Massive MIMO and Small Cells: Improving Energy Efficiency by Optimal Soft-Cell Coordination

Emil Björnson*†, Marios Kountouris‡, and Mérouane Debbah*
*Alcatel-Lucent Chair on Flexible Radio, SUPELEC, Gif-sur-Yvette, France
‡Department of Telecommunications, SUPELEC, Gif-sur-Yvette, France
†ACCESS Linnaeus Centre, Signal Processing Lab, KTH Royal Institute of Technology, Stockholm, Sweden
Emails: {emil.bjornson, marios.kountouris, merouane.debbah}@supelec.fr

***Abstract*—To improve the cellular energy efficiency, without sacrificing quality-of-service (QoS) at the users, the network topology must be densified to enable higher spatial reuse. We analyze a combination of two densification approaches, namely "massive" multiple-input multiple-output (MIMO) base stations and small-cell access points. If the latter are operator-deployed, a spatial soft-cell approach can be taken where the multiple transmitters serve the users by joint non-coherent multiflow beamforming. We minimize the total power consumption (both dynamic emitted power and static hardware power) while satisfying QoS constraints. This problem is proved to have a hidden convexity that enables efficient solution algorithms. Interestingly, the optimal solution promotes exclusive assignment of users to transmitters. Furthermore, we provide promising simulation results showing how the total power consumption can be greatly improved by combining massive MIMO and small cells; this is possible with both optimal and low-complexity beamforming.**

I. Introduction

The classical macro-cell network topology is well-suited for providing wide-area coverage, but cannot handle the rapidly increasing user numbers and QoS expectations that we see today—the energy efficiency would be very low. The road forward seems to be a densified topology that enables very high spatial reuse. Two main approaches are currently investigated: *massive MIMO* [1], [2] and *small-cell networks* [3], [4].

The first approach is to deploy large-scale antenna arrays at existing macro base stations (BSs) [1]. This enables precise focusing of emitted energy on the intended users, resulting in a much higher energy efficiency. The channel acquisition is indispensable for massive MIMO, which requires the exploitation of channel reciprocity using time-division duplex (TDD). This mode makes the channel estimation accuracy limited by the number of users and not the number of BS antennas [1].

The second approach is to deploy an overlaid layer of small-cell access points (SCAs) to offload traffic from BSs, thus exploiting the fact that most data traffic is localized and requested by low-mobility users. This approach reduces the average distance between users and transmitters, which translates into lower propagation losses and higher energy efficiency [4]. This comes at the price of having a highly heterogeneous network topology where it is difficult to control and coordinate inter-user interference. To meet this challenge, industry [3] and academia [4] are shifting focus from user-deployed femtocells to operator-deployed SCAs. The latter can rely on reliable backhaul connectivity and joint control/coordination of BS and SCAs; the existence of SCAs can even be transparent to the users, as in the soft-cell approach proposed for LTE in [3].

The total power consumption can be modeled with a static part that depends on the transceiver hardware and a dynamic part which is proportional to the emitted signal power [5]–[7]. Massive MIMO and small-cell networks promise great improvements in the dynamic part, but require more hardware and will therefore increase the static part. In other words, dense network topologies must be properly deployed and optimized to actually improve the overall energy efficiency.

This paper analyzes the possible improvements in energy efficiency when the classical macro-cell topology is modified by employing massive MIMO at the BS and/or overlaying with SCAs. We assume perfect channel acquisition and a backhaul network that supports interference coordination; we thus consider an ultimate bound on what is practically achievable. The goal is to *minimize the total power consumption while satisfying QoS constraints at the users and power constraints at the BS and SCAs*. We show that this optimization problem has a hidden convex structure that enables finding the optimal solution in polynomial time. The solution is proved to automatically/dynamically assign each user to the optimal transmitter (BS or SCA). A low-complexity algorithm based on classical regularized zero-forcing (RZF) beamforming is proposed and compared with the optimal solution. The potential merits of different densified topologies are analyzed by simulations.

II. System Model

We consider a single-cell downlink scenario where a macro BS equipped with N_{BS} antennas should deliver information to K single-antenna users. In addition, there are $S > 0$ SCAs that form an overlay layer and are arbitrarily deployed. The SCAs are equipped with N_{SCA} antennas each, typically $1 \leq N_{\text{SCA}} \leq 4$, and characterized by strict power constraints that limit their coverage area (see below). In comparison, the BS has generous power constraints that can support high QoS targets in a large coverage area. The number of antennas, N_{BS}, is anything from 8 to several hundred—the latter means that $N_{\text{BS}} \gg K$ and is known as massive MIMO. This scenario is illustrated in Fig. 1.

E. Björnson is funded by the International Postdoc Grant 2012-228 from The Swedish Research Council. This research has been supported by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering).

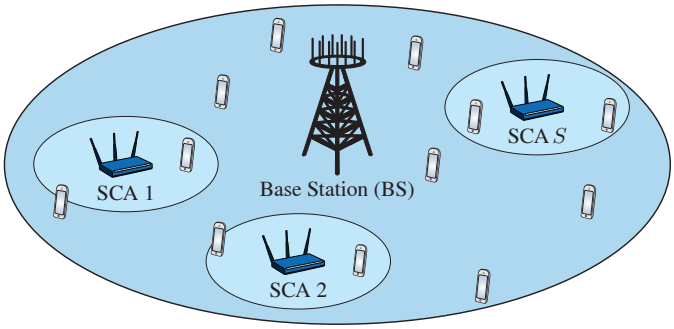

Fig. 1. Illustration of a downlink macro-cell overlaid with S small cells. The BS has N_{BS} antennas and the SCAs have N_{SCA} antennas. The K single-antenna users (e.g., smartphones) can be served (non-coherently) by any combination of transmitters, but the circles indicate typical coverage areas.

The channels to user k are modeled as block fading. We consider a single flat-fading subcarrier where the channels are represented in the baseband by $\mathbf{h}_{k,0}^H \in \mathbb{C}^{1\times N_{\text{BS}}}$ and $\mathbf{h}_{k,j}^H \in \mathbb{C}^{1\times N_{\text{SCA}}}$ for the BS and jth SCA, respectively. These are assumed to be perfectly known at both sides of each channel; extensions with robustness to channel uncertainty can be obtained as in [8]. The received signal at user k is

$$y_k = \mathbf{h}_{k,0}^H \mathbf{x}_0 + \sum_{j=1}^{S} \mathbf{h}_{k,j}^H \mathbf{x}_j + n_k \tag{1}$$

where $\mathbf{x}_0, \mathbf{x}_j$ are the transmitted signals at the BS and jth SCA, respectively. The term $n_k \sim \mathcal{CN}(0,\sigma_k^2)$ is the circularly-symmetric complex Gaussian receiver noise with zero-mean and variance σ_k^2, measured in milliwatt (mW).

The BS and SCAs are connected to a backhaul network that enables joint spatial soft-cell resource allocation but only linear non-coherent transmissions; that is, each user can be served by multiple transmitters but the information symbols will be coded and emitted independently. We call it *spatial multiflow transmission* [9] and it enables users barely covered by a SCA to receive extra signals from the BS or other SCAs.

The information symbols from the BS and the jth SCA to user k are denoted $x_{k,0}$ and $x_{k,j}$, respectively, and originate from independent Gaussian codebooks with unit power (in mW); that is, $x_{k,j} \sim \mathcal{CN}(0,1)$ for $j = 0,\ldots,S$. These symbols are multiplied with the beamforming vectors $\mathbf{w}_{k,0} \in \mathbb{C}^{N_{\text{BS}}\times 1}$ and $\mathbf{w}_{k,j} \in \mathbb{C}^{N_{\text{SCA}}\times 1}$ to obtain the transmitted signals

$$\mathbf{x}_j = \sum_{k=1}^{K} \mathbf{w}_{k,j} x_{k,j}, \quad j = 0,\ldots,S. \tag{2}$$

The beamforming vectors are the optimization variables in this paper. Note that $\mathbf{w}_{k,j} \neq \mathbf{0}$ only for transmitters j that serve user k. This transmitter assignment is obtained automatically and optimally from the optimization problem solved herein.

A. Problem Formulation

This paper considers minimization of the total power consumption while satisfying QoS constraints for each user. We will define both concepts before formulating the problem.

The QoS constraints specify the information rate [bits/s/Hz] that each user should achieve in parallel. These are defined as $\log_2(1+\text{SINR}_k) \geq \gamma_k$, where γ_k is the fixed QoS target and

$$\text{SINR}_k = \frac{|\mathbf{h}_{k,0}^H \mathbf{w}_{k,0}|^2 + \sum_{j=1}^{S} |\mathbf{h}_{k,j}^H \mathbf{w}_{k,j}|^2}{\sum_{\substack{i=1 \\ i\neq k}}^{K} \left(|\mathbf{h}_{k,0}^H \mathbf{w}_{i,0}|^2 + \sum_{j=1}^{S} |\mathbf{h}_{k,j}^H \mathbf{w}_{i,j}|^2 \right) + \sigma_k^2} \tag{3}$$

is the aggregate signal-to-interference-and-noise ratio (SINR) of the kth user. The information rate $\log_2(1+\text{SINR}_k)$ is achieved by applying successive interference cancellation on the own information symbols and treating co-user symbols as noise. Observe that this rate is obtained without any phase-synchronization between transmitters, contrary to coherent joint transmission that requires very tight synchronization [10].

The power consumption (per subcarrier) can be modeled as $P_{\text{dynamic}} + P_{\text{static}}$ [5]–[7] with the dynamic and static terms

$$P_{\text{dynamic}} = \rho_0 \sum_{k=1}^{K} \|\mathbf{w}_{k,0}\|^2 + \sum_{j=1}^{S} \rho_j \sum_{k=1}^{K} \|\mathbf{w}_{k,j}\|^2, \tag{4}$$

$$P_{\text{static}} = \frac{\eta_0}{C} N_{\text{BS}} + \sum_{j=1}^{S} \frac{\eta_j}{C} N_{\text{SCA}}, \tag{5}$$

respectively. The dynamic term is the aggregation of the emitted powers, $\sum_{k=1}^{K} \|\mathbf{w}_{k,j}\|^2$, each multiplied with a constant $\rho_j \geq 1$ accounting for the inefficiency of the power amplifier at this transmitter. The static term, P_{static}, is proportional to the number of antennas and $\eta_j \geq 0$ models the power dissipation in the circuits of each antenna (e.g., in filters, mixers, converters, and baseband processing). P_{static} is normalized with the total number of subcarriers $C \geq 1$. Representative numbers on these parameters are given in Table I, [6], and [11]

Each BS and SCA is prone to L_j power constraints

$$\sum_{k=1}^{K} \mathbf{w}_{k,j}^H \mathbf{Q}_{j,\ell} \mathbf{w}_{k,j} \leq q_{j,\ell}, \quad \ell = 1,\ldots,L_j. \tag{6}$$

The weighting matrices $\mathbf{Q}_{0,\ell} \in \mathbb{C}^{N_{\text{BS}}\times N_{\text{BS}}}$, $\mathbf{Q}_{j,\ell} \in \mathbb{C}^{N_{\text{SCA}}\times N_{\text{SCA}}}$ for $j = 1,\ldots,S$, are positive semi-definite. The corresponding limits are $q_{j,\ell} \geq 0$. The parameters $\mathbf{Q}_{j,\ell}, q_{j,\ell}$ are fixed and can describe any combination of per-antenna, per-array, and soft-shaping constraints [10]. We typically have $q_{0,\ell} \gg q_{j,\ell}$ for $1 \leq j \leq S$, because the BS provides coverage. Our numerical evaluation considers per-antenna constraints of q_j [mW] at the jth transmitter, given by $L_0 = N_{\text{BS}}$, $L_j = N_{\text{SCA}}$, $q_{j,\ell} = q_j\ \forall \ell$, and $\mathbf{Q}_{j,\ell}$ with one at ℓth diagonal element and zero elsewhere.

We are now ready to formulate our optimization problem. We want to minimize the total power consumption while satisfying the QoS constraints and the power constraints, thus

$$\begin{aligned} \underset{\mathbf{w}_{k,j}\ \forall k,j}{\text{minimize}} \quad & P_{\text{dynamic}} + P_{\text{static}} \\ \text{subject to} \quad & \log_2(1+\text{SINR}_k) \geq \gamma_k \quad \forall k, \\ & \sum_{k=1}^{K} \mathbf{w}_{k,j}^H \mathbf{Q}_{j,\ell} \mathbf{w}_{k,j} \leq q_{j,\ell} \quad \forall j,\ell. \end{aligned} \tag{7}$$

In the next section, we will prove that (7) can be reformulated as a convex optimization problem and thus is solvable

in polynomial time using standard algorithms. Moreover, the optimal power-minimizing solution is self-organizing in the sense that only one or a few transmitters will serve each user.

Remark 1. *The static part, P_{static}, of the power consumption depends on the number of SCAs and antennas. From an energy efficiency perspective, it therefore makes sense to put inactive SCAs and antenna elements into sleep mode. On the other hand, such adaptive sleep mode techniques make the sensing of user mobility and new users complicated. There is also a non-negligible transient behavior when switching from sleep mode to active mode [5]. Since these problems are outside the scope of this paper, we will instead compare setups with different values on N_{BS}, N_{SCA}, and S by using simulations.*

III. Algorithms for Non-Coherent Coordination

This section derives algorithms for solving the optimization problem (7). The QoS constraints in (7) are complicated functions of the beamforming vectors, making the problem non-convex in its original formulation. However, we will prove that it has an underlying convex structure that can be extracted using semi-definite relaxation. We generalize the original approach in [12] to spatial multiflow transmission.

To achieve a convex reformulation of (7), we use the notation $\mathbf{W}_{k,j} = \mathbf{w}_{k,j}\mathbf{w}_{k,j}^H \ \forall k,j$. This matrix should be positive semi-definite, denoted as $\mathbf{W}_{k,j} \succeq \mathbf{0}$, and have $\mathrm{rank}(\mathbf{W}_{k,j}) \leq 1$. Note that the rank can be zero, which implies that $\mathbf{W}_{k,j} = \mathbf{0}$. By including the BS and SCAs in the same sum expressions, we can rewrite (7) compactly as

$$\begin{aligned} \underset{\mathbf{W}_{k,j}\succeq\mathbf{0}\ \forall k,j}{\text{minimize}} \quad & \sum_{j=0}^{S}\rho_j\sum_{k=1}^{K}\mathrm{tr}(\mathbf{W}_{k,j}) + P_{\text{static}} \qquad (8)\\ \text{subject to} \quad & \mathrm{rank}(\mathbf{W}_{k,j}) \leq 1 \quad \forall k,j, \\ & \sum_{j=0}^{S}\mathbf{h}_{k,j}^H\left(\left(1+\frac{1}{\tilde{\gamma}_k}\right)\mathbf{W}_{k,j} - \sum_{i=1}^{K}\mathbf{W}_{i,j}\right)\mathbf{h}_{k,j} \geq \sigma_k^2 \quad \forall k, \\ & \sum_{k=1}^{K}\mathrm{tr}(\mathbf{Q}_{j,\ell}\mathbf{W}_{k,j}) \leq q_{j,\ell} \quad \forall j,\ell, \end{aligned}$$

where the QoS targets have been transformed into SINR targets of $\tilde{\gamma}_k = 2^{\gamma_k} - 1 \ \forall k$. The problem (8) is convex except for the rank constraints, but we will now prove that these constraints can be relaxed without losing optimality.

Theorem 1. *Consider the semi-definite relaxation of* (8) *where the rank constraints* $\mathrm{rank}(\mathbf{W}_{k,j}) \leq 1$ *are removed. This becomes a convex semi-definite optimization problem. Furthermore, it will always have an optimal solution* $\{\mathbf{W}_{k,j}^* \ \forall k,j\}$ *where all matrices satisfy* $\mathrm{rank}(\mathbf{W}_{k,j}^*) \leq 1$.

Proof: The proof is given in the appendix. ■

This theorem shows that the original problem (7) can be solved as a convex optimization problem. This means that the optimal solution is guaranteed in polynomial time [13]; for example, using the interior-point toolbox `SeDuMi` [14].

Further structure of the optimal solution can be obtained.

Corollary 1. *Consider the optimal solution* $\{\mathbf{W}_{k,j}^* \ \forall k,j\}$ *to* (8)*. For each user k there are three possibilities:*

1) *It is only served by the BS (i.e., $\mathbf{W}_{k,j}^* = \mathbf{0}$, $1 \leq j \leq S$);*
2) *It is only served by the jth SCA (i.e., $\mathbf{W}_{k,0}^* = \mathbf{0}$ and $\mathbf{W}_{k,i}^* = \mathbf{0}$ for $i \neq j$);*
3) *It is served by a combination of BS and SCAs, whereof at least one transmitter j has an active power constraint ℓ (i.e., $\sum_{k=1}^{K}\mathrm{tr}(\mathbf{Q}_{j,\ell}\mathbf{W}_{k,j}^*) = q_{j,\ell}$).*

Proof: The proof is given in the appendix. ■

This corollary shows that although users can be served by multiflow transmission, it is usually optimal to assign one transmitter per user. Users that are close to a SCA are served exclusively by it, while most other users are served by the BS. There are transition areas around each SCA where multiflow transmission is utilized since the SCA is unable to fully support the QoS targets. Corollary 1 is a positive result since a reduced transmission/reception complexity is often optimal.

If the power constraints are removed, then the transition areas disappear. We refer to [15] for prior work on dynamic transmitter assignment by means of convex optimization.

A. Low-Complexity Algorithm

The optimal beamforming for spatial soft-cell coordination can be computed in polynomial time using Theorem 1. This complexity is relatively modest, but the algorithm becomes infeasible for real-time implementation when N_{BS} and S grow large. In addition, Theorem 1 provides a centralized algorithm that requires all channel knowledge to be gathered at the BS. Distributed algorithms can certainly be obtained using primal/dual decomposition techniques [8], but these require iterative backhaul signaling of coupling variables—thus they are also infeasible for real-time implementations.

Theorem 1 should be seen as the ultimate benchmark when evaluating low-complexity algorithms for non-coherent coordination. To demonstrate the usefulness, we propose the low-complexity non-iterative **Multiflow-RZF beamforming**:

1) Each transmitter $j = 0,\ldots,S$ computes

$$\mathbf{u}_{k,j} = \frac{\left(\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\mathbf{h}_{i,j}\mathbf{h}_{i,j}^H + \frac{K}{\tilde{\gamma}_k q_j}\mathbf{I}\right)^{-1}\mathbf{h}_{k,j}}{\left\|\left(\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\mathbf{h}_{i,j}\mathbf{h}_{i,j}^H + \frac{K}{\tilde{\gamma}_k q_j}\mathbf{I}\right)^{-1}\mathbf{h}_{k,j}\right\|} \ \forall k,$$

$$g_{i,k,j} = |\mathbf{h}_{i,j}^H\mathbf{u}_{k,j}|^2 \ \forall i,k, \quad Q_{j,\ell,k} = \mathbf{u}_{k,j}^H\mathbf{Q}_{j,\ell}\mathbf{u}_{k,j} \ \forall \ell,k.$$

2) The jth SCA sends the scalars $g_{i,k,j}, Q_{j,\ell,k} \ \forall k,i,\ell$ to the BS. The BS solves the convex optimization problem

$$\begin{aligned} \underset{p_{k,j}\geq 0\ \forall k,j}{\text{minimize}} \quad & \sum_{j=0}^{S}\rho_j\sum_{k=1}^{K}p_{k,j} + P_{\text{static}} \qquad (9)\\ \text{subject to} \quad & \sum_{k=1}^{K}Q_{j,\ell,k}p_{k,j} \leq q_{j,\ell} \quad \forall j,\ell, \\ & \sum_{j=0}^{S}p_{k,j}g_{k,k,j}\left(1+\frac{1}{\tilde{\gamma}_k}\right) - \sum_{i=1}^{K}p_{i,j}g_{k,i,j} \geq \sigma_k^2 \quad \forall k. \end{aligned}$$

3) The power allocation $p_{k,j}^* \ \forall k$ that solves (9) is sent to the jth SCA, which computes $\mathbf{w}_{k,j} = \sqrt{p_{k,j}^*}\mathbf{u}_{k,j} \ \forall k$.

This algorithm applies the heuristic RZF beamforming (see e.g., [2]) to transform (7) into the power allocation problem

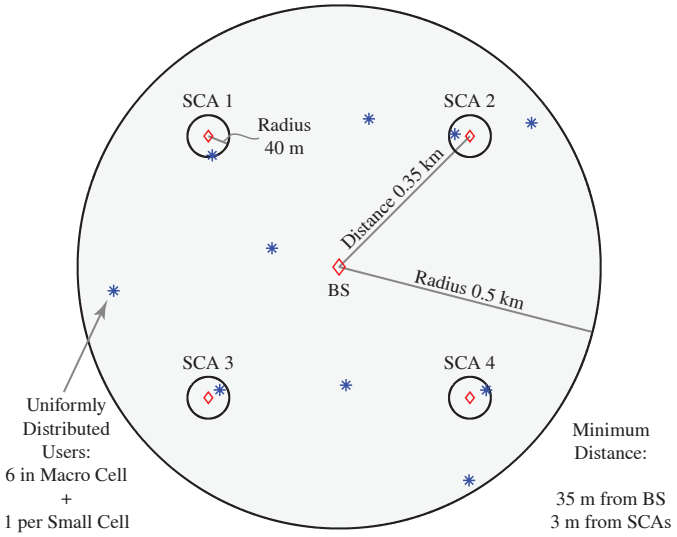

Fig. 2. The single-cell scenario analyzed in Section IV. The BS and SCAs are fixed, while the 10 users are randomly distributed as described above.

TABLE I
HARDWARE PARAMETERS IN THE NUMERICAL EVALUATION

Parameters	Values
Efficiency of power amplifiers	$\frac{1}{\rho_0} = 0.388$, $\frac{1}{\rho_j} = 0.052$ $\forall j$
Circuit power per antenna	$\eta_0 = 189$ mW, $\eta_j = 5.6$ mW $\forall j$
Per-antenna constraints	$q_{0,\ell} = 66$, $q_{j,\ell} = 0.08$ mW $\forall j, \ell$

(9), which has the same low complexity irrespectively of the number of antennas. The algorithm is non-iterative, but some scalar parameters are exchanged between the BS and SCAs to enable coordination. In practice, only users in the vicinity of an SCA are affected by it, thus only a few parameters are exchanged per SCA while all other parameters are set to zero.

IV. NUMERICAL EVALUATIONS

This section illustrates the analytic results and algorithms of this paper in the scenario depicted in Fig. 2. This figure shows a circular macro cell overlaid by 4 small cells. There are 10 active users in the macro cell, whereof 6 users are uniformly distributed in the whole cell and each SCA has one user uniformly distributed within 40 meters. We evaluate the average performance over user locations and channel realizations. Table I shows the hardware parameters that characterize the power consumption and is based on [6, Table 7] and [11].

The channels are modeled similarly to Case 1 for Heterogeneous deployments in the 3GPP LTE standard [16], but the small-scale fading is modified to reflect recent works on massive MIMO. We assume Rayleigh small-scale fading: $\mathbf{h}_{k,j} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{k,j})$. The correlation matrix is spatially uncorrelated, $\mathbf{R}_{k,j} \propto \mathbf{I}$, between the jth SCA and each user k. The correlation matrix between the BS and each user is modeled according to the physical channel model in [2, Eq. (34)], where the main characteristics are antenna correlation and reduced-rank channels. Note that the propagation loss is different for BS and SCAs; see Table II for all channel model parameters.

We first analyze the impact of having different number of antennas at the BS and SCAs: $N_{\text{BS}} \in \{20, 30, \ldots, 100\}$,

TABLE II
CHANNEL PARAMETERS IN THE NUMERICAL EVALUATION

Parameters	Values
Macro cell radius	0.5 km
Carrier frequency / Number of subcarriers	$F = 2$ GHz / $C = 600$
Total bandwidth / Subcarrier bandwidth	10 MHz / 15 kHz
Small-scale fading distribution	$\mathbf{h}_{k,j} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{k,j})$
Standard deviation of log-normal shadowing	7 dB
Path and penetration loss at distance d (km)	$148.1{+}37.6\log_{10}(d)$ dB
Special case: Within 40 m from SCA	$127{+}30\log_{10}(d)$ dB
Noise variance σ_k^2 (5 dB noise figure)	-127 dBm

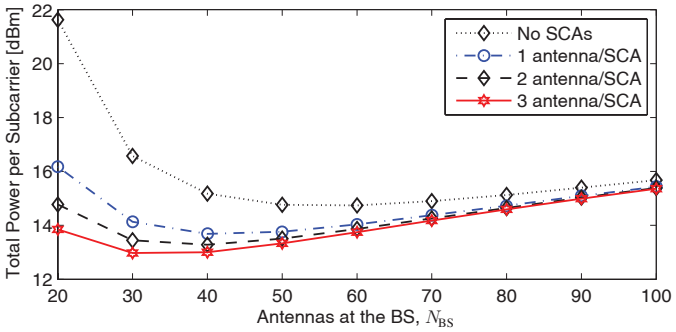

Fig. 3. Average total power consumption in the scenario of Fig. 2. We consider different N_{BS} and N_{SCA}, while the QoS constraints are 2 bits/s/Hz.

$N_{\text{SCA}} \in \{0, 1, 2, 3\}$. Fig. 3 shows the average total power consumption (per subcarrier) in a scenario where the 10 users have QoS constraints of 2 bits/s/Hz. The optimal spatial multiflow transmission is obtained using Theorem 1 and the convex optimization problems were solved by the algorithmic toolbox `SeDuMi` [14], using the modeling language `CVX` [17].

Fig. 3 demonstrates that adding more hardware can substantially decrease the total power consumption $P_{\text{dynamic}} + P_{\text{static}}$. This means that the decrease in the dynamic part, P_{dynamic}, due to better energy-focusing and less propagation losses clearly outweigh the increase in the static part, P_{static}, from the extra circuitry. Massive MIMO brings large energy efficiency improvements by itself, but the same power consumption can be achieved with half the number of BS antennas (or less) by deploying a few single-antenna SCAs in areas with active users. Further improvements in energy efficiency are achieved by having multi-antenna SCAs; a network topology that combines massive MIMO and small cells is desirable to achieve high energy efficiency with little additional hardware. However, there are saturation points where extra hardware will not decrease the total power anymore. Note that the power is shown in dBm, thus there are 10-fold improvements in Fig. 3.

Although the system allows for multiflow transmission, the simulation shows only a 0–3% probability of serving a user by multiple transmitters. This is in line with Corollary 1. The main impact of increasing N_{SCA} is that each SCA is likely to being allocated more than one user to serve exclusively; the probability is 20–45% for $N_{\text{SCA}} = 3$ but decreases with N_{BS}.

Next, Fig. 4 considers $N_{\text{BS}} = 50$ and $N_{\text{SCA}} = 2$ for different QoS constraints. Three beamforming algorithms are compared: 1) Optimal beamforming using only the BS; 2) Multiflow-RZF proposed in Section III-A; and 3) Optimal spatial soft-cell

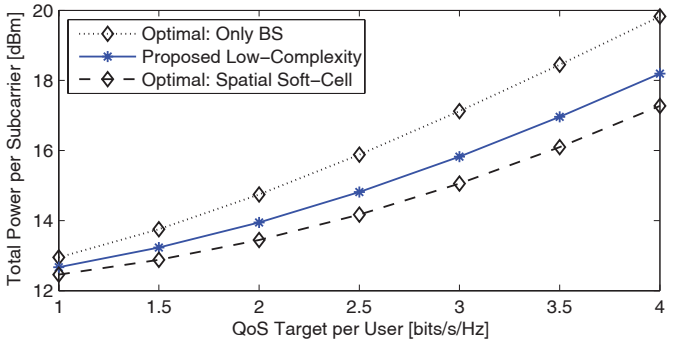

Fig. 4. Average total power consumption in the scenario of Fig. 2 with $N_{\text{BS}} = 50$ and $N_{\text{SCA}} = 2$. We consider different QoS constraints and beamforming.

coordination from Theorem 1. As in the previous figure, we observe great improvements in energy efficiency by offloading users to the SCAs. The proposed Multiflow-RZF beamforming gives promising results for practical applications, because a majority of the energy efficiency improvements is achievable by judicious low-complexity beamforming techniques.

V. Conclusion

The energy efficiency of cellular networks can be improved by employing massive MIMO at the BSs or overlaying current infrastructure by a layer of SCAs. This paper analyzed a combination of these concepts based on soft-cell coordination, where each user can be served by non-coherent beamforming from multiple transmitters. We proved that the *power-minimizing spatial multiflow transmission under QoS constraints* is achieved by solving a convex optimization problem. The optimal solution dynamically assigns users to the optimal transmitters, which usually is only the BS or one of the SCAs.

The analysis considered both the dynamic emitted power and static hardware consumption. We provide promising results showing that the *total power consumption* can be greatly improved by combining massive MIMO and small cells. Most of the benefits are also achievable by low-complexity beamforming, such as the proposed Multiflow-RZF beamforming.

Appendix

Proof of Theorem 1. The relaxed problem is a semi-definite optimization problem on standard form [13]. As shown in [10, Example 1], there might exist high-rank solutions. However, there always exist a solution with $\text{rank}(\mathbf{W}_{k,j}^*) \leq 1 \ \forall k,j$. To prove this, suppose there exist an optimal solution $\{\mathbf{W}_{k,j}^{**} \ \forall k,j\}$ with $\text{rank}(\mathbf{W}_{k,j}^{**}) > 1$ for some k,j. We can replace $\mathbf{W}_{k,j}^{**}$ by any $\mathbf{V} \succeq \mathbf{0}$ that maximizes $\mathbf{h}_{k,j}^H \mathbf{V} \mathbf{h}_{k,j}$ subject to $\text{tr}(\mathbf{V}) \leq \text{tr}(\mathbf{W}_{k,j}^{**})$, $\text{tr}(\mathbf{Q}_{j,\ell}\mathbf{V}) \leq \text{tr}(\mathbf{Q}_{j,\ell}\mathbf{W}_{k,j}^{**}) \ \forall \ell$ (i.e., not using more power than $\mathbf{W}_{k,j}^{**}$) and $\mathbf{h}_{i,j}^H \mathbf{V} \mathbf{h}_{i,j} \leq \mathbf{h}_{i,j}^H \mathbf{W}_{k,j}^{**} \mathbf{h}_{i,j} \ \forall i \neq k$ (i.e., not causing more interference than $\mathbf{W}_{k,j}^{**}$). One solution is $\mathbf{V} = \mathbf{W}_{k,j}^{**}$, but [10, Lemma 3] shows that problems of this form always have rank-one solutions.

Proof of Corollary 1. For convenience, let $\mathbf{A}_k = \frac{1}{\sigma_k^2}\text{diag}(\frac{1}{\rho_0}\mathbf{h}_{k,0}\mathbf{h}_{k,0}^H, \ldots, \frac{1}{\rho_S}\mathbf{h}_{k,S}\mathbf{h}_{k,S})$ be a block-diagonal matrix and $\mathbf{w}_k = [\sqrt{\rho_0}\mathbf{w}_{k,0}^T \ \cdots \ \sqrt{\rho_S}\mathbf{w}_{k,S}^T]^T$ be the aggregate beamforming vectors. Furthermore, let $\tilde{\mathbf{Q}}_{j,\ell}$ be the block-diagonal matrix that makes $\mathbf{w}_k^H \tilde{\mathbf{Q}}_{j,\ell} \mathbf{w}_k = \mathbf{w}_{k,j}^H \mathbf{Q}_{j,\ell} \mathbf{w}_{k,j}$.

Suppose $\mathbf{w}_k^* = \sqrt{p_k}\mathbf{u}_k$ is the optimal solution to (7), where $\mathbf{u}_k$ is unit-norm. By the uplink-downlink duality in [10, Lemma 4], we have

$$\tilde{\gamma}_k = \frac{p_k \mathbf{u}_k^H \mathbf{A}_k \mathbf{u}_k}{\sum_{i \neq k} p_i \mathbf{u}_i^H \mathbf{A}_k \mathbf{u}_i + 1} = \frac{\lambda_k \mathbf{u}_k^H \mathbf{A}_k \mathbf{u}_k}{\mathbf{u}_k^H \mathbf{B}_k \mathbf{u}_k} \tag{10}$$

where $\mathbf{B}_k = \left(\sum_{i \neq k} \lambda_i \mathbf{A}_i + \sum_{j,\ell} \mu_{j,\ell} \tilde{\mathbf{Q}}_{j,\ell} + \mathbf{I}\right)$ and $\lambda_k, \mu_{j,\ell}$ are the optimal Lagrange multipliers for the QoS and power constraints, respectively. The last expression in (10), the uplink SINR, takes its largest value when $\mathbf{u}_k$ is the dominating eigenvector of $\mathbf{B}_k^{-1/2} \mathbf{A}_k \mathbf{B}_k^{-1/2}$. Since $\mathbf{B}_k$ and $\mathbf{A}_k$ are block-diagonal, the dominating eigenvalue originates from one of the blocks and the corresponding eigenvector is only non-zero for this block. As each block corresponds to either the BS or one of the SCAs, this means that we ideally should serve user k by only one transmitter. The only reason to have another $\mathbf{u}_k$ is when there is multiplicity in the dominating eigenvalue and none of the single-transmitter solutions are supported by the power constraints; that is, when at least one power constraint is active. This proves the three cases stated in the corollary.

References

[1] F. Rusek, D. Persson, B. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, 2013.

[2] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, 2013.

[3] S. Parkvall, E. Dahlman, G. Jöngren, S. Landström, and L. Lindbom, "Heterogeneous network deployments in LTE – the soft-cell approach," *Ericsson Review*, no. 2, 2011.

[4] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Technol. Mag.*, vol. 6, no. 1, pp. 37–43, 2011.

[5] S. Cui, A. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2349–2360, 2005.

[6] G. Auer and *et al.*, *D2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown*. INFSO-ICT-247733 EARTH, ver. 2.0, 2012.

[7] D. Ng, E. 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, 2012.

[8] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Foundations and Trends in Communications and Information Theory*, vol. 9, no. 2-3, pp. 113–381, 2013.

[9] H. Holma and A. Toskala, *LTE Advanced: 3GPP Solution for IMT-Advanced*, 1st ed. Wiley, 2012.

[10] E. Björnson, N. Jaldén, M. Bengtsson, and B. Ottersten, "Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6086–6101, 2011.

[11] R. Kumar and J. Gurugubelli, "How green the LTE technology can be?" in *Proc. Wireless VITAE*, 2011.

[12] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming," in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed. CRC Press, 2001.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[14] J. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999.

[15] M. Bengtsson, "Jointly optimal downlink beamforming and base station assignment," in *Proc. IEEE ICASSP*, 2001, pp. 2961–2964.

[16] *Further advancements for E-UTRA physical layer aspects (Release 9)*. 3GPP TS 36.814, Mar. 2010.

[17] CVX Research Inc., "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, 2012.