





Radio & BH Load-Aware Multi-Objective Clustering in Multi-Cell MIMO Cooperative Networks

Selcuk Bassoy , Mona Jaber , *Member, IEEE*, Oluwakayode Onireti , *Senior Member, IEEE*, and Muhammad A. Imran , *Senior Member, IEEE*

Abstract—Coordinated Multipoint (CoMP) is one of the key technologies identified for future wireless networks to mitigate inter-cell interference, especially in a dense deployment scenario. However, CoMP can't be realized for the whole network due to its computational complexity, synchronization between coordinating base stations (BSs) and high backhaul (BH) capacity requirement. BSs need to be clustered into smaller groups and CoMP can be activated within these smaller clusters. In this paper, we develop a multi-objective, dynamic clustering model for multi-user, joint-transmission CoMP to jointly optimize spectral efficiency (SE), radio access network (RAN) load and BH load. We formulate our load-aware model as two coalitional sub-games for small cell and user equipment clustering, respectively. Merge/split/transfer actions for each sub-game are defined and a complexity and stability analysis is provided. Extensive simulation results show that our model provides as good SE in low load when compared to a greedy model, and significantly better load balancing with a reduced number of unsatisfied users and increased throughput in high load scenario. On average 49% increase in the overall system throughput is observed in our simulations when compared to the greedy model.

Index Terms—Backhaul-aware, coalitional sub-games, coordinated multipoint system, load-aware, multi-objective clustering.

I. INTRODUCTION

THE fifth generation (5G) cellular systems are being deployed aiming at 1000 times more capacity than the fourth generation (4G) to cope with increasing mobile data traffic [1]. Interference mitigation plays an important role in improving the much needed overall capacity, especially in highly interference-limited 5G dense deployment scenarios [2].

Manuscript received April 3, 2020; revised August 21, 2020 and January 15, 2021; accepted March 15, 2021. Date of publication April 5, 2021; date of current version June 9, 2021. This work was supported by EPSRC Global Challenges Research Fund the DARE Project under Grant EP/P028764/1. The review of this article was coordinated by Dr. S. Misra. (*Corresponding author: Selcuk Bassoy.*)

Selcuk Bassoy is with Telefonica U.K. Ltd, Slough SL1 4DX, U.K. (e-mail: selcuk.bassoy@telefonica.com).

Mona Jaber is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4FZ London, U.K. (e-mail: m.jaber@qmul.ac.uk).

Oluwakayode Onireti is with the James Watt School of Engineering, University of Glasgow, G12 8QQ Glasgow, U.K. (e-mail: Oluwakayode.Onireti@glasgow.ac.uk).

Muhammad A. Imran is with the James Watt School of Engineering, University of Glasgow, G12 8QQ Glasgow, U.K., and also with Artificial Intelligence Research Center, College of Engineering and Information Technology, Ajman University, Ajman, United Arab Emirates (e-mail: Muhammad.Imran@glasgow.ac.uk).

Digital Object Identifier 10.1109/TVT.2021.3070992

Coordinated Multipoint (CoMP) is identified as a promising interference mitigation technique in which multiple base stations (BSs) cooperate for joint transmission/reception. This is achieved by exchanging user/control data, thus, realizing joint signal processing which mitigates inter-cell interference and even exploits it as a useful signal. CoMP is already a key feature of long term evolution-advanced (LTE-A) [3] and is an essential function for 5G [1], [4]. 5G test-bed results from Qualcomm demonstrate the ability of CoMP to increase capacity by exploiting spatial multiplexing and to provide ultra-reliable connectivity by exploiting spatial diversity (i.e., transmitting the same data from each transmission point) [5]. Furthermore, new network architectures such as centralized radio access network (C-RAN) [6] and ultra-dense small cell (SC) networks facilitate the deployment of CoMP and enhance its benefits [7]. However, coordination among a high number of BSs necessitates high capacity, low latency backhaul (BH) links for sharing the required signaling and user data. On the other hand, multi BSs coordination requires the computation of precoding matrices which get larger as the number of BSs increases. Moreover, channel estimation relies on pilot channels and the resulting overhead also increases as the number of coordinating BSs increases [8], [9]. Due to these bottlenecks, CoMP is only feasible within small BS clusters which limits the potential gain. Consequently, BSs need to be intelligently grouped into small clusters within which CoMP can be operational while the gain is maximized.

A. Literature Review

The problem of network clustering to maximize CoMP efficiency has been extensively studied in the literature [10]. A comprehensive CoMP clustering solution needs to jointly optimize multiple key objectives, e.g., spectral efficiency (SE), RAN load and BH availability. However, most of the current works adopt SE as a single primary objective with network-centric clustering solutions [11], [12]. User-centric solutions are proposed in [13]–[15] with double objectives: SE and throughput at the cell edge, however do not consider BH limitations or other objectives. There are other solutions in the literature which optimize RAN load and BH availability, however, these objectives are studied in isolation, lacking a comprehensive multi-objective clustering approach.

BH capacity and latency are some of the biggest challenges for the realization of CoMP in future networks [16], [17]. The

impact of BH limitations, clock synchronization, and imperfect channel state information (CSI) on CoMP performance are evaluated in [18]. The resulting field tests show a significant impact on the achievable SE under these conditions. Realistic network clustering solutions will need to take BH availability into consideration for network clustering to maximize CoMP gain. Required BH capacity is taken as one of the key objectives in [19] where soft frequency reuse (SFR) and CoMP are employed together to improve cell edge user performance. An analytical framework is driven to optimize SFR parameters to maximize the overall cluster capacity and cell edge user throughput while minimizing the required BH capacity. In [20], the feasibility of deploying coordinated scheduling CoMP (CS-CoMP) under different BH infrastructures is analyzed in terms of convergence delay when exchanging scheduling information between SCs. The same authors further enhance this work in [21] and propose a bandwidth allocation scheme to prioritize inter-SC (X2) traffic for CS-CoMP and, hence, reduce scheduling information exchange latency in a BH limited 5G network. Limited fronthaul availability is studied in [22] for C-RAN architecture where user-centric clusters of remote radio heads are optimized to minimize the total transmission power while maintaining user's quality of service (QoS). More recently, limited BH capacity and per-BS power constraints are taken into account to optimize user-centric clusters and design transmit precoding for maximizing the sum rate in [23]. Both of these works present a user-centric clustering model but the proposed methods rely on high precoding complexity and tight BS synchronization.

Emerging mobile edge computing (MEC) and popular data caching at the BS is a promising concept for reducing the CoMP related BH requirements [24]. Caching data on the MEC servers eliminates the need for transmitting popular data from the core network over the BH. Consequently, during high load traffic, the BH capacity is available to support CoMP without compromising latency. In [25], authors propose to utilize user data caching at the BS to reduce BH load and improve CSI knowledge accuracy with improved BH availability. In [26], all BSs in the same cluster aim to cache identical data at BS, and an opportunistic joint transmission (JT) CoMP is employed for users where user-data is available at each BS. Otherwise, coordinated beamforming (CB) CoMP is employed where only CSI is shared between BSs for joint precoding. A number of studies in the literature utilize cached data at the BS to optimize user-centric CoMP clusters to reduce BH traffic demand in isolation [27], [28]. A further user-centric clustering is studied in [29] where cached data at SCs are utilized to form optimum user-centric clusters to reduce BH traffic and increase network throughput for a given maximum cluster size (CS). In these works, BH limitation is studied in isolation for CoMP clustering, without considering other network metrics i.e. SE, RAN load, etc. Furthermore, as pointed out earlier, precoding and synchronization complexity increases as the network size increases for user-centric clustering solutions, and hence these solutions are not scalable for larger networks. Realistic CoMP deployment will require network-centric clustering solutions to reduce these complexities and deploy user-centric solutions within each network-centric cluster to optimize gain.

RAN load is another key dependency that needs to be taken into account for CoMP clustering. CoMP is likely to be deployed in interference-limited, highly dense deployment scenarios where hotspot areas will form at certain times. CoMP clusters need to dynamically adjust to balance the load and shift traffic from highly loaded BSs to lightly-loaded BSs. In our previous work [30], we proposed a user-centric clustering algorithm where RAN load is taken into consideration for user-centric clusters. To form load-aware clusters, UEs at the cell edge of congested BSs are dynamically moved to relatively lightly-loaded BSs, thus, shifting traffic from highly loaded BSs to lightly loaded BSs. In [31], RAN load-aware user-centric clusters are formed by utilizing game theory for non-coherent CoMP in an ultra-dense heterogeneous network (HetNet) scenario. In both solutions, user-centric clusters are presented but these are not scalable for large networks due to their inherent increased complexity. To avoid the complexity of user-centric clusters, we proposed a novel, low-complexity, merge-split coalition game model to form RAN load-aware network-centric clusters in our previous work [32]. However this solution also lacks BH capacity awareness.

In this paper, we present a dynamic CoMP clustering algorithm that jointly optimizes BH load, RAN load and SE based on changing network conditions. We consider our solution as an improved mobility load balancing functionality within self-organizing networks (SON) framework [33]. SON is an important concept which aims to provide automated self-configuration, self-optimization and self-healing functions to dynamically adapt the network to changing conditions. Some SON features are currently deployed in existing networks whilst the 3rd Generation Partnership Project (3GPP) foresees a key role for SON in 5G deployments [34], [35]. Dynamic CoMP clustering function [10] is a typical application that would benefit from SON functionality, as recently shown in a 3GPP report on SON-based CoMP clustering with BH latency limitations in [36]. Similarly, the novel solution proposed in this manuscript for a multi-objective dynamic CoMP clustering algorithm could as well be deployed within a SON platform.

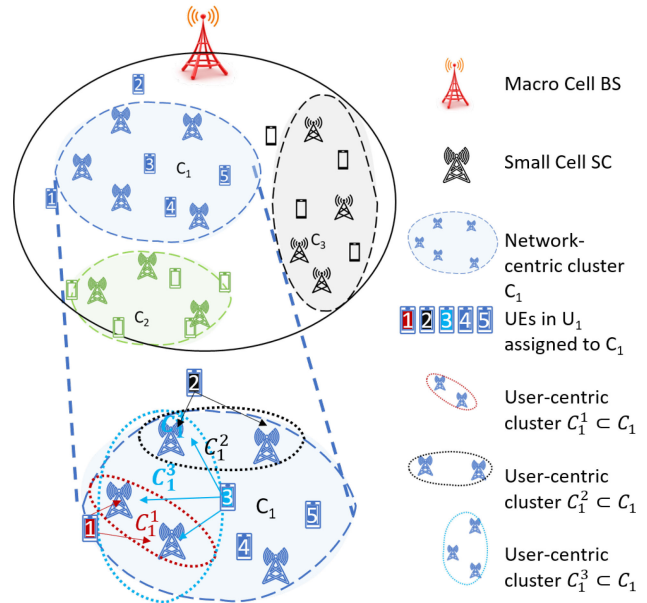
B. Contributions

BH limitation, RAN load and SE objectives have been studied for CoMP clustering but each objective studied in isolation. There is no CoMP clustering solution in the literature that jointly optimizes and analyzes the trade-off between SE and BH/RAN load. Furthermore, most of the BH-aware models utilize user-centric clustering models which are not scalable for larger networks. To this end, we propose a scheme that jointly accounts for BH/RAN load and SE objectives whilst offering a coalition game-based scalable implementation. The contributions of this paper can be summarised as follows:

- 1) Based on the authors' knowledge, this is the first scheme for the design of a comprehensive CoMP clustering framework that jointly optimizes multiple objectives, i.e. SE, RAN load and BH load. Each of these objectives has been studied in isolation, but there is no work in the literature which jointly optimizes all three objectives.

TABLE I
 TABLE OF ACRONYMS

3GPP	3rd Generation Partnership Project
AWGN	Additive White Gaussian Noise
BH	Backhaul
BS	Base Station
CB	Coordinated Beamforming
CoMP	Coordinated Multipoint
C-RAN	Centralised Radio Access Network
CS	Cluster Size
CS-CoMP	Coordinated Scheduling - CoMP
CSI	Channel State Information
EPA	Extended Pedestrian-A
GBR	Guaranteed Bit Rate
HetNet	Heterogeneous Network
HN	Homogeneous Network
JT	Joint-transmission
JT-CoMP	Joint Transmission CoMP
LBH-GA	RAN and BH load-aware game
L-GA	RAN load-aware game
LTE-A	Long Term Evolution - Advanced
MBS	Macro base station
MEC	Mobile Edge Computing
MIMO	Multiple Input Multiple Output
MU	Multi-User
NLOS	None Line of Sight
PPP	Poisson Point Process
PRB	Physical Resource Block
QoS	Quality of Service
RAN	Radio Access Network
RN	Random Network
SC	Small Cell
SE	Spectrum Efficiency
SE-GR	Greedy model employing SE based utility
SFR	Soft Frequency Reuse
SINR	Signal to Interference Noise Ratio
UE	User Equipment
VDSL2	Very high-speed Digital Subscriber Line 2


 Fig. 1. System model showing a heterogeneous network with network-centric SC clusters and corresponding UE clusters. The bottom diagram shows the formation of user-centric clusters within the network-centric cluster C_1 .

load-aware model, respectively, in the case of all SCs having limited BH.

The rest of the paper is organized as follows. In Section II, we introduce the system model. Key CoMP performance factors are defined in Section III and the optimization problem formulation is presented in Section IV. In Section V, we describe our clustering model as SC clustering and UE transfer sub-games and discuss its stability and complexity. Simulation results with insights are presented in Section VI and finally, we summarize the findings and conclude the paper in Section VII. Table I represents a list of all acronyms used in this paper.

II. SYSTEM MODEL

We consider a HetNet scenario, as shown in Fig. 1, where a group of SCs, $SC = \{SC_1, \dots, SC_n\}$, are distributed within the coverage area of one macro base station (MBS). We assume distinct and non-overlapping frequency bandwidth for SC and MBS layers, hence no inter-layer interference is expected. All SCs in the network are grouped into network-centric CoMP clusters such as cluster C_i which comprises a number of SCs, i.e., $C_i = \{SC_{i1}, SC_{i2}, \dots, SC_{iz}\}$. Therefore, the complete list of SCs in the network can be expressed as a set of clusters $\mathcal{C} = \{C_1, \dots, C_s\}$. Each user is assigned to a cluster C_i , and all users that are assigned to the same cluster C_i form a user cluster \mathcal{U}_i . Thus, we define the set of user clusters $\mathcal{U} = \{\mathcal{U}_1, \dots, \mathcal{U}_s\}$ which as assigned to the SC clusters $\mathcal{C} = \{C_1, \dots, C_s\}$ respectively. An example of network-centric clustering of SCs in a HetNet and corresponding user clusters is shown in Fig. 1. Moreover, we define a user-centric SC cluster C_i^k as a sub-cluster of C_i (i.e., $C_i^k \subseteq C_i$) for a user UE_k whose best server is $SC_{im} \in C_i$. We define best server SC_{im} for UE_k where the best average received signal at UE_k is from SC_{im} . We consider a larger time window (seconds, minutes) for clustering decisions to respond

- 2) All BH-aware clustering models in the literature propose user-centric solutions. The downside of such solutions becomes dominant when the network size increases as the computational complexity hinders their scalability. On the other hand, current network-centric solutions overcome the scalability issues but do not account for constrained BH. In this paper, we provide the first network-centric (hence scalable) clustering model that optimizes BH load alongside RAN load and SE.

In our proposed model, we design two coalition sub-games: 1) a SC clustering sub-game to form RAN/BH load-aware SC clusters by merge/split/transfer actions, 2) a novel user transfer sub-game to move users between SC clusters to improve load balancing further. Extensive simulation results for multiple scenarios are presented to show the performance of the proposed method under different BH availability conditions. Results are benchmarked against an improved version of our previous work on RAN load-aware clustering model presented in [32] and a greedy algorithm in [37]. We show that our multi-objective model provides an average of 49.9% increase in overall system throughput when compared to a greedy model across all different BH availability scenarios. This results in 41.7% and 18.4% less unsatisfied users when compared to a greedy model and RAN

TABLE II
SIMULATION PARAMETERS

Parameter Name	Parameter Value
Simulation Enviroment	Urban Microcell [53]
Frequency Carrier	5 GHz
Channel Bandwidth	20 MHz
PRB Bandwidth (B_{PRB})	180kHz
Number of PRBs/SC (R_{tot})	100
Shadow fading std	4 dB [53]
UE Antenna Gain	0 dBi
UE Thermal Noise Density	-174 dBm/Hz
TP Total Transmit Power (P_{Tx})	41dBm [53]
UE Noise Figure	7dB
TP Noise Figure (inc cable loss)	5dB
SC antenna gain (boresight)	17dBi
User-centric cluster: Min RX Power (P_{min})	-110dBm
User-centric cluster: Max RX power offset (P_{Δ})	20dB
Min RX power for Neighbor Def. (P_{min}^{nei})	-110dBm
Max RX power offset for Neighbor Def. (P_{Δ}^{nei})	-20dB
Min payoff gain for user transfer operation (δ_{Δ})	10
RN Simulation Area Radius	0.5km
RN SC deployment Area Radius	0.4km
RN Hotspot Area Radius	0.1km
SC Density for RN (λ_C)	80SC/km ²
UE Density within hotspot in RN Scenario ($\lambda_{U_{high}}$)	6000UE/km ²
UE Density outside hotspot in RN Scenario ($\lambda_{U_{low}}$)	200UE/km ²
UE Density in RN Scenario without Hotspot ($\lambda_{U_{low}}$)	200UE/km ²
GBR for UEs in the hotspot in RN Scenario	2048 kbps
GBR for UEs outside the hotspot in RN Scenario	256 kbps
GBR for UEs in RN Scenario without hotspot	256 kbps
GBR for UEs in HN Scenario	2048 kbps

to spatio-temporal changes in the network and user profiles, consequently average received signal power and average signal to-interference-plus-noise ratio (SINR) are considered for clustering decisions as detailed in Section II-D. Let p_{kj} and p_{km} be the average signal power values received at UE_k from $SC_{ij} \in C_i$ and SC_{im} , respectively. Then, for all $SC_{ij} \in C_i$, we consider $SC_{ij} \in C_i^k$ if $p_{kj}/p_{km} > P_{\Delta}$ and $p_{kj} > P_{min}$ (P_{Δ} and P_{min} are user-defined parameters as described in Table II). An example of the formation of user-centric clusters is presented in the bottom diagram of Fig. 1, which shows user-centric clusters C_1^1, C_1^2 , and C_1^3 , all subsets of the network-centric cluster C_1 and corresponding to users UE_1, UE_2 , and UE_3 , respectively.

In this work, we consider multi-user joint transmission (MU) JT-CoMP where multiple users within the same cluster are scheduled to the same physical resource block (PRB). In other words, user-data for UE_k is made available at each SC within C_i^k . In the following sections, we further elaborate on the BH last mile considerations, the BH-aware CoMP gain computation, and the cluster formation that we propose.

A. Backhaul Considerations

The MBS is assumed to have an ideal BH connection to the core network and to act as an aggregation point for the SC BH links. Thus, the SC BH last mile is the link from the MBS to the SC. In this work, we assume two possible technologies for the BH last mile: VDSL2 (Very high-speed Digital Subscriber Line 2) or fiber-based. Fiber technology offers quasi-ideal BH performance in terms of capacity > 10 Gbps and latency < 1 msec. However, VDSL2 technology offers limited performance where the capacity is capped at 100 Mbps and the latency is at least 3 msec [17]. Both BH technologies are considered to be robust,

hence the outage probability can be ignored. For each SC, the BH throughput demand is calculated based on the radio access user throughput. An additional overhead of 30% is added to the user throughput to account for BH specific control plane traffic [38], [39]. Thus, the overall cell load value is derived by accounting for both radio access capacity and BH capacity limitation (particularly for VDSL2-based last mile).

B. CoMP Gain With Ideal Backhaul

Consider a group of UEs \mathcal{U}_i^k which are assigned a user-centric cluster C_i^k and scheduled in the same PRB at each SC in C_i^k . We assume one antenna for both UE and SCs for simplicity, however, our coalitional game model is applicable to a network with multiple antennas at the SC and UE. Similar one antenna assumption has been made in other CoMP clustering studies [31], [40], [41]. With the assumption of one antenna for UE and SCs, a virtual MIMO system is formed with $|C_i^k| = T$ transmitters and $|\mathcal{U}_i^k| = R$ receivers. For each UE in \mathcal{U}_i^k , the received signal can be expressed as:

$$\mathbf{y} = \mathbf{H}\mathbf{W}\mathbf{x} + \mathbf{n}, \quad (1)$$

where $\mathbf{H} \in \mathbb{C}^{R \times T}$, $\mathbf{W} \in \mathbb{C}^{T \times R}$. The channel matrix can be expressed as $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \dots \mathbf{h}_R]^T$ while the channel vector at UE_k is given by:

$$\mathbf{h}_k = [h_{k1} h_{k2} \dots h_{kT}] \quad (2)$$

Further, the precoding matrix $\mathbf{W} = [\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_R]$ and beamforming vector for UE_k can be expressed as:

$$\mathbf{w}_k = [w_{1k} w_{2k} \dots w_{Tk}]^T \quad (3)$$

Moreover, the received signal y_k at UE_k can be expressed as:

$$\begin{aligned} y_k &= \mathbf{h}_k^{C_i^k} \mathbf{w}_k^{C_i^k} x_k + \sum_{i \in \mathcal{U}_i^k/k} \mathbf{h}_k^{C_i^k} \mathbf{w}_i^{C_i^k} x_i \\ &+ \sum_{j \in \mathcal{U}/\mathcal{U}_i^k} \hat{\mathbf{h}}_k^{C/C_i^k} \mathbf{w}_j x_j + n_k \end{aligned} \quad (4)$$

In (4), the first term represents the desired signal from each of the SCs within C_i^k , the second term represents the interference from within the cluster C_i^k , followed by interference from outside of C_i^k and the final term is the additive white Gaussian noise (AWGN). The dimension for \mathbf{h}_k is $1 \times T$ as it represents the channel vector from all SCs within C_i^k to UE_k and the dimension for $\hat{\mathbf{h}}_k$ is $1 \times (N - T)$ as this term represents the channel matrix from all SCs outside C_i^k to UE_k where N is the total number of SCs in the system. Consequently, the SINR at UE_k can be obtained as:

$$SINR_k$$

$$= \frac{|\mathbf{h}_k^{C_i^k} \mathbf{w}_k^{C_i^k} x_k|^2}{\sum_{i \in \mathcal{U}_i^k/k} |\mathbf{h}_k^{C_i^k} \mathbf{w}_i^{C_i^k} x_i|^2 + \sum_{j \in \mathcal{U}/\mathcal{U}_i^k} |\hat{\mathbf{h}}_k^{C/C_i^k} \mathbf{w}_j x_j|^2 + |n_k|^2} \quad (5)$$

Let the total transmit power for each SC P_{Tx} be the same and that for each PRB be equal, then (5) can be simplified to:

$$SINR_k = \frac{P_{Tx} \sum_{i \in \mathcal{C}_i^k} |h_{ki}|^2}{P_{Tx} \sum_{j \in \mathcal{C}_i^k} |h_{kj}|^2 + N_0 B_{tot}} \quad (6)$$

where N_0 is the noise spectral density, B_{tot} is the total system bandwidth. The channel coefficient h_{ki} is made up of 2 terms, the static distance-based path loss component with shadow fading, g_{ki} , and the fast fading complex coefficients f_{ki} such that $h_{ki} = g_{ki} f_{ki}$. In an ideal BH scenario that assumes fiber BH for each SC within \mathcal{C}_i^k , intra-cluster interference would be negligible with highly accurate knowledge of the CSI and very low latency at the MBS.

It is common and best practice to assume equal power distribution among PRBs and the same power setting to all SCs when conducting network-level simulations for CoMP clustering [42], [43]. Indeed, where link-level simulations necessitate accurate representation of actual power distribution, network-level simulations often assume simplified link-level results to limit the complexity level of the problem. Unequal power distribution would alter the computation of SE and impact RAN/BH load, thus, we anticipate that the algorithm would respond with clustering changes to reflect the updated SE.

C. CoMP Gain With Constrained Backhaul

In reality, not all SCs would afford a fiber-based last mile during deployment, hence some would have an alternative constrained BH. In our model, copper-based VDSL2 technology is considered for the alternative last mile. In addition to the throughput limit of this technology, the high latency (3 msec) causes imperfect CSI, hence intra-cluster interference does not get canceled completely resulting, thus, in degraded SINR. The impact of various latency values is analyzed in [44] for downlink JT-CoMP where an average 15% throughput loss is observed for 3 msec latency. As such, we consider 15% loss in SE when compared to perfect CSI (very low latency fiber-based last mile) for UE_k when \mathcal{C}_i^k contains at least one SC with VDSL2 last-mile link to the MBS.

D. CoMP Clustering With Fading Considerations

We propose that clustering decisions are made based on average SINR to respond to spatio-temporal changes in the network and user profiles (in seconds, minutes), but not to fast fading changes (in milliseconds). This provides additional resilience for incorrect clustering decisions due to imperfect CSI knowledge and prevents additional signaling overhead incurred from frequent re-clustering decisions [45]. For average SINR, the term h_{ki} in (6) can be simplified to the distance-based path-loss and shadow fading component only, i.e., $\hat{h}_{ki} = g_{ki}$ where fast fading component f_{ki} is averaged out over time.

III. CoMP PERFORMANCE FACTORS

In this section, we define the main CoMP performance metrics and utilize these metrics later in our coalitional game model.

Assume UE_k is assigned a network-centric cluster \mathcal{C}_i and user-centric cluster \mathcal{C}_i^k where $\mathcal{C}_i^k \subseteq \mathcal{C}_i$ and let d_k be the guaranteed bit rate (GBR) requirement for UE_k . The required number of PRBs for UE_k in no CoMP scenario would be $r_k = d_k / (y_k B_{PRB})$ where B_{PRB} is the user-data bandwidth in a single PRB, $y_k = \log_2(1 + SINR_k)$ and $SINR_k$ is as defined in (6) with the special case of one SC only in the CoMP cluster i.e. $|\mathcal{C}_i^k| = 1$. In MU-JT CoMP, a number of UEs (\mathcal{U}_i^k) are scheduled on the same PRB at each cell in \mathcal{C}_i^k so we define an estimated dedicated PRB count for UE_k at each SC in \mathcal{C}_i^k as $\hat{r}_k = r_k / n_k$, assuming $|\mathcal{C}_i^k| = |\mathcal{U}_i^k| = n_k$ [32].

A. RAN and BH Load

The main aim for CoMP is to improve SE, hence provide the required throughput with less radio resources and reduce RAN load for the cell. For MU JT-CoMP, increasing CoMP CS improves inter-cell interference cancellation and, therefore SE. However, as CS increases, additional pilot channels are required for CSI estimation which occupy parts of the bandwidth otherwise used for user data. As the available bandwidth for user data reduces, RAN load for the cell increases. So RAN load is one of the key metrics to measure CoMP performance where it implicitly reflects on SE improvement and also the CoMP pilot overhead. We define the RAN load metric for SC_m for MU JT-CoMP scenario as [32]:

$$\hat{l}_{im}^{RAN} = \frac{\sum_{k \in \mathcal{U}_{im}} \hat{r}_k}{R_{tot}} \quad (7)$$

where \mathcal{U}_{im} is the associated active UEs in SC_m and R_{tot} is the total number of PRBs for each SC, assuming all SCs have same total bandwidth.

A more realistic load metric should also consider BH load alongside RAN load. In a network where some SCs have constrained BH links, the overall cell load may be limited by the BH and not the radio access. In MU JT-CoMP scenario, user data for all users within \mathcal{U}_i^k needs to be available at all SCs within \mathcal{C}_i^k . As such, it is expected that an increase in CS results in an increase in BH load. Moreover, additional latency due to non-ideal BH will introduce delay in CSI estimation for precoding and hence reduce SE gain and increase RAN load. In summary, alongside RAN load, BH load is another key metric that needs to be considered in CoMP clustering.

To define the BH load \hat{l}_{im}^{BH} , firstly, we define the RAN throughput demand on SC_m in \mathcal{C}_i as $d_{im}^{RAN} = \sum_{k \in \mathcal{U}_{im}} d_k$. Similar throughput demand and cell load definitions are adopted in [32], [46]. The BH throughput demand d_{im}^{BH} is then computed with an average overhead factor of 1.3 to account for additional traffic on BH for X2 user/control plane and transport and security overheads [38], [39], i.e., $d_{im}^{BH} = d_{im}^{RAN} \times 1.3$. Once d_{im}^{BH} is known, \hat{l}_{im}^{BH} can then be defined as:

$$\hat{l}_{im}^{BH} = \frac{d_{im}^{BH}}{f_{im}^{BH}} \quad (8)$$

where f_{im}^{BH} is the BH capacity. When BH gets congested i.e. $d_{im}^{BH} > f_{im}^{BH}$, then the effective capacity f_{im}^{BH} goes down further due to retransmissions [39]. In the case of VDSL2 link congestion, we consider 10% retransmission rate, i.e., the effective

capacity of the VDSL2 link $f_{im}^{BH} = 90 \text{ Mbps}$. This is in-line with the assumptions made in [39].

A more realistic SC load definition needs to consider both the BH and RAN loads. Effectively, the overall load is the limiting one which is the highest of the two, defined as:

$$\hat{l}_{im} = \max(\hat{l}_{im}^{RAN}, \hat{l}_{im}^{BH}) \quad (9)$$

B. Cell Throughput

In MU JT-CoMP, the user data is transmitted from all of the SCs in C_i^k . Consequently, the total RAN throughput demand d_{im}^{RAN} , as defined in Section III-A, accounts for user UE_k throughput multiple times (in all SCs in C_i^k). As such, an estimated dedicated RAN throughput demand is defined for SC_m in C_i to reflect the actual cumulative throughput as perceived by end-users: $\hat{d}_{im}^{RAN} = \sum_{k \in \mathcal{U}_{im}} d_k/n_k$ where $|C_i^k| = n_k$. Based on estimated dedicated RAN throughput demand \hat{d}_{im}^{RAN} for SC_m , the estimated dedicated cell throughput \hat{l}_{im} for each SC_m in C_i can then be defined as:

$$\hat{l}_{im} = \begin{cases} \hat{d}_{im}^{RAN} / \hat{l}_{im} < 1 \\ \hat{d}_{im}^{RAN} / \hat{l}_{im} \geq 1 \end{cases} \quad (10)$$

C. Unsatisfied Users

Metrics that quantify the level of dissatisfaction of users as a result of high load are used in the literature as a means of user-centric performance indicators [30], [46]. In this work, we adopt the unsatisfied users metric as defined in [32] for MU JT-CoMP scenario as follows:

$$\hat{z}_{im} = \max\left(0, \hat{u}_{im} \left(1 - \frac{1}{\hat{l}_{im}}\right)\right) \quad (11)$$

where \hat{u}_{im} is the estimated dedicated user count at SC_m as:

$$\hat{u}_{im} = \sum_{k \in \mathcal{U}_{im}} 1/n_k \quad (12)$$

The estimated dedicated user count at each cell is driven from the total number of users connected at each cell \mathcal{U}_{im} to account for users that are connected to multiple SCs in MU JT-CoMP.

D. Pilot Overhead

To account for the additional pilot channel overhead, we adopt the pilot overhead estimation for multi-antenna channels in [47], as follows:

$$\alpha = \sqrt{(1 + SNR) \frac{\dot{C}(SNR)}{C(SNR)} 2n_T f_D} - \left((1 + SNR) \frac{\ddot{C}(SNR)}{C(SNR)} + 2 + \frac{1}{2SNR} \int_{-1}^{+1} \frac{d\xi}{\tilde{S}_H(\xi)} \right) n_T f_D + O(f_D^{3/2}) \quad (13)$$

where:

$$\begin{aligned} C(SNR) &= \mathbb{E}[\log_2(1 + SNR|H|^2)], \\ \dot{C}(SNR) &= \frac{1}{SNR} (\log_2 e - \frac{C(SNR)}{SNR}), \\ \ddot{C}(SNR) &= \frac{1}{SNR^2} [\log_2 e + \dot{C}(SNR) - 2 \frac{C(SNR)}{SNR}], \\ SNR &\text{ is the signal to noise ratio on the pilot channel,} \\ \tilde{S}_H(\xi) &\text{ is the Doppler spectrum of the wireless channel,} \end{aligned}$$

f_D is the normalised Doppler frequency and n_T is the number of transmit antennas.

We assume Extended Pedestrian-A (EPA-A) wireless channel from 3GPP [48] for Clarke-Jakes spectrum, where $f_D = 0.000357$ and the term $\int_{-1}^{+1} \frac{d\xi}{\tilde{S}_H(\xi)}$ simplifies to $\pi^2/2$. We assume one antenna per SC, i.e. $n_T = |C_i|$ and SNR=10 dB for pilot overhead estimation.

As CS $|C_i|$ increases, the pilot overhead increases and hence the bandwidth for user data is reduced on each PRB. Thus, the PRB bandwidth available for user data can be defined as $b_{PRB} = B_{PRB}(1 - \alpha)$.

IV. PROBLEM FORMULATION

Our optimization problem is to find the best clustering structure to maximize SE and also balance the RAN/BH load. We maximize user satisfaction by moving traffic from highly loaded clusters to lightly loaded ones. As discussed, in Section III-A, RAN/BH load is a key metric that implicitly includes SE improvement and CoMP pilot overhead as CS increases. We define a utility function as the main objective function of our optimization problem. The utility function captures the overall CoMP gain including the SE improvement, the RAN/BH load balance and the CoMP overheads.

The utility function for SC_m is defined as:

$$v_1(SC_m, C_i) = \begin{cases} \frac{-\hat{l}_{im}}{1-c(|C_i|)} \hat{u}_{im} \hat{l}_{im} < 1 \\ \frac{-\hat{l}_{im}^3}{1-c(|C_i|)} \hat{u}_{im} \hat{l}_{im} \geq 1 \end{cases} \quad (14)$$

where \hat{l}_{im} is the overall cell RAN/BH load at SC_m as defined in (9), \hat{u}_{im} is the estimated dedicated user count at SC_m as defined in (12), and $c(|C_i|)$ is the complexity function defined as $\frac{1}{1+e^{-(|C_i|-C_{max}^n)}}$.

Complexity function $c(|C_i|)$ represents the additional overhead for CoMP, such as precoding processing complexity, synchronization issues and additional BH capacity requirement. As the additional overheads for CoMP increase when CS increases, the complexity function is designed to introduce a soft limit to the maximum CS, C_{max}^n , based on the requirements of the network for the right trade-off between additional SE/load gain and CoMP overheads.

Our utility function in (14) is inversely proportional to SC load, i.e. SC utility gain is reduced as the SC load increase, and it is further penalized for any SC load increase in the high load range ($\hat{l}_{im} \geq 1$). The overall system utility function encourages load distribution from highly loaded SCs into lightly loaded SCs. This is enforced as the utility gain for reducing the load in the high load range ($\hat{l}_{im} \geq 1$) is higher than the utility loss for increasing load in the low load range ($\hat{l}_{im} < 1$). In other words, the overall system utility gain is increased when the load is shifted from highly loaded SCs to lightly loaded SCs. Furthermore, (14) is directly proportional to \hat{u}_{im} , the estimated dedicated user count at SC_m . The utility reduction due to load increase for SCs with higher user count is more than the SCs with lower user count. This gives priority to SCs with higher user count, i.e. better utility gain is achieved in the case of reducing the load on cells serving a higher number of users.

The overall system utility function for a given set of SC clusters \mathcal{C} and associated user clusters of \mathcal{U} is then defined as the sum of all SC utility gain in the system such that:

$$v_1(\mathcal{C}, \mathcal{U}) = \sum_{i=1}^n v_1(SC_i, \mathcal{C}) \quad (15)$$

The objective of our clustering problem is to find the best SC/user clusters, i.e. SC clusters $\mathcal{C}^f = \{C_1^f, \dots, C_s^f\}$ and associated user clusters $\mathcal{U}^f = \{U_1^f, \dots, U_s^f\}$ where the overall system utility is maximized. Therefore our optimization problem can be formulated as:

$$\max_{(\mathcal{C}, \mathcal{U})} v_1(\mathcal{C}, \mathcal{U}) \quad (16a)$$

$$\text{subject to } \forall i \neq j, C_i \cap C_j = \emptyset, \quad (16b)$$

$$\cup_{i=1}^s C_i = \mathcal{C}, \quad (16c)$$

$$\forall i \neq j, U_i \cap U_j = \emptyset, \quad (16d)$$

$$\cup_{i=1}^s U_i = \mathcal{U}. \quad (16e)$$

As (16b) refers, we consider non-overlapping clusters, so each SC can be part of one cluster only and all SCs in the system should be included in a cluster as referred to in (16c). Similarly, each user can only be in one user cluster, and all users should be part of a user cluster as referred to in (16d) and (16e), respectively.

The presented optimization problem increases in complexity as the number of SCs and the number of UEs in the system increase. The number of all possible clusters for a given set of SCs is given by the Bell number¹ [49] which increases exponentially as the number of SCs increases. As an example, the number of all possible cluster sets for a network with SC count = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 are 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, 678570, 4213597, respectively. Therefore, the number of possible cluster sets increases to more than four million for a network of 12 SCs. Additionally, a similar complexity arises in finding the best user-centric cluster where the number of users is expected to be much higher than the number of SCs. To overcome this complexity, we propose a novel coalition game theoretical framework to find the near-optimal clustering solution with significantly reduced complexity. We discuss the details of our coalition game model in the following section.

V. COALITION GAME FOR MULTI-OBJECTIVE CLUSTERING

Applications of coalitional game theory have recently become popular in cooperative wireless networks for self-organizing techniques to form CoMP clusters [50], [51]. A merge/split coalition formation game is employed in forming user clusters for uplink time division multiple access cooperative network scenario in [40]. A similar merge/split game is utilized in forming BS clusters in the downlink CoMP for the CRAN

¹Bell number is defined as the number of all the possible partitions for a given set of players. For example, for a given set of 3 players $\mathcal{C} = \{a, b, c\}$, total number of all possible partitions are 5 i.e. $B_3=5$ and all possible partitions are as follows: $\mathcal{C}^1 = \{\{a, b, c\}\}$, $\mathcal{C}^2 = \{\{a\}, \{b\}, \{c\}\}$, $\mathcal{C}^3 = \{\{a\}, \{b, c\}\}$, $\mathcal{C}^4 = \{\{b\}, \{a, c\}\}$, $\mathcal{C}^5 = \{\{c\}, \{a, b\}\}$.

scenario in [41]. A transfer game is employed alongside a collage admission game for the uplink user association problem in the HetNet scenario in [52]. In our previous work, we presented a merge/split game model to form load-aware clusters where both SE and RAN loads are jointly optimized [32]. In this paper, we formulate two coalitional sub-games to jointly optimize the overall load (BH and RAN) and SE. First, we extend our coalitional game model from our previous work in [32] to combine merge/split and transfer games into a single SC clustering sub-game to form clusters of SCs. Secondly, we drive an additional user transfer sub-game for user groups to transfer users between SC clusters for load distribution. In this section, we formulate and discuss the properties of each sub-game and analyze the overall stability and complexity of the proposed solution.

A. Coalitional Game Model for SC Clustering Sub-Game

In this section, we formulate the SC clustering sub-game where SC clusters are formed and dynamically updated based on spatio-temporal changes in the network and/or user profiles. Let $\mathcal{C} = \{SC_1, \dots, SC_n\}$ be the set of players of our coalition game, i.e. small cells in the network, and assume that they are grouped into clusters $\mathcal{C} = \{C_1, \dots, C_s\}$. A coalition is defined as the groups of players in the same cluster, i.e., $C_i = \{SC_{i1}, SC_{i2}, \dots, SC_{iz}\}$ and a partition is defined as the set of coalitions $\{C_1^a, C_2^a, \dots, C_k^a\}$ where $\forall i \neq j, C_i^a \cap C_j^a = \emptyset$ and $\cup_{i=1}^k C_i^a = \mathcal{C}$. The players in \mathcal{C} dynamically move between coalitions, forming different partitions. Different partitions of the same set of players \mathcal{C} are represented as C^a, C^b, \dots, C^n . The payoff for any coalition C_i in partition \mathcal{C} is defined by the utility function $v(C_i, \mathcal{C})$ and the overall SC clustering sub-game is defined by the pair (\mathcal{C}, v) . The utility function reflects the overall gain for cooperation including multiple objectives of CoMP deployment (e.g. SE and BH/RAN load balancing) and also the various cost factors of cooperation (e.g. additional pilot requirement, signal processing complexity).

We employ the main objective function (14) of our optimization problem defined in Section IV as the main utility function of our coalitional game model. We name (14) as our load-aware utility in the rest of the paper. We also introduce a SE-based utility which is adopted in a greedy clustering algorithm presented in [37] for benchmarking purposes. This utility does not consider cell load but aims to maximize SE only [32]. The SE-based utility function is defined as follows:

$$v_2(SC_m, C_i) = \sum_{k \in \hat{U}_{im}} y_k (1 - c(|C_i|)) \quad (17)$$

where \hat{U}_{im} is the list of users where SC_m is the best serving cell based on average received signal power, i.e. a subset of the associated users U_{im} at the SC_m , and y_k is the SE achieved at UE_k , i.e. $y_k = \log_2(1 + SINR_k)$.

The presented load-aware and SE-based utility functions provide sample utilities aiming to optimize SE and load jointly, and SE in isolation, respectively. Further adjustments can be made in these utilities to favor one of the objectives. Other network objectives, such as energy efficiency, can also be accounted

for by the utility function with weights that reflect the specific network priorities. Our novel game-theoretical clustering model can be utilized with any utility function for an optimal CoMP clustering solution.

To compare the utility of two different partitions $\mathcal{C}^a = \{\mathcal{C}_1^a, \mathcal{C}_2^a, \dots, \mathcal{C}_k^a\}$ and $\mathcal{C}^b = \{\mathcal{C}_1^b, \mathcal{C}_2^b, \dots, \mathcal{C}_z^b\}$, we define a comparison relation \triangleright where $\mathcal{C}^a \triangleright \mathcal{C}^b$ states that partition \mathcal{C}^a is preferable to \mathcal{C}^b . Several comparison relations are discussed in [50]. We employ the utilitarian comparison relation which aims to maximize the overall utility of all players (SCs) regardless of any utility reduction for some of the players. Therefore, partition \mathcal{C}^a is defined as preferable to partition \mathcal{C}^b i.e. $\mathcal{C}^a \triangleright \mathcal{C}^b$ if $\sum_{i=1}^k v(\mathcal{C}_i^a) > \sum_{i=1}^z v(\mathcal{C}_i^b)$ where the utility of any coalition $v(\mathcal{C}_i)$ is defined as the sum of all SC utilities within that coalition. In other words, \mathcal{C}^a is preferable to \mathcal{C}^b only when the total utility of all SCs in the system is increased as a result of this change i.e. $\sum_{i=1}^n v(SC_i, \mathcal{C}^a) > \sum_{i=1}^n v(SC_i, \mathcal{C}^b)$, regardless of possible utility reduction for any individual SC_m i.e. $v(SC_m, \mathcal{C}^a) < v(SC_m, \mathcal{C}^b)$ [32], [50].

In the proposed scheme, SC coalitions are formed and dynamically adapted to changing network/user profile conditions by three different clustering actions:

- **Merge:** Players (SCs) in any two or more coalitions $\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_z\}$ prefer to merge into one coalition $\mathcal{F} = \cup_{i=1}^z \mathcal{C}_i$ i.e. $\cup_{i=1}^z \mathcal{C}_i \triangleright \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_z\}$, if $v(\mathcal{F}) > \sum_{i=1}^z v(\mathcal{C}_i)$ following the utilitarian order.
- **Split:** Players (SCs) prefer to split from any coalition \mathcal{C}_i into smaller coalitions $\{\mathcal{C}_{i1}, \mathcal{C}_{i2}, \dots, \mathcal{C}_{iy}\}$ where $\mathcal{C}_i = \cup_{j=1}^y \mathcal{C}_{ij}$ i.e. $\{\mathcal{C}_{i1}, \mathcal{C}_{i2}, \dots, \mathcal{C}_{iy}\} \triangleright \mathcal{C}_i$ if $\sum_{j=1}^y v(\mathcal{C}_{ij}) > v(\mathcal{C}_i)$ following utilitarian order.
- **Transfer:** Any player in \mathcal{C}_i , i.e. $SC_{ix} \subseteq \mathcal{C}_i$ prefers to transfer from coalition \mathcal{C}_i to \mathcal{C}_j i.e. $\{\mathcal{C}_i \setminus SC_{ix}, \mathcal{C}_j \cup SC_{ix}\} \triangleright \{\mathcal{C}_i, \mathcal{C}_j\}$ if $(v(\mathcal{C}_i \setminus SC_{ix}) + v(\mathcal{C}_j \cup SC_{ix})) > (v(\mathcal{C}_i) + v(\mathcal{C}_j))$.

Assume $\mathcal{C}^a = \{\mathcal{C}_1^a, \mathcal{C}_2^a, \dots, \mathcal{C}_k^a\}$ is a partition of \mathcal{C} , i.e. the current network clustering structure. We propose to start with a split operation, followed by a merge operation and then a transfer operation afterwards. Split/merge/transfer operations are repeated until there is no more re-clustering action possible to improve overall utility. To explain the SC clustering subgame with an example, we look into the possible game actions for a sample network of nine SCs $\mathcal{C} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ where there are three coalitions in the partition \mathcal{C} i.e. $\mathcal{C}_1 = \{1, 2, 3\}$, $\mathcal{C}_2 = \{4, 7\}$, $\mathcal{C}_3 = \{5, 6, 8, 9\}$, as shown in Fig. 2. In the sample network, we have a high density of users within \mathcal{C}_3 and low density of users in other coalitions.

Split operation checks possible split options for $\forall \mathcal{C}_i^a$ in \mathcal{C}^a , and implements the split operation when it finds a suitable split option based on utilitarian order i.e. $(\sum_{j=1}^y v(\mathcal{C}_{ij}^a) > v(\mathcal{C}_i^a))$. For example, in our sample network, there are four split options for \mathcal{C}_1 i.e. $\mathcal{C}_{11} = \{\{1, 2\}, \{3\}\}$, $\mathcal{C}_{12} = \{\{1, 3\}, \{2\}\}$, $\mathcal{C}_{13} = \{\{1\}, \{2, 3\}\}$ and $\mathcal{C}_{14} = \{\{1\}, \{2\}, \{3\}\}$. Split options are checked and once any split option with additional payoff is found, it will be implemented without checking the rest of the split options. Other coalitions are then checked for possible split options and this operation is repeated until no further split is possible, as detailed in Algorithm 1.

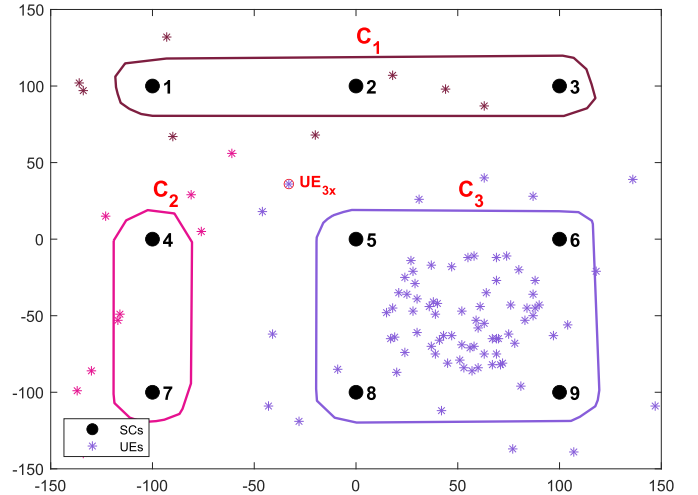


Fig. 2. Sample network for example CoMP clustering .

A new partition \mathcal{C}^b is formed after the split operation. \mathcal{C}^b is then subject to merge operation as detailed in Algorithm 2. Merge operation starts with coalition \mathcal{C}_i^b with the maximum absolute payoff value and looks for merge options to its neighbor coalitions. We avoid the exhaustive search of possible merge with every other coalition in the network which reduces the algorithm complexity significantly. The merge operation is implemented for $(\mathcal{C}_i^b, \mathcal{C}_j^b)$ coalition pair where \mathcal{C}_j^b is the neighbor coalition for \mathcal{C}_i^b with maximum additional payoff in the case of a possible merge operation. Neighbor coalitions are defined based on the reported average received signal power from the users. For any user UE_k within the serving area of $SC_m \subseteq \mathcal{C}_m^b$, a neighbor rank value is incremented for $\{\mathcal{C}_m^b, \mathcal{C}_j^b\}$ pair if $p_{kj}/p_{km} > P_{\Delta}^{nei}$ and $p_{kj} > P_{min}^{nei}$, where p_{km} and p_{kj} are the average signal power values received from UE_k for $SC_m \subseteq \mathcal{C}_m^b$ and $SC_j \subseteq \mathcal{C}_j^b$, respectively. A similar neighbor cluster concept is used in our previous work in [32]. Merge operation continues for $\forall \mathcal{C}_i^b$ in \mathcal{C}^b and is repeated for the whole partition until no other merge is possible. In our sample network, assuming the same partition in Fig. 2, the merge operation is likely to start with \mathcal{C}_3 as it will have the maximum absolute payoff value. As absolute payoff value is directly proportional to the number of users served in both utility functions, merge operation prioritizes the coalitions with the highest users. The possible merge operation is checked with neighbor coalitions, i.e., \mathcal{C}_1 and \mathcal{C}_2 and merge operation is implemented on the coalition pair with maximum additional payoff. In a typical larger network, the total number of SCs/coalitions is much higher, however, our proposed merge algorithm only looks for the neighbor coalitions for a possible merge which is likely to be few coalitions around the main coalition, rather than checking all coalitions. Assuming a merge operation on $(\mathcal{C}_3, \mathcal{C}_1)$ pair, the total number of coalitions in the sample network will reduce to two and the resultant two coalitions are checked for any further merge operations until no other merge is possible.

Once the merge stage is completed, the transfer operation starts with the resulting new partition \mathcal{C}^d . For $\forall \mathcal{C}_i^d \in \mathcal{C}^d$, each $SC_{ix} \in \mathcal{C}_i^d$ are checked for a possible transfer to one of the neighbor coalition \mathcal{C}_j^d i.e. $T(SC_{ix}, \mathcal{C}_i^d, \mathcal{C}_j^d)$. Within each

Algorithm 1: Split Operation.

For any given network clustering state
 $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s\}, \forall \mathcal{C}_i \in \mathcal{C}$, set $\mathcal{C}_i.\text{splitpossible}=1$
Split-ongoing=1
while Split-ongoing **do**
 Split-ongoing=0
 for all \mathcal{C}_i where $(\mathcal{C}_i.\text{split-possible}=1$ and $|\mathcal{C}_i| > 1)$ **do**
 Update $\mathcal{C}_i.\text{Split-options}$
 $\mathcal{C}_i.\text{split-possible}=0$
 for all $\mathcal{C}_i.\text{Split-Options}$ **do**
 if Any split option is possible i.e.
 $(\sum_{j=1}^y v(\mathcal{C}_{ij}) > v(\mathcal{C}_i))$ **then**
 Split(\mathcal{C}_i to $\{\mathcal{C}_{i1}, \mathcal{C}_{i2}, \dots, \mathcal{C}_{iy}\}$)
 Split-ongoing=1
 $\forall \mathcal{C}_{ij}$, set $\mathcal{C}_{ij}.\text{split-possible}=1$
 Break for-loop and continue with next \mathcal{C}_i
 end if
 end for
end for
end while

coalition \mathcal{C}_i^d , all possible transfer operations are ranked and transfer operation $T(SC_{ix}, \mathcal{C}_i^d, \mathcal{C}_j^d)$ is implemented for the one with the maximum additional payoff. Transfer operation continues for all $\forall \mathcal{C}_i^d \in \mathcal{C}^d$ and is repeated for the newly formed partition until there is no further transfer possible with additional payoff, as detailed in Algorithm 3. In our sample network, assuming the same partition in Fig. 2, each coalition is checked to find the SC with maximum additional payoff gain for a possible coalition transfer. For example, in \mathcal{C}_1 , all three SCs are checked for a possible transfer operation to neighbor coalitions i.e. \mathcal{C}_2 and \mathcal{C}_3 . In a typical larger network, there would be a higher number of coalitions but these are not checked for each transfer operation unless they are neighbor coalitions. Transfer operation is implemented for the SC with the maximum additional payoff if it exists. For example, SC_3 in \mathcal{C}_1 may prefer to transfer to \mathcal{C}_3 and form new coalitions $\mathcal{C}_1 = \{1, 2\}$ and $\mathcal{C}_3 = \{3, 5, 6, 8, 9\}$. SC transfer operations are then repeated in each coalition until no further SC transfer operation is possible.

Once, SC transfer operation is completed, split, merge and transfer operations are then repeated until there are no further SC coalition actions possible. The order of game actions is arbitrarily selected as split/merge/transfer, as these actions are re-iterated within the SC clustering sub-game until there is no further game action possible. In other words, the algorithm controls which game action will be utilized more than others depending on the existing clustering structure and the re-clustering changes required to adapt to spatio-temporal changes in user/network profile.

B. Coalitional Game Model for User Transfers Sub-Game

Assume $\mathcal{C}^e = \{\mathcal{C}_1^e, \mathcal{C}_2^e, \dots, \mathcal{C}_p^e\}$ be the SC partition of \mathcal{C} resulting from the SC clustering sub-game (\mathcal{C}, v) . The list of users $\mathcal{U} = \{UE_1, \dots, UE_q\}$ can be expressed as coalitions of users assigned to each SC cluster, i.e., $\mathcal{U}^e = \{\mathcal{U}_1^e, \dots, \mathcal{U}_p^e\}$, where

Algorithm 2: Merge Operation.

For any given network clustering state
 $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s\}, \forall \mathcal{C}_i \in \mathcal{C}$, set $\mathcal{C}_i.\text{clustered}=0$
Merge-ongoing=1
while Merge-ongoing **do**
 Merge-ongoing=0
 Sort $\forall \mathcal{C}_i \in \mathcal{C}$ based on $|v(\mathcal{C}_i)|$ in descending order
 for all \mathcal{C}_i where $\mathcal{C}_i.\text{clustered}=0$ **do**
 Update $\mathcal{C}_i.\text{nei}$
 for all \mathcal{C}_j in $\mathcal{C}_i.\text{nei}$ where $\mathcal{C}_j.\text{clustered}=0$ **do**
 Update payoff gain for possible merge($\mathcal{C}_i, \mathcal{C}_j$) i.e.
 $\delta_{v_{ij}} = v(\mathcal{C}_i \cup \mathcal{C}_j) - \{v(\mathcal{C}_i) + v(\mathcal{C}_j)\}$
 end for
 Find $\mathcal{C}_m \in \mathcal{C}_i.\text{nei}$ where $\delta_{v_{im}} = \max_{\mathcal{C}_j \in \mathcal{C}_i.\text{nei}} (\delta_{v_{ij}})$ and $\delta_{v_{im}} > 0$
 while \mathcal{C}_m exist **do**
 Merge($\mathcal{C}_i, \mathcal{C}_m$)
 $\mathcal{C}_m.\text{clustered}=1$
 Update $\mathcal{C}_i.\text{nei}$
 for all \mathcal{C}_j in $\mathcal{C}_i.\text{nei}$ where $\mathcal{C}_j.\text{clustered}=0$ **do**
 Update payoff gain for possible merge($\mathcal{C}_i, \mathcal{C}_j$) i.e.
 $\delta_{v_{ij}} = v(\mathcal{C}_i \cup \mathcal{C}_j) - \{v(\mathcal{C}_i) + v(\mathcal{C}_j)\}$
 end for
 Find $\mathcal{C}_m \in \mathcal{C}_i.\text{nei}$ where $\delta_{v_{im}} = \max_{\mathcal{C}_j \in \mathcal{C}_i.\text{nei}} (\delta_{v_{ij}})$ and $\delta_{v_{im}} > 0$
 end while
 $\mathcal{C}_i.\text{clustered}=1$
 if Any merge operation with \mathcal{C}_i **then**
 Break for-loop and continue with while-loop
 Merge-ongoing=1
 end if
 end for
 end while

users in \mathcal{U}_i^e are assigned to SC coalition \mathcal{C}_i^e . We formulate here a user transfer sub-game (\mathcal{U}, v) which distributes SC clusters' loads by transferring users between user coalitions.

Transfer operation introduced in SC clustering sub-game in Section V-A is deployed for the user transfer sub-game, i.e., any user $UE_{ix} \subseteq \mathcal{U}_i^e$ prefer to transfer from coalition \mathcal{U}_i^e to \mathcal{U}_j^e i.e. $\{\mathcal{U}_i^e \setminus UE_{ix}, \mathcal{U}_j^e \cup UE_{ix}\} \triangleright \{\mathcal{U}_i^e, \mathcal{U}_j^e\}$ if $v(\{\mathcal{U}_i^e \setminus UE_{ix}\} + v(\mathcal{U}_j^e \cup UE_{ix})) > \{v(\mathcal{U}_i^e) + v(\mathcal{U}_j^e)\}$ following utilitarian order. We utilize the load-aware utility in (14) for user transfer sub-game and transfer users to re-assign to another cluster if the overall utility is improved. The neighbor concept introduced in the SC clustering sub-game is employed in the user transfer sub-game too at the user level, so that each user only looks for the neighbor coalitions instead of all coalitions for a possible transfer. A list of SC clusters is kept as neighbors for UE_k based on the received average reference signal level. For any user UE_k within the serving area of $SC_m \subseteq \mathcal{C}_m$, \mathcal{C}_j is included in the neighbor list if $p_{kj}/p_{km} > P_{\Delta}^{nei}$ and $p_{kj} > P_{min}^{nei}$ where p_{km} and p_{kj} are the average signal power values received at UE_k from $SC_m \subseteq \mathcal{C}_m$ and $SC_j \subseteq \mathcal{C}_j$, respectively.

For each user coalition $\mathcal{U}_i^e \in \mathcal{U}^e$, users are checked for possible user transfer operation to all of its neighbor coalitions.

Algorithm 3: Transfer Operation.

For any given network clustering state $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s\}$
 Transfer-ongoing=1
while Transfer-ongoing **do**
 Transfer-ongoing=0
for all $\mathcal{C}_i \in \mathcal{C}$ **do**
 Update $\mathcal{C}_i.\text{nei}$
for all $SC_{ix} \subset \mathcal{C}_i$ **do**
for all \mathcal{C}_j in $\mathcal{C}_i.\text{nei}$ **do**
 Update payoff gain for possible
 Transfer($SC_{ix}, \mathcal{C}_i, \mathcal{C}_j$) i.e.
 $\delta_{v_{ixj}} = \{v(\mathcal{C}_i \setminus SC_{ix}) + v(\mathcal{C}_j \cup SC_{ix})\} -$
 $\{v(\mathcal{C}_i) + v(\mathcal{C}_j)\}$
end for
end for
 Find ($SC_{ix}, \mathcal{C}_i, \mathcal{C}_j$) where $\delta_{v_{ixj}} = \max_{\substack{\mathcal{C}_j \in \mathcal{C}_i.\text{nei} \\ SC_{ix} \subset \mathcal{C}_i}} (\delta_{v_{ixj}})$
 and $\delta_{v_{ixj}} > 0$
if ($SC_{ix}, \mathcal{C}_i, \mathcal{C}_j$) exist **then**
 Transfer($SC_{ix}, \mathcal{C}_i, \mathcal{C}_j$)
 Transfer-ongoing=1
end if
end for
end while

The best transfer option with maximum additional payoff is implemented for UE_{ix} from \mathcal{U}_i^e to \mathcal{U}_j^e and user coalitions are updated. All other user coalitions are then checked for any possible user transfer and single user from each coalition with maximum payoff gain is transferred in a similar way. User transfers are limited to the ones with certain additional payoff δ_{Δ} . This is introduced as an input parameter in the algorithm for the right balance between the number of user transfers and additional overall system payoff. User transfer operation is repeated for all user coalitions until no further user transfer is possible, as detailed in Algorithm 4. In our sample network, assuming the same SC partition in Fig. 2, users in each coalition are checked for a possible transfer to other coalitions. Users at the cluster boundary are likely to be transferred to other coalitions if the additional payoff created in the current coalition is more than the payoff loss in the destination coalition. Based on load-aware utility (14), moving users from highly loaded coalitions generates more payoff than the payoff loss in lightly loaded destination coalitions. Consequently, the transfers of users located at the coalition edge are encouraged from highly loaded coalitions to lightly loaded coalitions. For example, UE_{3x} in Fig. 2 is located at the coalition edge in between the coalitions \mathcal{C}_3 and \mathcal{C}_1 where \mathcal{C}_3 is highly loaded and the destination coalition \mathcal{C}_1 is lightly loaded. If UE_{3x} is the best candidate for user transfer with the maximum additional payoff in \mathcal{C}_3 , then this user transfer is implemented. Other coalitions \mathcal{C}_1 and \mathcal{C}_2 are then checked for any possible user transfers and this is repeated for each coalition until no further user transfers are possible.

At the end of the user transfer sub-game, a new user partition $\mathcal{U}^f = \{\mathcal{U}_1^f, \dots, \mathcal{U}_p^f\}$ is formed where user coalition \mathcal{U}_j^f represents the associated users in SC cluster \mathcal{C}_j^e . After forming the new

user partition \mathcal{U}^f , SC clustering sub-game is re-deployed for further merge/split/transfer operations where both SC and user partitions are updated. For any SC merge operation, $\mathcal{C}_x = \cup_{i=1}^z \mathcal{C}_i$, the associated user coalitions are also merged $\mathcal{U}_x = \cup_{i=1}^z \mathcal{U}_i$. In the case of a SC cluster split operation of \mathcal{C}_i into smaller coalitions $\{\mathcal{C}_{i1}, \mathcal{C}_{i2}, \dots, \mathcal{C}_{iy}\}$, then associated user coalition \mathcal{U}_i is also split to $\{\mathcal{U}_{i1}, \mathcal{U}_{i2}, \dots, \mathcal{U}_{iy}\}$ based on each user's best serving SC within the cluster (not necessarily the best serving SC in the network as the user may have been transferred to non-best serving SC coalition during user transfer sub-game). For example, assume $SC_{ix} \in \mathcal{C}_i$ is the best serving SC within \mathcal{C}_i for $UE_k \in \mathcal{U}_i$, then in the case when \mathcal{C}_i splits and SC_{ix} falls in the new coalition \mathcal{C}_{ix} , then user coalition \mathcal{U}_i is split similarly where $UE_k \in \mathcal{U}_{ix}$. Similarly, for transfer operation of $SC_{ix} \subset \mathcal{C}_i$ transferring from coalition \mathcal{C}_i to \mathcal{C}_j , users in \mathcal{C}_i where SC_{ix} is the best serving SC within \mathcal{C}_i are transferred from \mathcal{U}_i to \mathcal{U}_j .

Both SC clustering and user-transfer sub-games are repeated until there is no further SC cluster or user cluster changes. As the utility for both sub-games is the same, each SC/UE coalition change improves the overall utility and converges to a final SC/user partition. The final partition is the clustering solution for the current network/user status. To adapt to the dynamic spatio-temporal changes in the network and user profiles, the algorithm is proposed to run regularly in set time intervals and adapt SC/user clusters to these changes accordingly. As discussed in Section II, re-clustering changes are proposed in seconds/minutes as opposed to milliseconds to avoid too frequent clustering decisions based on fast fading changes. In the next sub-section, we discuss the stability and complexity of our algorithm.

C. Algorithm Stability

In this subsection, we prove that both SC clustering and user transfers sub-games always converge to a final partition and analyze the overall game stability.

Assume that the current state of the SC partition is $\mathcal{C}^1 = \{\mathcal{C}_1^1, \mathcal{C}_2^1, \dots, \mathcal{C}_s^1\}$. In SC clustering sub-game, partition \mathcal{C}^1 is subject to merge-split-transfer operations which will transfer the network partition to \mathcal{C}^n following a sequence of partitions.

$$\mathcal{C}^1 \rightarrow \mathcal{C}^2 \rightarrow, \dots, \rightarrow \mathcal{C}^n \quad (18)$$

Any merge/split/transfer operation between coalitions \mathcal{C}_i and \mathcal{C}_j increases the overall utility of the involved SCs/coalitions following utilitarian preference order, i.e., $v(\text{Merge/Split/Transfer}(\mathcal{C}_i^1, \mathcal{C}_j^1)) > (v(\mathcal{C}_i^1) + v(\mathcal{C}_j^1))$. As detailed in Section II, we assume that clustering decisions are made in longer time intervals (seconds, minutes), fast fading component of the signal is averaged out for clustering decision and hence the interference created from any $SC \in (\mathcal{C}_i^1 \cup \mathcal{C}_j^1)$ to the rest of the network is the same regardless of any merge/split/transfer changes within $(\mathcal{C}_i^1 \cup \mathcal{C}_j^1)$. Hence, $v(\mathcal{C}^1 \setminus (\mathcal{C}_i^1 \cup \mathcal{C}_j^1))$ is unchanged when there is any merge/split/transfer operation between coalitions \mathcal{C}_i^1 and \mathcal{C}_j^1 . As $v(\text{Merge/Split/Transfer}(\mathcal{C}_i^1, \mathcal{C}_j^1)) > (v(\mathcal{C}_i^1) + v(\mathcal{C}_j^1))$, and there is no change for the rest of the network as a result of this

Algorithm 4: User Transfer Operation.

For any given network clustering state $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s\}$
and corresponding user coalitions $\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_s\}$
UserTransfer-ongoing=1
while UserTransfer-ongoing **do**
 UserTransfer-ongoing=0
 for all $\mathcal{U}_i \in \mathcal{U}$ **do**
 for all $UE_{ix} \subset \mathcal{U}_i$ **do**
 for all \mathcal{U}_j in $UE_{ix}.nei$ where $i \neq j$ **do**
 Update payoff gain for possible
 Transfer($UE_{ix}, \mathcal{U}_i, \mathcal{U}_j$) i.e.
 $\delta_{v_{xij}} = \{v(\mathcal{U}_i \setminus UE_{ix}) + v(\mathcal{U}_j \cup UE_{ix})\} -$
 $\{v(\mathcal{U}_i) + v(\mathcal{U}_j)\}$
 end for
 end for
 Find ($UE_{ix}, \mathcal{U}_i, \mathcal{U}_j$) where
 $\delta_{v_{xij}} = \max_{\substack{\mathcal{U}_j \in UE_{ix}.nei \\ UE_{ix} \in \mathcal{U}_i}} (\delta_{v_{xij}})$ and $\delta_{v_{xij}} > \delta_\Delta$
 if ($UE_{ix}, \mathcal{U}_i, \mathcal{U}_j$) exist **then**
 Transfer($UE_{ix}, \mathcal{U}_i, \mathcal{U}_j$)
 UserTransfer-ongoing=1
 end if
end for
end while

operation, then the overall system utility always increases with every partition in sequence (18), i.e.

$$v(\mathcal{C}^n) > v(\mathcal{C}^{n-1}) \dots v(\mathcal{C}^2) > v(\mathcal{C}^1) \quad (19)$$

where $\mathcal{C}^i \neq \mathcal{C}^j, i \neq j$. As the overall system utility is always increased with every partition in the sequence, i.e., the same partition is never visited again and there is a finite number of partitions limited by the Bell number, then the sequence in (18) is guaranteed to converge to a final SC partition.

For a given fixed SC partition $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s\}$, the associated user coalitions $\mathcal{U}^1 = \{\mathcal{U}_1^1, \dots, \mathcal{U}_s^1\}$ are subject to user transfers which will transform the user coalitions into \mathcal{U}^t and the overall system utility of SC partition \mathcal{C} will increase with every user partition change as per the definition of user transfer rule following utilitarian order, i.e.,

$$v(\mathcal{U}^t) > v(\mathcal{U}^{t-1}) \dots > v(\mathcal{U}^2) > v(\mathcal{U}^1) \quad (20)$$

where $\mathcal{U}^i \neq \mathcal{U}^j, i \neq j$. Similar to SC partition convergence, as there is a finite number of user partitions limited by the Bell number, and user partitions will always evolve to a better utility, then the user partition sequence is guaranteed to converge to a final partition. When both sub-games are employed jointly, the overall system utility is always increased with every SC/user partition changes, and hence the same SC and user partition will never be re-visited. There will be a finite number of possible SC/user partitions and therefore the overall SC/user partition will always converge to a final SC/user partition. As such, the proposed coalition-based multi-objective approach is bound to improve the CoMP performance and always converge to a final partition which is an equilibrium state with respect to defined game actions. However, until a tractable and precise system-level analytical modeling becomes feasible, it is not possible

to demonstrate the existence of a Nash equilibrium state that guarantees the optimum setting.

D. Algorithm Complexity

An exhaustive search for the optimum SC clustering with multi-objective considerations is a highly complex task where the number of possibilities increases exponentially as the network size increases. We propose a coalition-game approach to allow for a practical clustering method that outperforms existing methods yet with bounded complexity. The approach is composed of two sub-games: the first relates to clustering of the SCs while the second relates to clustering of users.

1) *SC Clustering Sub-Game:* The SC clustering sub-game consists of three different steps: Split, Merge, and Transfer. The Split algorithm is first conducted over all the clusters in the network. For each cluster, all split options are considered. As seen in Algorithm 1, the Split operation requires the evaluation of a constant multiple of $|\mathcal{C}| = n$ possibilities, where \mathcal{C} is the set of SCs in the network and n is the number of SCs. This upper bound is only reached in case no split option is found before the last evaluation. Thus, the asymptotic complexity of the split operation is linear in the order of $\mathcal{O}(n)$.

Once the split step is completed, the merge operation is initiated, as described in Algorithm 2. All clusters resulting from the split step are evaluated for a possible merge, hence an upper bound of n/\mathcal{C}_{\max} clusters are visited, where \mathcal{C}_{\max} is the soft max user-defined value to limit the size of clusters. For each cluster, the possible merge evaluations are limited to its neighboring clusters, as defined in the neighboring list. As the number of allowed neighbors is user-controlled, the asymptotic complexity of the Merge step also leads to $\mathcal{O}(n)$. If it were not for the neighbor list that limits the search for merge options, the merge algorithm would have had quadratic complexity (i.e., $\mathcal{O}(n^2)$) instead of linear.

The last step is the transfer operation, as described in Algorithm 3. The transfer operation checks each cell in each cluster for a possible transfer to one of the neighbor coalitions. Thus, there is a total of n cells that are evaluated for options within the neighbor coalition list, leading to an asymptotic algorithm complexity in the order of $\mathcal{O}(n)$ as opposed to quadratic $\mathcal{O}(n^2)$ (if no neighbor restriction were implemented). The three operations are repeated until convergence, which is reached within a finite number of iterations, as shown in Section V-C (see Fig. 8). However, the algorithmic complexity of the Split operation is further reduced with every iteration, as detailed in Section V-C. Indeed, any split operation for \mathcal{C}_i does not depend on the structures of other coalitions. Thus, any coalition \mathcal{C}_i that is found to not have a split possibility in one iteration will not be checked again in following iterations unless the Merge or Transfer operations resulted in changes in that same coalition \mathcal{C}_i .

In summary, the algorithm complexity of the first sub-game is reduced from quadratic to linear owing to the neighbor SC/coalition concept. The user-defined neighbor thresholds can be adjusted for a more relaxed/tight neighbor definition and increased/reduced merge/transfer options for the right balance between additional CoMP gain and complexity.

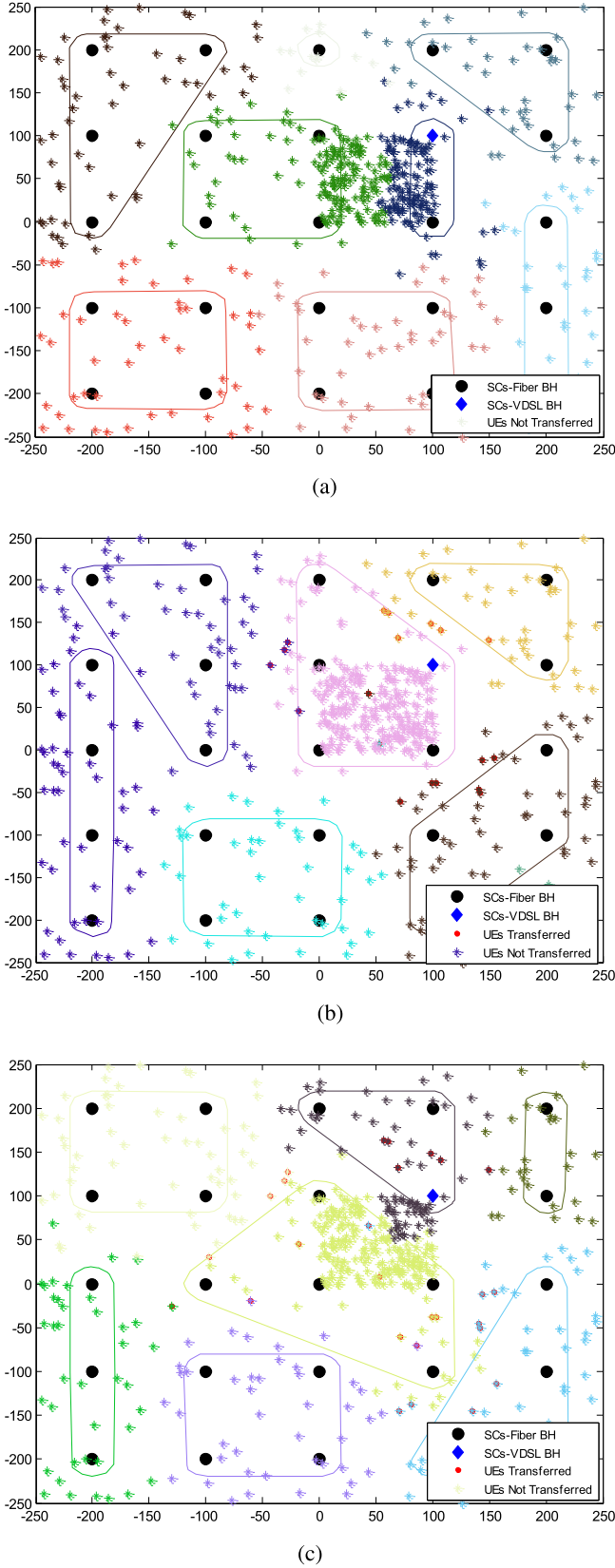


Fig. 3. Snapshot of SE-GR, L-GA and LBH-GA clusters in HN with hotspot scenario.

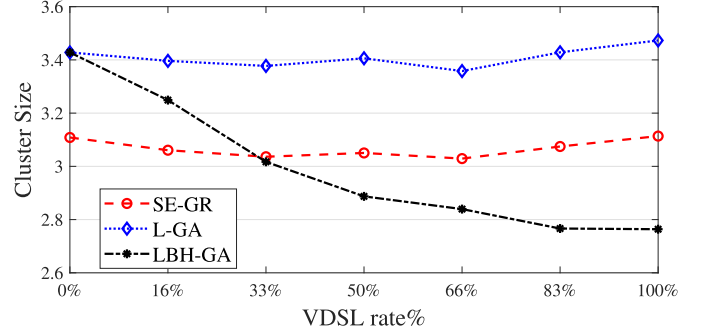


Fig. 4. Cluster size comparison for all BH cases.

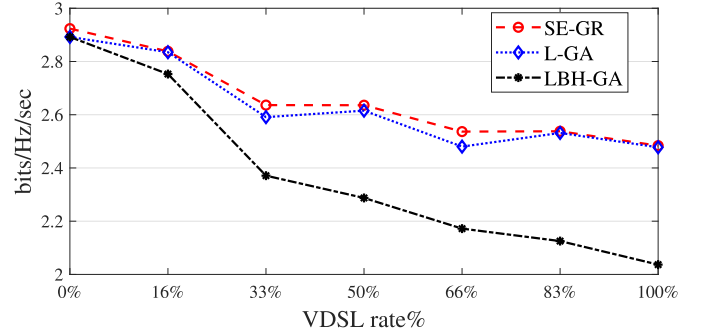


Fig. 5. SE comparison for all BH cases.

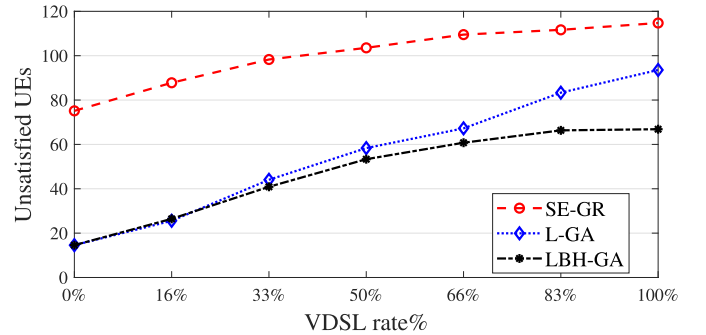


Fig. 6. Unsatisfied UEs comparison for all BH cases.

2) *Users Transfer Sub-Game*: The user transfer sub-game looks at each user in each user cluster and evaluates its transfer options within the neighbor list. Thus, the algorithm has a linear complexity which is a function of the number of users $\mathcal{O}(|\mathcal{U}|)$, where \mathcal{U} is the set of users and $|\mathcal{U}|$ is the number of users. Similar to the first sub-game, the neighbor concept in the user transfer coalition game reduces the algorithm complexity from quadratic $\mathcal{O}(|\mathcal{U}|^2)$ to linear $\mathcal{O}(|\mathcal{U}|)$. Once the user transfer sub-game is completed, the SC clustering sub-game is revisited, followed by the user transfer sub-game until convergence. Thus the overall complexity of the proposed scheme can be expressed as $\mathcal{O}(n + |\mathcal{U}|)$. Accordingly, the proposed scheme is scalable and can be implemented in large networks, thus realizing the true potential of CoMP.

VI. SIMULATION RESULTS

In this section, we present the simulation results to evaluate the performance of the novel clustering model (LBH-GA) introduced in this work that jointly optimizes RAN and BH

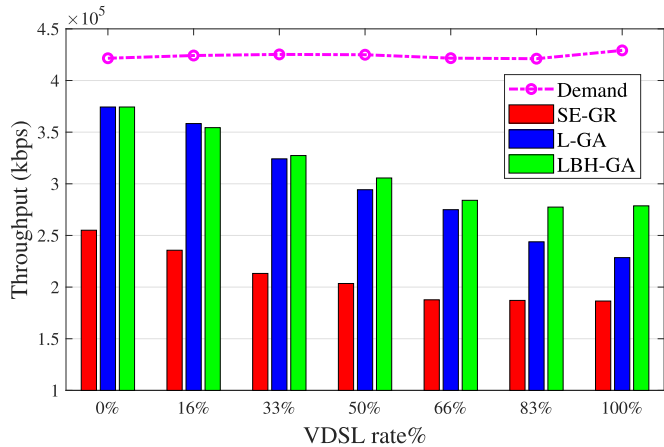


Fig. 7. System throughput comparison for all BH cases.

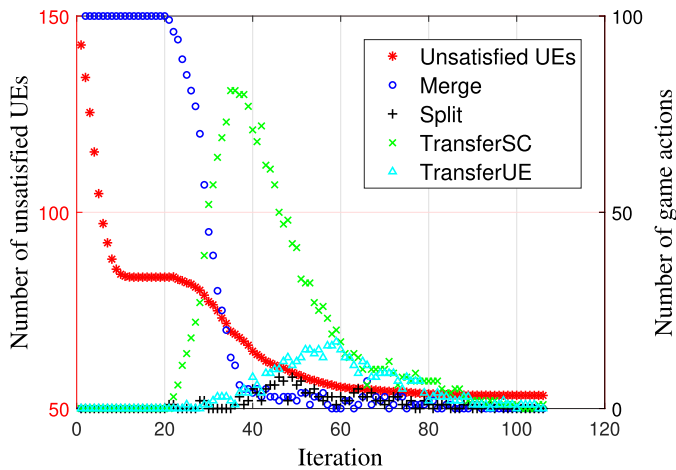


Fig. 8. LBH-GA Game Actions vs. Unsatisfied UEs for 50% VDSL2 rate.

loads while maximizing SE. The novel model is benchmarked against two solutions: 1) L-GA, an improved version of the RAN load-aware solution presented in [32] and 2) SE-GR, a greedy model from [37]. The first model L-GA represents an improved version of the RAN-only load-aware clustering solution previously presented in [32]. In this case, the novel two-stage coalitional game model is followed but the load-aware utility function (14) is based on RAN-only considerations (not BH). For a fair comparison with the greedy solution (SE-GR), we adapt our SE-based utility function (17) in the greedy model and lift the hard CS limit where an implicit soft CS limit is employed via the cost function in the utility (17). Additionally, the neighbor concept introduced for our RAN/BH load-aware model (LBH-GA) is also employed in the greedy algorithm. The SE-based greedy algorithm (SE-GR) used in this work for benchmarking is presented in detail in our previous work [32]. In the rest of the paper, the following abbreviations are used for the presented clustering models:

- SE-GR: Greedy model employing SE based utility (17).
- L-GA: RAN load-aware game-theoretic model with load based utility (14) considering RAN load only.
- LBH-GA: RAN and BH load-aware game-theoretic model with load-based utility (14) considering combined RAN and BH load.

For each of the three listed algorithms, we consider two scenarios: Homogeneous network (HN) and Random network (RN). The HN scenario considers hotspots and investigates the formation of clusters by each algorithm in these conditions. We then run extensive simulations for the RN scenario with and without hotspots. As described in Section II, we assume a HetNet composed of one MBS overlaid with SCs, where each SC is a single cell with an omni-directional antenna.

Our simulation platform is built in MatLab and each scenario for each clustering model is repeated for 100 snapshots. The simulations are run on a machine with Windows 10 Enterprise 64-bit operating system, Inter(R) Core(TM) i5-8350 U CPU @ 1.70 GHz 1.90 GHz processor and 8.00 GB RAM. For a RN scenario without hotspots, the average time for the clustering solution to converge to the final clusters (starting from a no-clustering state) is 90.54 and 11.13 seconds for LBH-GA and SE-GR models, respectively. The additional time required for LBH-GA to converge is due to the additional SC merge/split/transfer actions and user transfer actions which provide the additional capacity gain compared to the greedy model, as discussed in the rest of this section. The time required to achieve the final cluster state is expected to be lower when the model is applied to an existing clustering solution which would require minimal changes in comparison to a no-clustering state starting point. Additionally, processing capacity is expected to be much higher in a real network scenario which will reduce the processing time. As discussed in Section II-D, we aim to respond to user profile/network changes rather than fast fading changes, and, hence, the time scale between two consecutive re-clustering operations is expected to be in the order of several minutes to reflect pertinent dynamics in user/network profiles. As presented in Section V-D, the complexity of our model is reduced significantly by avoiding exhaustive search as a result of the introduction of the neighbor concept.

Firstly, we run simulations in HN deployment with a hotspot scenario to illustrate the clusters formed by each algorithm. In HN scenario, 25 SCs are deployed in $500 \text{ m} \times 500 \text{ m}$ simulation area with 100 m inter-site distance. 300 UEs are distributed in the whole area, following a uniform random distribution. In addition, 200 UEs are also uniformly distributed with a $100 \text{ m} \times 100 \text{ m}$ area to simulate a hotspot scenario. All SCs are assumed to have fiber BH connection to the MBS except one with a VDSL2 BH link. Each UE is assumed to have a fixed GBR requirement of 2048 kbps. The pathloss model is adapted from ITU-R microcell urban non-line-of-site (NLOS) model [53] as follows: $PL = 36.7 \log_{10}(d) + 22.7 + 26 \log_{10}(f_c)$, where d is the distance in meters and f_c is the carrier frequency in GHz. The rest of the simulation parameters are summarized in Table II.

Fig. 3(a) depicts the clusters formed by the SE-GR algorithm in HN deployment scenario with hotspot. As SE-GR clustering starts from a random SC, it fails to achieve a cluster around the loaded cells. As shown in Fig. 3 b, L-GA algorithm forms the cluster around the hotspot area as the algorithm utility takes cell load into account, and gives priority to loaded SCs for clustering. Furthermore, L-GA CS is increased around the hotspot, giving better SE and hence reduced load. Fig. 3 c shows clusters formed by the LBH-GA model where a cluster is formed around the

TABLE III
AVERAGE SE AND CS IN RN WITHOUT HOTSPOT SCENARIO WHEN VDSL2
RATE = 33%

	SE-GR	L-GA	LBH-GA
SE (bits/Hz/sec)	2.6783	2.6339	2.5036
Cluster Size	2.9950	3.0743	2.8426

hotspot, but the only one VDSL2 site is excluded from this cluster as BH capacity limitation introduces a higher BH load than RAN load, thus reducing the utility gain for forming a cluster.

We performed extensive simulations in a more realistic RN scenario with and without hotspots. In our simulation setup, we deployed SCs randomly following the Poisson point process (PPP) distribution with density parameter (λ_c) within a circle of 0.4 m radius. UEs are also randomly distributed following PPP distribution. In RN with hotspot scenario, we simulate a hotspot area in an inner circle with 0.1 m radius. A high density $\lambda_{U_{high}}$ of UEs are deployed in the inner circle and a lower density of UEs $\lambda_{U_{low}}$ are deployed in the outer ring where the radius is set to 0.5 m. UE deployment area is set to a bigger radius than the SC deployment area to make sure that UEs are distributed to the whole coverage area of the SCs. The GBR for UEs within the hotspot is set to 2048 kbps and for UEs outside of the hotspot ring to 256 kbps. For RN without hotspot scenario, UE density is set to $\lambda_{U_{low}}$ for both inner and outer ring areas and GBR is set to 256 kbps for all UEs.

We first analyze the results in RN without hotspot scenario. We ran our simulation for 100 snapshots where 33% of the SCs are assumed to have VDSL2 BH and the remaining have fiber BH. Table III shows the achieved SE and CS, respectively, for the three algorithms. We observe that L-GA performs similar to SE-GR when there is no hotspot with a marginal difference in achieved SE and CS. LBH-GA achieves a slightly lower CS value when compared to L-GA as it accounts for the BH limitations on some sites. As observed in HN clustering scenario, LBH-GA tends to exclude sites with VDSL2 connection. For SCs with VDSL2, RAN load is the limiting factor in low CS, and as the CS increases, BH load becomes the limiting factor in our simulation setup with 20 MHz channel bandwidth. Unlike RAN load, any CS increase for the VDSL2 site will always increase the BH load. Indeed, when additional users are scheduled within VDSL2 site, the user-data for the additional users will be added to the BH load regardless of the SE improvement. Once BH load is higher than RAN load, any CS increase will increase the overall load for VDSL2 sites which introduces extra cost in the utility function i.e. reduction in payoff for the VDSL2 site. When BH load is the limiting factor, VDSL2 site only enters into a CoMP set when the additional payoff for other SCs with fiber is greater than the payoff loss for the VDSL2 site. In other words, when BH load is taken into account i.e. for LBH-GA model, it is harder to get BH-limited SCs within CoMP clusters. Overall, without hotspots, L-GA achieves similar results to SE-GR and LBH-GA achieves marginally less CS due to not promoting CoMP on sites with VDSL2.

We run further simulations in RN with hotspot scenario for different rates of fiber connection available in the network. Seven different fiber/VDSL2 availability rates are considered and 100

snapshots of simulations are run for each scenario. Fig. 4 shows the average CS for each VDSL2 rate in a hotspot scenario where L-GA CS is consistently higher than SE-GR. This is in-line with HN simulations and the clustering snapshot shown in Fig. 3 where L-GA CS is increased when there is high load to improve SE and reduce the load. LBH-GA starts with the same CS as L-GA with 0% VDSL2 availability and average CS is reduced as the VDSL2 rate increases. LBH-GA tends to form clusters without the SCs with VDSL2 for the same reasons we discussed in RN without hotspot scenario. As shown in Fig. 5, a similar trend is observed in average SE, following average achieved CS as expected. Intuitively, increased CS helps in eliminating further inter-cell interference and hence improve SE. Fig. 6 depicts the unsatisfied UE count for each of the algorithms at different VDSL2 rate scenarios. L-GA model reduces the unsatisfied users by 80.6% when compared to SE-GR model when there is no SC with VDSL2 connection. As the VDSL2 rate increases, unsatisfied users increase in all models as expected, however LBH-GA model results in the lowest unsatisfied users with 41.7% and 18.4% less unsatisfied users when compared to SE-GR and L-GA, respectively, in the case when all SCs are connected with VDSL2. LBH-GA model achieves a better load-balanced network with less unsatisfied users while CS is kept low and hence low computational complexity for CoMP deployment. Similar to unsatisfied UEs, system throughput is also significantly improved in LBH-GA model when compared to SE-GR model as depicted in Fig. 7. An average of 49.9% increase in overall system throughput is observed with LBH-GA when compared to SE-GR across all BH scenarios. As the VDSL2 rate increase, LBH-GA throughput gets better when compared to L-GA as LBH-GA model clustering takes BH availability into account where SCs with VDSL2 is not preferred in clusters of highly loaded cells. LBH-GA achieves 21.9% higher overall throughput when compared to L-GA in the case when all SCs have VDSL2 BH.

We further look at an example scenario of 50% VDSL2 rate in RN with hotspot and analyze the details of each sub-game actions (i.e. SC merge/split/transfer actions and UE transfers) and the changes in SE, unsatisfied UEs and game payoff during the iterations. Fig. 8 shows the changes in the average number of unsatisfied UEs and the total number of each game action at each iteration for the 100 snapshots run in this scenario. At the start of the game, the average number of unsatisfied UEs are sharply reduced as the initial clusters are formed and SE is improved with merge operations. Later iterations of merge operations only give marginal improvements and other game actions start increasing. SC transfer actions are significantly high at the next stage where unsatisfied users are further reduced significantly. It can be noted that the number of split operations is relatively low when compared to other SC game actions. UE transfer actions are also relatively high numbers and can be controlled with δ_{Δ} parameter to allow only the most significant UE transfer actions, as discussed in Section V-D. Fig. 9 shows the average changes in SE and the number of unsatisfied UEs at each iteration. SE is increased sharply with the initial merge actions but reduced marginally in the later actions. This is due to the high associated priority on load balancing actions which may not be necessarily the best action for increasing SE. The number of unsatisfied UEs

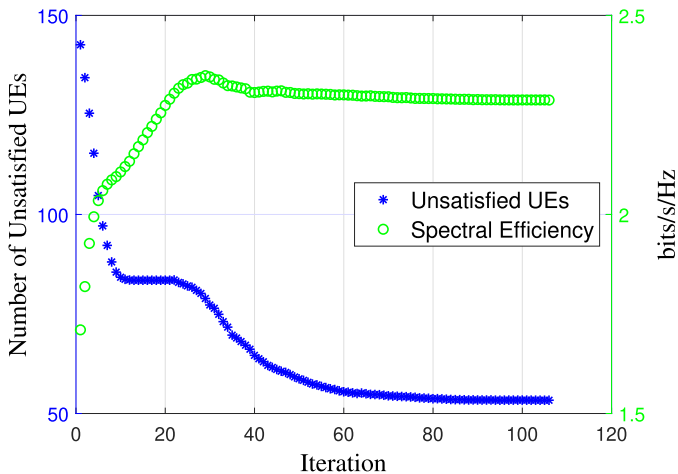


Fig. 9. LBH-GA Unsatisfied UEs vs SE for 50% VDSL2 rate.

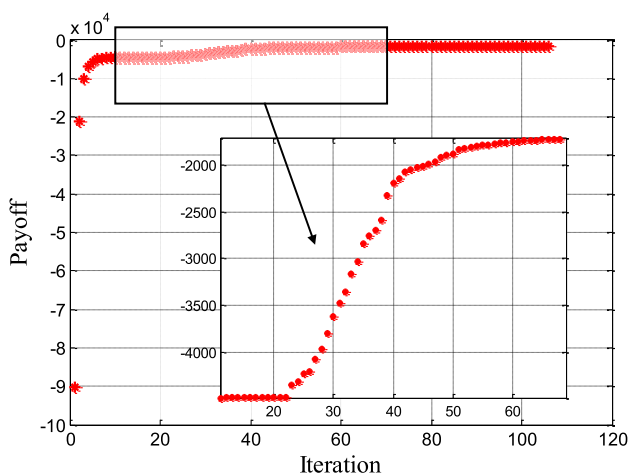


Fig. 10. LBH-GA Payoff for 50% VDSL2 rate.

continues to reduce at each game action. Overall system payoff is depicted in Fig. 10 where a similar pattern to the number of unsatisfied UEs is observed where a sharp improvement is observed in the initial merge actions and it continues to improve in smaller intervals in following game actions.

The resulting overall load distribution of all SCs in all three algorithms is shown in Fig. 11 a for the 50% VDSL2 rate case. LBH-GA model clearly achieves better load distribution owing to the traffic transfer to lightly loaded SCs. Fig. 11 b shows the BH load distribution for all SCs, and it is clear that BH load increases sharply when CoMP is enabled as user data needs to be available in multiple SCs in our JT-CoMP scenario. A significantly better BH load distribution with a low number of SCs with high load is achieved with LBH-GA resulting in better system throughput and higher capacity.

VII. CONCLUSION

We have presented a novel low-complexity, multi-objective clustering model in the MU JT-CoMP scenario where SE, RAN load and BH load are optimized collectively. An SC merge/split/transfer coalitional sub-game and a UE transfer coalitional sub-game are designed. Game properties, complexity

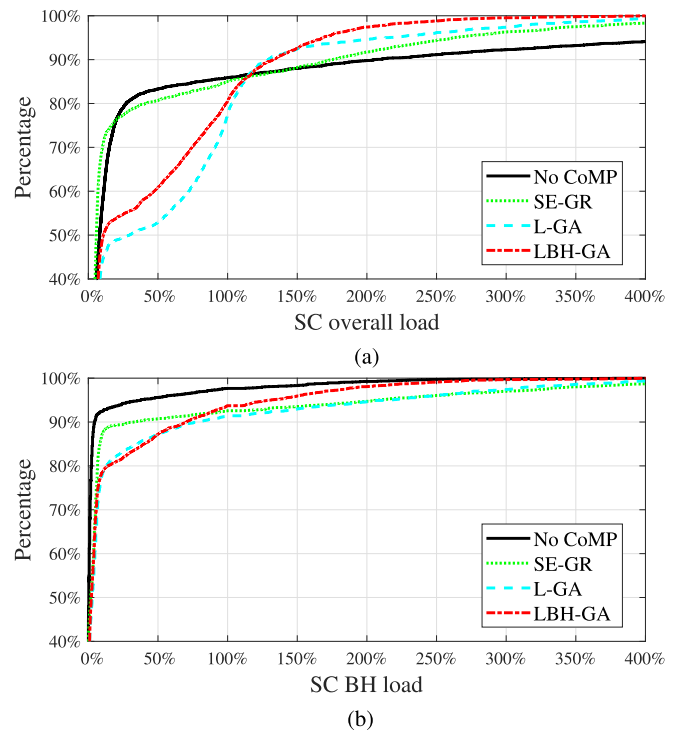


Fig. 11. SC load distribution for 50% VDSL2 rate.

and stability analysis are presented. It is shown that our novel LBH-GA algorithm is a low complexity model that is scalable and always converges to a final optimum cluster. Simulation results are compared to a RAN load-aware model (L-GA) and an SE based greedy (SE-GR) algorithm to show the impact of BH awareness. We show that LBH-GA successfully forms clusters dynamically around the hotspots and excludes BH limited SCs when possible to improve the SE and reduce overall load. In a hotspot scenario where throughput demand is higher than the overall capacity, the average system throughput is increased by 49.9% with LBH-GA when compared to the SE-GR model. The average throughput is also increased by 21.9% when compared to the L-GA model in the case of all SCs being BH-limited (VDSL2). LBH-GA model is also effective in scenarios without hotspots, dynamically adjusting the CS based on BH availability and load conditions. Our presented model provides a low complexity, stable framework where it can be enhanced further with improved utility functions to include additional network objectives and provide the right balance between CoMP overhead costs and various objectives based on network requirements.

REFERENCES

- [1] Q. C. Li, H. Niu, A. T. Papanthassiou, and G. Wu, "5G network capacity: Key elements and technologies," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 71–78, Mar. 2014.
- [2] S. W. H. Shah, A. N. Mian, S. Mumtaz, and J. Crowcroft, "System capacity analysis for ultra-dense multi-tier future cellular networks," in *IEEE Access*, vol. 7, pp. 5050503–50512512, 2019.
- [3] The 3rd Generation Partnership Project (3GPP), "Coordinated multi-point operation for LTE physical layer aspects," TR 36.819 R11 v11.2.0, Sep. 2013.
- [4] J. A. P. Pérez, F. Riera-Palou, and G. Femenias, "Multi-objective optimization of coordinated multipoint-aided MIMO-OFDMA systems with frequency reuse," *IEEE Access*, vol. 5, pp. 1515448–15467467, 2017.

- [5] Qualcomm, "How CoMP Can Extend 5G NR to high capacity and ultra-reliable communications," 2018. [Online]. Available: <https://www.qualcomm.com/documents/how-comp-can-extend-5g-nr-high-capacity-and-ultra-reliable-communications>
- [6] 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.
- [7] W.-S. Liao, M. G. Kibria, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Coordinated multi-point downlink transmission for dense small cell networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 431–441, Jan. 2019.
- [8] R. Irmer *et al.*, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, *IEEE*, vol. 49, no. 2, pp. 102–111, Feb. 2011.
- [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] S. Bassoy, H. Farooq, M. A. Imran, and A. Imran, "Coordinated multi-point clustering schemes: A survey," *IEEE Commun. Surv. Tut.*, vol. 19, no. 2, pp. 743–764, Apr.–Jun. 2017.
- [11] R. Karmakar, S. Chattopadhyay, and S. Chakraborty, "A learning-based dynamic clustering for coordinated multi-point (CoMP) operation with carrier aggregation in LTE-advanced," in *Proc. 10th Int. Conf. Commun. Syst. Netw.*, 2018, pp. 283–290.
- [12] M. Karavolos, V. I. Tatsis, D. N. Skoutas, N. Nomikos, D. Vouyioukas, and C. Skianis, "A dynamic hybrid clustering scheme for LTE-A networks employing CoMP-DPS," in *Proc. 22nd Int. Workshop Comput. Aided Model. Des. Commun. Links Netw.*, 2017, pp. 1–5.
- [13] Z. Zhang, N. Wang, J. Zhang, X. Mu, and K. M. Wong, "Cooperation resource efficient user-centric clustering for QoS provisioning in uplink CoMP," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun.*, 2017, pp. 1–5.
- [14] T. M. Shami, D. Grace, A. Burr, and M. D. Zakaria, "User-centric JT-CoMP clustering in a 5G cell-less architecture," in *Proc. 29th IEEE Annu. Int. Symp. Pers. Indoor Mobile Radio Commun.*, 2018, pp. 177–181.
- [15] L. Li, C. Yang, M. E. Mkiramweni, and L. Pang, "Intelligent scheduling and power control for multimedia transmission in 5G CoMP systems: A dynamic bargaining game," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 7, pp. 1622–1631, Jul. 2019.
- [16] T. Biermann, L. Scalia, C. Choi, W. Kellerer, and H. Karl, "How backhaul networks influence the feasibility of coordinated multipoint in cellular networks," *IEEE Commun. Mag.*, vol. 51, no. 8, pp. 168–176, Aug. 2013.
- [17] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G backhaul challenges and emerging research directions: A survey," *IEEE Access*, vol. 4, pp. 1743–1766, 2016.
- [18] G. Song, W. Wang, D. Chen, and T. Jiang, "KPI/KQI-driven coordinated multipoint in 5G: Measurements, field trials, and technical solutions," *IEEE Wireless Commun.*, vol. 25, no. 5, pp. 23–29, Oct. 2018.
- [19] J. A. Pastor-Pérez, F. Riera-Palou, and G. Femenias, "Analytical network-wide optimization of CoMP-aided MIMO-OFDMA irregular networks with frequency reuse: A multiobjective approach," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2552–2568, Mar. 2019.
- [20] A. Marotta *et al.*, "Performance evaluation of CoMP coordinated scheduling over different backhaul infrastructures: A real use case scenario," in *Proc. IEEE Int. Conf. Sci. Elect. Eng.*, 2016, pp. 1–5.
- [21] A. Marotta *et al.*, "Reducing comp control message delay in pon backhauled 5G networks," in *Proc. 23th Eur. Wireless Conf.*, 2017, pp. 1–5.
- [22] V. N. Ha, L. B. Le and N. Dào, "Coordinated multipoint transmission design for cloud-RANs with limited fronthaul capacity constraints," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7432–7447, Sep. 2016.
- [23] K.-G. Nguyen, Q.-D. Vu, M. Juntti, and L.-N. Tran, "Globally optimal beamforming design for downlink CoMP transmission with limited backhaul capacity," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 3649–3653.
- [24] S. Chen, T. Zhao, H.-H. Chen, and W. Meng, "Downlink coordinated multipoint transmission in ultra-dense networks with mobile edge computing," *IEEE Netw.*, vol. 33, no. 2, pp. 152–159, Mar./Apr. 2018.
- [25] M. Deghel, E. Bastug, M. Assaad, and M. Debbah, "On the benefits of edge caching for MIMO interference alignment," in *Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun.*, 2015, pp. 655–659.
- [26] A. Liu and V. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.
- [27] Y.-J. Yu, W.-C. Tsai, and A.-C. Pang, "Backhaul traffic minimization under cache-enabled CoMP transmissions over 5G cellular systems," in *Proc. IEEE Global Commun. Conf.*, 2016, pp. 1–7.
- [28] R. Hou, J. Cai, and K.-S. Lui, "Distributed cache-aware CoMP transmission scheme in dense small cell networks with limited backhaul," *Comput. Commun.*, vol. 138, pp. 11–19, 2019.
- [29] S. Chen, Y. Wang, J. Yu, N. Wang, and Y. Yan, "User association in cache-enabled ultra dense network with JT CoMP," in *Proc. IEEE 3rd Adv. Inf. Technol., Electron. Automat. Control Conf.*, 2018, pp. 964–968.
- [30] S. Bassoy, M. J. Jaber, M. A. Imran, and P. Xiao, "Load aware self-organising user-centric dynamic comp clustering for 5G networks," *IEEE Access*, vol. 4, pp. 2895–2906, 2016.
- [31] L. Liu, Y. Zhou, V. Garcia, L. Tian, and J. Shi, "Load aware joint CoMP clustering and inter-cell resource scheduling in heterogeneous ultra dense cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2741–2755, Mar. 2018.
- [32] S. Bassoy, M. A. Imran, S. Yang, and R. Tafazolli, "A load-aware clustering model for coordinated transmission in future wireless networks," *IEEE Access*, vol. 7, pp. 9292693–92708708, 2019.
- [33] The 3rd Generation Partnership Project (3GPP), "Evolved universal terrestrial radio access network (E-UTRAN); self-configuring and self-optimizing network (SON) use cases and solutions," TR 36.902 R9 v9.3.1, Mar. 2011.
- [34] The 3rd Generation Partnership Project (3GPP), "Self-Organizing Networks (SON) for 5G Networks-Release 16," TS 28.313 v0.2.0, Nov. 2019.
- [35] The 3rd Generation Partnership Project (3GPP), "Telecommunication Management; study on the Self-Organizing Networks (SON) for 5G Networks (Release 16)," TR 28.861 v16.0.0, Dec. 2019.
- [36] The 3rd Generation Partnership Project (3GPP), "Study on SON for eCoMP," TR 36.742 R15 v2.0.0, Jun. 2017.
- [37] A. Papadogiannis, D. Gesbert, and E. Hardouin, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," *IEEE Int. Conf. Commun.*, 2008, pp. 4033–4037.
- [38] Next Generation Mobile Networks Alliance (NGMN) Alliance, "Small Cell Backhaul Requirements," White Paper, Jun., 2012.
- [39] M. Jaber, M. Imran, R. Tafazolli, and A. Tukmanov, "An adaptive backhaul-aware cell range extension approach," in *Proc. IEEE Int. Conf. Commun. Workshop*, 2015, pp. 74–79.
- [40] W. Saad, Z. Han, M. Debbah, and A. Hjørungnes, "A distributed coalition formation framework for fair user cooperation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4580–4593, Sep. 2009.
- [41] Z. Zhao, M. Peng, Z. Ding, C. Wang, and H. V. Poor, "Cluster formation in cloud-radio access networks: Performance analysis and algorithms design," in *Proc. IEEE Int. Conf. Commun.*, 2015, pp. 3903–3908.
- [42] E. Katranaras, M. A. Imran, and M. Dianati, "Energy-aware clustering for multi-cell joint transmission in LTE networks," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2013, pp. 419–424.
- [43] F. Guidolin, L. Badia, and M. Zorzi, "A distributed clustering algorithm for coordinated multipoint in LTE networks," in *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 517–520, Oct. 2014.
- [44] S. Brueck, L. Zhao, J. Giese, and M. A. Amin, "Centralized scheduling for joint transmission coordinated multi-point in LTE-Advanced," in *Proc. Int. ITG Workshop Smart Antennas*, 2010, pp. 177–184.
- [45] A. Papadogiannis and G. C. Alexandropoulos, "The value of dynamic clustering of base stations for future wireless networks," in *Proc. Int. Conf. Fuzzy Syst.*, 2010, pp. 1–6.
- [46] I. Viering, M. Döttling, and A. Lobinger, "A mathematical perspective of self-optimizing wireless networks," in *Proc. IEEE Int. Conf. Commun.*, 2009, pp. 1–6.
- [47] N. Jindal and A. Lozano, "A unified treatment of optimum pilot overhead in multipath fading channels," *IEEE Trans. Commun.*, vol. 58, no. 10, pp. 2939–2948, Oct. 2010.
- [48] The 3rd Generation Partnership Project (3GPP), "Evolved universal terrestrial radio access (E-UTRA); base station (BS) radio transmission and reception," TS 36.104, Jun. 2011.
- [49] L. Lovász, *Combinatorial Problems and Exercises*. American Mathematical Soc., vol. 361, 2007.
- [50] W. Saad, Z. Han, M. Debbah, A. Hjørungnes, and T. Basar, "Coalitional game theory for communication networks," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 77–97, Sep. 2009.
- [51] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, pp. 171–220, 2012.
- [52] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2014, pp. 1096–1104.
- [53] ITU-R M.2135.1, "Guidelines for evaluation of radio interface technologies for IMT-Advanced," *ITU-R*, Geneva, Switzerland, Dec. 2009.