Performance Analysis of Dynamic Re-Clustering and Resource Allocation in Ultra Dense Network

WENLE BAI[ID][1], **YAOMIN LI**[1], **TONG YAO**[1], **AND HAIJUN ZHANG**[ID][2], **(Senior Member, IEEE)**

[1]North China University of Technology, Beijing, China
[2]University of Science and Technology Beijing, Beijing, China

Corresponding author: Wenle Bai (bwl@ncut.edu.cn)

ABSTRACT Coordinated multi-point (CoMP) is a key technology to mitigate inter-cell interference, which can significantly improve the cell edge performance and system throughput. In addition, the choice of the cells within cluster will directly affect the effect of CoMP, and especially, it may cause the increase of the number of dissatisfied users in the system. To minimize dissatisfied users in the ultra-dense network (UDN) application scenario, based on the control-data separation architecture model, a clustering algorithm is studied to improve the SINR and throughout. Then, load balancing is developed and the overall percentage of unsatisfied users are reduced obviously. Furthermore, a dynamic resource allocation scheme is proposed by defining a factor α and applied to optimize the performance of system, which is compared with load balancing under the conditions of a different clustered size and density of users. The experiments prove that the performance of load balancing is more sensitive to the change of clustered size than resource allocation, and the load balancing has overwhelming advantage compared with resource allocation in the dense deployment scenario. However, when the density of users becomes sparser, resource allocation performs better, which gives important meanings for future UDN.

INDEX TERMS CoMP, UDN, load balancing, resource allocation.

I. INTRODUCTION

With the rapid development of wireless communication networks, the number of mobile terminal users are exponentially increasing, and huge data demand in cellular networks will be one of the biggest challenges in the next decade. The next 5G deployment requires that the density of cities and hotspots with heavy traffic will reach $20Tbps/km^2$. Besides other requirements such as low latency and lower interrupt rates are also proposed in ITU-R Report M.2320 [1], as one of the key technology UDN(Ultra-Dense Network) trends to satisfy the 5G requirements for solving above problems. Compared to the traditional cellular networks, the density of APs(Wireless Access Points) in UDN is more denser. The coverage of each AP becomes smaller with the dense deployment of APs, which leads to higher interruption rates and more frequent handover [2]. So how to save bandwidth and expand capacity in the case of ensuring user perception has become a major problem that 5G needs to solve.

The concept of cellular is a breakthrough in improving the spectrum resource shortage and user capacity limitation without changing any major technology when the available spectrum resources are limited. Using separated wireless channels over the same carrier frequency to cover different ranges can make the interference between the channels no longer annoying [4]. Since then, cellular communication has entered our field of vision, and the architecture of cellular communications has also not changed any more [5].

However, with the development of the physical layer, the service data rate of mobile communications will increase by 100 times every six to seven years. It is expected that by 2020, the expected data rate will be 100 to 1000 times than now [6]. Next-generation mobile communications promise to provide faster rates and blocky communication access services for wide-area regions [7]. Correspondingly, there are still many challenges. First, because of the limited transmission power, the cover distance of traditional cells can't meet the required data rate. Second, because the frequency of switching is too fast in the high-speed mobile scenario, so the traditional network can't provide service normally. Last but not the least, the future wireless communication system
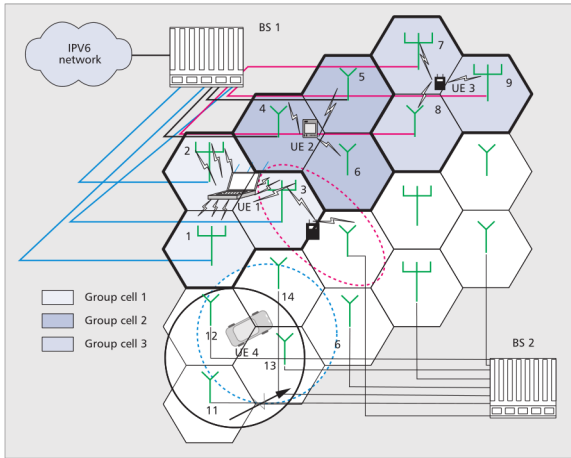
FIGURE 1. Community group architecture.

is carried in a frequency band higher than 2 GHz, and the wireless signal of a large-scale antenna becomes an important factor causing serious edge effects [8].

In order to solve the above problems, some cooperative technologies have been emerging, such as relay, DAS(distributed antenna system) and multi-cell cooperation. All of these technologies transform traditional cellular network systems into collaborative network systems [9], [10]. As shown in Figure 1, collaborative communication dynamically changes the abstraction of wireless networks and provides important potential benefits for wireless communication networks. In contrast to unicast transmission, coordinated communication may join multiple points of transmission at the receiver.

CoMP as one of the new technologies is especially important in UDN(ultra-dense network) deployments. Because CoMP utilizes the shared data between coordinated TPs(transmission points) such as CSI(channel state information), resource allocation and serviced data [11]. So the interference within cells can be mitigated and even converted into useful signals at terminal. However, CoMP within the cells requires huge signal processing, complex resource allocation and beamforming design among BSs (Base Stations), which are very complicated and all users' data exchange between BSs requires high backhaul bandwidth. This results in costly overhead.

For the purpose of reducing these overhead, small-scale cells need be performed only within the cluster. Because the size of coordinated clusters that are too small can't get complete CoMP gain, while too large coordinated clusters will result in increased CSI and backhaul bandwidth for users and base station switching [12]. So the size of the cluster should be kept at an optimal level while being dynamically adjusted according to channel state and characteristics of users.

There are a large amount of literature studying about load balancing for improving the user's perception, but there is no few research about exploring the resource allocation with the

aim of improving the clustering. The purpose of this paper is looking for how to improve the performance under UDN. Firstly, a self-organized user-centered clustering algorithm is achieved in more detail according to precious research [13]. Secondly, load balancing is developed and applied to reduce the number of unsatisfied users. Furthermore a dynamic resource allocation scheme is proposed by defining a factor α and employed to optimize the performance of system, which is compared with load balancing under the conditions of different clustered size and density of users. Finally, the performance is analyzed in detail and several important conclusions are obtained, which are meaningful for the deployment of UDN in the future 5G.

This paper is arranged as follows: section II introduces the CDSA system framework. Section III implements the clustering algorithm. Section IV develops the load balancing algorithm. Section V proposes a dynamic resource allocation algorithm. Section VI presents performance comparison between load balance and resource allocation. The conclusion is followed in section VII.

II. SYSTEM MODEL

In conventional cellar system, a wireless channel can ensure full coverage regardless of how the spatial and temporal demand of service change [14]. However, it will bring some problems because of requiring the tight coupling between data access points and network, heterogeneous deployments and network desertification. To overcome these issues, the control and data planes are separated logical as a promising technology by performing data services under a coverage layer [15].

The CDSA (control-data separation architecture) model used in this paper is an emerging model for the RAN(Radio Access Network) architecture in 5G networks [15]. The key idea is that the signals for full coverage is separated from the signals to support high speed transmission. The MBS(Macro Base Station) provides coverage of wide area and handles of most control signaling, and SCs(Small Cells) under the macro station provide the services of required data. The Multipoint CCU(Coordinated Control Unit) enhances the functionality of the MBS. All SCs within the coverage of the MBS are connected to the macro station via the wireless backhaul network as shown in Figure 2. The CCU on the MBS can be deployed in the SON(Self-Organization Network) framework to provide a central clustering option on the SC side [17]. Supposed that the CCU can also handle central precoding design and baseband processing based on the selected cluster. Since every SCs is connected to the related MBS, high bandwidth backhaul is not required among the SCs [18], which is also applied to the Cloud-RAN architecture [19], [21].

As shown in Figure 2, the signals received by each user in the cluster can be written as:

$$y = HWx + n, \quad H \in C^{R \times M}, \quad W \in C^{T \times R} \quad (1)$$

$$h_k = [h_{k1} h_{k2} \ldots \ldots h_{kT}] \quad (2)$$

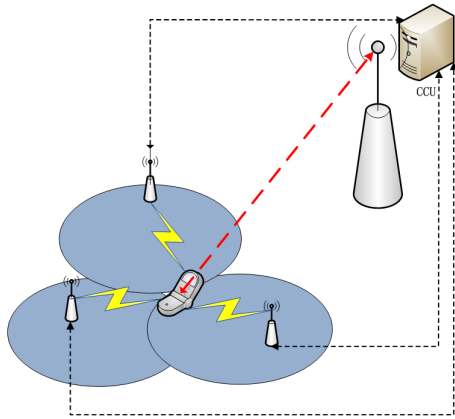$$H = [h_1 h_2 \ldots \ldots h_R]^T \quad (3)$$

FIGURE 2. Control data separation architecture.

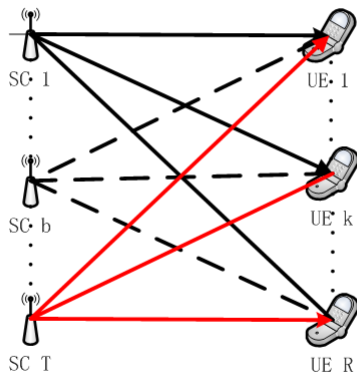**FIGURE 3.** Downlink multi-user join-type multi-point collaboration.

In the figure 3, the beamforming vector at the user are:

$$w_k = [w_{1k}w_{2k}......w_{Tk}]^T \tag{4}$$

$$W = [w_1w_2.......w_R] \tag{5}$$

The received signal of the user can be represented by:

$$y_k = h_k^{C_M^k}w_k^{C_M^k}x_k + \sum_{i \in C_M^k/k} h_k^{C_M^k}w_i^{C_M^k}x_i$$
$$+ \sum_{j \in K/C_K^k} h_k^{M/C_M^k}w_j^{C_M^k}x_j + n_k \tag{6}$$

SINR at each UE is represented by:

$$SINR_k = \frac{\left| h_k^{C_M^k}w_k^{C_M^k}x_k \right|^2}{\left| \sum_{j \in K/C_K^k} h_k^{M/C_M^k}w_j^{C_M^k}x_j \right|^2 + |n_k|^2} \tag{7}$$

Assume obtaining complete channel knowledge, all the PRBs allocated to the cell are assigned the same transmit power, and each SCs is assigned the same total transmit power. A ZF(Zero-Forced Precoding Algorithm) is applied to the Multipoint Coordination Control Center, which can eliminate the interference within the cluster. Therefore, the SINR

at the user can be written as[16]:

$$SINR_k = \frac{P_{Tx}\sum_{i \in C_M^k}|l_{ki}|^2}{P_{Tx}\sum_{j \in M/C_M^k}|l_{ki}|^2 + N_0B_{tot}} \tag{8}$$

$$PL = 36.7\log_{10}(d) + 22.7 + 26\log_{10}(f_c) \tag{9}$$

N_0 denotes the noise power spectral density and B_{tot} represents the bandwidth of the system. The channel coefficient l_{ki} consists of two parts, one is the static distance according to path loss and shadow fading, the other is the fast fading complex coefficient.

$$l_{ki} = p_{ki}s_{ki} \tag{10}$$

In the equation (10), p_{ki} is the distance part of path loss and shadow fading, s_{ki} is the fast fading complex coefficient. As it will be mentioned later in this paper, cluster selection is based on long-term average receive levels. Therefore, the part of the fast fading effect in equation(8) can be almost eliminated. Hence, equation(8) can further simplify the part of the cluster that eliminates fast fading as:

$$SINR_k = \frac{P_{Tx}\sum_{i \in C_M^k}|P_{ki}|^2}{P_{Tx}\sum_{j \in M/C_M^k}|P_{kj}|^2 + N_0B_{tot}} \tag{11}$$

$$PL = 36.7\log_{10}(d) + 22.7 + 26\log_{10}(f_c) \tag{12}$$

III. CLUSTERING ALGORITHM

On the basis of a user-centric clustering algorithm in [13], each UE_k is allocated its own clustering group according to the average received power. In this algorithm, The UEs transmit their received signal power level from SCs to the SC of optimal services, then transfer it to the CCU of MBS by the fiber backhaul. The CCU processes these signals and assigns SCs to each UE for cooperation.

For each UE_k, the average received power level, which is transmitted by all SCs in the MBS, can be acknowledged by CCU. The received power levels are averaged in time domain to eliminate the fading part. Therefore, p_{km} is divided into two parts: path loss and shadow fading:

$$p_{km}^{rx} = p_{km}^{tx}|p_{km}|^2, \quad m \in M \tag{13}$$

Where p_{km}^{tx} denotes the transmit power assigned to the cluster. p_{km} represents the average received power of UE_k in the cluster SC_m. And then every UE_k is sorted by p_{km}^{rx}:

$$P_{KM}^{RX} = \arg\max_m p_{km}^{rx}, \quad m \in M \tag{14}$$

Figure 4 shows the procedure of the multi-point CCU(cooperative control unit) in the central MBS performing small cell clustering based on the user's information.

IV. LOAD BALANCING

Considering the user-centered clustering algorithm in last section, a load balancing algorithm in [13] is further studied in detail, a mathematical framework in [20] is used for computing "cell load" and "unsatisfied users" under traditional networks. For convenience, it is implemented without considering the energy loss here.
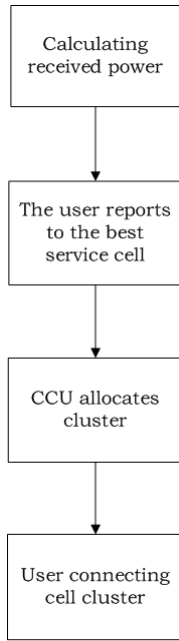
FIGURE 4. Procedure of the multi-point CCU performing small cell clustering.

FIGURE 5. Load balancing algorithm flow chart.

It is assumed that each SC has a total number of R_{tot} assigned to a PRB(physical resource block), here the bandwidth of each PRB is B_{RB}. The maximum throughput achievable by a PRB can is expressed as follows:

$$y_k = B_{RB}\log_2\left(1 + SINR_k\right) \qquad (15)$$

Here, it is assumed that a stable bit rate each UE_k need to require is denoted by d_k, so the average number of PRBs that each user need to require without a CoMP can be written as:$r_k^{NoCoMP} = d_k / y_k$. However, in the case of MU JT-CoMP (multi-user join multi-point cooperation), user data of UE_k can also be transmitted from other SCs in the cluster C_M^k. Therefore, resource blocks from each SC in its cluster are required. On the other hand, the same resources assigned to UE_k can be used by other $UE_s \in C_K^k$, which are arranged within the same cluster. Provided that the number of UEs sharing the same PRB in the same cluster is equal to the cluster size of UE_k, i.e. $\left|C_M^k\right| = \left|C_K^k\right| = n_m$. Therefore, the average number of "virtual" dedicated PRBs will be $r^{CoMP} = r^{NoCoMP} / n_m$. The actual dedicated PRB required for all UEs within the cluster is expressed as:

$$r^{CoMP} = \frac{d_k}{y_k n_m} \qquad (16)$$

$$L_m = \frac{\sum_{k \in S_K^m} r_k}{R_{tot}} \qquad (17)$$

In Equation 17, L_m is utilization efficiency of PRBs. if $L_m < 1$, the total PRBs can support users to achieve d_k, vice versa. It can be used to compute a term "dissatisfied users" under the SCs. Since every UE repeats connection with all SCs within the related cluster, the virtual number of UE associated with SCs within the cluster can be computed by
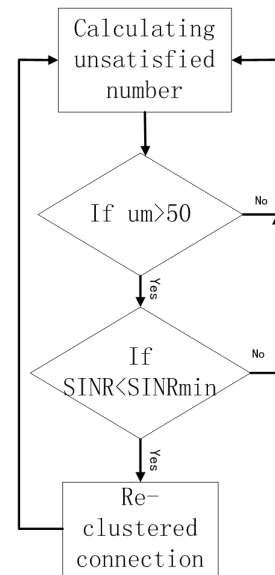
accumulating all number of associated UEs by the factor $\frac{1}{n_m}$. So the users' number s_m in the cluster can be expressed as:

$$s_m = \sum_{k \in S_K^m} \frac{1}{n_m} \qquad (18)$$

Therefore, the number of dissatisfied users u_m can be calculated by the equation(19):

$$u_m = \max\left(0, s_m\left(1 - \frac{1}{L_m}\right)\right) \qquad (19)$$

The procedure of load balancing reforming clusters is shown in figure 5. Here, a minimum threshold u_m of dissatisfied users' number is set to 50, and if its value over 50; then computing the SINR of UEs by equation (11), if the unsatisfied users' SINR are lower than SINRmin of the overloaded SCs, they will be reconnected to the light-load SCs to balance the load.

V. RESOURCE ALLOCATION ALGORITHM

The above load balancing algorithm is effective for reducing system dissatisfaction, but it requires the PRBs allocated by all SCs are same, which may cause obvious difference in users' perception among SCs. On the other hand, it needs CoMP to be complexity. However, in the case of constant bit rate, the SCs with more loads need to be allocated more PRBs. Motivated by this, a dynamic resource allocation algorithm is proposed as follows:

Based on the load of SCs, for convenience, the dynamic allocation factor α is defined as equation (20), moreover, allocated numbers of PRBs to SCs can be computed by equation (21).

$$\alpha = \frac{l_c}{l_{mtot}} \qquad (20)$$
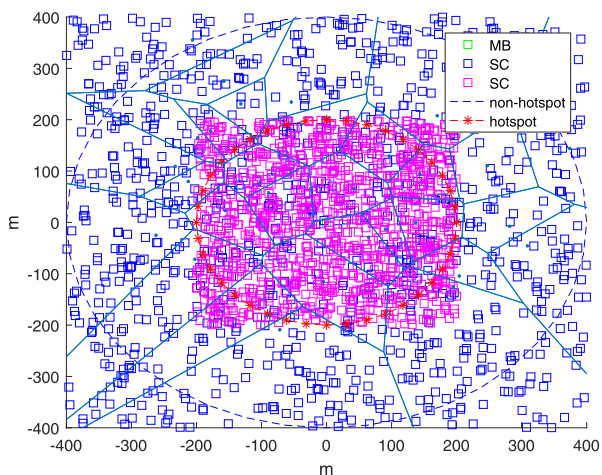
$$PRB_{fen} = PRB_{tot} \times \alpha \qquad (21)$$
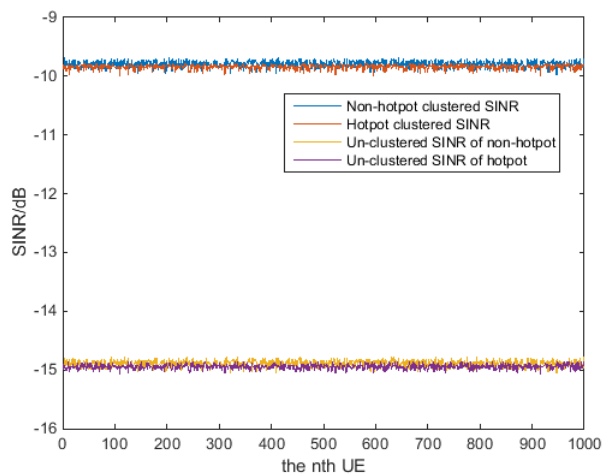
FIGURE 6. Simulation environment of the UDN.

FIGURE 8. PDF of unclustered SINR.

FIGURE 7. Comparison of SINR before and after clustering deployment.

FIGURE 9. PDF of clustered SINR.

Here, l_c is the number of load users connected to SCs, l_{mtot} is the number of total users in this system, PRB_{fen} is the number of PRBs allocated to SCs by MBS, PRB_{tot} is the number of total PRBs in MBS.

VI. NUMERICAL RESULTS

The simulation scenario is illustrated in figure 6, one MBS station is in the center of $400m^2$, 30 SCs are randomly distributed around the MBS. The boundaries of each SC are subject to Taylor polygons, regardless of overlapping coverage. UEs are randomly distributed in the area, which are divided into hot-spots and non-hot-spots. Compared with [18], the users' distribution is more denser for UDN. The hot-spot is in the central $200m^2$ range, the user density is $7.5 people/m^2$ and the user density outside the hot spot area is $2.5 people/m^2$.

Figure 7 displays the change of SINR of UDN when the clustering is applied. As seen in results, the overall user's receiving SINR both in hot-spots and non-hot-spots are significantly improved about 5dB. There are more detailed
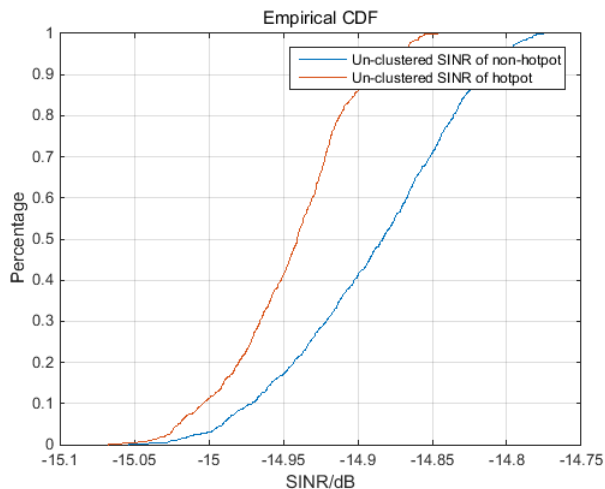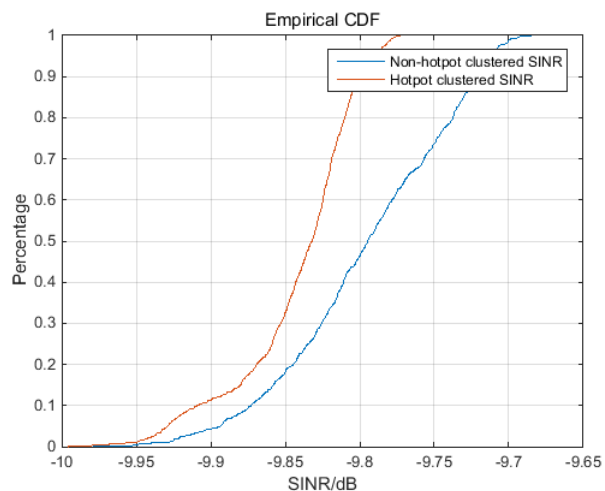
results shown in figure 8 and 9 respectively. Before clustering, most users' SINR is mainly concentrated at -15dB, but it's increased at -10dB after clustering. On the other hand, throughout of system are computed according to the equation (15), the results are depicted in figure 10 and 11. Unclustered throughout of system is mostly distributed ranging from 2.2×10^8 to $2.3 \times 10^8 kbps$, however, clustered throughout of system is mainly concentrated ranging from 6.9×10^8 to $7.2 \times 10^8 kbps$. The performance of system is improved obviously.

Simulation of Load balancing is computed according to equation (19) and it's demonstrated in figure 12 and figure 13 respectively. As shown in Figure 12, the overall number of unsatisfied users are relatively high before using the load balancing, the average number of unsatisfied users are more than 200, and the biggest number of unsatisfied users is more than 500. In more detail, there are a few SCs with less than 100 unsatisfied users in non-hot-spot, but most SCs have more than 200 unsatisfied users in hot-spot. Here, load balancing is applied to transfer the unsatisfied users from hot-spot to non-hot-spot to improve the performance of
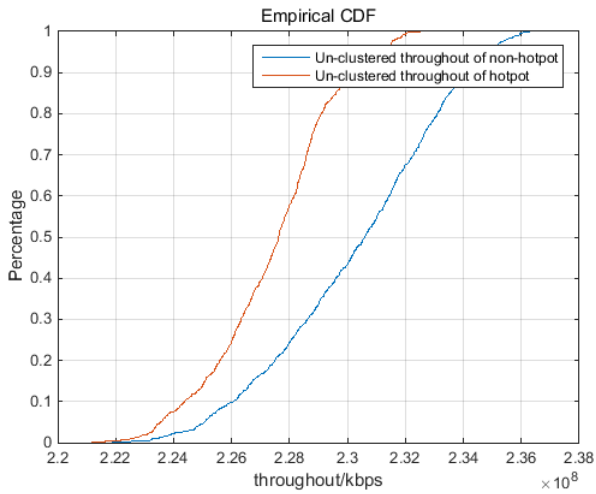
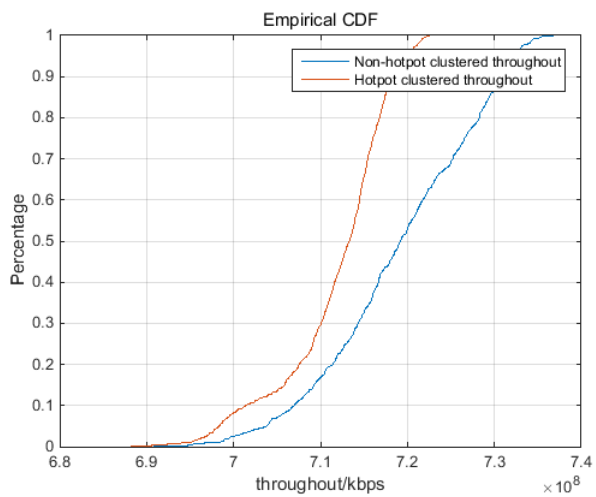**FIGURE 10.** PDF of unclustered throughout of system.
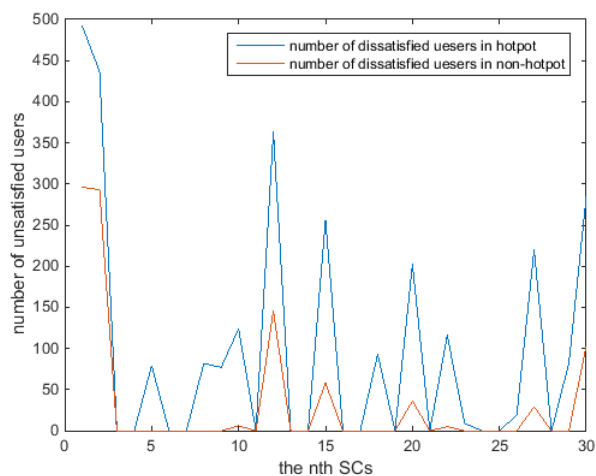

FIGURE 11. PDF of clustered throughout of system.

FIGURE 12. Dissatisfied users before load balance.

FIGURE 13. Dissatisfied users after load balance.
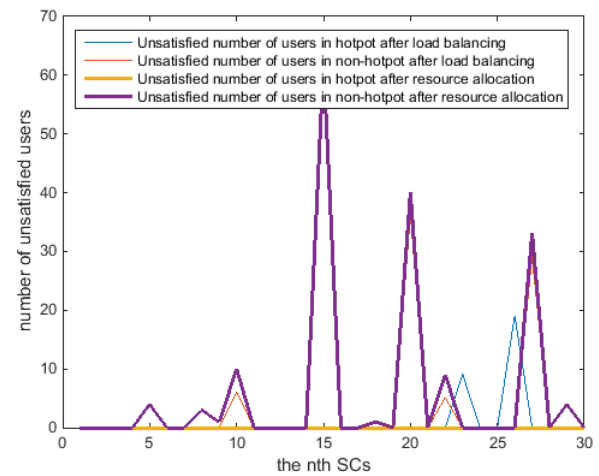
FIGURE 14. Resource allocation VS load balancing–One cluster of three SCs, 4000 UEs/400m^2.

non-hot-spot, the performance of hot-spot is improved more obviously than that in non-hot-spot.

In moreover work, resource allocation is applied to the above system according to equation (21), there are 5 simulated experiments to study and compare about unsatisfied users with that of load balancing under the conditions of different clustered size and density of users, the results are shown in figure 14, figure 15, figure 16 and figure 17 respectively. For being fairly compared with performance, one of the clustered size and density of users is varied in each experiment. The comparison of simulation works are discussed as follows:

1) COMPARISON 1- CLUSTERED SIZE OF 3 AND 4 SCs, 4000 UEs/400m^2

Figure 14 depicts the results under the condition of one cluster of 3 SCs. It is shown that the resource allocation can reduce the number of unsatisfied users to 0, which performs better than load balancing in hot-spot, but it has poorer performance than that of load balancing in non-hot-spot. When the size of one cluster increases to 4 SCs, the result is demonstrated in figure 15. Compared with figure 14 and

system. Compared figure 13 with figure 12, the dissatisfied users' numbers are reduced greatly, the biggest number is less than 60, some unsatisfied users in hot-spot are transferred to
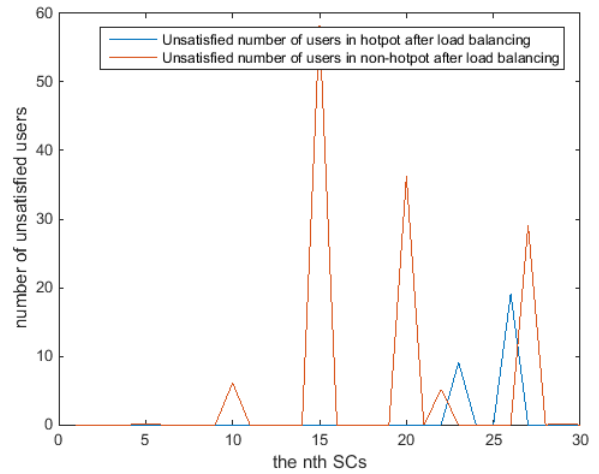
TABLE 1. Comparison of performance between load balancing and resource allocation on different clustered size.

	Load Balancing							
Clustered size	non-hot-spot				hot-spot			
	dissatisfied users		SCs with dissatisfied users		dissatisfied users		SCs with dissatisfied users	
	numbers	percentage	numbers	percentage	numbers	percentage	numbers	percentage
3	10	0.025%	2	6.6%	30	0.75%	2	6.6%
4	110	2.75%	6	20.6%	50	1.25%	3	10%

	Resource Allocation							
Clustered size	non-hot-spot				hot-spot			
	dissatisfied users		SCs with dissatisfied users		dissatisfied users		SCs with dissatisfied users	
	numbers	percentage	numbers	percentage	numbers	percentage	numbers	percentage
3	155	3.8%	7	23.3%	0	0	0	0
4	200	5%	8	26.7%	0	0	0	0

TABLE 2. Comparison of performance between load balancing and resource allocation on different users' density.

	Load Balancing							
Density of users	non-hot-spot				hot-spot			
	dissatisfied users		SCs with dissatisfied users		dissatisfied users		SCs with dissatisfied users	
	numbers	percentage	numbers	percentage	numbers	percentage	numbers	percentage
4000 UEs/400m^2	10	0.25%	2	6.7%	30	0.75%	2	6.7%
1500 UEs/400m^2	8	0.53%	4	13.3%	24	1.6%	2	6.7%
1000 UEs/400m^2	0	0	0	0	26	2.6%	2	6.7%

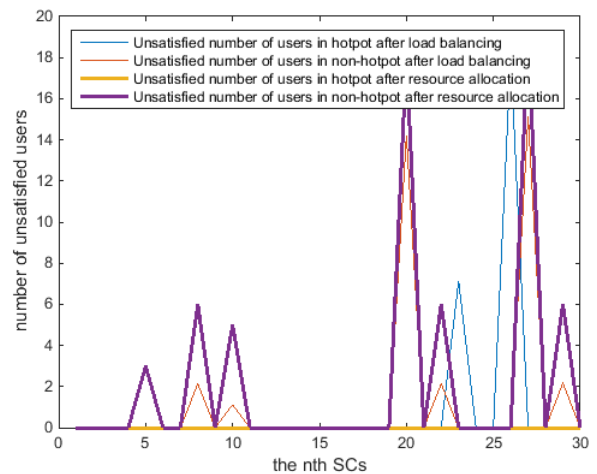|	Resource Allocation							
Density of users	non-hot-spot				hot-spot			
	dissatisfied users		SCs with dissatisfied users		dissatisfied users		SCs with dissatisfied users	
	numbers	percentage	numbers	percentage	numbers	percentage	numbers	percentage
4000 UEs/400m^2	155	3.87%	7	23.3%	0	0	0	0
1500 UEs/400m^2	58	2.86%	6	18.3%	0	0	0	0
1000 UEs/400m^2	0	0	0	0	0	0	0	0
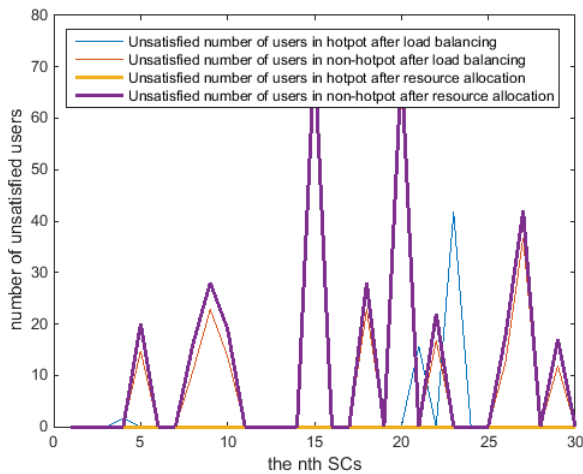
FIGURE 15. Resource allocation VS load balancing–One cluster of four SCs, 4000 UEs/400m^2- clustered size of 3 and 4 SCs, 4000 UEs/400m^2.

FIGURE 16. Resource allocation VS load balancing–One cluster of three SCs, 1500 UEs/400m^2.

figure 15, the unsatisfied number of users of load balancing are increased obviously both in hot-spot and non-hot-spot. However, the number of unsatisfied users of resource allocation keep constant in hot-spot and has a little increasing in non-hot-spot. The increasing of clustered size will cause more impact on load balancing than resource allocation. See more detail in Table 1.

2) COMPARISON 2- CLUSTERED SIZE OF 3 SCs, DENSITY OF 4000 UEs/400m^2,1500 UEs/400m^2,1000 UEs/400m^2

In this group of experiments, the clustered size is set to 3, the two algorithm are applied into dense/medium/sparse

deployment scenarios with the users' density equaling to 4000/400m^2, 1500/400m^2 and 1000/400m^2 respectively. The corresponding results are shown in figure 14, figure 16 and figure 17. In figure 14, for the system, there are about 40 unsatisfied users for load balancing, but there are about 150 unsatisfied users for resource allocation. In figure 16, when the users' density reduce to 1500/400m^2, there are about 32 unsatisfied users for load balancing, but there are about 58 unsatisfied users for resource allocation, the percentage of unsatisfied users decreased by 20% after load balancing, however, it is decreased by 63% after resource allocation.
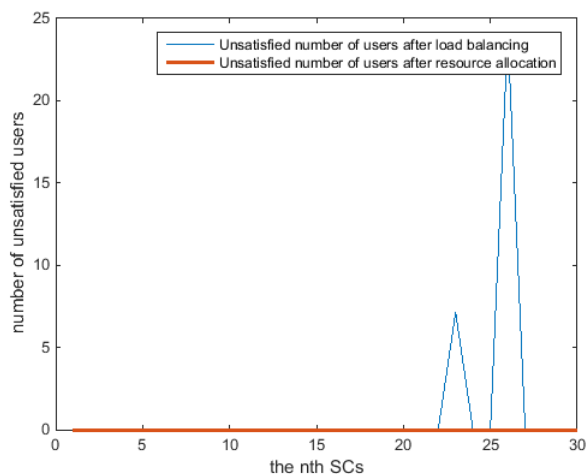
FIGURE 17. Resource allocation VS load balancing–One cluster of three SCs, 1000 UEs/ $400m^2$.

In figure 17, the users' density reduce to $1000/400m^2$, there are about 26 unsatisfied users for load balancing, but there are no unsatisfied users for resource allocation, the number of unsatisfied users has little change for load balancing but it is a surprising change for resource allocation. When the density of users become sparser, resource allocation performs better. The detail percentage of unsatisfied users and SCs are listed in Table 2.

VII. CONCLUDE

This paper focuses on the performance of load balance and resource allocation with different density of users and clustered size in UDN. Based on the CDSA model, a clustering algorithm is studied to improve the SINR by 5dB and the throughout increase from 2.2×10^8kbps to 6.9×10^8kbps. Load balancing is developed and the overall percentage of unsatisfied users are reduced by 80%. Furthermore, a dynamic resource allocation scheme is proposed and applied to optimize the system performance, which is compared with load balancing under the conditions of different clustered size and density of users. The experiments prove that the density of users and clustered size have more impact on the performance of load balancing than that of resource allocation. As the size of clustered increasing, the performance of load balancing become weaker than that of resource allocation, the former is more sensitive to the size of cluster than the latter. In the dense deployment scenario, the load balancing has overwhelming advantage compared to resource allocation. However, when the density of users become sparser, resource allocation performs better, which are likely cases for the deployment of UDN in the future 5G.

REFERENCES

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[2] D. Wubben *et al.*, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[3] P. Rost *et al.*, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[4] A. G. Armada, M. Sánchez-Fernández, and R. Corvaja, "Waterfilling schemes for zero-forcing coordinated base station transmission," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Honolulu, HI, USA, Nov./Dec. 2009, pp. 1–5.

[5] *Self-Configuring and Self-Optimising Network Use Cases and Solutions (Release 9) Version 9.1.0*, document TR 32.902, 3GPP, Mar. 2010.

[6] P. Marsch and G. Fettweis, "Static clustering for cooperative multi-point (CoMP) in mobile communications," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2011, pp. 1–6.

[7] R. Irmer *et al.*, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[8] X.-J. Wang, H. Tian, F. Jiang, X.-Y. Li, X.-J. Hong, and T.-R. Li, "Cell-cluster based traffic load balancing in cooperative cellular networks," in *Proc. 7th IEEE Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2010, pp. 1–5.

[9] X. Lu, E. Kunnari, J. Leinonen, O. Piirainen, M. Vainikka, and M. Juntti, "LTE uplink power control and base station antenna down tilt in a 3D channel model," in *Proc. Eur. Wireless Conf. (EW)*, 2010, pp. 377–381.

[10] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," *Wireless Netw.*, vol. 14, pp. 103–120, Jan. 2008.

[11] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-advanced [coordinated and distributed MIMO]," *IEEE Wireless Commun.*, vol. 17, no. 3, pp. 26–34, Jun. 2010.

[12] W. Feng, Y. Wang, Y. Li, X. Xu, and J. Wang, "Coordinated power allocation for generalized multi-cluster distributed antenna systems," *IEICE Trans. Commun.*, vol. E94-B, no. 6, pp. 2656–2659, Sep. 2011.

[13] S. Bassoy, M. 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.

[14] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, "FluidNet: A flexible cloud-based radio access network for small cells," *IEEE/ACM Trans. Netw.*, pp. 915–928, 2016.

[15] P. Baracca, F. Boccardi, and V. Braun, "A dynamic joint clustering scheduling algorithm for downlink CoMP systems with limited CSI," in *Proc. ISWCS*, Aug. 2012, pp. 830–834.

[16] *Guidelines for Evaluation of Radio Interface Technologies for IMT-Advanced*, document ITU-R M.2135.1, ITU-R, Geneva, Switzerland, Dec. 2009.

[17] J. Hoydis, M. Kobayashi, and M. Debbah. (Mar. 2010). "On the optimal number of cooperative base stations in network MIMO systems." [Online]. Available: https://arxiv.org/abs/1003.0332

[18] G. Cili, H. Yanikomeroglu, and F. R. Yu, "Cell switch off technique combined with coordinated multi-point (CoMP) transmission for energy efficiency in beyond-LTE cellular networks," in *Proc. Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 5931–5935.

[19] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to empower SON with big data for enabling 5G," *IEEE Netw.*, vol. 28, no. 6, pp. 27–33, Nov. 2014.

[20] I. Viering, M. Dottling, and A. Lobinger, "A mathematical perspective of self-optimizing wireless networks," in *Proc. Int. Conf. Commun. (ICC)*, Jun. 2009, pp. 1–6.

[21] J. Zhu, M. Zhao, and S. Zhou, "An optimization design of ultra dense networks balancing mobility and densification," *IEEE Access*, vol. 6, pp. 32339–32348, 2018.

WENLE BAI was born in Shanxi, China, in 1967). He received the M.S. degree in automation engineering from Chongqing University in 1997 and the Ph.D. degree in communication engineering from the Beijing University of Posts and Telecommunication of China in 2006. He is currently an Associate Professor with the Department of Communication Engineering, North China University of Technology. His research interests are in the areas of wireless communication systems, statistical signal processing, and multi-user communications. He has published about 20 papers in the correlative fields.
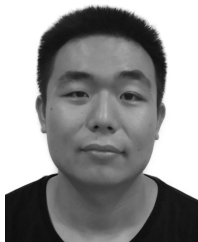
YAOMIN LI was born in Anhui, China, in 1996. She received the B.E. degree in communication engineering from the North China University of Technology. She is about to pursue a master's degree in information and Communication Engineering at CATT. Her future research direction is 5G core algorithm.

TONG YAO was born in Beijing, China. He received the B.E. degree in EE from the North China University of Technology in 2012, where he is currently pursuing the M.E. degree in information and communication engineering. His research interests include wireless network and wireless communications.

HAIJUN ZHANG (M'13–SM'17) was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada. From 2011 to 2012, he visited the Centre for Telecommunications Research, King's College London, London, U.K., as a Visiting Research Associate. He is currently a Full Professor with the University of Science and Technology Beijing, China.

● ● ●