Energy Optimization for Full-Duplex Self-Backhauled HetNet with Non-Orthogonal Multiple Access

Lei Lei[1], Eva Lagunas[1], Sina Maleki[1], Qing He[2], Symeon Chatzinotas[1], and Björn Ottersten[1]

[1]Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg
[2]Department of Science and Technology, Linköping University, Sweden
Emails: {*lei.lei; eva.lagunas; sina.maleki; symeon.chatzinotas; bjorn.ottersten@uni.lu; qing.he@liu.se*}

Abstract—**Small cell densification and advanced multi-user access schemes are promising approaches to dramatically improve 5G system performance. Towards efficient spectrum usage in ultra-dense heterogeneous networks, spectrum reuse between backhauling and access links combined with full duplex is applied. This forms a full-duplex self-backhauled heterogeneous network (FS-HetNet). Considering co-channel interference caused by frequency reuse, and residual self interference due to imperfect interference cancellation in full duplex, interference management becomes a major issue in boosting network performance. In this paper, motivated by the emerging non-orthogonal multiple access (NOMA) for 5G, we consider a NOMA-based scheme to mitigate co-channel interference and achieve efficient spectrum utilization for FS-HetNet. We address an energy-saving problem for the considered network, aiming to satisfy all users' data demand within a limited transmit duration by consuming minimum energy. In addition to the energy consumption in transmission, the consumed decoding energy due to signal processing in successive interference cancellation is also taken into account. We propose an energy-efficient and delay-constrained scheduling algorithm to jointly optimize transmit power, user clustering, and transmission duration. Numerical results demonstrate that the proposed approach outperforms previous schemes.**

Index Terms—**Non-orthogonal multiple access, wireless backhaul, full duplex, energy minimization, user clustering, heterogeneous network.**

I. INTRODUCTION

In ultra-dense heterogeneous networks, wireless backhauling for small cells, as an alternative to fiber, is an effective solution to reduce the cost and capital expenditure in network deployment. In this scenario, backhaul links for densely deployed small-cell base stations (SBSs) are wirelessly provided by macro base stations (MBSs), and user equipments (UEs) are served by the SBSs. Due to the scarcity of the frequency resources, the backhaul and UEs' access links are designed to reuse the same channel. This type of the network with in-band self-backhauling for base stations (BSs) and with capabilities of full-duplex and self-interference suppression, is known as a full-duplex self-backhauled heterogeneous network (FS-HetNet), and has attracted considerable research attention [1]–[3]. Compared to half-duplex operations, the results in [1]–[3] have shown the advantages of FS-HetNet in spectrum usage and performance improvement. Practically, imperfect

self-interference cancellation in full duplex and co-channel deployment will result in coexistence of cross-tier, co-tier, and self interference [1]. Performance is thus limited by such mutual-influenced interference which also brings bottlenecks and challenges in achieving maximum performance gains. Interference management is then addressed, mainly from two aspects, power control and interference cancellation.

To improve spectrum efficiency and mitigate co-channel interference, power-domain non-orthogonal multiple access (NOMA), has been proposed for 5G systems [4], [5]. In downlink, NOMA is enabled by superposition coding at the transmitter, and relying on multi-user detection (MUD) and successive interference cancellation (SIC) at the receivers to remove some of the co-channel interference [6]. Many works, e.g., [4], [7], [8], have shown that NOMA can improve performance in terms of throughput, fairness, power, and outage probability.

In this paper, we aim to minimize network energy consumption via optimizing power control and exploiting synergies of NOMA and SIC. Motivations and contributions are summarized below: 1) we investigate a NOMA based scheme for FS-HetNet to jointly optimize the energy consumption in data transmission and in SIC decoding operations, which has not been fully investigated in the literature. 2) in previous works, the UE grouping/clustering schemes in NOMA are typically heuristic, e.g., randomly grouping UEs to a channel, paring two UEs with best-worst channel gains [4], [5]. The performance of NOMA largely depends on the decisions of which UEs are scheduled to the same channel simultaneously and their power allocation. Therefore, in this work, based on power optimization, we investigate optimal UE-grouping schemes for NOMA. 3) we formulate an energy-efficient and delay-constrained scheduling problem for FS-HetNet. We characterize several bottlenecks in limiting performance and investigate trade-offs in power control and energy-efficient scheduling. The derived analyses are served as the basis for the algorithm development in this work. 4) although it is intuitive to observe that application of NOMA can mitigate interference and maintain efficient spectrum usage, it is not immediately clear how much energy-saving gain can be expected, and how this performance gain varies in different FS-HetNet

instances. We provide algorithmic solutions to jointly optimize power control in base stations, UE grouping in NOMA, and transmission time in scheduling, and further we use numerical results to answer these questions.

II. System Model

We consider downlink transmission in a two-tier FS-HetNet. As an illustration, a network layout is shown in Fig. 1, where an MBS with access to the core network provides wireless backhaul service to one or multiple SBSs, and each SBS directly provides data service to its associated K UEs. The set of UEs $1, \ldots, K$ is denoted as $\mathcal{K}$. The deployed SBS
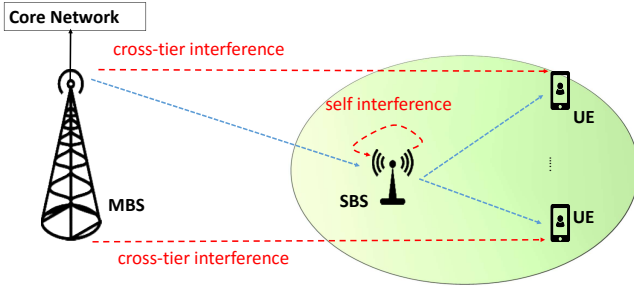

Figure 1. An illustration: network layout and interference environment of the considered two-tier FS-HetNet

is within the coverage area of the MBS, but the UEs may be outside of the MBS's coverage. We consider all the UEs' traffic to be routed through the SBS in order to provide better service for them. The link between an MBS and an SBS is referred to as a backhaul link, and the link between an SBS and a UE is referred to as an access link. With in-band self-backhauling, the access links and the backhaul for the same SBS share a common frequency channel to achieve efficient spectrum usage. In this work, for simplicity, we focus on a simple MBS-SBS-UEs scenario with one MBS, one SBS, and multiple UEs. However, we remark that our analysis and derived results are applicable to multiple MBSs and SBSs. For example, if frequency resources are orthogonally allocated among in-band self-backhauled SBSs, there is no interference in present among the SBSs, then the optimization task for the whole network can be divided to multiple independent MBS-SBS-UEs networks. Each of them is corresponding to the considered scenario in this work.

At the backhaul link in Fig. 1, the deployed self-backhauled SBS can operate at full-duplex mode, thus data transmission to the UEs and data received from the MBS can be done simultaneously in the same channel. Practically, we consider that the SBS has imperfect full-duplex operations, that is, due to imperfect self-interference cancellation, we introduce a coefficient $0 \leq \beta \leq 1$ to reflect the SBS's ability in suppressing its self interference. For $\beta = 1$, it means the SBS is not able to eliminate any self interference and $\beta = 0$ indicates residual self interference is completely removed. The

signal-to-interference-plus-noise ratio (SINR) at the SBS after self-interference cancellation can be modeled as [2],

$$\text{SINR}_s = \frac{P_m |h_{ms}|^2}{\beta P_s |h_{ss}|^2 + \sigma^2} \tag{1}$$

where P_m is the transmit power of the MBS, h_{ss} is the channel coefficient for the self-interference channel, and σ^2 is the power of the noise. Also, we use h_{ms}, h_{sk}, and h_{mk} to denote the channel coefficients for the links MBS→SBS, SBS→UE k, and MBS→UE k, respectively, which are complex Gaussian random variables with zero mean and unit variance.

At the access links, we apply NOMA to cancel part of the co-tier and co-channel interference among UEs. Multiple UEs can be simultaneously scheduled to the same channel for data transmission. We use "cluster" c to represent a set of UEs, denoted by $\mathcal{K}_c$. It can also be referred to as UE groups/sets presented in other works. In total there are $2^K - 1$ possible clusters which can be selectively and sequentially scheduled to deliver UEs' demand. Let $c \in \mathcal{C}$ index the cth cluster, and $\mathcal{C}$ is the set of all candidate clusters.

During the transmission period of a cluster c, say $\mathcal{K}_c = \{1, 2\}$ for example, the MSB, by applying superposition coding, transmits a superposed signal x_s to the SBS, mixed with information symbols x_1 and x_2 for UE 1 and 2, respectively,

$$x_s = \sqrt{P_m \alpha_1^c} x_1 + \sqrt{P_m \alpha_2^c} x_2 \tag{2}$$

where $\alpha_1^c + \alpha_2^c = 1$, and in general α_k^c, $k \in \mathcal{K}_c$, is the power allocation factor of UE k in cluster c. Each UE is able to perform MUD and SIC, in order to remove part of the co-channel interference. Note that the information segment containing the superposed signal x_s is received twice at each UE, first as a weak signal over the MBS→UE link, second as a strong signal forwarded from the SBS to the UE. If no signal combining is performed or the weak signal is undecodable or undesirable for the UEs, the first weak signal simply causes cross-tier interference [1]. If the two signals are cooperatively combined, performance can be improved. To mitigate the complexity of signal processing at the UE side, we consider no cooperative communications at the UEs, and treat the first weak signal from the MBS→UE link as noise. According to the basis of NOMA [4]–[6], the decoding order in cluster c is predefined as the descending order of $\frac{|h_{sk}|^2}{P_m |h_{mk}|^2 + \sigma^2}$ [4], [5]. We use $b_c(k)$ to indicate the position of UE $k \in \mathcal{K}_c$ in the sorted sequence in cluster c. Thus, UE k in cluster c before decoding its desired signal x_k, first decodes and subtracts the received signals $x_{j'}$ in $\{j' \in \mathcal{K}_c \backslash \{k\} : b_c(j') > b_c(k)\}$. UE k treats other signals x_j with $\{j \in \mathcal{K}_n \backslash \{k\} : b_c(j) < b_c(k)\}$ as noise. The SINR for UE k in cluster c is given as [2],

$$\text{SINR}_k^c = \frac{\alpha_k^c P_s |h_{sk}|^2}{\sum\limits_{\substack{j \in \mathcal{K}_c \backslash \{k\}: \\ b_c(j) < b_c(k)}} \alpha_j^c P_s |h_{sk}|^2 + P_m |h_{mk}|^2 + \sigma^2} \tag{3}$$

where P_s is the transmit power of the SBS.

Once a cluster is scheduled, we also take decoding energy into account. Decoding energy consumption is incurred at the

UEs during the SIC decoding attempts. We adopt a linear model used in [9] to quantify the decoding energy for a cluster c. Each decoding operation consumes a fixed value of power P_{fix}^d on the circuit. Let $P_c = P_{fix}^d A_c$ denote the total decoding power for a cluster c, where A_c indicates the total number of decoding attempts in cluster c. For scheduling a cluster c, its decoding energy is proportional to P_{fix}^d, t_c, and A_c. The cumulative energy consumption for SIC is non-negligible when the MBS and the SBS operate at low power levels, and a large number of UEs need to be served and more decoding operations have to be processed.

III. PROBLEM FORMULATION FOR ENERGY-EFFICIENT SCHEDULING

We aim to provide optimal solutions for performing energy-efficient scheduling in FS-HetNet such that all UEs' data requests are satisfied within a limited transmission time. The optimization task consists of two parts: power control and cluster scheduling. The former is to decrease and balance the transmit power P_m and P_s at the MBS and the SBS. The latter is to reduce the total scheduling time by optimizing which clusters should be scheduled, and how long of their individual duration. To this end, we use two sets of optimization variables, power variables P_m, P_s and time variables t_c, $\forall c \in \mathcal{C}$. Note that transmit power P_s is uniform over all clusters. In this work, we focus on investigating optimal clustering solutions, thus the set $\mathcal{C}$ has to enumerate all possible clusters as an input for optimization. This can lead to a large size of $\mathcal{C}$. In practice, some criteria or heuristic methods can be applied to largely reduce the number of candidate clusters and control the size of $\mathcal{C}$ in a manageable scale [10]. Each cluster c is associated with a variable t_c to indicate its transmission duration, and decoding power P_c is precalculated for each cluster.

$$\text{P1:} \quad \min_{P_m,P_s,t_c} \ (P_m + P_s)\sum_{c\in\mathcal{C}} t_c + \sum_{c\in\mathcal{C}} P_c t_c \tag{4a}$$

$$\text{s.t.} \ \sum_{c\in\mathcal{C}} t_c \log(1+ \frac{\alpha_k^c P_s |h_{sk}|^2}{\sum\limits_{\substack{j\in\mathcal{K}_c\setminus\{k\}:\\ b_c(j)<b_c(k)}} \alpha_j^c P_s|h_{sk}|^2 + P_m|h_{mk}|^2+\sigma^2})$$
$$\geq D_k, \ \forall k \in \mathcal{K} \tag{4b}$$

$$\sum_{c\in\mathcal{C}} t_c \leq T \tag{4c}$$

$$\sum_{k\in\mathcal{K}_c} \log(1+ \frac{\alpha_k^c P_s |h_{sk}|^2}{\sum\limits_{\substack{j\in\mathcal{K}_c\setminus\{k\}:\\ b_c(j)<b_c(k)}} \alpha_j^c P_s|h_{sk}|^2 + P_m|h_{mk}|^2+\sigma^2})$$
$$\leq \log(1+\frac{P_m|h_{ms}|^2}{\beta P_s|h_{ss}|^2+\sigma^2}), \ \forall c\in\mathcal{C} \tag{4d}$$

$$t_c \geq 0, \ \forall c \in \mathcal{C} \tag{4e}$$
$$0 \leq P_s \leq P_s^{\max} \tag{4f}$$
$$0 \leq P_m \leq P_m^{\max} \tag{4g}$$

We formulate the energy-efficient scheduling problem in P1. Objective (4a) is to minimize the total energy in scheduling.

Once a cluster c is scheduled with duration t_c, an amount of energy $(P_m+P_s)t_c$ is consumed in data transmission, and $P_c t_c$ stands for the consumed decoding energy in SIC processing. Constraints (4b) ensure that each UE's data demand D_k in bits is delivered. It implies that the data traffic for the same UE can be flexibly transmitted in multiple clusters. Constraints (4c) restrict that all the data transmission for delivering UEs' demand has to be done within a time limit T to maintain an appropriate level of quality of service (QoS). In constraints (4d), since the capacity of the wireless backhaul is limited, thus when a cluster is scheduled, the aggregated rate of the access links should be no larger than the rate of backhauling. Constraints (4e) to (4g) confine the boundary of the feasible region for the variables. Power P_m and P_s are constrained to their maximum power limits P_m^{max} and P_s^{max} at the MBS and the SBS, respectively. Typically $P_m > P_s$, however, in this work, we do not impose this constraint in optimization since P_m can be seen as a portion of the total transmit power of the MBS allocated to serve a specific SBS if the MBS has to serve its own mobile UEs and other SBSs. For this case, P_m may not necessarily dominate the performance.

IV. COMPLEXITY CHARACTERIZATION AND ALGORITHMIC SOLUTION

To develop an algorithm to solve P1, in this section we investigate the problem's structure and the complexity first, then we propose the algorithmic solution.

A. Complexity and Property Characterizations

P1 is a non-linear problem, however, as an intuitive observation, this non-linearity can be immediately removed when P_m and P_s are fixed. Based on this fact, we derive the following analysis. Let $z_c \in \{0,1\}, \forall c \in \mathcal{C}$ be binary indicators to show whether cluster c is scheduled with positive t_c.

Lemma 1. *At the optimum of P1, $\sum\limits_{c\in\mathcal{C}} z_c \leq K+1$.*

Proof: Once the power variables P_m and P_s are fixed, the remaining problem is equivalent to solving the following linear programming (LP) problem P2.

$$\text{P2:} \quad \min_{t_c} \ (P_m + P_s)\sum_{c\in\mathcal{C}} t_c + \sum_{c\in\mathcal{C}} P_c t_c \tag{5a}$$

$$\text{s.t.}\sum_{c\in\mathcal{C}} t_c \log(1+\frac{\alpha_k^c P_s |h_{sk}|^2}{\sum\limits_{\substack{j\in\mathcal{K}_c\setminus\{k\}:\\ b_c(j)<b_c(k)}}\alpha_j^c P_s|h_{sk}|^2+P_m|h_{mk}|^2+\sigma^2})\geq D_k, \forall k\in\mathcal{K} \tag{5b}$$

$$\sum_{c\in\mathcal{C}} t_c \leq T \tag{5c}$$

Given P_m and P_s, the throughput for both access and backhaul links has been known. For some clusters, constraints (4d) may not be satisfied. Then these infeasible clusters will be excluded before the optimization procedure. At the optimum, the equalities in (5b) and (5c) hold, then in fact P2 is in an LP standard form. By applying the fundamental optimality theory of LP [11], [12], we can conclude that at most $K+1$

clusters will be used at the optimum of P2. The rationale is that, for a basic feasible solution in LP, the number of variables in the base matrix equals the number of constraints. Then at an optimal basic solution, the number of positive t_c is at most $K+1$. The result holds for any given feasible P_m and P_s, including the feasible optimal power of P1. Hence the conclusion. ∎

From P2, we can observe that the optimization problem is still in the domain of classical scheduling problems which are in general hard to solve [12]. To further investigate P2's complexity, one can construct a special instance that the transmit power dominates the energy consumption in the objective, and the decoding power is negligible such that it has no influence on determining cluster selections. Then approximately, determining no more than $K+1$ clusters to be scheduled and their duration is reduced to solving a minimum-time scheduling problem (MTSP): "$\min_{t_c} \sum_{c \in \mathcal{C}} t_c$, s.t. (5b)". Note that in this case, the constraint (5c) is redundant. Violation of (5c) directly identify the infeasibility of the problem. In the complexity proof of [12], the MTSP has been proved to be NP-hard by constructing a polynomial reduction from the weighted fractional coloring problem which is NP-complete. We also observe that, to optimally solve P1, searching optimal power P_m and P_s for P1 and deciding optimal scheduling in P2 is typically not easier than solely solving P2, Thus optimally solving P1 is inherently difficult in general.

B. Algorithm Development

To optimally solve P1, there are several aspects need to be considered in the optimization. The first aspect is the problem's non-linearity. As a result, it is difficult to optimize the power and time variables jointly. Our previous analysis enables a possible and reasonable solution to handle this issue, that is, one can iteratively search power, and at each iteration, solve the corresponding LP for the fixed power. The second aspect is that, in general, the optimal transmit power for the MBS and the SBS may not be immediately observed due to the mutual influence and the dependence between P_m and P_s. As can be verified in (4d), when we increase P_m or decrease P_s to enhance the backhauling rate, an opposite effect is always imposed to the access links. Also, since the sum rate of the access links is constrained by the backhaul capacity, P_m and P_s will depend on each other.

Next, we proposed an NOMA based energy-efficient scheduling (NES) algorithmic framework to solve P1. The operations of NES are presented in Algorithm 1. NES consists of two components, power control and cluster scheduling. As an outer loop of NES, a systematic search for P_m and P_s is iteratively carried out in Line 1. This can be done by two methods: either continuous or discrete. The former applies dual-bisection method to search P_m and P_s, which enables optimal power solutions for P1. The latter adjusts power values from a finite set of discrete power levels. This is practically relevant to realistic systems, and leads to an approximated solution to the optimum of P1.

Algorithm 1 NES for P1

Input: K, T, β, D_k, $\mathcal{C}$, ϵ, $\mathcal{E}^* = \emptyset$, P_c, $\forall c \in \mathcal{C}$
Output: E^*

1: **while** not converge, then iteratively optimize P_m and P_s **do**
2:    **for** $c = 1, \ldots, C$ **do**
3:      with P_m, sort UEs and determine the decoding order in cluster c
4:      with P_s, calculate α_k^c for all UEs in cluster c
5:      **if** $\log(1 + \text{SINR}_s) < \sum_{k \in \mathcal{K}_c} \log(1 + \text{SINR}_k^c)$ **then**
6:        $\mathcal{C} = \mathcal{C} \backslash \{c\}$
7:    Solve P2 and obtain the optimal solution $t_1^*, \ldots, t_C^*$
8:    **if** P2 is feasible **then**
9:      $\mathcal{E}^* \leftarrow (P_m + P_s) \sum_{c \in \mathcal{C}} t_c^* + \sum_{c \in \mathcal{C}} P_c t_c^*$
10:    **else**
11:      adjust and update P_m or P_s
12: **until** no improvement is larger than tolerance ϵ
13: $E^* \leftarrow \min \mathcal{E}^*$

For cluster scheduling from Line 2 to 11, all clusters are pre-processed to provide inputs for solving P2. In Line 4 for allocating power P_s to the UEs within a cluster, with practical considerations, several schemes from literature can be used to decide coefficients α_k^c, for example, in [13], to mitigate the implementation complexity in deciding UEs' power allocation, power coefficients are simply selected from multiple predefined discrete sets. Then optimal UE clustering under the different power allocation schemes can be addressed by solving P2. In Line 6, if the sum rate of users in a cluster exceeds the backhaul capacity, the cluster will not be considered in P2. Note that this excluded cluster may lead to good performance, but it is invalid under the current power allocation. In the later iterations, by optimizing power P_s and P_m, this invalid cluster may become feasible and thus can be reconsidered in the optimization procedure. In Line 7, P2 can be efficiently solved in general by applying standard optimization tools. Algorithm terminates when the energy consumption between iterations is less than a predefined tolerance ϵ. In the end of NES, the minimum energy E^* over all iterations is obtained.

V. PERFORMANCE EVALUATION

In this section, we present numerical results to illustrate the performance of the proposed solutions. There are $K = 10$ UEs randomly and uniformly distributed in the coverage area of the SBS. Channel models for path loss, shadowing, and fast fading follow ITU Urban macrocell model (UMa) and Urban microcell model (UMi) [13]. For power coefficients α_k^c, we adopt the method "fractional transmit power allocation" used in [4], and scale a parameter $0 \le \gamma \le 1$ to optimize the power allocation among the UEs of each cluster, where $\gamma = 0$ enables equal power allocation, and increasing γ results in more power allocated to the UE with poor channel condition. The maximum transmit power is limited by $P_m^{max} = 40$ dBm and $P_s^{max} = 30$ dBm. In performance comparisons, NES is

used to provide the optimal solutions for P1. The results are averaged over 1000 channel realizations.
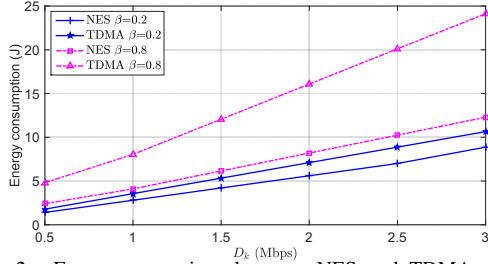

Figure 2. Energy comparison between NES and TDMA with respect to demand and residual self interference, $\gamma = 0.5$.

First in Fig. 2, we set the time limit T sufficiently large such that no infeasible cases in the simulation, then we evaluate the consumed energy between NES and an orthogonal scheme time division multiple access (TDMA) with successively increasing UEs' uniform demand D_k. In simulations, optimal TDMA can be simply implemented by solving P1 with using K single-UE clusters only. From the results, the energy, i.e., the objective of P1, for both NES and TDMA grows in D_k. Note that if T becomes tight, TDMA can be infeasible due to the lack of diversity in cluster selections. Also, the energy in NES will increase steeply with D_k. NOMA based scheme NES consistently consumes lower energy than TDMA in particular for the high demands, with approximately 20% to 30% performance improvement in the case of $\beta = 0.2$. Also, It can be observed that larger residual self interference, $\beta = 0.8$, results in higher energy consumption in both schemes, and leads to larger performance gaps, around 30% to 40%.
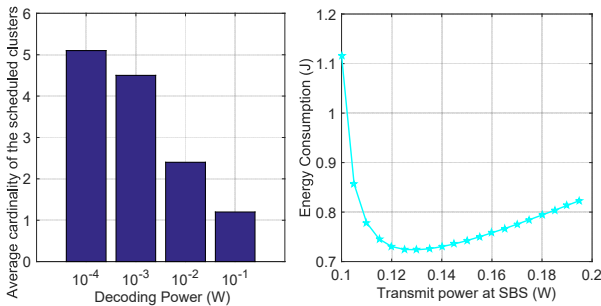

Figure 3. The results of UE clustering (left) and the monotonicity of energy in P_s (right).

Next, in Fig. 3, we reveal the impacts of decoding power to optimal cluster selections and transmit power P_s to energy consumption. For the former, when the consumed power per decoding operation increases, the smaller-size clusters tend to be favorable to the optimum. When the cumulative decoding power completely dominates, TDMA is optimal if the time limit T is not exceeded. This result is consistence with the conclusion in [10]. In the right-side figure, it verifies that with a fixed P_m, e.g., 30 dBm used in simulations, the monotonicity of the resulting energy in P_s is not fixed. When P_s increases, once the aggregated rate of the access links exceeds the backhaul capacity, the problem can become infeasible, then P_m, in its turn, needs to be optimized. This fact also verifies

the necessity of the developed algorithm for searching the optimal transmit power for P1.

VI. CONCLUSIONS

We considered reducing energy consumption in self-backhauled heterogeneous networks with full-duplex capabilities. A NOMA based scheme is applied to mitigate co-channel interference at the access links. We formulated an energy-efficient scheduling problem with a set of QoS constraints. We then developed an algorithmic framework, NES, to provide optimal solutions for power control and cluster scheduling. Numerical results showed that NOMA based NES is able to significantly reduce energy compared to TDMA based scheduling, particularly when the residual self interference is high.

VII. ACKNOWLEDGMENTS

REFERENCES

[1] R. A. Pitaval, O. Tirkkonen, R. Wichman, K. Pajukoski, E. Lahetkangas and E. Tiirola, "Full-duplex self-backhauling for small-cell 5G networks," in *IEEE Wireless Communications*, vol. 22, no. 5, pp. 83-89, Oct. 2015.

[2] Y. Liu, L. Lu, G. Y. Li, Q. Cui and W. Han, "Joint User Association and Spectrum Allocation for Small Cell Networks With Wireless Backhauls," in *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 496-499, Oct. 2016.

[3] Z. Zhang, X. Chai, K. Long, A. V. Vasilakos and L. Hanzo, "Full duplex techniques for 5G networks: self-interference cancellation, protocol design, and relay selection," in *IEEE Communications Magazine*, vol. 53, no. 5, pp. 128-137, May 2015.

[4] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *IEEE PIMRC*, Sept. 2013, pp. 611–615.

[5] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in lte and 5G networks," in *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[6] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," in *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, Sept. 2015.

[7] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," in *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8580- 8594, Dec. 2016.

[8] L. Lei, D. Yuan, and P. Vrbrand, "On power minimization for non-orthogonal multiple access (NOMA)," in *IEEE Communications Letters*, vol. 20, no. 12, pp. 2458–2461, Dec. 2016.

[9] G. Geraci, M. Wildemeersch and T. Q. S. Quek, "Energy Efficiency of Distributed Signal Processing in Wireless Networks: A Cross-Layer Analysis," in *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 1034-1047, Feb. 2016.

[10] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Optimal cell clustering and activation for energy saving in load-coupled wireless networks," in *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6150–6163, Nov. 2015.

[11] K. Murty, *Linear programming*, Wiley, 1983.

[12] V. Angelakis, A. Ephremides, Q. He and D. Yuan, "Minimum-Time Link Scheduling for Emptying Wireless Systems: Solution Characterization and Algorithmic Framework," in *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1083-1100, Feb. 2014.

[13] 3GPP, R1-153332, "Evaluation methodologies for downlink multiuser superposition transmissions, May 2015.