLIOTT[1] (Senior Member, IEEE), WITOLD A. KRZYMIEŃ[1] (Fellow, IEEE), CHINTHA TELLAMBURA[1] (Fellow, IEEE), AND JORDAN MELZER[2] (Member, IEEE)

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada

[2]Department of Technology Stategy, TELUS Communications, Ottawa, ON K1P 0A6, Canada

CORRESPONDING AUTHOR: W. A. KRZYMIEŃ (e-mail: krzymien@ualberta.ca)

This work was supported in part by the Rohit Sharma Professorship, by TELUS, and by the Natural Sciences and Engineering Research Council of Canada.

ABSTRACT Due to the dramatic increase in wireless data traffic and the associated increase in energy consumption, designing energy-efficient wireless networks with improved spectral efficiency is a pressing concern. The focus of this article is the design of a green, highly energy-efficient cellular heterogeneous network (HetNet) by taking advantage of multiple-input-multiple-output (MIMO) structure and deployment of small cells. We consider the downlink of a two-tier HetNet, in which multiple-antenna small cells are coordinated to serve users. Even though the deployment of MIMO together with small cells improves the communication system's performance in terms of data rate and reliability, circuit energy consumption in such a network is a critical issue. To address this, an energy-efficient antenna selection and radio resource block assignment algorithm is proposed for the small cells, and a single radio-frequency (RF) chain structure is considered for the massive MIMO macro base station. Then, while coordinating transmissions between cells subject to user-centric clustering, an energy-efficient beamforming design and power allocation optimization problem with respect to the quality of service requirement of users, transmit power budget of base stations, and fronthaul capacity is formulated; the problem is solved using the Dinkelbach method. Simulation results demonstrate the performance potential of our proposed algorithm in terms of energy efficiency and spectral efficiency.

INDEX TERMS Multiple-input-multiple-output (MIMO) system, small cells, energy efficiency, interference management, radio resource allocation, heterogeneous cellular networks (HetNets), coordinated transmission.

I. INTRODUCTION

FIFTH GENERATION (5G) cellular systems are expected to answer increasing capacity demands and quality of service (QoS) requirements of mobile users. For example, a seven-fold increase in global mobile data traffic is forecast between 2017 and 2022 [1]. Moreover, in order to make the global telecommunication network greener, enhancing energy efficiency (EE) is also of very high interest. 5G networks are expected to provide increases in EE commensurate with their improvements to spectral efficiency (SE) [2]. Multiple-input multiple-output (MIMO) transceiver structures, small cell (SC) deployment, and advanced interference mitigation techniques are the key candidates to help enable the greener 5G network [3].

In particular, large-scale antenna arrays, also known as massive MIMO, are of interest for 5G systems, due to several beneficial features that arise from having many antenna elements. These include increased SE and EE for no additional transmitted power, enabling the use of very simple linear precoding methods, and robustness to fading and interference [4], [5]. Massive MIMO transceivers are supported in the 3rd Generation Partnership Project (3GPP) standards in Release 15 and above [6], [7]. Moreover, massive MIMO transceivers have been commercialized

and have already been implemented in practical cellular systems [6], [8].

The combination of MIMO and small cells overlaying larger ones forms a two-tier MIMO heterogeneous cellular network (HetNet), which can increase both SE and EE significantly and meet 5G requirements, but also brings new challenges. In practice, densification of cells causes severe inter-cell interference, which restricts performance gains and the commercial deployment of HetNets. To mitigate interference, increase the cell-edge throughput, and enable the potential gains of HetNets, it is crucial to utilize advanced signal processing techniques; coordinated multipoint (CoMP) transmission and reception is a potential solution [9], [10]. This technique has been introduced for Long Term Evolution - Advanced (LTE-A) and standardized by the 3GPP since Release 11 [11]. CoMP is considered to be a fundamental technique in 5G networks [12], [13]. It is also essential in the implementation of cell-free massive MIMO [4, Ch. 7.4.3], [8], [14]. Qualcomm has implemented a 5G CoMP testbed for high capacity and ultra-reliable communications, the results of which have indicated that CoMP is an important technology for 5G cellular networks [15]. The combination of small cells, CoMP, and massive MIMO has been investigated for 5G in [16], where its effect on SE has been discussed.

Deploying CoMP in HetNets adds complexity and signaling overhead and heavily depends on the backhaul constraints and density of SCs [17]. A cloud radio access network (CRAN) or heterogeneous CRAN (H-CRAN) design is a potential solution to handle these issues [18], [19]. The CRAN architecture is composed of a baseband unit (BBU) that performs baseband processing, connected by radio-over-fiber to remote radio heads (RRHs) that act as distributed transceivers. Radio resource allocation can be performed jointly for the connected RRHs at the BBU. At the same time, coordinated downlink transmission in a cluster of SCs can mitigate the inter-cell interference within the cluster, while precoding/beamforming of signals can mitigate intra-cell interference between users. It is well accepted that coordination should be localized to nearby cells/antennas, as there are diminishing returns for increasing the antenna set coordinated for each user, resulting in increasing processing costs and fronthaul traffic approaching its capacity limits.

Although CRANs are commonly envisioned under the assumption of a fiber-based backhaul and/or fronthaul, some network operators have also considered the use of a wireless backhaul/fronthaul instead, such as a millimeter-wave mesh backhaul [13]. Even though a fiber-based fronthaul is more reliable and has a much larger capacity, in some environments it is not possible to deploy it. The deployment of fiber may also incur substantial costs for installation or leasing, which smaller operators may be unwilling or unable to contend with. In comparison, a wireless fronthaul is cheaper and more flexibly deployed (which also aids cell densification), but has a much smaller and variable capacity. Hence, when optimizing the performance of a network, a limited fronthaul capacity should be accounted for [20]. In [18] and [19], the available research and challenges of massive MIMO-enabled H-CRANs are surveyed, and the issues of system architecture, spectral and energy efficiency performance, and promising key techniques are discussed.

In a HetNet, the joint optimization problem of allocating resources (e.g., antennas, power, etc.) under constraints (e.g., on available transmit power, minimum user rates, etc.) while also designing the precoding strategy is complex. Most commonly, the allocation strategy approach is intended to maximize the system sum rate or SE; [21]–[23] provide just a few examples of different approaches to it. The EE maximization (EEmax) problem is less commonly considered, but is of increasing importance in consideration of the energy consumption and environmental impact of 5G networks.

A. RELATED WORKS

In [24], an energy-efficient resource allocation problem has been formulated for the downlink of an orthogonal frequency division multiple access (OFDMA) H-CRAN. By considering the power constraints of RRHs and QoS requirements of users, a non-convex optimization problem has been formulated there to maximize EE. For inter-tier interference mitigation, an enhanced soft fractional frequency reuse method has been used. In [25], the authors have proposed a joint resource block (RB) and power allocation algorithm for OFDMA-based femtocell HetNets, where they aim to maximize the weighted sum of the individual energy efficiencies and the network energy efficiency. In [26], an EEmax problem has been formulated as a multi-objective optimization problem for subchannel assignment and power allocation in an OFDMA HetNet. In [27], energy-efficient beamforming design and power allocation for both macro cells and SCs has been addressed considering user QoS requirements and transmit power budget constraints. In [28], base stations (BSs) cooperate with each other to jointly design their linear precoders to maximize the network's EE. Block diagonalization (BD) [29] has been used for the precoders, and both centralized and decentralized approaches have been considered for the EEmax optimization problem. To further increase the EE, the problem of joint BS selection/muting and precoder design has been considered as well. The authors of [30] have studied energy-efficient transmit power control for both cooperative and non-cooperative regimes in MIMO HetNets, incorporating BS and antenna activation control schemes. Reference [31] has investigated joint antenna selection and spatial switching for QoS-constrained EE maximization in a MIMO simultaneous wireless information and power transfer (SWIPT) system. It has considered a non-convex joint optimization problem of eigenchannel assignment, power allocation, and active receive antenna set selection. The authors of [32] have considered a problem of joint antenna selection and power

allocation for a massive MIMO transmitter to maximize EE. An effective iterative algorithm based on a Lagrangian dual method was also proposed to solve the non-convex problem.

The authors of [33] have proposed a framework to evaluate the spectral and energy efficiency for massive MIMO HetNets that guarantees user QoS, mitigates interference, and provides sufficient capacity for a wireless backhaul. To solve the non-convex optimization problem, an algorithm based on Lagrange duality and successive convex approximation has been proposed. In [34], interference management and the power allocation problem have been jointly considered for a MIMO non-orthogonal multiple access (NOMA) HetNet. First, to cancel the inter-cluster and co-tier interference, an interference alignment and coordinated beamforming technique has been proposed for both macro cells and SCs. Then, the cross-tier interference is managed by optimizing the allocated power to the macro BS and SC BSs to maximize the sum rate. Reference [35] has investigated the resource optimization problem of NOMA HetNets with SWIPT. By decoupling subchannel allocation and power control, a low-complexity subchannel matching algorithm has been designed. Then, an EEmax problem was solved for optimal power allocation using Lagrangian duality. The authors of [36] have considered the problem of joint user association, carrier allocation, antenna selection, and power control in the uplink of a MIMO HetNet to maximize the data rate of small cell users, by imposing a maximum threshold on the cross-tier interference. By decomposing the original problem into two subproblems and finding an iterative solution, a locally optimal solution has been obtained. Reference [37] has investigated the EEmax problem via a joint design of sub-channel assignment, power control, and antenna selection for the uplink of a multi-cell network. The problem was first formulated as a multi-objective optimization problem and then converted into a single objective optimization problem via the weighted Tchebycheff method. To tackle the intractability, a suboptimal resource allocation algorithm based on the majorization minimization approach has also been proposed.

For the resource allocation problem in HetNets, designing efficient clustering methods and user association strategies is essential for obtaining good system capacity and achieving interference management [38]. However, most of the articles listed above assume that users have already been associated with BSs. The authors of [39] have considered the joint optimization problem of user association, subchannel allocation, and power allocation for downlink transmission in a multi-cell multi-association (i.e., where users may associate with more than one BS) OFDMA HetNet, with single-input single-output (SISO) and single-input multiple-output (SIMO) scenarios. By dividing the weighted sum-rate maximization problem into two subproblems, a locally optimal solution has been obtained by alternating between solving these two subproblems. A similar approach has been followed by the authors of [40], for which a deep reinforcement learning method has been utilized to tackle the user association and resource allocation optimization problem. In [41], a joint user association and power allocation optimization problem has been tackled using non-cooperative game theory in a relay-based ultra-dense HetNet. To maximize the total rate of the users while guaranteeing QoS requirements and throughput balance, the proposed game has been divided into two sub-games and an iterative algorithm has been implemented to perform the sub-games in sequence and guarantee convergence.

B. PROBLEM DESCRIPTION

In this article, we study the network EE for the downlink of a two-tier multi-carrier MIMO coordinated HetNet. The network structure is quite similar to an H-CRAN, in that a central processing node is assumed to collect CSI and perform calculations related to precoding and resource allocation; however, the transmit nodes are BSs and not simpler RRHs. Frequency selectivity of the broadband channel is addressed through the use of OFDMA, splitting up the broadband channel into narrowband frequency-flat subchannels. Throughout most of this article, perfect CSI is assumed to be available at the central processing node; a detailed examination of imperfect CSI is left for future work. The macro BS is equipped with a large-scale antenna array (i.e., massive MIMO), whereas each SC BS is equipped with a few antennas. All users are assumed to have a single receive antenna. Minimum data rate constraints are imposed to ensure a degree of fairness of resource allocation to users. Furthermore, since one of the performance limiting factors of the network may be its limited-capacity fronthaul to send user data and allocation decisions to the BSs, we consider the capacity of the fronthaul as a constraint, as well as the standard transmit power limits per BS. Since we have a two-tier network, in which frequency is reused densely, inter-tier and inter-cell interference may be severe without BS coordination. Therefore, we use coordinated beamforming within our HetNet. The CoMP clusters are user-centric; that is, the group of BSs chosen to coordinate for each user is customized for that user (as opposed to, for example, several fixed sets of nearby coordinated transmit nodes). To mitigate interference from the SC tier, null-space projection beamforming similar to BD [29] is applied at the SCs, whereas zero-forcing (ZF) beamforming is performed at the macro BS to mitigate interference from it. The overall precoder design problem is incorporated into the power allocation problem when maximizing the EE.

In an EEmax problem, besides the radiated power used to transmit data, additional power that is independent of the data also contributes to the transmitter energy usage. For example, the latter category includes power used by baseband processing, mixers, digital-to-analog converters, filters, etc. Also, power is consumed by the radio frequency (RF) chain of each antenna used for the transmission. This means that utilizing the most appropriate number of antennas for the transmission can lead to higher EE. Antenna selection

has also been used to improve the EE in [31], [32], [42]. Moreover, with the use of OFDMA, there arises the problem of assigning users to different RBs on the various subchannels; most related work only considers single-carrier systems. Hence, overall we propose a low-complexity energy-efficient joint antenna selection and RB allocation scheme for the SCs. The selection is based on the Frobenius norm (F-norm) of the user's channel gains from the antennas of its serving cell, taken over all antennas when doing antenna selection, and over the selected active antennas when doing RB allocation.

At the macro BS, due to the use of massive MIMO, the power dissipated in antenna RF chains could be significantly higher in a traditional architecture with one RF chain per antenna. Some works in the literature have proposed to reduce massive MIMO hardware complexity by employing hybrid analog-digital structures with fewer RF chains [43]. However, in this article, we instead adopt a single-RF-chain transmitter structure that is based on load modulation instead of voltage modulation [44], [45]. This structure requires no mixer and only a single power amplifier (PA) outputting a constant-envelope sinusoid, yet can still obtain the full spatial multiplexing gain as the traditional transceiver structure. We refer the reader to [44], [45] for more details.

5G and beyond cellular systems are evolving and getting more complex. Hence, in this work, we incorporate the techniques described earlier in the introduction together, so as to provide a reasonable approximation to a realistic system. Our reasons for the choice of these techniques to be combined are the following:

- Massive MIMO is a key enabler for enhanced SE and EE and part of the 5G standards;
- HetNets are commonly considered for improving the network performance and reflect modern cellular network layouts;
- CoMP reduces cell-edge and inter-cell interference caused by cell densification and HetNet layouts;
- H-CRAN is a potential architecture to help enable CoMP, and it is included in the standards;
- OFDMA is the preferred approach to enable high bit-rate transmission on broadband frequency-selective channels;
- Antenna selection can reduce the power consumption by RF chains in the small cells;
- Antenna selection is not desirable in a massive MIMO system (doing so may cause the transceiver to leave the so-called "massive MIMO regime" [4]), so we consider the load-modulated transceiver architecture instead for its potential EE savings.

C. MAIN CONTRIBUTIONS

Our focus in this article is to maximize the EE of a MIMO-enabled H-CRAN, which is a candidate architecture for 5G systems [18], [46]. In an H-CRAN, we need to consider two factors to achieve an acceptable and energy-efficient system performance. First, due to the potential capacity constraint of the fronthaul, to manage the interference CoMP should be limited to the BSs near a given user. Hence, we consider energy-efficient user-centric clustering and CoMP precoding is performed at both the macro cell and SCs. Second, radio resources (i.e., RBs and power) should be optimally allocated to maximize EE. Therefore, in this article, a joint RB allocation and antenna selection algorithm is proposed and power allocation optimization is performed. To further reduce power consumption, in addition to antenna selection at the SCs, a load-modulated single-RF-chain structure is also considered for the massive MIMO macro BS.

Overall, we can summarize our contributions as follows.

- While previous work in the literature considers the EEmax problem in various scenarios, in our examination we combine the factors of massive and small-scale MIMO in a two-tier HetNet, OFDMA, coordinated beamforming, user-centric clustering, antenna selection and RB assignment, transmit power constraints and allocation, minimum data rate constraints, and fronthaul capacity constraints, altogether at the same time. To our knowledge, the examination of a system combining all these factors simultaneously has not been well investigated in the literature. For example, [24]–[27] do not consider antenna selection, [27], [28], [30]–[33] consider single-carrier systems, while [31] and [32] are also single-tier. A constrained fronthaul seems a particularly rare consideration for EEmax problems; none of the above papers includes it. (While [33] does ensure sufficient data rates are allocated to the backhaul for the SCs, it does not set an explicit maximum constraint on those rates.) Table 1 summarizes some of the related work in the literature and compares it to the work in this article.
- Our use of the single-RF-chain transceiver structure for the massive MIMO macro BS is also relatively novel, especially in the EEmax context. We formulate a power consumption model for this type of transceiver, and examine the EE both of the macro cell and the overall network in comparison with a traditional transceiver structure with one RF chain per antenna.
- We also propose a novel joint antenna selection and RB allocation algorithm, for which we compare its complexity and EE performance with other algorithms. For the EEmax power optimization problem itself, our use of coordinated beamforming with user-centric clustering together with our choice of precoding allows the system to be approximated by one in which users do not experience interference. This further allows the non-convex problem to be reformulated into an equivalent feasible convex one, which can be solved using the commonly-used Dinkelbach method [47], [48]. We derive closed-form expressions for the optimal power allocation using the Lagrange dual decomposition method. Simulations examine the no-interference approximation to show when and how well it holds. We furthermore examine the effect of cell association bias on the EE in our system. The results demonstrate that our overall

TABLE 1. Summary of related work and comparison with our proposed approach.

Ref.	Type of System & Communication	Objective Function	Constraints* (per cell)	Optimization Variables	Interference Management	CoMP and User Association (UA) Considered?
[24]	HetNet, downlink, SISO, multi-carrier	EE	Min. user rates, max. inter-tier interference	RB allocation, power allocation	Enhanced soft fractional frequency reuse, inter-tier interference constraint	No
[25]	HetNet, downlink, SISO, multi-carrier	Weighted EE	Min. user rates, proportionally fair rates for delay-tolerant users, max. one user per RB per cell	RB allocation, power allocation	One user per RB per cell, ∴ no intra-cell interference	No
[27]	HetNet, downlink, MIMO, single-carrier	EE	Min. user rates	Precoding vector, power allocation	ZF/BD precoding; multiple cases considered, cancelling only intra-tier or cancelling cross-tier interference	Network-wide CoMP in some cases; no UA
[32]	Homogeneous network, downlink, MIMO, single-carrier	EE	Min. user rates	Antenna selection, power allocation	ZF	No
[33]	HetNet, downlink, MIMO, single-carrier	Multiobjective SE-EE tradeoff	Min. user rates, incoming wireless backhaul rates to SCs greater than outgoing aggregate rates to users	Power allocation to users and backhaul, backhaul bandwidth allocation	Regularized ZF	No
[35]	HetNet, downlink, SISO, NOMA, multi-carrier	EE	Min. user rates, max. cross-tier interference, max. two users per subchannel	Subchannel allocation, power allocation	Cross-tier interference constraint	No
[37]	Homogeneous network, uplink, MIMO, multi-carrier	EE	Min. user rates, max. one user per subchannel, max. one antenna per user	Antenna selection, subchannel allocation, power allocation,	No intra-cell interference	No
[39]	HetNet, downlink, SIMO, multi-carrier	Weighted SE	One user per BS and one BS per user on each subchannel	User association, subchannel allocation, power allocation	One user per subchannel per cell, ∴ no intra-cell interference	Only UA (multi-assoc. users receive different data from each BS)
[41]	HetNet w/ relays, downlink, SISO/SIMO, multi-carrier	SE	Min. user rates, max. one user per RB	User association, power allocation (all RBs are equivalent per user)	One user per subchannel per cell, ∴ no intra-node interference	Only UA
Our Paper	HetNet, downlink, MIMO, multi-carrier	EE	Min. user rates, fronthaul capacity, max. number of users per RB equals number of active antennas	Antenna selection, RB allocation, power allocation	ZF/BD for intra-cell; CoMP for inter-cell within clusters	Both (user-centric CoMP clusters)

* All the listed papers also have a maximum transmit power constraint.

proposed scheme is more energy efficient than the reference schemes.

The remainder of this article is organized as follows. In Section II we present the system model, describe the cell association and user-centric clustering methodology used, and outline the performance metrics of interest. This section also contains the proposed power model for the single-RF-chain transmitter. Section III describes the proposed antenna selection and RB allocation algorithm, the precoding design, and the EEmax problem formulation. The methodology and algorithm to optimize the solution for the power allocation problem are given in Section IV, and the simulation setup and results are presented in Section V. Finally, we conclude this article in Section VI.

Notation: Variables in italics denote scalars, whereas boldface uppercase and lowercase variables denote matrices and vectors, respectively. A calligraphic variable denotes a set. $(\cdot)^T$, $(\cdot)^H$, and $(\cdot)^\dagger$ denote the transpose, Hermitian transpose, and Moore-Penrose pseudoinverse of a matrix, respectively. $\mathbf{I}_n$ means the $n \times n$ identity matrix. $||\mathbf{A}||_F$ is the Frobenius norm of a matrix $\mathbf{A}$. $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor and the ceiling operators, respectively.

II. SYSTEM MODEL

A. CELLULAR NETWORK MODEL

We consider the downlink of a HetNet where a macro cell containing a massive MIMO enabled macro BS with N_M antennas is densely overlaid with S SC BSs each equipped with N_S antennas. The total number of single-antenna users served by all cells is N_U. For simplicity, we assign $s = 0$ to the macro BS; then, the set of all BSs can be denoted as $\mathcal{S} = \{0, \ldots, S\}$.

OFDMA is utilized in the network in order to convert frequency-selective MIMO channels into a series of RBs on

parallel frequency-flat fading subchannels. Each subchannel is assumed to be quasi-static in time, meaning that the channel gains stay constant during each transmission, then change independently for the next transmission. The total bandwidth of the system is W and B RBs each with a bandwidth of $W_0 = W/B$ are available in each scheduling interval.

In a dense HetNet deployment, interference is the primary bottleneck that limits the system performance, especially for users near the edges of cells. To address this problem, we use a structure similar to a CRAN as a solution, with a centralized processing structure that is inherently suited to mitigate interference through the use of CoMP. Specifically, we consider coordinated beamforming, as it requires less overhead data transfer on the capacity-constrained fronthaul. In CoMP, a set of BSs that coordinate transmissions is often called a cluster. Clusters can be fixed or dynamically changing over time, but in both cases, they can potentially end up shifting the interference problem from the cell edge to the cluster edge (i.e., they cause users near the cluster edge to experience the worst interference). To avoid this, we form user-centric clusters; that is, the set of BSs that coordinate for each user is customized for that user. Only nearby SCs coordinate for any given user; it is unnecessary to cluster distant ones, since they cause little interference, and doing so would contribute needlessly to the overhead on the fronthaul. The macro BS may also be part of the cluster. We discuss our clustering method in more detail in the following subsection.

B. CELL ASSOCIATION AND USER-CENTRIC CLUSTERING

The layout of our proposed clustered HetNet is shown in Fig. 1. Joint optimization of user association to cells and resource allocation would lead to an overall optimal solution when aiming to maximize the EE of the system. For example, the authors of [49] have proposed an algorithm based on game theory to solve the joint problem of user association, power allocation, and frequency subband assignment. Their goal was to maximize a system utility metric with typical constraints on the maximum transmit power and the total number of RBs assigned. Unfortunately, in problems like ours in which many parameters need to be optimized while satisfying several constraints, the non-convex joint optimization problem will become so complicated that the optimal solution can only be obtained by exhaustive search with an infeasibly large computational load [50]. Thus, we instead decompose the mixed-integer nonlinear programming problem into a series of subproblems [51].

The problem of energy-efficient user association has been recently discussed in the literature and several analytical and heuristic algorithms have been proposed. However, most of these schemes have made some simplifying assumption that is not applicable to our problem. For example, the authors of [52] have examined a user association optimization problem to maximize both EE and SE along with a heuristic solution method; however, they have assumed single-antenna

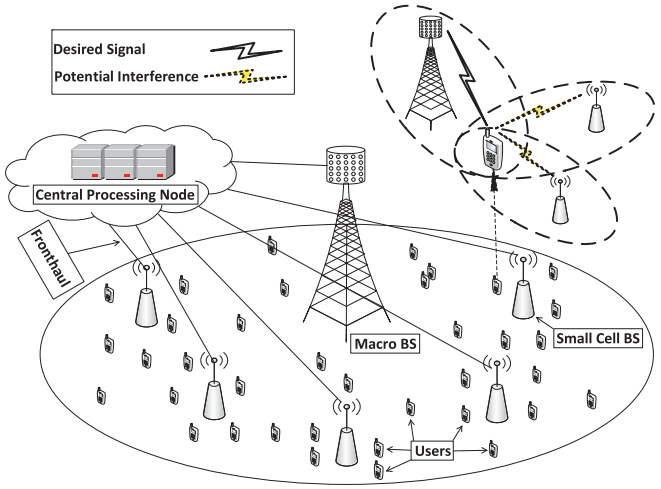

FIGURE 1. System layout of coordinated HetNet with clustered MIMO transceivers. In this CRAN-based network, there is a massive MIMO macro BS with N_M antennas whose coverage is overlaid with several densely-deployed SC BSs each equipped with N_S antennas. Fronthaul links with limited capacity connect all BSs to the central processing node, which coordinates user clustering, precoding, and resource allocation for its connected BSs.

transmitters and equal power allocation between subcarriers. By modifying the method of [52] for a multiple-antenna scenario, though, it can be used as a benchmark for comparison with our proposed scheme, as discussed more in Section V. In this article, we adopt a biased cell association policy [53] along with user-centric clustering. Biased user association leads to better load balancing [54], [55], but the improvement in performance due to load balancing may not completely compensate for the SINR degradation resulting from associating with weaker BSs. Thus, the choice of the bias factor to maximize EE is an important problem, which will be examined in Section V.

Let the average received signal strength or channel quality (in dB) received from BS s by user n be denoted by $\gamma_{s,n}$. The channel quality (along with the CSI in general) can be measured using a reference signal. In FDD systems, the users would measure the downlink channel and feed back this information to the BSs. In TDD systems, the BSs could measure uplink channels using reference signals sent by users and then assuming channel reciprocity obtain CSI for the downlink. $\gamma_{s,n}$ is proportional to the maximum transmit power available at BS s, and accounts for path loss and shadowing, but not small-scale fading. We then define a bias Υ_s in dB in favor of BS s for load balancing purposes [56]. User n then associates with and is served by cell s_n^* for which

$$s_n^* = \arg\max_s \{\gamma_{s,n} + \Upsilon_s\} \tag{1}$$

For our user-centric clustering, each user will be served by one BS in its cluster, while the remaining BSs in the cluster perform coordinated beamforming for that user. We follow a similar approach as in [23] to select BSs for the cluster of each user. The cluster is based on the difference of the average received signal strength. BSs whose signal

strength is within ζ_c dB of the signal strength of the serving BS s_n^* for user n are selected for the cluster $\mathcal{S}_n$ for user n:

$$\mathcal{S}_n = \left\{ l \in \mathcal{S} | \gamma_{s_n^*,n} - \gamma_{l,n} \leq \zeta_c \right\} \tag{2}$$

The value of the clustering threshold ζ_c should be chosen such that the interference from BSs that are not included in the cluster for user n is negligible. From (2) it is clear that in general cell-center users will have fewer BSs in their clusters than cell-edge users. There is some similarity between our work and [27], in which an "interference zone" around each SC is considered when forming clusters, to determine whether the BS should be in a given cluster. The difference is that the clusters in [27] are therefore BS-centric, whereas we perform user-centric clustering. Both the association bias and the clustering threshold can potentially also be used for load balancing, to offload users to lightly-loaded cells and/or avoid overloading in other cells. However, this aspect is outside the scope of this work.

Let $\mathcal{K}_s$ (with cardinality K_s) denote the set of users that are associated with and receive information from BS s, and let $\mathcal{I}_s$ (with cardinality I_s) denote the users that are associated with other BSs, but that have BS s as a member of their cluster. Hence, BS s coordinates with other BSs to design its precoding vectors such that it does not interfere with the users in $\mathcal{I}_s$. Then, we define the set $\mathcal{L}_s = \mathcal{K}_s \cup \mathcal{I}_s$, including all the users that have BS s in their cluster, with cardinality $L_s = K_s + I_s$.

C. PERFORMANCE METRICS

In a MU-MIMO system, several users can receive their data at the same time in each RB, which causes different levels of interference. The complex-valued baseband signal $y_{s,n,b}$ received by user n served by BS s in RB b is expressed as

$$\begin{aligned} y_{s,n,b} = &\sqrt{\Gamma_{s,n}}\mathbf{h}_{s,n,b}\mathbf{f}_{s,n,b}x_{s,n,b} \\ &+ \sum_{i \in \mathcal{K}_s \setminus \{n\}} \sqrt{\Gamma_{s,n}}\mathbf{h}_{s,n,b}\mathbf{f}_{s,i,b}x_{s,i,b} \\ &+ \sum_{r \in \mathcal{S} \setminus \{s\}} \sum_{j \in \mathcal{K}_r} \sqrt{\Gamma_{r,n}}\mathbf{h}_{r,n,b}\mathbf{f}_{r,j,b}x_{r,j,b} + n_{s,n,b}, \end{aligned} \tag{3}$$

where $\mathbf{f}_{s,n,b} \in \mathbb{C}^{N_s \times 1}$ and $x_{s,n,b}$ are the complex-valued beamforming vector and data symbol from BS s to user n in RB b, respectively (cf. [4]). We also define $\mathbf{t}_{s,b} = \sum_n \mathbf{f}_{s,n,b}x_{s,n,b}$ as the transmitted signal vector from BS s on RB b. $\Gamma_{s,n}$ is the large-scale signal power gain/attenuation between BS s and user n, which includes path loss and log-normal shadowing. $n_{s,n,b} \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) at user n on RB b.

$\mathbf{h}_{s,n,b} \in \mathbb{C}^{1 \times N_s}$ denotes the small-scale fading of the MIMO channel vector between user n and the BS s for RB b. $\mathbf{h}_{s,n,b}$ is modeled as $\sim \mathcal{CN}(0, \mathbf{R}_{s,n})$, which represents frequency-flat spatially-correlated Rayleigh fading on each subchannel, where $\mathbf{R}_{s,n}$ is the spatial correlation matrix between BS s and user n. The channel gains are independent between users and for each RB b, though we assume each RB has the same spatial correlation matrix for a given s and n. For SCs, the antennas are assumed to be located closer to the ground, so that SC channels experience a rich scattering environment. As such, there is assumed to be no spatial correlation between antennas, so $\mathbf{R}_{s,n} = \mathbf{I}_{N_s}$ for SCs. In contrast, the macro BS antennas are assumed to be higher up, such that scatterers are located only near the users (i.e., localized scattering). In this case, we model element (l, m) of $\mathbf{R}_{0,n}$ by [4, eq. (2.24)]:

$$\exp\left[-2\pi j d_H (l-m)\sin(\phi) - \frac{\sigma_\phi^2}{2}(2\pi d_H (l-m)\cos(\phi))^2 \right] \tag{4}$$

Equation (4) models a Gaussian spread of angles of departure of paths from the antenna array.[1] In (4), d_H is the antenna spacing (in number of wavelengths), ϕ is the angle (in radians) between the user and the antenna array, and σ_ϕ (in radians) is the angular standard deviation (ASD) of the multipath angles around the nominal angle ϕ. Equation (4) is valid when the ASD is small, e.g., below $\frac{\pi}{12}$ radians (15°).

In (3), the summation in the second term is intra-cell interference between users served by BS s, whereas the double summation in the third term represents inter-cell interference from BSs other than the one sending data to user n. The goal of clustering is to reduce the magnitude of the third term as much as possible, ideally such that it becomes negligible.

In our network, resources are pooled and allocated centrally, and CSI can be shared among connected BSs. Since all BSs of any given cluster are connected to the central node where processing is done, the channel and data information of that cluster's user is available at the central node and resource allocation and precoding vector design can be performed collaboratively for all clustered BSs.

The total sum data rate (in bits/s) is calculated as

$$C = \sum_{s=0}^{S} \sum_{n=1}^{N_U} \sum_{b=1}^{B} \delta_{s,n,b} c_{s,n,b}, \tag{5}$$

where $c_{s,n,b}$, the throughput of user n that receives data from BS s on RB b, is

$$c_{s,n,b} = W_0 \log_2(1 + \sigma_{s,n,b}) \tag{6}$$

$\delta_{s,n,b}$ is a binary user-BS association and RB assignment indicator, which is equal to 1 if user n receives data from BS s on RB b, and 0 otherwise. $\sigma_{s,n,b}$ is the signal-to-interference-plus-noise ratio (SINR) and $p_{s,n,b}$ is the power allocated to the nth user from the sth BS on the bth RB. We also denote N_0 as the spectral density of the AWGN. The SINR is expressed as in (7), shown at the bottom of the next page (cf. [4]).

1. Note that we have inserted an additional minus sign at the start of the exp term in (4), in comparison to [4, eq. (2.24)]. The equation in [4] is for signals on the uplink. For the downlink, we add the minus sign to obtain the phase reversal experienced by signals traveling in the opposite direction.

Since our goal is to maximize EE, defined as the sum rate achieved per unit power consumed by the network equipment (in bits/s/W, or equivalently bits/J) [4], it is also necessary to define an accurate power model. The power consumed by each transmission node s includes (radiated) transmit power, dynamic circuit power, and static circuit power [57]:

$$P_s = \frac{1}{\eta_s} \sum_{n \in \mathcal{K}_s} \sum_{b=1}^{B} \delta_{s,n,b} p_{s,n,b} + N_s P_s^{dyn} + P_s^{sta} \tag{8}$$

η_s is the efficiency of the PA, P_s^{dyn} is dynamic circuit power, and P_s^{sta} is static circuit power. Dynamic power refers to the power consumed in the RF chains connected to the antennas. This includes the power consumption of circuitry such as mixers, digital-to-analog converters, filters, etc. Static circuit power is a constant term that includes power consumption by other transceiver circuitry, e.g., baseband processing, etc.

In a traditional digital transceiver structure, there is an RF chain connected to each of the transmit antennas. Thus, dynamic circuit power is proportional to the number of transmit antennas. Our goal is to develop strategies to decrease consumed power and potentially increase the EE of the network. For SC BSs, as they have relatively few transmit antennas compared to the macro BS, by optimal transmit antenna selection the number of antennas can be decreased and higher EE is achievable. However, for massive MIMO, antenna selection is not as viable of an option. Massive MIMO systems need to have a large number of active antennas to be in the so-called "massive MIMO regime" and obtain its benefits (e.g., very narrow spatial beams, near-orthogonal channels, etc. [4]). This number is typically around at least an order of magnitude larger than the number of served users per RB. Therefore, we instead utilize an alternative structure called a single-RF-chain load-modulated transceiver introduced in [44], [45].

In the load-modulated transceiver structure, as shown in Fig. 2, each antenna is connected to a single common PA via a load modulator. Each load modulator is a lossless, reciprocal, two-port network with adjustable complex impedance parameters. Adjusting these parameters changes the complex-valued current that flows to each antenna, and thus determines the complex-valued symbol that is sent from that antenna. Since the load modulator parameters can be set independently for each antenna, this allows the transceiver to support any arbitrary type of modulation and achieve the full spatial multiplexing gain of the array. The PA outputs a constant-envelope sinusoid. As such, the transceiver can use a Class F PA, which can reach an efficiency of about 80% [60]. In order to protect the PA against reflected power, a circulator and matching network are added between the PA

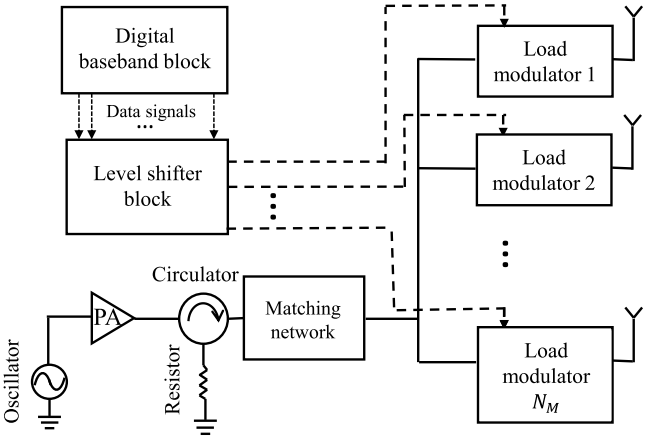

FIGURE 2. Single-RF-chain load-modulated massive MIMO transceiver (cf. [45], [58], [59]).

and load modulators. Some slight variations in the design are presented in [58] and [59].

Despite having only one RF chain, the load-modulated transceiver can support an arbitrary number of users or streams.[2] All digital processing, such as channel coding and precoding, is done at baseband. The output of the digital processing block (e.g., the precoded signals $\mathbf{t}_{s,b}$ from BS s) is used to adjust the levels of the load modulators [59]. The load modulators themselves can be implemented in various ways, such as soft tuning through variable capacitors and/or varactor diodes, or discrete tuning with PIN or Schotkky diodes, micro-electro-mechanical systems, or distributed microstrip transmission lines and switches [45], [59]. The authors of [61] have compared a load-modulated transmitter with another single-RF structure implemented using electronically steerable parasitic array radiator (ESPAR) antennas. It was shown that the load-modulated design reduced the power consumption by $50-81\%$ and yielded $5-42\%$ smaller bit error rates than the ESPAR-based scheme. A physical implementation of a load-modulated transmitter with 4 antennas was demonstrated in [62]. Overall, the single-RF-chain transceiver has the potential for addressing the issues of hardware complexity and EE of massive MIMO transmitters.

With a slight adjustment to the definition in (8), the consumed power for downlink transmission of the single-RF-chain massive MIMO macro BS transmitter can be expressed as

$$P_0 = \frac{1}{\eta_0} \sum_{n \in \mathcal{K}_0} \sum_{b=1}^{B} p_{0,n,b} + P_0^{dyn} + P_0^{sta} \tag{9}$$

2. Restrictions on the number of supported users/streams therefore come from the number of antennas and/or the precoding method, rather than the number of RF chains.

$$\sigma_{s,n,b} = \frac{\Gamma_{s,n} p_{s,n,b} \left| \mathbf{h}_{s,n,b} \mathbf{f}_{s,n,b} \right|^2}{W_0 N_0 + \sum_{i \in \mathcal{K}_s \setminus \{n\}} \Gamma_{s,n} p_{s,i,b} \left| \mathbf{h}_{s,n,b} \mathbf{f}_{s,i,b} \right|^2 + \sum_{r \in \mathcal{S}_n \setminus \{s\}} \sum_{j \in \mathcal{K}_r} \Gamma_{r,n} p_{r,j,b} \left| \mathbf{h}_{r,n,b} \mathbf{f}_{r,j,b} \right|^2} \tag{7}$$

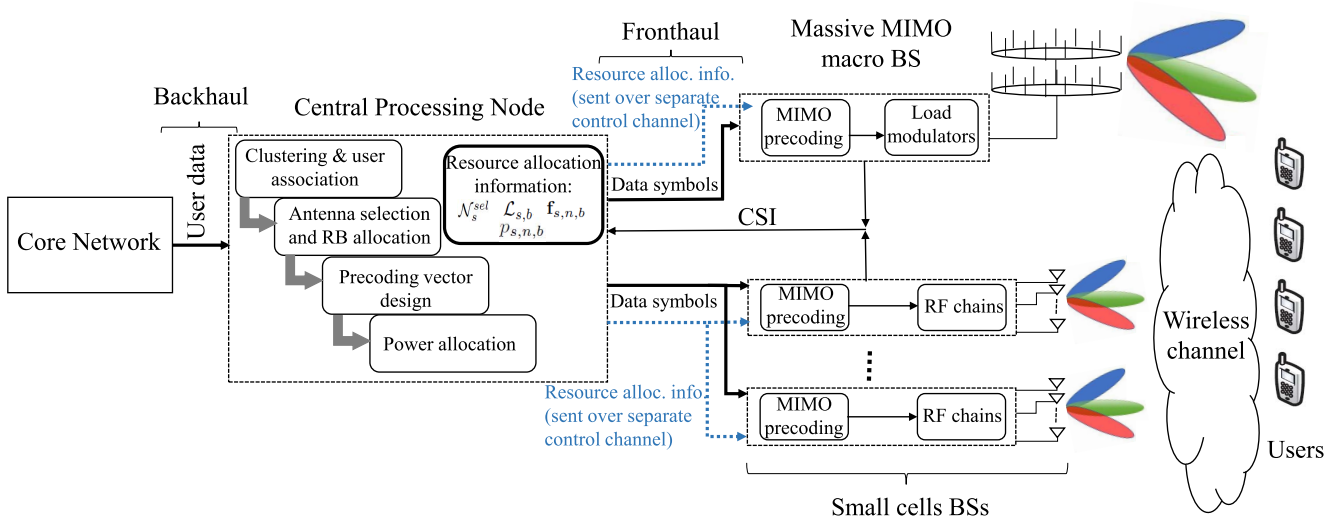

FIGURE 3. The overall block diagram of the system illustrating system architecture, components, and key functionalities.

TABLE 2. Ratio (in dB) of total average reflected power P_R to power amplifier output power P_a, for single-RF-chain transceiver as a function of number of antennas N_M and allowable distortion D_N.

Distortion \ N_M	1	10	100	200
$D_N = 10^{-6}$	-0.45	-2.33	-6.48	-8.16
$D_N = 10^{-5}$	-0.61	-2.86	-7.75	-9.79
$D_N = 10^{-4}$	-0.86	-3.83	-10.48	-13.70

Similarly to as in (8), P_0^{dyn} is the average consumed non-radiated power per antenna, whereas P_0^{sta} is independent of the number of antennas. The key difference is that P_0^{dyn} is dependent on N_M rather than a constant. By using [44, Fig. 8], we derive Table 2 to determine the dynamic consumed power of the macro BS. The table and the referenced figure show the total average power P_R reflected back from the load modulators for all antennas (due to impedance mismatch) relative to the power P_a output by the PA. The reflected power is dissipated by the resistor in the circulator. This ratio is dependent on the number of antennas and the amount of distortion allowed (due to clipping of the signal). Interestingly, the reflected power decreases with more antennas; this results from the impedance being more likely to be matched due to the law of large numbers. We assume any calculations required to adjust the load modulator impedance parameters are included in P_0^{sta}. We also note again that η_0 in (9) will be better in general than η_s in (8) due to the use of a Class F PA.

Finally, the total EE for our proposed scenario can be written as

$$\varepsilon \triangleq \frac{C}{\sum_{s=0}^{S} P_s}, \tag{10}$$

where C is given in (5) and P_s is given in (8) and (9) for the SC BSs and macro BS, respectively.

Fig. 3 shows the overall block diagram of the system.

III. RESOURCE ALLOCATION AND PRECODING DESIGN

A. RESOURCE BLOCK ASSIGNMENT AND ANTENNA SELECTION FOR SMALL CELLS

The optimal RB assignment and antenna selection approach is by exhaustive search, which rapidly grows in complexity with the number of RBs, antennas, and users served per SC. Hence, to reduce the computational complexity, we propose a suboptimal low-complexity RB and antenna selection strategy. The main idea is to decouple the RB and antenna selection into a two-part selection approach. First, we investigate the transmit antenna selection strategy in our MIMO-OFDMA system. Since different antennas may be selected for different RBs, causing all antennas to be activated, the selection cannot be conducted in a per-subchannel manner. Moreover, if any given antenna is selected, it should be used for all RBs as it will result in higher rates with no additional dynamic circuit power consumption. Hence, transmit antenna selection should be performed for all RBs together. RB assignment can then be conducted for each user for the selected antenna set. We take a similar approach as in [31], [42] and use an F-norm-based method to select antennas for SCs. First we define $\arg\operatorname{sort}_{\downarrow j}\{X_j\}$ to return the sorted arguments/indices $\{j\}$ corresponding to when the values $\{X_j\}$ are sorted in descending order. Then, the antennas are sorted according to

$$\underset{j\in\{1,2,\ldots,N_S\}}{\arg\operatorname{sort}_{\downarrow}} \left\{ \left\| \mathbf{g}_{s,j} \right\|_F^2 \right\}, \tag{11}$$

where $\mathbf{g}_{s,j}$ is the jth column of a $K_s B \times N_s$ channel matrix, which represents the channel gain (including small-scale fading, shadowing, and path loss) of the jth transmit antenna for all users served by SC s across all RBs. A similar approach has been used in [31], in which acceptable performance has been shown at lower complexity in comparison to an exhaustive search. The idea behind using channel F-norms for the selection is geared somewhat towards SE rather than EE, in that the selection represents the antenna

with the best mean squared channel gain, averaged over all users and RBs. If the same single symbol was sent to all users on all RBs, the chosen antenna would provide the best SE. Nevertheless, it is also related to maximizing EE, in that the SE would be the highest for a given amount of transmit power (divided equally across users and RBs), or conversely the lowest transmit power would be needed to achieve a certain SE.

After sorting the antennas in descending order using the F-norm-based method, the active transmit antenna set $\mathcal{N}_s^{sel}$ for cell s is selected from the first $N_s^{sel} = \lceil L_s \hat{B}/B \rceil$ antennas, where $\hat{B}$ is the maximum number of RBs that can be allocated to a user. The reasoning behind this value of N_s^{sel} is that N_s^{sel} is also the number of degrees of freedom for spatial multiplexing per RB. Thus, the maximum number of single-antenna users that can be supported per RB by linear precoding is N_s^{sel}, or at most $N_s^{sel}B$ users in total. BS s has L_s users to support (either to serve data or to mitigate interference for), requiring $N_s^{sel}B \geq L_s$, or $N_s^{sel} \geq L_s/B$. If users are to be allocated up to $\hat{B}$ RBs each, that would require at minimum $N_s^{sel} \geq L_s\hat{B}/B$, or $N_s^{sel} = \lceil L_s\hat{B}/B \rceil$.

With the transmit antenna set selected for each SC, RBs can then be allocated to the users. This is done jointly for all BSs in light of coordinated beamfoming — if user n is allocated an RB at its serving cell, user n must also be accounted for on that RB at the other BSs in cluster n for interference mitigation. Let the subset of users from $\mathcal{L}_s$ that are allocated to RB b (either to receive data from BS s or to mitigate interference for other cells) be denoted $\mathcal{L}_{s,b}$, with cardinality $L_{s,b}$. We can similarly define subsets of $\mathcal{K}_s$ and $\mathcal{I}_s$ as $\mathcal{K}_{s,b}$ and $\mathcal{I}_{s,b}$ respectively, with cardinality $K_{s,b}$ and $I_{s,b}$. Similar to the antenna selection, we use an F-norm-based approach. First, we sort the BS-user-RB index triplets according to

$$\underset{(s,n,b),\ \forall s\in\mathcal{S},\ \forall n\in\mathcal{K}_s,\ \forall b\in\{1,\ldots,B\}}{\arg\,\mathrm{sort}_{\downarrow}} \left\{ \frac{\left|\left|\tilde{\mathbf{h}}_{s,n,b}\right|\right|_F^2}{N_s^{sel}} \right\}, \qquad (12)$$

where $\tilde{\mathbf{h}}_{s,n,b} \in \mathbb{C}^{1\times N_s^{sel}}$ is the small-scale fading channel vector to user n on RB b from the selected set of antennas at its serving BS s. In the case of the macro BS, which does not perform antenna selection, $\tilde{\mathbf{h}}_{0,n,b} = \mathbf{h}_{0,n,b}$ and $N_s^{sel} = N_M$. We emphasize that the channel vector in (12) includes only the small-scale fading component of the channel gains, i.e., the path loss and shadowing are normalized out. This is to provide fairness[3] to all users in an effort to meet the minimum data rate constraints, so that users near BSs are not allocated a disproportionately large number of RBs to the detriment of cell-edge users. Each RB potentially can be selected to serve data to any user, as long as the total number of users $L_{s,b}$ sharing any given RB b at any SC s is at most $\hat{U}_s^{alc} = N_s^{sel}$, and the number of RBs $N_{RB,n}$ allocated to user n is at most $\hat{B}$. For the macro BS, however, we impose an additional constraint. Due to the abundance of transmit antennas for massive MIMO, with $N_M \gg K_0$, there is essentially no limit to the number of users that can be served on any one RB. However, allocating too many users to one RB could put pressure on the resources for interference mitigation at the SCs on that RB. Hence, at the macro BS, we shall attempt to divide users among the RBs as evenly as possible, with no more than $\hat{U}_0^{srv} = \lceil (K_0\hat{B}/B) \rceil$ users served on any given RB. The two-part antenna and RB selection algorithm is outlined in Algorithm 1. After antenna selection, RBs are allocated in the order given by (12), but first ensuring that each user is given at least one RB from its serving BS in phase 1. Then, if $\hat{B} > 1$ and sufficient resources remain, additional RBs will be allocated to the users in phase 2, in the order given by (12).

Occasionally, the algorithm may run into corner cases when assigning users to RBs. Users with large clusters require the resource allocation at numerous BSs to be sufficiently coordinated in order to avoid turning on antennas at SC BSs. Specifically, an allocation spot must be available at all BSs in the cluster on the same RB. In other words, each user must be a member of $\mathcal{L}_{s,b}$ for some same value of b for all BSs s in its cluster. If these large-cluster users end up being allocated RBs near the end of the process, there may be insufficient remaining resources at all BSs in the user's cluster on any one RB to allow for coordination in the cluster. To deal with such corner cases if they occur, we restart the allocation process (at line 6 in Alg. 1) with these users at the start of the allocation order. The remaining users are ordered as normal afterwards; since they require fewer resources to be coordinated, they are easier to slot into the remaining positions.

Obviously, there is a trade-off between the achievable sum rate and power usage. As we have discussed, switching off antennas will affect the spatial degrees of freedom in the users' channels (and thus the degrees available to the precoder), as well as the maximum number of users that can be served simultaneously per RB, which will lead to a reduction in the achievable sum rate of the system. However, activating fewer antennas will lead to less power consumption, which is beneficial in terms of EE. This trade-off will be investigated through simulations.

B. PRECODING VECTOR DESIGN

The design of the precoding scheme used for transmission in a MU-MIMO system is an important factor in the resulting EE of the system. The objective of precoding in general is to mitigate intra-cell interference between users of that cell. When coordinated beamforming is used, it can also mitigate intra-tier and inter-tier interference between BSs of a cluster. To begin, we look at mitigating the interference caused by SCs. For lower computational complexity, we use linear precoding. Specifically, we consider the same

3. This has similarities to proportionally fair user scheduling [63], [64], in that both give highest priority to users who have the best channel relative to their average channel.

Algorithm 1 Antenna Selection and Resource Block Allocation

Initialize: $N_k^{\mathrm{RB}}=0, \forall k \in \{1,2,\ldots,N_U\}$;
  $\mathcal{L}_{s,b}=\varnothing$ and $L_{s,b}=0, \forall s \in \mathcal{S}, \forall b \in \{1,2,\ldots,B\}$;
  $\mathcal{K}_{s,b}=\varnothing$ and $K_{s,b}=0, \forall s \in \mathcal{S}, \forall b \in \{1,2,\ldots,B\}$;
Sort antennas based on decreasing F-norm across all subcarriers using (11);
Find best N_s^{sel} antennas and store in $\mathcal{N}_s^{sel}$, $\forall s \in \mathcal{S}\backslash\{0\}$;
Initialize: $\hat{U}_0^{srv}=\lceil K_0\hat{B}/B\rceil$; $\hat{U}_s^{alc}=N_s^{sel}, \forall s \in \mathcal{S}$; $i=1$; $phase=1$;
Using channel F-norms, sort all BS-user-RB triplets (s,n,b) into ordered set $\mathcal{F}$ as in (12);
while [$\exists(L_{s,b} < \hat{U}_s^{alc})$ for any (s,b) pair] AND
    [$\exists(N_k^{\mathrm{RB}} < \hat{B})$ for any k] AND $\mathcal{F}\neq\varnothing$ **do**
  $chk\_next=$ true;
  while chk_next **do**
    Get next best BS-user-RB triplet $(s^*,n^*,b^*)=\mathcal{F}(i)$;
    if $phase=1$ AND $N_{n^*}^{\mathrm{RB}}>0$ **then**
      $i=i+1$;
      if $i>|\mathcal{F}|$ **then**
        $i=1$; $phase=2$;
      end if
    else $chk\_next=$ false;
    end if
  end while
  $\mathcal{C}^*=\mathcal{S}_{n^*}$; $j=1$; $test\_alloc=$ true;
  while $test\_alloc$ AND $j\leq|\mathcal{C}^*|$ **do**
    $s=\mathcal{C}^*(j)$;
    if $s=s^*$ AND [$(s>0$ AND $L_{s,b^*}\geq\hat{U}_s^{alc})$ OR
        $(s=0$ AND $K_{s,b^*}\geq\hat{U}_s^{srv})$] **then**
      $test\_alloc=$ false;
    else if $s\neq s^*$ AND $L_{s,b^*}\geq\hat{U}_s^{alc}$ AND $K_{s,b^*}>0$ **then**
      $test\_alloc=$ false;
    end if
    $j=j+1$;
  end while
  if $test\_alloc$ **then**
    Assign user n^* to RB $b^*, \forall s \in \mathcal{C}^*$;
    Add n^* to $\mathcal{L}_{s,b^*}, \forall s \in \mathcal{C}^*$;
    $L_{s,b^*}=L_{s,b^*}+1, \forall s \in \mathcal{C}^*$;
    Add n^* to $\mathcal{K}_{s^*,b^*}$;
    $K_{s^*,b^*}=K_{s^*,b^*}+1$;
    $N_{n^*}^{\mathrm{RB}}=N_{n^*}^{\mathrm{RB}}+1$;
  end if
  $\mathcal{F}=\mathcal{F}\backslash\mathcal{F}(i)$;
  if $phase=1$ AND $\{N_k^{\mathrm{RB}}>0, \forall k\}$ **then**
    $phase=2$;
  end if
  for $j=1$ to $|\mathcal{C}^*|$ **do**
    $s=\mathcal{C}^*(j)$;
    if $[L_{s,b^*}=\hat{U}_s^{alc}]$ OR $[s=0$ AND $K_{0,b^*}=\hat{U}_0^{srv}]$ **then**
      $\mathcal{F}_{b^*}=\{(s,n,b^*)\}, \forall n \in \{1,2,\ldots,N_U\}$;
      $\mathcal{F}=\mathcal{F}\backslash\mathcal{F}_{b^*}$;
    end if
  end for
  if $N_{n^*}^{\mathrm{RB}}=\hat{B}$ **then**
    $\mathcal{F}_{n^*}=\{(s^*,n^*,b)\}, \forall b \in \{1,2,\ldots,B\}$;
    $\mathcal{F}=\mathcal{F}\backslash\mathcal{F}_{n^*}$;
  end if
end while
Output: $\mathcal{L}_{s,b}, \mathcal{N}_s^{sel}, \mathcal{K}_{s,b}, \forall s \in \mathcal{S}, \forall b \in \{1,2,\ldots,B\}$.

type of null-space projection precoding as is used in BD precoding [29] (described in more detail later). The BD technique can be considered as a generalization of zero-forcing (ZF) precoding to the case where users have multiple antennas. BD has been widely used in related literature [65]–[67]; the same null-space projection technique has also been used in [22], [27], [68] in the context of single-antenna users, as it is with our work herein.[4] We have chosen null-space projection since it has been shown (e.g., in [27]) that, while null-space projection and the channel inversion technique used in "classical" ZF precoding perform identically when maximizing the sum rate of a MU-MIMO HetNet, when considering EE instead, using channel inversion for SCs results in smaller EE in various cases, depending on how much coordination there is within and between tiers of the HetNet.

The precoding is performed on a per-RB basis. With null-space projection precoding, the transmit precoding vector of each user is designed to lie in the null space of the channels of all $L_{s,b}-1$ other users in $\mathcal{L}_{s,b}$. This means the precoding vectors must satisfy [29]:

$$\tilde{\mathbf{h}}_{s,n,b}\mathbf{f}_{s,i,b}=0,\ \ \forall(n,i)\in\mathcal{L}_{s,b}\ \text{such that}\ n\neq i,\ \ \forall(b,s) \tag{13}$$

Let $\tilde{\mathbf{H}}_{s,n,b}\in\mathbb{C}^{(L_{s,b}-1)\times N_s^{sel}}$ be a matrix that vertically concatenates the channel vectors for RB b of all users in $\mathcal{L}_{s,b}$ except for user n:

$$\tilde{\mathbf{H}}_{s,n,b}=\left[\tilde{\mathbf{h}}_{s,1,b}^T\cdots\tilde{\mathbf{h}}_{s,n-1,b}^T\ \ \tilde{\mathbf{h}}_{s,n+1,b}^T\cdots\tilde{\mathbf{h}}_{s,L_{s,b},b}^T\right]^T \tag{14}$$

We denote $\tilde{r}_{s,n,b}$ as the rank of that aggregate null space and $\tilde{\mathbf{V}}_{s,n,b}^0\in\mathbb{C}^{N_s^{sel}\times(N_s^{sel}-\tilde{r}_{s,n,b})}$ as a set of orthonormal basis vectors for that null space [29] (and thus the basis for $\mathbf{f}_{s,n,b}$). The equivalent channel $\ddot{\mathbf{h}}_{s,n,b}\in\mathbb{C}^{1\times(N_s^{sel}-\tilde{r}_{s,n,b})}$ for user n is $\ddot{\mathbf{h}}_{s,n,b}=\tilde{\mathbf{h}}_{s,n,b}\tilde{\mathbf{V}}_{s,n,b}^0$, with $\lambda_{s,n,b}=||\ddot{\mathbf{h}}_{s,n,b}||$ being its channel gain. Hence, the signal received by user n from SC s on RB b can be expressed as

$$y_{s,n,b}=\sqrt{\Gamma_{s,n}p_{s,n,b}}\lambda_{s,n,b}x_{s,n,b}+n_{s,n,b} \tag{15}$$

To summarize, through the use of null-space projection precoding, the precoding vectors are designed such that intra-cell and intra-cluster interference is completely canceled, and the MU-MIMO channels are decomposed into several equivalent non-interfering single-user MIMO channels. In other words, the first summation in both (3) and the denominator of (7) becomes equal to 0, and for BSs that coordinate with SC s, their contribution to the double summation in (3) and the denominator of (7) also becomes equal to 0. In what follows, we assume that through the design of the clusters, the remaining portion of the double summation (i.e., the remaining interference from uncoordinated BSs) is negligible. (We shall examine the effect of this assumption further in the simulations to verify how well it holds.) Under the assumption of no significant interference after precoding, the SINR in (7) reduces to

$$\sigma_{s,n,b}=\frac{\Gamma_{s,n}\lambda_{s,n,b}^2}{W_0N_0}p_{s,n,b}=\chi_{s,n,b}p_{s,n,b}, \tag{16}$$

4. As mentioned, BD has been defined for multiple-antenna users. The technique is still valid for single-antenna users, although in such a case, it is not particularly "block" diagonalization anymore, as the "block" ends up being a scalar (i.e., of size 1×1). For the single-antenna case, we shall therefore refer to it as null-space projection, to differentiate the technique from "classical" ZF precoding, which instead uses channel inversion.

where we define $\chi_{s,n,b}$ as the ratio of the equivalent subchannel power gain to the noise power on that subchannel.

At the macro BS, "classical" ZF precoding is performed to mitigate interference both between macro users and to users of SCs; the precoding is again done per RB. The relatively simpler precoding (compared to null-space projection) is sufficient for the massive MIMO BS, since the law of large numbers makes channel vectors to different users near-orthogonal even without precoding [4]. First, we define $\breve{\mathbf{H}}_{0,b} \in C^{L_{0,b} \times N_M}$ as the channel matrix that vertically concatenates the small-scale fading portion of the channel vectors for the users in $\mathcal{K}_{0,b}$, followed by the users in $\mathcal{I}_{0,b}$:

$$\breve{\mathbf{H}}_{0,b} = \Big[\mathbf{h}^T_{0,\mathcal{K}_{0,b}(1),b}\ \mathbf{h}^T_{0,\mathcal{K}_{0,b}(2),b} \cdots \mathbf{h}^T_{0,\mathcal{K}_{0,b}(K_{0,b}),b}\ \ \mathbf{h}^T_{0,\mathcal{I}_{0,b}(1),b} \cdots \mathbf{h}^T_{0,\mathcal{I}_{0,b}(I_{0,b}),b}\Big]^T \tag{17}$$

Then the ZF precoding vector can be found from [27]

$$\breve{\mathbf{F}}_{0,b} = \breve{\mathbf{H}}^{\dagger}_{0,b} = \breve{\mathbf{H}}^{H}_{0,b}\Big(\breve{\mathbf{H}}_{0,b}\breve{\mathbf{H}}^{H}_{0,b}\Big)^{-1} \tag{18}$$

Consider the first $K_{0,b}$ columns of $\breve{\mathbf{F}}_{0,b}$, and let the nth column of $\breve{\mathbf{F}}_{0,b}$ be denoted $\breve{\mathbf{f}}_{0,n,b}$. The precoding vector $\mathbf{f}_{0,n,b}$ for user $\mathcal{K}_{0,b}(n)$ (i.e., the nth user served by the macro BS on RB b) is

$$\mathbf{f}_{0,n,b} = \frac{\breve{\mathbf{f}}_{0,n,b}}{\left\|\breve{\mathbf{f}}_{0,n,b}\right\|} \tag{19}$$

In this way, the interference from the macro BS has been mitigated. Assuming as we did for the SCs that all remaining interference is negligible, then the SINR for macro users reduces to

$$\sigma_{0,n,b} = \frac{\Gamma_{0,n}\left|\mathbf{h}_{0,n,b}\mathbf{f}_{0,n,b}\right|^2}{W_0 N_0} p_{0,n,b} = \chi_{0,n,b} p_{0,n,b}. \tag{20}$$

C. POWER ALLOCATION

After antenna selection, RB allocation, and precoding vector calculation, the final stage to maximize EE is power allocation. For compactness of notation, let $\mathbf{p}$ be a vector containing all the power allocation variables $\{p_{s,n,b}\}, \forall s, n, b$. Then, the EEmax problem under minimum-rate constraints, maximum fronthaul capacity limitations, and total transmit power constraints is formulated as

$$\max_{\mathbf{p}} \ \frac{W_0 \sum_{s=0}^{S}\sum_{n=1}^{N_U}\sum_{b=1}^{B}\delta_{s,n,b}\log_2\left(1+\chi_{s,n,b}p_{s,n,b}\right)}{\sum_{s=0}^{S}\sum_{n=1}^{N_U}\sum_{b=1}^{B}\frac{1}{\eta_s}\delta_{s,n,b}p_{s,n,b} + \sum_{s=0}^{S}N_s^{sel}P_s^{dyn} + \sum_{s=0}^{S}P_s^{sta}} \tag{21a}$$

$$\text{s.t.} \ \sum_{b=1}^{B} c_{s,n,b} \geq \kappa_{min}, \quad \forall s \in \mathcal{S}, \forall n \in \mathcal{K}_s \tag{21b}$$

$$\sum_{n=1}^{N_U}\sum_{b=1}^{B}\delta_{s,n,b}c_{s,n,b} \leq c_{s,limit}, \quad \forall s \in \mathcal{S} \tag{21c}$$

$$\sum_{n=1}^{N_U}\sum_{b=1}^{B}\delta_{s,n,b}p_{s,n,b} \leq P_s^{\max}, \quad \forall s \in \mathcal{S} \tag{21d}$$

$$p_{s,n,b} \geq 0 \quad \forall s, n, b \tag{21e}$$

$c_{s,n,b}$ is given by (6). P_s^{max} is the maximum transmit power of BS s. κ_{min} is the minimum data rate guaranteed for users and $c_{s,limit}$ is the maximum data rate that can be transferred over the fronthaul links. Like in (12), $N_s^{sel} = N_M$ for the macro BS. The constraints given by (21b) characterize the minimum rate guaranteed for each user and the constraints given by (21d) and (21e) represent the maximum transmit power available at each BS. One of the performance limiting factors of the network can be its limited fronthaul capacity, which needs to be taken under consideration, and hence has been included in our problem as the constraints given by (21c). Our optimization problem is in fractional and nonconvex form, so we have used various methods adopted from related literature to convert the problem to a convex one and obtain optimal allocated power, as described in the next section.

IV. SOLUTION OF THE OPTIMIZATION PROBLEM

Since the optimization problem defined in (21) is classified as nonlinear fractional programming, which results in a nonconvex problem, there is no one standard method for solving it. Our first step is to simplify the objective function using techniques from nonlinear fractional programming. Given that antenna selection and RB assignment have been done, we will maximize EE by optimal power allocation. We denote by ε^* the maximum EE of the overall network, expressed as

$$\varepsilon^* = \frac{C(\mathbf{p}^*)}{P(\mathbf{p}^*)} = \max_{\mathbf{p}} \frac{C(\mathbf{p})}{P(\mathbf{p})}, \tag{22}$$

where $\mathbf{p}^*$ is the optimal power allocation vector, and $\mathbf{p}$ is any feasible solution of the problem in (21) that satisfies the constraints given by (21b)-(21e). Following [47], we can formulate an equivalent problem as follows:

$$\max_{\mathbf{p}} \ \left\{C(\mathbf{p}) - \varepsilon^* P(\mathbf{p})\right\} = C\left(\mathbf{p}^*\right) - \varepsilon^* P\left(\mathbf{p}^*\right) = 0 \tag{23}$$

In other words, for any optimization problem with an objective function in fractional form, there is an equivalent optimization problem with an objective function in subtractive form that leads to the same solution. This has been proven in [23]; see also [69, Appx. A]. Hence, we can concentrate on the equivalent problem in the rest of this article. The equivalent problem can be formulated as

$$\max_{\mathbf{p}} \ \{C(\mathbf{p}) - \varepsilon P(\mathbf{p})\} \quad \text{s.t.} \ \ (21b)\text{–}(21e) \tag{24}$$

Now we must find the optimal value of ε. Since ε^* cannot be obtained directly, an iterative algorithm (based on what is known as the Dinkelbach method [47], which is commonly used for EEmax problems [48]) is proposed, in which the obtained solution ensures (23) is satisfied. Pseudocode for the proposed algorithm is described in Algorithm 2.

The algorithm consists of an outer loop, which updates the value of ε, and an inner loop, which updates $C(\mathbf{p})$ and $P(\mathbf{p})$. Convergence to the optimal solution is guaranteed if

TABLE 3. Computational complexity.

Antenna selection for SCs	$\mathcal{O}\left(\sum_{s=1}^{S} N_s(K_s B + \log(N_s))\right)$
RB assignment for all cells	$\mathcal{O}\left(B[K_0 N_M + \sum_{s=1}^{S} K_s N_s^{sel}] \ + BN_U \log(BN_U)\right)$
Precoding design for SCs	$\mathcal{O}\left(\sum_{s=1}^{S} \sum_b K_{s,b}(N_s^{sel})^2 L_{s,b}\right)$
Precoding design for macro BS	$\mathcal{O}\left(\sum_b N_M L_{0,b}^2\right)$
Power optimization algorithm	$\mathcal{O}\left(n_{outer} n_{inner} B N_U\right)$

one is able to solve the inner problem. As $\varepsilon^{(i+1)}$ is updated in each iteration i in the outer loop with $C(\mathbf{p}^{(i)})$ and $P(\mathbf{p}^{(i)})$ obtained in the last iteration, the value of ε converges towards its maximum. Meanwhile, by solving the inner loop for a given $\varepsilon^{(i)}$, the optimal power policy needed for the next loop would be obtained, with the whole algorithm iterating until all the values converge or some other stopping criterion (e.g., a maximum number of iterations) is reached.

A. SOLUTION OF THE INNER LOOP PROBLEM

The transformed problem can be expressed as in (24), with ε replaced now by $\varepsilon^{(i)}$. The problem is now concave with respect to optimization variable $\mathbf{p}$. We derive the Lagrangian function [70] of the problem as follows:

$$\begin{aligned}
\mathrm{L}(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}) &= \sum_{s=0}^{S}\sum_{n=1}^{N_U}\sum_{b=1}^{B} W_0 \delta_{s,n,b} \log_2\left(1 + \chi_{s,n,b} p_{s,n,b}\right) \\
&- \varepsilon^{(i)}\left(\sum_{s=0}^{S}\sum_{n=1}^{N_U}\sum_{b=1}^{B} \frac{1}{\eta_s}\delta_{s,n,b} p_{s,n,b} + \sum_{s=0}^{S} N_s^{sel} P_s^{dyn} + \sum_{s=0}^{S} P_s^{sta}\right) \\
&+ \sum_{s=0}^{S}\sum_{n\in\mathcal{K}_s} \alpha_{s,n}\left(\sum_{b=1}^{B} c_{s,n,b} - \kappa_{min}\right) \\
&+ \sum_{s=0}^{S} \beta_s\left(c_{s,limit} - \sum_{n=1}^{N_U}\sum_{b=1}^{B} W_0 \delta_{s,n,b} \log_2\left(1 + \chi_{s,n,b} p_{s,n,b}\right)\right) \\
&+ \sum_{s=0}^{S} \mu_s\left(P_s^{\max} - \sum_{n=1}^{N_U}\sum_{b=1}^{B} \delta_{s,n,b} p_{s,n,b}\right) \qquad (25)
\end{aligned}$$

The vector $\boldsymbol{\mu}$ contains the Lagrangian multipliers μ_s corresponding to the maximum transmit power limit for BS s in (21d). $\boldsymbol{\alpha}$ contains the Lagrangian multipliers $\alpha_{s,n}$ associated with the minimum rate constraints in (21b). Finally, $\boldsymbol{\beta}$ contains the Lagrangian multipliers β_s accounting for the fronthaul capacity constraint for BS s in (21c).

For fixed $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, and ε, the problem can be solved utilizing the Karush-Kuhn-Tucker (KKT) conditions [70], [71]. The optimal value of $p_{s,n,b}$ is then obtained by making the partial derivatives of L with respect to $p_{s,n,b}$ equal to zero, which yields a "water-filling" type of solution. We derive the optimal power as:

$$p_{s,n,b}^{*} = \max\left(w_{s,n}^{*} - \frac{1}{\chi_{s,n,b}}, 0\right), \qquad (26)$$

Algorithm 2 Dinkelbach Method for EE Maximization

1: Initialize the convergence threshold λ_{th};
2: Set $i = 1$ and $\varepsilon^{(1)} = 0$;
3: For initial $\varepsilon^{(1)}$, obtain $C(\mathbf{p}^{(1)})$ and $P(\mathbf{p}^{(1)})$ by solving the problem in (24) using (26) and (27);
4: **while** $[C(\mathbf{p}^{(i)}) - \varepsilon^{(i)} P(\mathbf{p}^{(i)}) < \lambda_{th}]$ **do**
5: $\quad i = i + 1$;
6: $\quad$ (Inner Loop): Solve the resource allocation problem in (24) using (26) and (27) with $\varepsilon^{(i-1)}$ to obtain the optimal solution $\mathbf{p}^{*(i)}$;
7: $\quad \varepsilon^{(i)} = \frac{C(\mathbf{p}^{*(i)})}{P(\mathbf{p}^{*(i)})}$;
8: **end while**
9: Output: $\mathbf{p}^{*}$, ε^{*}, C, P.

where

$$w_{s,n}^{*} = \frac{W_0\left(1 + \alpha_{s,n} - \beta_s\right)}{\ln 2\left(\mu_s + \varepsilon^{(i)}/\eta_s\right)}, \qquad (27)$$

and $\chi_{s,n,b}$ is given in (16) and (20). The values of $p_{s,n,b}^{*}$ are used to update the value of $\varepsilon^{(i)}$. Then, the subgradient method can be used for updating the values of the Lagrange multipliers in the outer loop; this method is guaranteed to converge to the optimal Lagrange multipliers, as long as the step size values are chosen to be sufficiently small [72].

When $\hat{B} \geq 1$ and a user has been allocated more than one RB, it can occur that after power optimization, some of the user's RBs may be allocated zero power. (This may occur, for instance, if the user's channel conditions on the RB are particularly bad.) In such an event, we remove that user's assignment to that RB, then go back and recalculate the precoding basis vectors for that RB in the user's cluster. The removal of a user from an RB means that the null space becomes less constrained (i.e., it has higher rank), potentially allowing the rates of other users to be increased. After, Algorithm 2 is re-run to update the power allocation for the users.

B. COMPUTATIONAL COMPLEXITY

The computational complexity of our scheme is given in Table 3. For the first two rows, the complexity comes from the F-norm calculations and sorting in (11) and (12). For the antenna selection, finding the F-norm of a $K_s B \times 1$ vector requires $\mathcal{O}(K_s B)$ floating point operations (flops) [73], and there are N_s such vectors per SC. Sorting the resulting N_s F-norm values has a complexity of $\mathcal{O}(N_s \log(N_s))$ [74]. This is done for each of the S SCs. Similarly, for RB selection,

an F-norm is calculated for each {BS, user, RB} triplet of the channel vectors for that triplet, which have length N_s^{sel} for SC s and length N_M for the macro cell. (The one extra flop for normalization by N_s^{sel} or N_M can be neglected.) This has a total complexity of $B[K_0 N_M + \sum_{s=1}^{S} K_s N_s^{sel}]$. The total BN_U F-norm values are then sorted with complexity $BN_U \log(BN_U)$.

The main complexity of the null-space projection precoding at the SCs is from the singular value decomposition used to find $\tilde{\mathbf{V}}_{s,n,b}^{0}$ [29]. On each RB b, for each of the $K_{s,b}$ users served on that RB, singular value decomposition is done on an $(L_{s,b} - 1) \times N_s^{sel}$ matrix. With $N_s^{sel} > (L_{s,b} - 1)$, this takes $\mathcal{O}((N_s^{sel})^2 L_{s,b})$ flops [73]. We note, though, that the complexity may be smaller with a different choice of precoding. For the macro BS, the complexity of precoding comes from calculating the matrix pseudoinverse in (18), for each RB. The matrix inversion in (18) can be found efficiently by first performing an LU decomposition [73]. However, the whole psuedoinverse operation for the $L_{0,b} \times N_M$ matrix can also be performed via LU [75], Cholesky, or LDL^H decompositions [76]. In each of these cases, the decomposition causes the highest order complexity of the operations involved; all three have complexity $\mathcal{O}(N_M L_{0,b}^2)$ [73].

Finally, the optimization problem acts on scalar values; its complexity comes mostly from the triple sums in (21a) to calculate ε, combined with the number of iterations of the outer and inner loops, n_{outer} and n_{inner}, respectively. In calculating the complexity of ε, it is clear that there are $(S+1)N_U B$ terms in the triple sums in (21a) that must be found. However, each cell s only has to deal with K_s out of the N_U users; the remaining terms will be zeros. In total, there will be $B\sum_{s=0}^{S} K_s = BN_U$ non-zero terms. Since each non-zero term operates on scalars, each takes $\mathcal{O}(1)$ complexity to calculate. Then, the BN_U non-zero terms must be added together. Hence, updating the value of ε has complexity $\mathcal{O}(BN_U)$. For the power waterfilling of (26)-(27), there are a total of N_U terms of $w_{s,n}^{*}$ and BN_U power terms $p_{s,n,b}^{*}$ that are relevant to be calculated; the system does not need to calculate power allocated by BS s to users not served by that BS. These terms again operate on scalar values. Hence, the power waterfilling update also has complexity order $\mathcal{O}(BN_U)$. These updates are performed over a total number $n_{outer} \times n_{inner}$ iterations of the outer and inner loops combined, making the total complexity of the power optimization $\mathcal{O}(n_{outer} n_{inner} BN_U)$. The number of loop iterations depends on the stopping criteria of the loops (for example, the value of λ_{th}). However, it is known that the Dinkelbach method converges superlinearly to find the optimal value of ε, and the inner loop problem, being convex, can be solved in polynomial time with the number of variables and constraints [48].

For comparison with our scheme, the optimal antenna/RB selection method would be an exhaustive search. Each possible permutation of RB assignments to users and number of active antennas per BS would be examined, with precoding and power allocation calculations done for each permutation the same as for our scheme. Unfortunately, the complexity of an exhaustive search is quite difficult to enumerate analytically. The number of permutations for the joint antenna/RB selection depends on the number of antennas at each BS and which BSs are in the cluster for each user. However, we have been able to determine upper and lower bounds for the number of permutations.

Upper Bound: Each user has only its own serving BS as its cluster. This upper bound is tight; it is the exact maximum possible permutations that an exhaustive search might ever potentially have to check. This scenario is equivalent to no coordination occurring between BSs, and results in the largest possible number of permutations. In this case, the antenna/RB selection at each BS is independent. Hence, the number of permutations for the network is the product of the permutations at each BS. If BS s serves K_s users, then by using generating functions [77], the total possible number of antenna/RB selections is:

$$\left\{ K_0! \left[z^{K_0}\right] \left(\sum_{i=0}^{N_M} \frac{z^i}{i!} \right)^B \right\} \times \prod_{s=1}^{S} \left\{ \sum_{N_s^{sel}=1}^{N_s} \binom{N_s}{N_s^{sel}} K_s! \left[z^{N_s}\right] \left(\sum_{i=0}^{N_s^{sel}} \frac{z^i}{i!} \right)^B \right\} \tag{28}$$

In the above, the notation $[z^n]f(z)$ represents the coefficient on the z^n term in the polynomial $f(z)$ [77]. We can also express $\sum_{i=0}^{N} z^i/i!$ as a formal power series $a(z) = \sum_{i=0}^{\infty} a_i z^i$, where $a_i = 1/i!$ for $0 \le i \le N$ and $a_i = 0$ for $i > N$. If we then define the power series $c(z) = \sum_{i=0}^{\infty} c_i z^i = (a(z))^B$, the coefficients c_i can be found in terms of a_i. Specifically, the desired coefficient c_m for $m \ge 1$ is given recursively by [78]:

$$c_0 = a_0^B, \tag{29}$$

$$c_m = \left[z^m\right] \left(\sum_{i=0}^{N} \frac{z^i}{i!} \right)^B = \frac{1}{a_0 m} \sum_{i=1}^{m} ((B+1)i - m) a_i c_{m-i}. \tag{30}$$

Lower Bound: Every BS is in every user's cluster. This lower bound is loose, as here we also ignore the cases where a BS can sometimes serve more users on a given RB, provided that no other BS serves users on that RB. (Hence, there would be no interference on that RB that requires coordination.) The number of permutations in this scenario is the same as for joint transmission, where each BS sends a useful signal to its coordinated users instead of just performing coordinated beamforming. For this scenario, it is equivalent to consider that RBs for all N_U users are assigned at one location, since assigning an RB to a user at one BS removes the same resource from being available at all other BSs. It therefore ends up being the BS with the fewest active antennas that determines the number of possible permutations for the whole network; the fewer antennas that are active,

TABLE 4. Simulation parameters.

Parameter	Symbol	Value
Number of small cells	S	10
Cell radius	R_s // R_0	40 m // 289 m
Position of SC BSs	-	On circle 249 m from macro BS, equally spaced in angle
Static power	P_s^{sta} // P_0^{sta}	0.1 W // 10 W
Dynamic power	P_s^{dyn}	0.1 W
Maximum transmit power	$P_s^{\max}$ // $P_0^{\max}$	30 dBm // 46 dBm
PA efficiency	η_s // η_0	0.5 // 0.8
Fronthaul capacity	$c_{s,limit}$ // $c_{0,limit}$	150 Mbits/s // 300 Mbits/s
Number of RBs	B	50
Maximum RBs allocated per user	$\hat{B}$	1
Total bandwidth	W	10 MHz
Bandwidth per RB	W_0	200 kHz
Carrier frequency	f_s	2 GHz
Minimum QoS	κ_{min}	128 kbits/s
Path loss exponent	α_s // α_0	3.67 // 3.76
Shadowing std. dev.	σ_{shadow}	10 dB
Noise power spectral density	N_0	-174 dBm/Hz
Number of antennas	N_S // N_M	8 // 100
Antenna spacing (multiple of wavelengths)	d_H	$\frac{1}{2}$
Macro BS spatial correlation ASD	σ_ϕ	$\frac{\pi}{18}$ radians (10°)
Macro BS distortion	D_N	10^{-6}
Total number of users	N_U	Variable, default 200
User distribution	-	Uniform in cell outside exclusion radius
Exclusion radius	$R_{x,s}$ // $R_{x,0}$	10 m // 35 m
Association bias	Υ_s // Υ_0	3 dB // 0 dB (defaults)
Clustering threshold	ζ_c	Determined by simulation (see Fig. 5)
Optimization loop iterations	n_{outer} // n_{inner}	8 // 20

the fewer permutations there are. Let $\check{N}^{sel} = \min_{s\in\mathcal{S}} N_s^{sel}$ denote the smallest number of active antennas at any BS. For the smallest number of permutations, this BS must also have the smallest total number of antennas of any BS, i.e., $\check{N} = \min_{s\in\mathcal{S}} N_s$. Then, the number of permutations is [77]:

$$\sum_{\check{N}^{sel}=1}^{\check{N}} \binom{\check{N}}{\check{N}^{sel}} N_U! \left[z^{N_U}\right] \left(\sum_{i=0}^{\check{N}^{sel}} \frac{z^i}{i!}\right)^B \tag{31}$$

We lastly note that our scheme is scalable to an arbitrary network size, since the complexity is dependent only on the local coordinated transmit nodes rather than the entire network.

V. SYSTEM SETUP AND NUMERICAL RESULTS

A. SYSTEM SETUP AND PARAMETERS

In this section, we evaluate the performance of the proposed algorithm through simulations. The default parameters are as follows unless otherwise stated. Most of the system parameters are based on the recommendations in [79]. The radius of the macro cell R_0 is assumed to be 289 m (i.e., $500/\sqrt{3}$, which corresponds to an inter-site distance of 500 m in a hexagonal macro BS layout). $S = 10$ SC BSs are placed deterministically and evenly spaced in angle on a circle of radius 249 m centered at the macro BS. The radius R_s of each SC is assumed to be 40 m. The massive MIMO macro BS has $N_M = 100$ antennas, whereas each SC BS is equipped with $N_S = 8$ antennas. We assume half-wavelength spacing for the antenna elements, i.e., $d_H = \frac{1}{2}$ in (4). The total number of users N_U is 200, from which $\lfloor \frac{2N_U}{3S} \rfloor$ are uniformly distributed over the area of each SC, while the remainder are uniformly distributed over the entire area of the macro cell. Additionally, there is a circular exclusion zone around each BS, inside of which no users may be placed. The radius of this zone is $R_{x,0} = 35$ m for the macro BS and $R_{x,s} = 10$ m for the SCs. The path loss from the macro BS to a user is given by $128.1 + 37.6\log_{10}(d)$ dB and the path loss from the SC BSs to a user is determined by $140.7 + 36.7\log_{10}(d)$ dB (with distance d in km). In particular, the propagation parameters are from [79, Tab. A.2.1.1.2-3], the cell radii are from [79, Tab. A.2.1.1-1], the user distribution setup is from [79, Tab. A.2.1.1.2-4], and the power values and bandwidth are from [24], [25], [80]. For small-scale spatial correlation for the macro BS, we assume the ASD $\sigma_\phi = \frac{\pi}{18}$ radians (10°) in (4) [4], [81]. It should be noted that these system parameters are merely chosen to demonstrate the EE performance in an example and can easily be modified to any other values depending on the specific scenario under consideration.

The association bias Υ_s for each SC is set to 3 dB, while the bias for the macro BS is 0 dB. These values make a user who is located on an SC edge the nearest to the macro BS approximately equally likely to associate with the macro BS or with that SC. Each user is assigned a maximum of $\hat{B} = 1$ RB. The Monte Carlo method with 500 drops of users and 1000 channel realizations for each user drop is used to obtain numerical results. Other parameters used in the simulations are listed in Table 4; these values are similar to those used in related work.

B. ANALYTICAL RESULTS

In this subsection, we analytically approximate the gains in EE that are possible by some of the methods used in our system. Due to all the various methods combined in the system, it is unfortunately infeasible to analytically characterize the system as a whole. We therefore focus on two of the tractable portions of the system. First, we characterize the effect of utilizing the load-modulated single-RF chain architecture at the macro BS on the macro tier EE. Then, we approximate the effect of antenna selection at the SCs on the EE.

1) EFFECT OF TRANSCEIVER STRUCTURE ON EE

To begin, we consider two similar macro BS transceivers, one with a traditional voltage-modulated massive MIMO structure, and one with the load-modulated single-RF chain structure. Since the transceivers otherwise are similar, from (5)-(7), if the channel gains, resource allocation, and radiated transmit power $P^{tx} = \sum_{n\in\mathcal{K}}\sum_{b=1}^{B} p_{n,b}$ are the same, the resulting sum rate will be the same in both cases. Thus, only the non-radiated dynamic per-antenna power consumption P^{dyn} and the PA efficiency η of the macro BS change between the two structures. To show the effect of the load-modulated structure on the macro tier EE, we calculate $\frac{\varepsilon_{lm}}{\varepsilon_{vm}}$ as

$$\frac{\varepsilon_{lm}}{\varepsilon_{vm}} = \frac{\frac{C_{lm}}{P_{lm}}}{\frac{C_{vm}}{P_{vm}}} = \frac{\frac{1}{\eta_{vm}}P^{tx} + N_M P^{dyn}_{vm} + P^{sta}}{\frac{1}{\eta_{lm}}P^{tx} + P^{dyn}_{lm} + P^{sta}} \tag{32}$$

We note that, based on Table 2, the value of P^{dyn}_{lm} is proportional to the output power of the power amplifier P_a (or equally, the input power to the load modulator array), and thus can be seen as a function of the transmit power. (That is, $P_a = P_R + P^{tx}$, where $P_R = P^{dyn}_{lm}$ is the power dissipated in the circulator resistor.) Hence, we can express P^{dyn}_{lm} as $c \times P^{tx}$, where c is some constant. Since $\frac{P_R}{P_a}$ depends on both N_M and D_N, the value of c differs for each number of antennas and distortion amount. For example, for $N_M = 100$ and $D_N = 10^{-6}$, $c = 0.2902$.

By using Table 2, assuming various numbers of transmit antennas, and utilizing the parameters given in previous subsection, we derive Fig. 4. This figure depicts the benefit of deploying the load-modulated single-RF chain massive MIMO macro BS on the EE of the system. As can be seen, by increasing the number of antennas, the gain $\frac{\varepsilon_{lm}}{\varepsilon_{vm}}$ increases significantly, which demonstrates the benefit of the load-modulated structure for systems with many antennas. Moreover, for P^{tx} larger than about 20 W, the ratio for the EE gain converges to $\frac{\eta_{lm}}{\eta_{vm}(1+c\eta_{lm})}$ and the curves become almost flat, such that the asymptotic value depends on the value of c.

2) EFFECT OF SC ANTENNA SELECTION ON EE

To examine the effect of antenna selection at the SCs on the EE, we focus solely on the SC tier. In this case, we assume that the macro BS is either part of the cluster for the SC users so that it provides no interference to those users, or its interference power is below the clustering threshold, and so its interference may be neglected when considered in aggregate with the other interfering SC BSs.

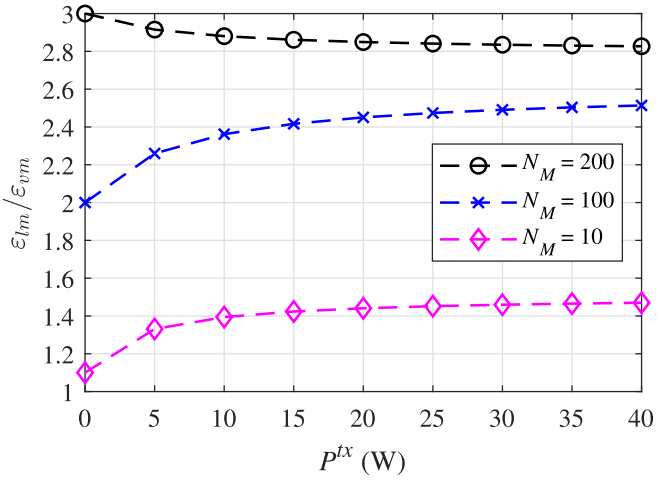

FIGURE 4. Gain in EE from using load-modulated structure at macro BS (relative to EE with voltage-modulated structure) vs. transmitted power P^{tx}, with varying antenna array sizes N_M; distortion $D_N = 10^{-6}$.

For the analysis, we will draw on the some of the results using stochastic geometry modeling of cellular networks (e.g., [55], [82]–[84]). For simplicity and tractability, we assume that the SC BS layout can be modeled by a homogeneous Poisson point process (PPP) Φ_s having intensity λ_s, wherein the BS locations are distributed uniformly over the plane. The users are located according to an independent homogeneous PPP Φ_u with intensity λ_u. We further assume a) that every served user at every BS is assigned the same power p, with $K_s p \leq P_s^{\max}$, b) that supported users are distributed approximately evenly across the RBs, so $K_{s,b}$, $I_{s,b}$, and $L_{s,b}$ are the same for every cell and RB, and c) that the system is interference-limited, so that the noise power is negligible. This provides a lower bound on the performance; naturally, the performance would be better by optimizing the power allocation and by better assigning users to RBs. (For example, in the latter case, users that are part of $\mathcal{I}_s$ for BS s could be largely assigned to their own separate RBs, to avoid removing degrees of freedom for beamforming for the users in $\mathcal{K}_s$.) The inclusion of log-normal shadowing with zero mean and standard deviation σ_{shadow} has the effect of scaling the intensity λ_s by a factor of $\exp(2(\frac{\sigma_{shadow}\ln(10)}{10\alpha_s})^2)$ [83], where α_s is the path loss exponent. However, it is interesting to note that the SINR and SE (and thereby the EE) of the system under the stated assumptions is known to be independent of the value of λ_s [82], [83].

Under the above conditions, the SINR for one user n served on a given RB b by BS s from (7), when null-space

projection precoding is used at the BSs, can be expressed as:

$$\sigma_{s,n,b} = \frac{\Gamma_{s,n} p \left|\mathbf{h}_{s,n,b}\mathbf{f}_{s,n,b}\right|^2}{\sum_{r\in\mathcal{S}_n\setminus\{s\}}\sum_{j\in\mathcal{K}_r}\Gamma_{r,n} p \left|\mathbf{h}_{r,n,b}\mathbf{f}_{r,j,b}\right|^2} = \frac{d_{s,n}^{-\alpha_s} g_{s,n}}{\sum_{r\in\Phi_s\setminus\{s\}} d_{r,n}^{-\alpha_s} g_{r,n}}, \quad (33)$$

where $d_{s,n}$ ($d_{r,n}$) denotes the distance between BS s (r) and user n, α_s is the path loss exponent, and $g_{s,n} = |\mathbf{h}_{s,n,b}\mathbf{f}_{s,n,b}|^2$ and $g_{r,n} = \sum_{j\in\mathcal{K}_r} |\mathbf{h}_{r,n,b}\mathbf{f}_{r,j,b}|^2$ represent the precoded channel power gain for user n from the serving BS s and an interfering BS r, respectively. As seen, the specific value of p cancels out.

It is known that if a BS has N_s^{sel} active antennas, the elements of $\mathbf{h}_{s,n,b}$ are independent and distributed $\sim \mathcal{CN}(0, 1)$ (i.e., Rayleigh fading), and the precoding vector $\mathbf{f}_{s,n,b}$ for user n is orthogonal to the channels of $L_{s,b} - 1$ users supported on RB b at BS s (such as with ZF or null-space projection precoding), then the precoded channel power gain is a Gamma-distributed random variable such that $g_{s,n} \sim \Gamma(N_s^{sel} - L_{s,b} + 1, 1)$ [84], [85]. Meanwhile, the precoding at other BSs that are not part of the cluster for user n is independent of the interfering channel to user n. As such, a single beam from the interfering BS r has a power gain distributed $\sim \Gamma(1, 1)$, and the sum of K_r beams from BS r, i.e., $g_{r,n}$, has a power gain distributed $\sim \Gamma(K_r, 1)$ [85].

Given the PPP model and the association scheme described by (1), it is straightforward to see that if a user associates with the SC tier, it will associate with the closest SC BS. Furthermore, from the clustering scheme described by (2), if a BS is part of a cluster for user n, its received reference signal strength will be within ζ_c dB of the received reference signal strength from the serving BS s, or numerically, $P_r^{ref} \geq 10^{-\zeta_c/10} P_s^{ref}$. As previously stated, the reference signal strength is assumed to be proportional to the maximum signal power available at the BS, which is the same for all SC BSs. Moreover, the path loss for BS r is given by $d_{r,n}^{-\alpha_s}$. Hence, if BS r is part of the cluster for user n, then $P_s^{\max} d_{r,n}^{-\alpha_s} \geq 10^{-\zeta_c/10} P_s^{\max} d_{s,n}^{-\alpha_s}$. Rearranging, we find that the cluster for a given user will consist of all BSs where

$$d_{r,n} \leq \Delta d_{s,n}, \text{ where } \Delta = 10^{\zeta_c/(10\alpha_s)}. \quad (34)$$

Lemma 1: In the PPP model, the mean number of users supported by a BS (either served or part of their cluster) is

$$\mathbb{E}[L_s] = \Delta^2 \lambda_u / \lambda_s. \quad (35)$$

Proof: See the Appendix. ∎

The achievable rate of a user is given by $W_0 \log_2(1 + \sigma_{s,n,b})$. Since the signals precoded at a given BS do not interfere between the users served by that BS (due to null-space projection precoding), nor do the signals interfere between RBs, the sum rate of SC s is simply $\sum_{n,b} W_0 \log_2(1 + \sigma_{s,n,b})$. The mean SE of a user, i.e., $\mathbb{E}_{\sigma_{s,n,b}}[\log(1 + \sigma_{s,n,b})]$, can theoretically be found following the methodology found in [83] for fading that follows a κ-μ distribution. (A Gamma distribution $\Gamma(k, \theta)$ with shape parameter k and scale parameter θ is a special case of the κ-μ distribution in [83], where $\kappa = 0$, $m = \mu = k$, and $\theta_1 = \theta_2 = \theta$.) Unfortunately, we have found that for our proposed system, the final integral for $\mathbb{E}_z[\log(1+z)]$ is non-convergent. Therefore, we will rely instead on a simulation of the PPP model.

For the PPP simulation, we simulate 200000 realizations of a PPP, within a square region containing on average 10000 points. In each realization, the distance of the points to the origin (center of the square) is measured. The closest point to the origin is set as the serving BS. Each point is assigned a Gamma-distributed value representing the precoded channel power gain for that BS. As described earlier, the shape parameter for the serving BS is $N_s^{sel} - L_{s,b} + 1$, whereas it is $K_{s,b}$ for interfering BSs; the scale parameter is 1 for all BSs. Using the point distances and power gain values, the SINR for a user located at the origin is calculated according to (33). The mean EE for a small cell is then found by $\sum_{n,b} W_0 \log_2(1 + \sigma_{s,n,b})/P_s$, where P_s is given by (8).

We consider a few different scenarios in terms of served and clustered user loads. To begin, based on our full cellular model, if there were the default 200 users distributed within the area of the macrocell, this would correspond on average to 13 users served per SC. For the PPP model, if ζ_c is set to 25 dB, from (35), with $\lambda_u/\lambda_s = 13$ users served per SC, there would on average be $L_s = 300$ supported users per SC. (In the PPP model, some of these would correspond to users located in other macrocells.) So, on average there would be $L_{s,b} = 300/50 = 6$ users supported per RB. 13 out of the 300 are served users, but there are 50 RBs they could be assigned to. We assume for this scenario that there is at most one served user per RB at all SCs. Hence, there is only a 13/50 probability that an interfering SC will be transmitting power on the same RB as the typical user at the origin. Consequently, in this scenario we scale the measured interference by a factor of 13/50.

In the remaining scenarios, we take a more generic approach. For the second scenario, we assume that there are 4 supported users on each RB, out of which 1 is served ($K_{s,b} = 1, L_{s,b} = 4$). In the third scenario, 2 users are assumed to be served per RB, with one additional user supported ($K_{s,b} = 2, L_{s,b} = 3$). Finally, we consider a no-clustering case where $K_{s,b} = L_{s,b} = 4$. For all scenarios, although the value of λ_s has no impact on the EE, for the purposes of the simulations, we set $\lambda_s = 1.62\times10^{-4}$ m^{-2}. This corresponds to an intensity yielding on average 10 SC BSs located within a hexagonal macrocell with an inter-site distance of 500 m (the same SC BS density as for our simulated network layout), scaled by a factor of 3.5 to account for log-normal shadowing with a standard deviation of $\sigma_{shadow} = 10$ dB and a path loss exponent of $\alpha_s = 3.67$. Other relevant parameters are the same as in Table 4.

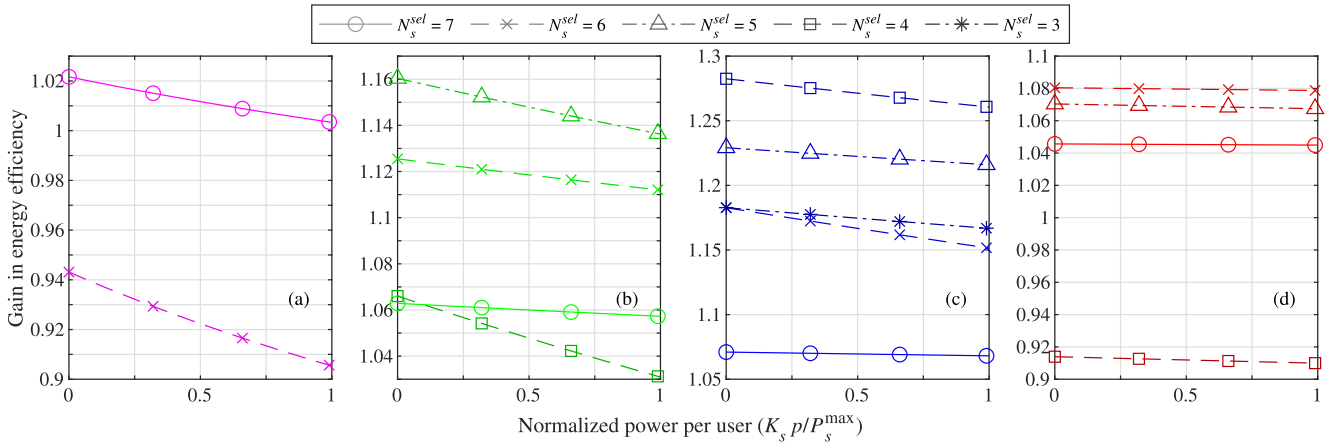

FIGURE 5. Gain in energy efficiency by antenna selection, relative to when all antennas are active, vs. normalized power allocated per user ($K_S p/P_S^{\max}$), with $N_S = 8$. (a) $K_S = 13$, $L_S = 300$. (b) $K_{S,b} = 1$, $L_{S,b} = 4$. (c) $K_{S,b} = 2$, $L_{S,b} = 3$. (d) $K_{S,b} = L_{S,b} = 4$.

In Fig. 5, we examine the gain in EE that is possible due to antenna selection. The gain is the EE achieved when the specified number N_s^{sel} of SC antennas are activated per cell, relative to when all $N_s = 8$ antennas are active. As seen, significant gains are possible; Fig. 5(c) in particular displays gains of over 25%. As expected, the gain is the highest when the amount of power allocated to each user is the lowest. In this situation, changing the dynamic potion of the consumed power has the relatively largest effect on the overall reduction in the value of P_s, and thus the gain in EE. We also note that there is in general an optimal number of antennas to activate to maximize the EE gain. If too few antennas are active, the loss of available spatial multiplexing gain can cause the drop in SE to become larger relative to the drop in consumed power. In some cases, the system can even see a loss in EE compared to when all antennas are active, as observed in Figs. 5(a) and 5(d). However, one cannot draw a specific conclusion on the exact optimal number of antennas to activate from these results. To start, the allocation of users to RBs also has a significant effect on how many antennas it is possible to turn off. Thus, the specific allocation of RBs to users would also impact the optimal number of active antennas. The relationship between the two factors is far from straightforward. Furthermore, these results are for the equivalent of activating N_s^{sel} antennas at random, whereas our scheme selects the best N_s^{sel} antennas out of N_s to activate.

C. SIMULATION RESULTS

1) EFFECT OF CLUSTERING THRESHOLD ON EE & SE

One of the key factors in user-centric cluster formation is the clustering threshold ζ_c. The value of ζ_c affects which BSs are clustered and sets the level where the interference from non-clustered BSs is considered negligible. We show the impact of ζ_c on the EE and SE (normalized to the total system bandwidth) of our proposed scheme in Fig. 6. We also modify the scheme proposed in [52] and compare its resulting EE and SE with those of our user association method. To this end, we define $P_{s,n}^{heur}$ as

$$P_{s,n}^{heur} = P_{s,n}^{load} + N_s P_s^{dyn} + P_s^{sta}, \tag{36}$$

where

$$P_{s,n}^{load} = \frac{P_s^{\max}}{B}\left[\frac{\kappa_{min}}{W_0 \log_2(1+\Psi_{s,n})}\right] \tag{37}$$

and

$$\Psi_{s,n} = \frac{\Gamma_{s,n} P_s^{max}}{B\left(\tilde{I}_{s,n}^{est} + W_0 N_0\right)} \tag{38}$$

$\Psi_{s,n}$ is the estimated average SINR (per RB) for user n if it associates with BS s, under the assumptions that all BSs transmit at maximum power, the power is divided equally among subcarriers, and user n is the only user served on its RB. In modifying the scheme in [52] for use with clusters, we define the cluster $\tilde{\mathcal{S}}_n$ for user n prior to association as the set of BSs that provide an average received signal strength within ζ_c dB of the strongest. $\tilde{I}_{s,n}^{est} = \sum_{\tilde{s}\in\mathcal{S}\setminus\tilde{\mathcal{S}}_n} \Gamma_{\tilde{s},n} P_{\tilde{s}}^{\max}/B$ is the estimated average total interference (per RB) experienced by UE n if it associates with BS s. $P_{s,n}^{load}$ is the estimated load-dependent power consumption of BS s per RB; if ideal RF circuitry with 100% power efficiency was assumed, this would be equivalent to the estimated transmitted power per RB if user n associates with BS s. User n is associated with and served by the BS s with the smallest value of $P_{s,n}^{heur}$.

Fig. 6 shows both the estimated EE and SE and the actual EE and SE for our proposed method. For the sake of brevity and legibility, only the actual EE and SE are depicted for the modified method from [52]. The estimated values are calculated by the optimization algorithm assuming no interference, using the SINRs in (16) and (20). The actual values use the power and precoding results from the optimization and the true SINR in (7), accounting for the residual interference.

From the figure it is clear that by increasing ζ_c, the difference between the estimated and the actual curves decreases, since more BSs are included in each cluster and more interference is mitigated. For ζ_c greater than 25 dB, the

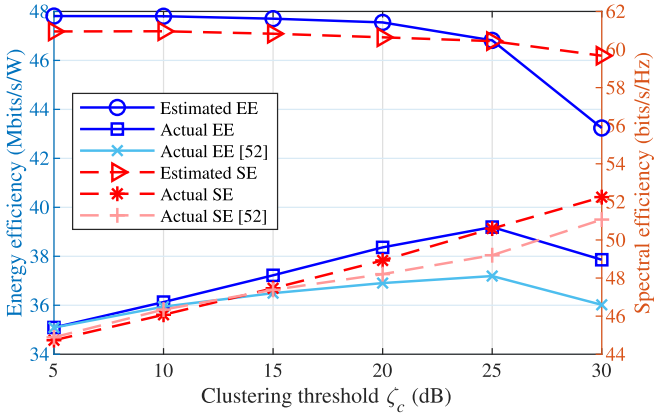

FIGURE 6. EE and SE (estimated and actual) of our proposed method vs. clustering threshold ζ_C, with $N_U = 200$ and $\Upsilon_S = 3$ dB for SC BSs equipped with $N_S = 8$ antennas (of which only a subset are active) and a macro BS equipped with $N_M = 100$ antennas. Actual EE and SE for modified method proposed in [52] are depicted for comparison.

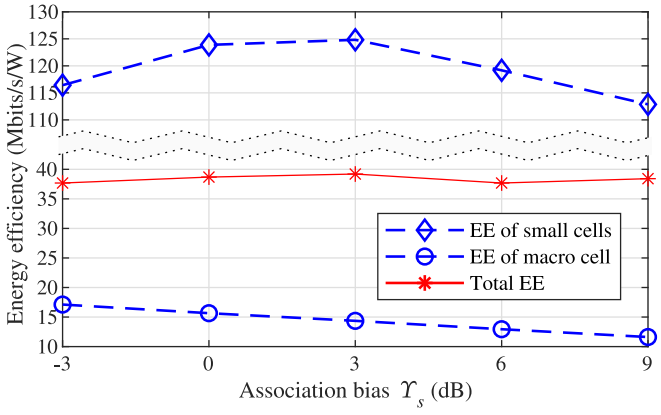

FIGURE 7. Total EE of HetNet and EE of macro and small cell HetNet tiers vs. association bias Υ_S for small cells for $\zeta_C = 25$ dB and with $N_U = 200$. SC BSs are equipped with $N_S = 8$ antennas (of which only a subset are active) and the macro BS is equipped with $N_M = 100$ antennas.

actual EE decreases, meaning including more BSs per cluster degrades the EE. This is due to more antennas needing to be activated at the SC BSs to support BS coordination for the larger clusters. Thus, even though the estimated EE and SE become closer to the actual values for higher ζ_c, increasing ζ_c further has a detrimental effect on the actual performance. Hence, we use ζ_c = 25 dB for the remainder of this article.

As can be seen, our proposed method outperforms the method modified from [52] for both EE and SE. As our proposed method performs better while having less computational complexity, it is a better solution for user association. Finally, please note that the EE and SE values reported hereafter are the actual values, not estimated ones.

2) PERFORMANCE COMPARISON OF OUR PROPOSED METHOD WITH EXHAUSTIVE SEARCH

Here, we compare the performance of our antenna selection and RB allocation algorithm with that of an exhaustive search. Since the exhaustive search complexity rapidly grows with the variables involved, we have considered a simpler network; besides the macro BS, only 1 SC BS equipped with 4 antennas is used, and 6 users (3 per BS) are served on 5 RBs. In this scenario, the EE of the exhaustive search is 15.2 Mbits/s/W, whereas our algorithm yields 14.1 Mbits/s/W, about 7% lower. However, the exhaustive search checks 125020 to 198125 permutations[5] of antenna/RB selections, the exact number depending on the size of each user's cluster. Precoding and power allocation calculations are done for each permutation, whereas our algorithm only does this once, with some additional calculation and sorting of channel F-norms. Therefore, our algorithm's complexity in this scenario is a tiny fraction of the exhaustive search complexity; that fraction would be even smaller in a larger network.

3) EFFECT OF ASSOCIATION BIAS ON EE OF MACRO AND SC TIERS AND ON TOTAL EE OF HETNET

Fig. 7 shows the comparison of the EE performance for the total HetNet, and for the macro and SC tiers separately, when the association bias Υ_s for the SCs changes. Υ_0, the association bias for the macro BS, stays constant at 0 dB, so increasing Υ_s means more emphasis is given to SCs and more users are served by SCs.

As expected, by increasing Υ_s up to 3 dB, the EE of the SCs increases. The cause stems from the higher priority given to SCs that results in more users being served by them, while activating a minimal necessary number of antennas. However, with further increases to Υ_s, even more users are served by SCs, such that more antennas need to be activated. This leads to EE degradation. On the other hand, since larger Υ_s means fewer users are served by the macro BS, the macro cell EE decreases.

Interestingly, changing Υ_s from -3 dB to $+9$ dB yields little change in the total EE; the EE changes by only about $\pm 2\%$ from its mean value of 38.3 Mbits/s/W over that range. This is due to the trade-off between the need to activate more antennas at the SCs, which leads to higher dynamic power consumption, versus serving users with lower total radiated transmit power, since in general the users will be closer to the SC BSs.

Although serving more users by the macro BS seems more energy-efficient for the macro cell, using an association bias in favor of the SCs leads to transferring the traffic load to the comparatively lightly-loaded SCs. Offloading users to the SCs leaves more resources available for macro users. Thus, if we were to consider a scenario with higher mobility users, who would prefer to be served by the macro BS, offloading the traffic to SCs would provide those users with more resources, so increasing Υ_s could also lead to an increase in

5. These numbers are exact empiric values calculated numerically, rather than the lower and upper bounds given by (28) and (31), respectively. (However, we also note (31), being a tight bound, gives the same upper value of 198125 permutations.)

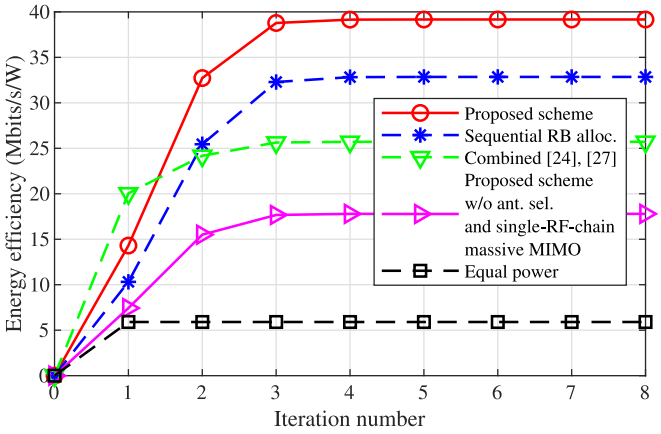

FIGURE 8. Comparison of convergence in (actual) EE between proposed scheme and other reference algorithms (shown in legend) vs. number of iterations. SC BSs are equipped with $N_S = 8$ antennas (of which only a subset are active) and the macro BS is equipped with $N_M = 100$ antennas; $\zeta_C = 25$ dB, $N_U = 200$ and $\Upsilon_S = 3$ dB.

the macro BS EE. However, the case of highly mobile users is outside the scope of this article.

4) COMPARISON OF CONVERGENCE RATE OF PROPOSED METHOD WITH REFERENCE SCHEMES

The convergence of the EE of our proposed scheme is illustrated in Fig. 8. Since there is no comparable scheme in the literature that combines all our considered factors and constraints, to evaluate the EE performance of our proposed scheme, an equal power allocation algorithm and a sequential RB assignment scheme are chosen as baseline algorithms for comparison. These algorithms are also used as benchmarks in [24]. For the equal power allocation algorithm, the maximum transmit power is used, which is equally allocated between users, and in the sequential RB assignment scheme, RBs are allocated to users sequentially. Furthermore, to compare with an algorithm similar to our own, we have also modified and combined the schemes proposed in [24] and [27]. This combined scheme follows a similar beamforming approach as in [27] (therein called EE ZF), while using the the power allocation and RB assignment algorithm from [24] modified to be applicable to multiple-antenna BSs. Just as in [24] and [27], no antenna selection is considered for the combined scheme, but for a fair comparison, the constraints have been modified to match ours and the single-RF-chain structure is assumed for massive MIMO. We also compare the EE performance of our scheme with no antenna selection for SCs and a conventional transceiver structure for the massive MIMO macro BS.

As can be seen, even though the computational complexity is higher in our proposed algorithm, it converges within 3 iterations of the outer loop, and the convergence speed is acceptable when compared with the reference algorithms. Moreover, as discussed in the computational complexity subsection (Section IV-B), the complexity of our proposed method is proportional to the number of outer loop iterations. Hence, the fast convergence rate of our method demonstrates acceptable complexity. We also note that by optimizing the initial points and step sizes used for updating the values of the Lagrangian multipliers in the subgradient method, the number of inner loop iterations can also be decreased, which can lead to a further decrease in the computational complexity. Furthermore, if in practice temporal correlation exists for the channel gains, the optimal power allocation will also be correlated in time. Thus, using the previous set of power values as an initialization point for the next calculation could reduce required number of iterations even more.

As expected, our proposed scheme outperforms the other algorithms: the EE is about 12% higher than the second-best scheme of sequential RB allocation. This result shows the impact of an effective and efficient RB allocation method on the total EE of the system. The worst performance is for equal power allocation, in which even though antenna selection, the single-RF-chain structure, and RB assignment are considered, power is equally allocated between users regardless of their channel quality and minimum rate requirements. Our proposed scheme provides about 6.4 times higher EE than the equal power case. Based on this figure, we can conclude that even though full power transmission increases the SE significantly, it performs the worst from the EE perspective. Moreover, our proposed scheme outperforms the combined schemes from [24] and [27], with about 37% higher EE.

Without antenna selection, even though increasing the number of active antennas enhances the SE, since more power is consumed by each SC BS, the EE decreases. Moreover, deploying a conventional transceiver structure at the macro BS (with one RF chain per antenna) will lead to higher consumption of dynamic power. Thus, the EE of our proposed scheme but with a conventional macro BS transceiver structure and with all SC antennas active is only about twice that of the equal power case.

5) EFFECT OF NUMBER OF USERS ON EE AND SE

To evaluate the impact of the number of users on EE and SE, we compare the performance of our proposed method with three of the baseline schemes for varying numbers of users N_U in Fig. 9.

In all cases, both the EE and the SE increase approximately linearly with the number of users. Our proposed scheme provides significantly higher EE than the combined schemes from [24] and [27]. There is gap in EE performance between our scheme above that provided by the reference combined schemes at lower N_U, which increases in magnitude as N_U increases. This is primarily a result of the decreased power consumption due to antenna selection, which is the main difference between the two schemes. At the same time, our proposed scheme provides only slightly lower SE than the combined schemes of [24] and [27]. However, we also note that the rate of growth in the gap appears to stabilize at larger N_U, with the size of the gap becoming

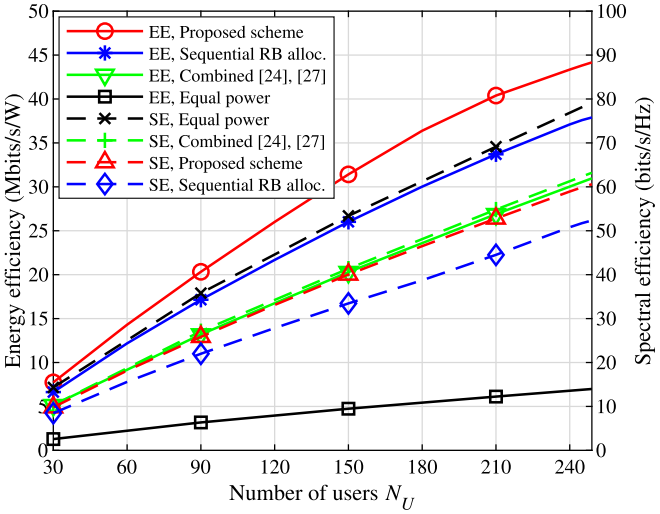

FIGURE 9. EE and SE of proposed scheme and other reference algorithms (shown in legend) vs. number of users N_U. SC BSs are equipped with $N_S = 8$ antennas (of which only a subset are active) and the macro BS is equipped with $N_M = 100$ antennas; $\zeta_C = 25$ dB and $\Upsilon_S = 3$ dB.

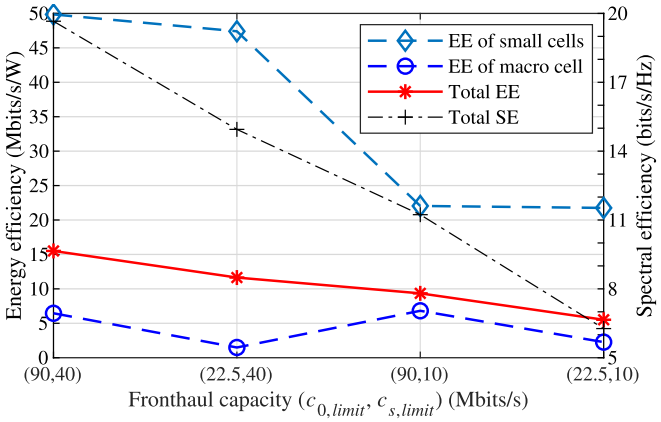

FIGURE 10. Impact of fronthaul capacity on total EE, on EE of small cell and macro tiers separately, and on SE. SC BSs are equipped with $N_S = 8$ antennas (of which only a subset are active) and the macro BS is equipped with $N_M = 100$ antennas; $N_U = 60$, $\kappa_{min} = 1.28$ Mbits/s, $\zeta_C = 25$ dB, and $\Upsilon_S = 3$ dB.

closer to constant. This effect is a result of the system gradually activating more antennas to enable resources for the increasing user load, and therefore losing some of its advantage in EE performance. Sequential RB allocation, having the same kind of antenna selection, also has a similar gradually decreasing slope for its EE.

As indicated earlier in Fig. 8, equal power allocation has the worst EE performance, and as seen here its rate of increase in EE with larger numbers of users is also very small. As expected, the rate of increase in EE for our proposed scheme with larger numbers of users is much larger than that for equal power allocation. Even though antenna selection has also been considered for equal power allocation, its large constant maximum transmit power dominates the total consumed power. This causes the effect of changes in the dynamic power to become negligibly small. Therefore, even though higher numbers of users need more active antennas at the SCs, almost no change in the slope of the EE and SE for equal power allocation is visible as N_U increases.

From the SE perspective, the SE of equal power allocation is higher than that of our scheme, indicating that the excess power is being used (somewhat inefficiently) to increase bit rates. For sequential RB allocation, the rates of change for both EE and SE are somewhat similar to those of our proposed scheme, though said changes, as well as the EE and SE themselves, end up being smaller. Moreover, while the combined schemes from [24] and [27] provide higher SE than our proposed method, the relatively small gain in SE of the combined schemes over our method is much lower than the gain in EE of our method over the combined schemes. Thus, our proposed method is only trading off a small amount of SE performance for a significant gain in EE performance, compared to the combined schemes from [24] and [27].

Since the reference combined schemes do not use antenna selection, they do not experience the same gradual decrease in the slope of their EE performance as our proposed scheme. We predict that the gap between our scheme and the reference combined schemes may be expected to decrease as N_U continues to grow. At some point at larger N_U, our scheme would be forced to activate all SC antennas. At this point, the performance (and its slope) of our proposed scheme and the combined schemes of [24] and [27] would likely become about the same.

6) EFFECT OF FRONTHAUL CAPACITY ON EE AND SE

We now evaluate the effect of the fronthaul capacity, $c_{s,limit}$, on the EE and SE in Figs. 10 and 11. In previous figures, we have assumed there was enough fronthaul capacity to serve the users, but in these figures we limit fronthaul capacity to investigate the effect of the constraints in (21c) in our EEmax problem. For this purpose, we change the number of users to $N_U = 60$ and κ_{min} to 1.28 Mbits/s.[6] In Fig. 10, EE is depicted vs. fronthaul capacity, where the pair of values on the x-axis indicate $(c_{0,limit}, c_{s,limit})$, i.e., the fronthaul capacity for the macro BS and for SC BSs, respectively.

First, at the far left we assign sufficient fronthaul capacity for all BSs, such that all users can get their minimum required rate. Next, the fronthaul capacity for the macro BS is decreased. This leads to the macro BS becoming overloaded and it hence drops some of its users. All EE and SE values decrease as a result. Third, we keep the sufficient fronthaul capacity for the macro BS and decrease the fronthaul capacity of the SC BSs. In this case, the total EE, SC EE, and SE all decrease even more, since the majority of users in the network are SC users, many of whom are now not receiving their minimum guaranteed rates. However, the EE of the macro BS remains about the same as it was initially. Finally, we decrease $c_{s,limit}$ for all BSs. As expected, the total EE and SE decreases even further with more users

6. These values are simply examples intended to ensure the constraints in (21b) are also active, so we can examine their interaction with the fronthaul capacity constraints.

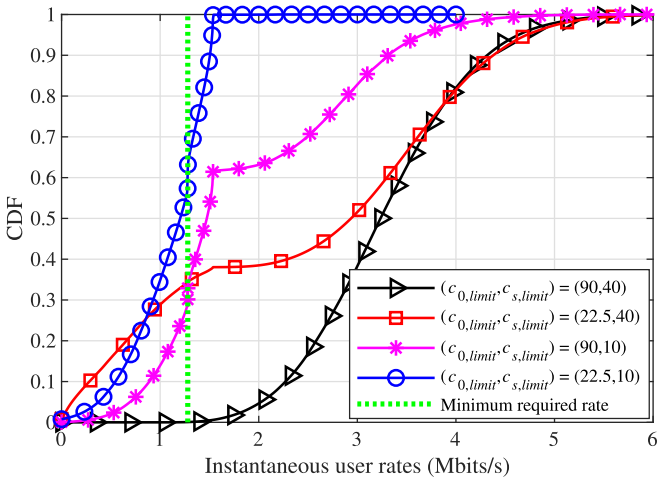

FIGURE 11. CDFs of instantaneous user rates for different pairs of $(c_{0,limit}, c_{s,limit})$ (Mbits/s). SC BSs are equipped with $N_S = 8$ antennas (of which only a subset are active) and the macro BS is equipped with $N_M = 100$ antennas; $N_U = 60$, $\kappa_{min} = 1.28$ Mbits/s, $\zeta_C = 25$ dB, and $\Upsilon_S = 3$ dB.

receiving even smaller rates. However, the EEs of the macro cell and SCs are about the same as for the (22.5, 40) case and the (90, 10) case, respectively. Hence, whether or not a tier is overloaded has little effect on the EE of the other tier.

Fig. 11 shows the CDFs of the instantaneous user rates for the same $(c_{0,limit}, c_{s,limit})$ pairs as used in the previous figure. The minimum required rate is indicated by the vertical dotted line. As seen, when $(c_{0,limit}, c_{s,limit}) = (90, 40)$ Mbits/s, all the users get their minimum required rates. In comparison, the worst performance is for the case when all fronthaul capacities are too low to support the users, for which over 57% of the users don't get their minimum required rates. In the remaining two cases, about 30% to 35% of the users don't receive their minimum required rates.

A "breakpoint" can be seen in several of the curves at about 1.5 Mbits/s, demonstrating a separation in performance between macro and SC users if the fronthaul capacity of a tier has been reached. Users of the overloaded tier contribute to the portion of the CDF left of the breakpoint, whereas users of the other tier contribute to the rightmost portion.

A vertical jump in the CDF can also be seen in some curves right at the minimum rate of 1.28 Mbits/s. This indicates that when a tier has not enough fronthaul capacity, the system often attempts to deliver only the minimum required rate to many users and no more than that minimum, due to insufficient resources being available.

7) EFFECT OF IMPERFECT CSI ON EE AND SE

Imperfect CSI can have a notable impact on the performance of cellular systems. Massive MIMO systems specifically may experience errors in CSI due to pilot contamination [5], [86]. Our examined network layout has only the single massive MIMO macro BS, so pilot contamination is not present. Nevertheless, we wish to conduct a brief numerical examination of imperfect CSI in general. The model we have adopted for imperfect CSI is given as [87], [88]:

$$\mathbf{h}_{s,n,b}^{est} = \varsigma \mathbf{h}_{s,n,b} + \sqrt{1-\varsigma^2}\bar{\mathbf{h}}_{s,n,b}, \tag{39}$$

where $\mathbf{h}_{s,n,b}^{est} \in \mathbb{C}^{1\times N_s}$ is the estimated (small-scale fading) channel vector, $\mathbf{h}_{s,n,b} \in \mathbb{C}^{1\times N_s}$ is the actual channel vector (corresponding to perfect CSI), and $\bar{\mathbf{h}}_{s,n,b} \in \mathbb{C}^{1\times N_s} \sim \mathcal{CN}(0, \mathbf{I}_{N_s})$ is an independent error vector. The parameter ς, where $0 \le \varsigma \le 1$, represents the reliability of the channel estimate. When $\varsigma = 0$, there is a complete failure in the channel estimation, whereas for $\varsigma = 1$ the estimation is perfect and the error component in (39) becomes zero.

In this model, since the small-scale fading vector is scaled by the large-scale fading value $\sqrt{\Gamma_{s,n}}$ in (3) (which is still assumed to be known perfectly), the error component will be scaled by the same amount. Hence, the power of the error will be around the same order of magnitude as the signal, but scaled down by $1-\varsigma^2$. In practice, in an interference-limited system, errors in CSI would likely be due to interference and/or pilot contamination, and variations in the error power can be reasonably expected to be handled within the range of values for ς in the model. (In comparison, the noise power in our simulations is many orders of magnitude smaller.) Meanwhile, the Gaussian distribution for the error component of the model holds due to the central limit theorem applying to the sum of many interfering signals in a practical system.

Precoding techniques that null interference, like those we use in this article, are known to be somewhat sensitive to errors in CSI. As one example, [89] examines the case of ZF precoding specifically within the context of massive MIMO. To overcome this issue, a large body of work has been done to design precoders that account for CSI error. References [86] and [90] survey techniques designed to combat pilot contamination; see also [4, Ch. 3]. More generally, channel estimators and precoders can attempt to account for the channel estimation error (e.g., [91]–[93]) and/or attempt to bound its magnitude (e.g., [94], [95]). However, a more in-depth examination of channel estimation and the design of precoders robust to imperfect CSI are outside the scope of the current work. Hence, herein we only examine the robustness of our proposed scheme to channel estimation error.

The EE/SE performance vs. $(1-\varsigma)$ is depicted in Fig. 12. We assume that the reliability of the estimate is the same for all BSs. As can be seen, by decreasing ς both the EE and SE decrease, which is expected as the decreased CSI reliability/increased channel estimation error naturally leads to the degradation of performance. The most significant decrease occurs for equal power allocation. Since the power allocation is not optimized, but rather divided equally between users, there is no opportunity in that regard to mitigate interference by, for example, reducing the transmitted power by only allocating enough power to certain users such that they receive just their minimum guaranteed rates. Thus, the primary source of interference mitigation is due to the

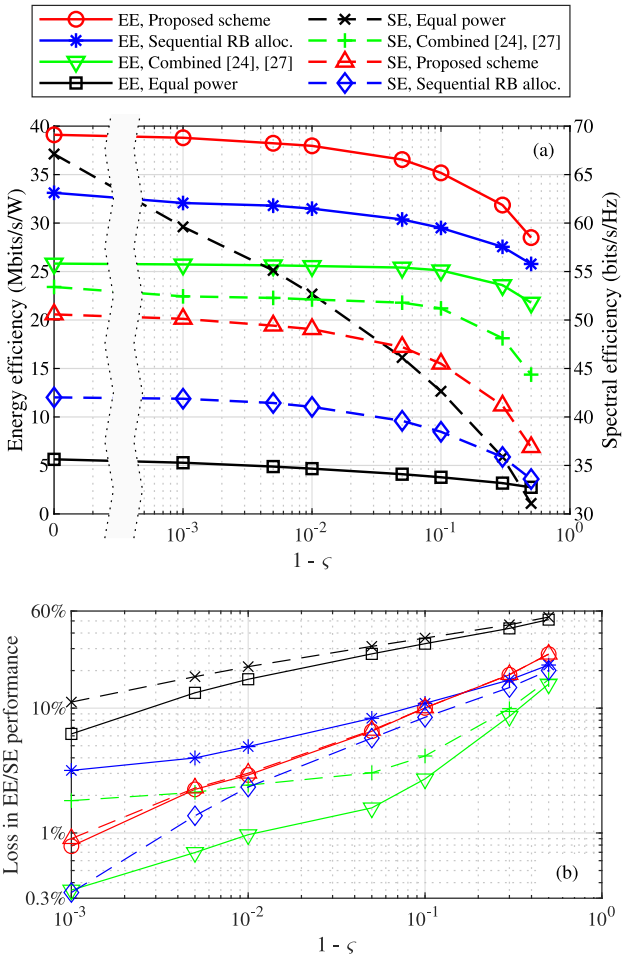

FIGURE 12. EE and SE of proposed scheme and other reference algorithms shown in legend vs. $1-\varsigma$, where ς represents the CSI reliability. SC BSs are equipped with $N_S = 8$ antennas (of which only a subset are active) and the macro BS is equipped with $N_M = 100$ antennas; $N_U = 200$, $\zeta_C = 25$ dB, and $\Upsilon_S = 3$ dB. (a) Absolute values of EE and SE. (b) Percentage loss in EE/SE performance, relative to having perfect CSI ($\varsigma = 1$).

precoding, and as already mentioned, the precoding can be sensitive to CSI errors. Hence, for equal power allocation, both the EE and SE drop by over half when the CSI reliability drops from $\varsigma = 1$ (perfect CSI) to $\varsigma = 0.5$.

The decrease in SE of our proposed method is sharper than for the remaining two reference schemes, indicating slightly less robustness in SE toward channel estimation error in our proposed method. However, the relative decrease in EE of our proposed scheme is less than that of sequential RB allocation for larger values of ς. Interestingly, the percentage drops in EE and SE for our proposed scheme as ς decreases are nearly identical, indicating a good trade-off between EE and SE. For sequential RB allocation, the relative drop in EE is faster than for SE, whereas for the combined schemes of [24] and [27], the reverse is true.

Overall, all the examined schemes (other than equal power allocation) see their degradation in performance start to happen faster once ς decreases past about 0.95 to 0.9. If the reliability of channel estimation can be kept above $\varsigma = 0.9$, then the degradation in performance will also be fairly limited. In the case of our proposed algorithm, when $\varsigma \geq 0.9$, both the EE and the SE lose no more than 10% in performance compared to having perfect CSI.

VI. CONCLUSION

In this article, the energy-efficient deployment of MIMO in SC HetNets has been considered. To achieve high EE, MIMO and SC deployments need to be integrated with well-designed interference mitigation and resource allocation methods. To this end, we have proposed and investigated the use of user-centric clustering and coordinated beamforming with null-space projection and ZF precoding to mitigate interference in a HetNet. A single-RF-chain massive MIMO transceiver design for the macro cell and antenna selection for the SCs have been proposed to reduce hardware power consumption. We have furthermore designed a joint antenna selection and RB allocation algorithm, followed by a power optimization algorithm, to maximize the system EE under the additional constraints of minimum guaranteed user rates and maximum fronthaul capacity limits. The power allocation problem has been solved using the Dinkelbach method. Simulation results have demonstrated that our proposed methodology and algorithms ensure higher EE than previously known benchmark algorithms, while being significantly less complex than an exhaustive search. The effect on the system EE and SE performance when varying the clustering threshold, number of users, cell association bias, fronthaul capacity, and reliability of CSI has been examined.

Future work on this topic may involve the addition of user scheduling for even larger numbers of users requesting service, as well as the impact of user mobility on the performance of the proposed scheme. Also, the current work has mostly assumed perfect channel estimation. However, in practical massive MIMO systems, imperfect CSI due to pilot contamination [5], [86] is quite common. Therefore, future work should account for its effects on the system and the proposed scheme.

APPENDIX
PROOF FOR LEMMA 1

One may consider the PPPs Φ_s and Φ_u to be uniformly distributed over the area A of a circle centered at the origin o of the plane and having radius R_A, where R_A tends to infinity. We consider the typical user $u \in \Phi_u$, who is served by BS $s \in \Phi_s$. Without loss of generality, we may consider this user to be located at the origin. The distribution of the distance $d_{s,u}$ from the user to its serving BS, i.e., the nearest point in Φ_s, is [82]

$$f_{d_{s,u}}(r) = 2\pi\lambda_s r \exp\left(-\pi\lambda_s r^2\right) \tag{40}$$

Meanwhile, the location of any other arbitrary BS $r \in \Phi_s \backslash \{s\}$ is independent of the location of s and uniformly distributed over A. Therefore, the distribution of the distance

$d_{r,u}$ between r and the origin is

$$f_{d_{r,u}}(r) = \frac{2\pi r}{\pi R_A^2} = \frac{2r}{R_A^2} \tag{41}$$

Since u is served by s, $d_{r,u}$ must be larger than $d_{s,u}$. However, to be part of the cluster for u, from (34) we have that $d_{r,u}$ can be no larger than $\Delta d_{s,u}$. Therefore, the probability that an arbitrary BS r will be part of the cluster for u (but not serve u), conditioned on $d_{s,u}$, is $\mathbb{P}_u^{supp}|d_{s,u} = \mathbb{E}_{d_{r,u}}[\mathbb{1}(d_{s,u} < d_{r,u} \leq \Delta d_{s,u})|d_{s,u}]$, where $\mathbb{1}(x)$ is an indicator function that equals 1 if x is true, and 0 otherwise.

$$\begin{aligned}\mathbb{P}_u^{supp}|d_{s,u} &= \int_0^{\infty} \mathbb{1}\left(d_{s,u} < d_{r,u} \leq \Delta d_{s,u}\right) \cdot f_{d_{r,u}}(r)\mathrm{d}r \\ &= \int_{d_{s,u}}^{\Delta d_{s,u}} \frac{2r}{R_A^2}\mathrm{d}r = \frac{d_{s,u}^2\left(\Delta^2 - 1\right)}{R_A^2}\end{aligned} \tag{42}$$

Then, deconditioning on $d_{s,u}$ to obtain $p_u^{supp} = \mathbb{E}_{d_{s,u}}[\mathbb{P}_u^{supp}|d_{s,u}]$:

$$\begin{aligned}p_u^{supp} &= \int_0^{\infty} \mathbb{P}_u^{supp}|d_{s,u}\, f_{d_{s,u}}(r)\mathrm{d}r \\ &= \int_0^{\infty} \frac{r^2\left(\Delta^2 - 1\right)}{R_A^2} 2\pi\lambda_s r \exp\left(-\pi\lambda_s r^2\right)\mathrm{d}r \\ &= \frac{\left(\Delta^2 - 1\right)}{\pi\lambda_s R_A^2}\end{aligned} \tag{43}$$

p_u^{supp} is the probability of supporting, but not serving, some typical user u. (This would be a user in the set $\mathcal{I}_s$.) However, there are a total of N_U users uniformly distributed over A, where N_U is a Poisson-distributed random variable with mean $\lambda_u \pi R_A^2$. Each user is independently placed, meaning the statistics for any given user are identical. (When considering some other user, without loss of generality one can relocate the origin of the infinite plane to that user's location and thus obtain the same probability.) Consequently, the probability of supporting (but not serving) n out of N_U users is a binomial-distributed random variable:

$$\mathbb{P}[I_s = n|N_U] = \binom{N_U}{n}\left(p_u^{supp}\right)^n\left(1 - p_u^{supp}\right)^{N_U - n} \tag{44}$$

The mean value of I_s is

$$\begin{aligned}\mathbb{E}[I_s] &= \mathbb{E}_{N_U,n}\{\mathbb{P}[I_s = n|N_U]\} = \mathbb{E}_{N_U}[\mathbb{E}_n\{\mathbb{P}[I_s = n|N_U]\}] \\ &= \mathbb{E}_{N_U}\left[N_U p_u^{supp}\right] = \lambda_u \pi R_A^2 p_u^{supp} \\ &= \lambda_u \pi R_A^2 \frac{\left(\Delta^2 - 1\right)}{\pi\lambda_s R_A^2} = \left(\Delta^2 - 1\right)\frac{\lambda_u}{\lambda_s}\end{aligned} \tag{45}$$

The mean value of the number of served users per BS, $\mathbb{E}[K_s]$, is known to be λ_u/λ_s [55]. Since $\mathcal{K}_s$ and $\mathcal{I}_s$ are disjoint, the mean number of supported users (served and clustered) $\mathbb{E}[L_s]$ in $\mathcal{L}_s$ is just the sum of $\mathbb{E}[K_s]$ and $\mathbb{E}[I_s]$. Therefore:

$$\begin{aligned}\mathbb{E}[L_s] &= \mathbb{E}[K_s] + \mathbb{E}[I_s] \\ &= \left(\Delta^2 - 1\right)\frac{\lambda_u}{\lambda_s} + \frac{\lambda_u}{\lambda_s} = \Delta^2\frac{\lambda_u}{\lambda_s},\end{aligned} \tag{46}$$

which completes the proof.

ACKNOWLEDGMENT

This work has been facilitated by the computing resources of Information Services and Technology at the University of Alberta, WestGrid, and Compute/Calcul Canada. The authors also thank Bitan Banerjee for his assistance with some stochastic geometry calculations.

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," Cisco, San Jose, CA, USA, White Paper, Feb. 2019. Accessed: Jul. 11, 2020. [Online]. Available: https://perma.cc/GZ53-WJKH

[2] "IMT vision—Framework and overall objectives of the future development of IMT for 2020 and beyond," Int. Telecommun. Union, Geneva, Switzerland, Recommendation ITU-R M.2083-0, Sep. 2015.

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

[4] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[5] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[6] Rysavy Research, "Global 5G: Implications of a transformational technology," 5G Amer., Bellevue, WA, USA, White Paper, Sep. 2019. Accessed: Jul. 11, 2020. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2019/09/2019-5G-Americas-Rysavy-Implications-of-a-Transformational-Technology-White-Paper.pdf

[7] *Multiplexing and Channel Coding (Release 15), v15.9.0*, 3GPP Standard TS 38.212, Jun. 2020.

[8] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next?: Five promising research directions for antenna arrays," *Digit. Signal Process.*, vol. 94, pp. 3–20, Nov. 2019.

[9] D. Lee *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.

[10] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. 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.

[11] "Coordinated multi-point operation for LTE physical layer aspects (release 11), v11.2.0" 3rd Gener. Partnership Project, Sophia Antipolis, France, Tech. Rep. 3GPP TR 36.819, Sep. 2013.

[12] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, "5G network capacity: Key elements and technologies," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 71–78, Mar. 2014.

[13] Y.-J. Yu, T.-Y. Hsieh, and A.-C. Pang, "Millimeter-wave backhaul traffic minimization for CoMP over 5G cellular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4003–4015, Apr. 2019.

[14] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[15] D. Malladi, *How Can CoMP Extend 5G NR to High Capacity and Ultra-Reliable Communications?*, Qualcomm Webinar, San Diego, CA, USA, Jul. 2018. Accessed: Jul. 11, 2020. [Online]. Available: https://www.qualcomm.com/media/documents/files/how-comp-can-extend-5g-nr-to-high-capacity-and-ultra-reliable-communications.pdf

[16] V. Jungnickel *et al.*, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 44–51, May 2014.

[17] P. Xia, C.-H. Liu, and J. G. Andrews, "Downlink coordinated multipoint with overhead modeling in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4025–4037, Aug. 2013.

[18] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A comprehensive survey of RAN architectures toward 5G mobile communication system," *IEEE Access*, vol. 7, pp. 70371–70421, 2019.

[19] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[20] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[21] W. Hao, O. Muta, H. Gacanin, and H. Furukawa, "Dynamic small cell clustering and non-cooperative game-based precoding design for two-tier heterogeneous networks with massive MIMO," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 675–687, Feb. 2018.

[22] W. Hao and S. Yang, "Small cell cluster-based resource allocation for wireless backhaul in two-tier heterogeneous networks with massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 509–523, Jan. 2018.

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

[24] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015.

[25] N.-T. Le, L.-N. Tran, Q.-D. Vu, and D. Jayalath, "Energy-efficient resource allocation for OFDMA heterogeneous networks," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7043–7057, Oct. 2019.

[26] M. R. Mili, A. Khalili, D. W. K. Ng, and H. Steendam, "A novel performance tradeoff in heterogeneous networks: A multi-objective approach," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1402–1405, Oct. 2019.

[27] L. D. Nguyen, H. D. Tuan, T. Q. Duong, O. A. Dobre, and H. V. Poor, "Downlink beamforming for energy-efficient heterogeneous networks with massive MIMO and small cells," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3386–3400, May 2018.

[28] Q.-D. Vu, L.-N. Tran, R. Farrell, and E.-K. Hong, "Energy-efficient zero-forcing precoding design for small-cell networks," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 790–804, Feb. 2016.

[29] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[30] Y. Kwon, T. Hwang, and X. Wang, "Energy-efficient transmit power control for multi-tier MIMO HetNets," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2070–2086, Oct. 2015.

[31] J. Tang, D. K. C. So, A. Shojaeifard, K.-K. Wong, and J. Wen, "Joint antenna selection and spatial switching for energy efficient MIMO SWIPT system," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4754–4769, Jul. 2017.

[32] H. Li, J. Cheng, Z. Wang, and H. Wang, "Joint antenna selection and power allocation for an energy-efficient massive MIMO system," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 257–260, Feb. 2019.

[33] B. Huang and A. Guo, "Spectral and energy efficient resource allocation for massive MIMO HetNets with wireless backhaul," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 552–555, Apr. 2019.

[34] A. Nasser, O. Muta, M. Elsabrouty, and H. Gacanin, "Interference mitigation and power allocation scheme for downlink MIMO–NOMA HetNet," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6805–6816, Jul. 2019.

[35] H. Zhang, M. Feng, K. Long, G. K. Karagiannidis, V. C. M. Leung, and H. V. Poor, "Energy efficient resource management in SWIPT enabled heterogeneous networks with NOMA," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 835–845, Feb. 2020.

[36] A. Khalili, S. Akhlaghi, H. Tabassum, and D. W. K. Ng, "Joint user association and resource allocation in the uplink of heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 804–808, Jun. 2020.

[37] A. Khalili, M. R. Mili, M. Rasti, S. Parsaeefard, and D. W. K. Ng, "Antenna selection strategy for energy efficiency maximization in uplink OFDMA networks: A multi-objective approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 595–609, Jan. 2020.

[38] Y. Chen, J. Li, W. Chen, Z. Lin, and B. Vucetic, "Joint user association and resource allocation in the downlink of heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5701–5706, Jul. 2016.

[39] F. Wang, W. Chen, H. Tang, and Q. Wu, "Joint optimization of user association, subchannel allocation, and power allocation in multi-cell multi-association OFDMA heterogeneous networks," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2672–2684, Jun. 2017.

[40] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.

[41] A. Khodmi, S. B. Rejeb, N. Agoulmine, and Z. Choukair, "A joint power allocation and user association based on non-cooperative game theory in an heterogeneous ultra-dense network," *IEEE Access*, vol. 7, pp. 111790–111800, 2019.

[42] J. Xu and L. Qiu, "Energy efficiency optimization for MIMO broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 690–701, Feb. 2013.

[43] I. Ahmed *et al.*, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart., 2018.

[44] M. A. Sedaghat, R. R. Müller, and G. Fischer, "A novel single-RF transmitter for massive MIMO," in *Proc. 18th Int. ITG Workshop Smart Antennas*, Erlangen, Germany, Mar. 2014, pp. 1–8.

[45] M. A. Sedaghat, V. I. Barousis, R. R. Müller, and C. B. Papadias, "Load modulated arrays: A low-complexity antenna," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 46–52, Mar. 2016.

[46] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[47] W. Dinkelbach, "On nonlinear fractional programming," *Manag. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[48] A. Zappone, L. Sanguinetti, G. Bacci, E. Jorswieck, and M. Debbah, "Energy-efficient power control: A look at 5G wireless technologies," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1668–1683, Apr. 1, 2016.

[49] N. Trabelsi, C. S. Chen, R. El Azouzi, L. Roullet, and E. Altman, "User association and resource allocation optimization in LTE cellular networks," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 2, pp. 429–440, Jun. 2017.

[50] X. Wu, Z. Ma, X. Chen, F. Labeau, and S. Han, "Energy efficiency-aware joint resource allocation and power allocation in multi-user beamforming," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4824–4833, May 2019.

[51] G. Dong, H. Zhang, S. Jin, and D. Yuan, "Energy-efficiency-oriented joint user association and power allocation in distributed massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5794–5808, Jun. 2019.

[52] A. Mesodiakaki, F. Adelantado, L. Alonso, M. Di Renzo, and C. Verikoukis, "Energy- and spectrum-efficient user association in millimeter-wave backhaul small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1810–1821, Feb. 2017.

[53] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[54] A. Damnjanovic *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, Jun. 2011.

[55] H.-S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.

[56] A. Liu and V. K. N. Lau, "Joint BS-user association, power allocation, and user-side interference cancellation in cell-free heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 335–345, Jan. 2017.

[57] D. Ha, K. Lee, and J. Kang, "Energy efficiency analysis with circuit power consumption in massive MIMO systems," in *Proc. IEEE 24th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 938–942.

[58] R. R. Müller, M. A. Sedaghat, and G. Fischer, "Load modulated massive MIMO," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Atlanta, GA, USA, Dec. 2014, pp. 622–626.

[59] M. A. Sedaghat, R. R. Müller, G. Fischer, and A. Ali, "Discrete load-modulated single-RF MIMO transmitters," in *Proc. 20th Int. ITG Workshop Smart Antennas*, Munich, Germany, Mar. 2016, pp. 260–266.

[60] S. C. Cripps, *RF Power Amplifiers for Wireless Communications*, 2nd ed. Boston, MA, USA: Artech House, 2006.
[61] S.-E. Hong and K.-S. Oh, "A comparison of ESPAR-MIMO and LMA-MIMO for single-RF transmission of spatially multiplexed QAM signals," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Las Vegas, NV, USA, Jan. 2017, pp. 133–134.
[62] J. Oh, H. Kim, S. Cho, and G. Jo, "A single RF-chain load modulation transmitter of simple structure for massive MIMO," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Jeju, South Korea, Oct. 2017, pp. 954–956.
[63] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.
[64] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Mar. 1998.
[65] Y. Zeng, E. Gunawan, and Y. L. Guan, "Modified block diagonalization precoding in multicell cooperative networks," *IEEE Trans. Veh. Technol.*, vol. 61, no. 8, pp. 3819–3824, Oct. 2012.
[66] D. H. N. Nguyen, H. Nguyen-Le, and T. Le-Ngoc, "Block-diagonalization precoding in a multiuser multicell MIMO system: Competition and coordination," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 968–981, Feb. 2014.
[67] K. Zu, R. C. de Lamare, and M. Haardt, "Generalized design of low-complexity block diagonalization type precoding algorithms for multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4232–4242, Oct. 2013.
[68] L. D. Nguyen, H. D. Tuan, and T. Q. Duong, "Energy-efficient signalling in QoS constrained heterogeneous networks," *IEEE Access*, vol. 4, pp. 7958–7966, 2016.
[69] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3292–3304, Sep. 2012.
[70] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[71] K. T. K. Cheung, S. Yang, and L. Hanzo, "Achieving maximum energy-efficiency in multi-relay OFDMA cellular networks: A fractional programming approach," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2746–2757, Jul. 2013.
[72] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.
[73] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. Baltimore, MD, USA: John Hopkins Univ. Press, 2013.
[74] D. E. Knuth, *The Art of Computer Programming: Vol. 3: Sorting and Searching*, 2nd ed. Upper Saddle River, NJ, USA: Addison-Wesley, 1998.
[75] P. S. Stanimirović and M. B. Tasić, "Computing generalized inverses using LU factorization of matrix product," *Int. J. Comput. Math.*, vol. 85, no. 12, pp. 1865–1878, Dec. 2008.
[76] I. P. Stanimirović and M. B. Tasić, "Computation of generalized inverses by using the *LDL** decomposition," *Appl. Math. Lett.*, vol. 25, pp. 526–531, Mar. 2012.
[77] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[78] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 8th ed., D. Zwillinger and V. Moll, Eds. Waltham, MA, USA: Academic, 2014.
[79] "Evolved universal terrestrial radio access (E-UTRA); Further advancements for E-UTRA physical layer aspects, v9.2.0," 3rd Gener. Partnership Project, Sophia Antipolis, France, Tech. Rep. 3GPP TR 36.814, Mar. 2017.
[80] Q. Wang, G. Lim, L. J. Cimini, L. J. Greenstein, D. S. Chan, and A. Hedayat, "Quantifying and comparing energy efficiencies on SU-MIMO and MU-MIMO downlinks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
[81] K. I. Pedersen, P. E. Mogensen, and B. H. Fleury, "Power azimuth spectrum in outdoor environments," *Electron. Lett.*, vol. 33, no. 18, pp. 1583–1584, Aug. 1997.
[82] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[83] Y. J. Chun, S. L. Cotton, H. S. Dhillon, F. J. Lopez-Martinez, J. F. Paris, and S. K. Yoo, "A comprehensive analysis of 5G heterogeneous cellular systems operating over κ-μ shadowed fading channels," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 6995–7010, Nov. 2017.
[84] V. Chandrasekhar, M. Kountouris, and J. G. Andrews, "Coverage in multi-antenna two-tier networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5314–5327, Oct. 2009.
[85] K. Hosseini, W. Yu, and R. S. Adve, "Large-scale MIMO versus network MIMO for multicell interference mitigation," *IEEE J. Sel. Topics. Signal Process.*, vol. 8, no. 5, pp. 930–941, Oct. 2014.
[86] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO—5G system," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 905–923, 2nd Quart., 2016.
[87] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
[88] J. Minango and C. de Almeida, "Low complexity zero forcing detector based on Newton–Schultz iterative algorithm for massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11759–11766, Dec. 2018.
[89] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, "Massive MIMO performance with imperfect channel reciprocity and channel estimation error," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3734–3749, Sep. 2017.
[90] Z. Gong, C. Li, and F. Jiang, "Pilot contamination mitigation strategies in massive MIMO systems," *IET Commun.*, vol. 11, no. 16, pp. 2403–2409, Nov. 2017.
[91] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, Jun. 2014.
[92] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4611–4624, Jul. 2016.
[93] A.-A. Lu, X. Gao, W. Zhong, C. Xiao, and X. Meng, "Robust transmission for massive MIMO downlink with imperfect CSI," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5362–5376, Aug. 2019.
[94] A. Tajer, N. Prasad, and X. Wang, "Robust linear precoder design for multi-cell downlink transmission," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 235–251, Jan. 2011.
[95] M. Soleimani, M. Mazrouei-Sebdani, R. C. Elliott, W. A. Krzymień, and J. Melzer, "Robust precoder design for massive MIMO with peak total power constrained single-RF-chain transmitters," *IET Commun.*, vol. 11, no. 17, pp. 2667–2672, Nov. 2017.

MAHTAB ATAEESHOJAI received the B.Sc. degree in electrical engineering from Shahid Beheshti University, Tehran, Iran, in 2014, and the M.Sc. degree in electrical engineering from the Amirkabir University of Technology, Tehran, in 2017. She is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Alberta, Edmonton, AB, Canada.

During her B.Sc. studies, she was an Intern with the Iran Telecommunication Research Center working on gigabit passive optical networks. Her current research interests include wireless communications, green communications, MIMO and massive MIMO, energy harvesting wireless networks, and radio resource allocation.

ROBERT C. ELLIOTT (Senior Member, IEEE) received the B.Sc. (cooperative) and M.Sc. degrees in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2000 and 2003, respectively, and the Ph.D. degree in communications from the Department of Electrical and Computer Engineering, University of Alberta in 2011.

During his B.Sc. studies, he held several cooperative work experience positions. In 1998, he was with Computing Devices Canada (currently, General Dynamics Mission Systems—Canada), Calgary, AB, and he was with Nortel Networks, Ottawa, ON, Canada, in 1999. From 2001 to 2016, he was also affiliated with Telecommunications Research Laboratories (TRTech), Edmonton. In 2005, he was a Visiting Researcher with the German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He has also done collaborative research with Huawei Technologies and TELUS Communications, in part as a Postdoctoral Fellow. He is currently a Research Associate with the Department of Electrical and Computer Engineering with the University of Alberta. His research interests include heterogeneous cellular networks, coordinated transmission techniques in broadband multiuser multiple-input multiple-output wireless systems, massive MIMO systems, radio resource management, and metaheuristic stochastic optimization methods.

Dr. Elliott received the Governor General's Silver Academic Medal and the APEGGA Medal in Electrical Engineering in 2000 for having the highest overall undergraduate academic standing at the University of Alberta. He has also held numerous scholarships and fellowships during his academic studies.

WITOLD A. KRZYMIEŃ (Fellow, IEEE) received the M.Sc. (Eng.) and Ph.D. degrees in electrical engineering from the Poznań University of Technology, Poznań, Poland, in 1970 and 1978, respectively.

Since April 1986, he has been with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada, where he currently holds the endowed Rohit Sharma Professorship in Communications and Signal Processing. In 1986, he was one of the key research program architects of the newly launched TRLabs, which for a long time was Canada's largest industry-university-government pre-competitive research consortium in the information and communication technology area. His research activity was closely tied to the consortium for the following three decades. Over the years, he has also done collaborative research work with TELUS Communications, Huawei Technologies, Nortel Networks, Ericsson, German Aerospace Center (DLR—Oberpfaffenhofen) and the University of Padova, Italy. His research is currently focused on radio resource management and transceiver signal processing for broadband heterogeneous cellular networks employing multiuser MIMO and massive MIMO antenna techniques.

From 1999 to 2005, Dr. Krzymień was the Chairman of Commission C (Radio Communication Systems and Signal Processing) of the Canadian National Committee of URSI (Union Radio Scientifique Internationale). He has been an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY since 2007. He has been elected as a Fellow of the IEEE for his contributions to radio resource management for cellular systems and networks. He is also a Fellow of the Engineering Institute of Canada, and a licensed Professional Engineer with the Province of Alberta, Canada. He has chaired or co-chaired technical program committees for numerous conferences in wireless communication systems and communication theory areas.

CHINTHA TELLAMBURA (Fellow, IEEE) received the B.Sc. degree (First-Class Hons.) from the University of Moratuwa, Sri Lanka, the M.Sc. degree in electronics from the King's College, University of London, U.K., and the Ph.D. degree in electrical engineering from the University of Victoria, Canada.

He was with Monash University, Australia, from 1997 to 2002. He is a Professor with the Department of Electrical and Computer Engineering, University of Alberta. He has authored or coauthored over 500 journal and conference papers with an h-index of 77 (Google Scholar). His current research interests include the design, modeling and analysis of cognitive radio, heterogeneous cellular networks, 5G wireless networks, and machine learning algorithms.

Prof. Tellambura has received the Best Paper Awards in the Communication Theory Symposium in 2012 IEEE International Conference on Communications (ICC) in Canada and 2017 ICC in France. He is the winner of the prestigious McCalla Professorship and the Killam Annual Professorship from the University of Alberta. He served as an Editor for both IEEE TRANSACTIONS ON COMMUNICATIONS from 1999 to 2011, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2001 to 2007, and for the latter he was an Area Editor for *Wireless Communications Systems and Theory*, from 2007 to 2012. In 2011, he was elected as an IEEE Fellow for his contributions to physical layer wireless communication theory. In 2017, he was elected as a Fellow of Canadian Academy of Engineering.

JORDAN MELZER (Member, IEEE) received the B.A.Sc. degree in engineering science from the University of Toronto, Toronto, ON, Canada, and the M.S.E.E. and Ph.D. degrees in communications from the University of Southern California, Los Angeles, CA, USA. He is currently a Senior Engineer with the Broadband Access Group, TELUS Communications, Ottawa, ON, Canada. He has co-authored several dozen academic and industry publications. Prior to joining TELUS, he co-developed TrellisWare Technology's F-LDPC error control code product and was an early member of Nortel Networks' all-IP Cellular Radio Research Group. His current research interest is in rural and high speed Internet access.