Load Balancing in UDN Networks by Migration Mechanism with Respect to the D2D Communications and E2E Delay

Mohamad Salhani and Markku Liinaharja
Aalto University, Espoo, 02150, Finland
Email: mohamad.salhani@aalto.fi; markku.liinaharja@gmail.com

Abstract —Ultra-Dense Network (UDN) is considered as one of the key technologies of 5G. However, the dense deployment of small cells in UDN hotspots generates an uneven traffic distribution. To address this problem, this paper proposes a load migration mechanism to transfer the extra users from the small cells to the macrocells. In addition, this paper employs the design structure matrix (DSM) method with different approaches in order to balance the load among the small cells and to reduce the inter-communications between the access points. Once the load balancing and the user transfer are achieved, the DSM method is capable of taking the device-to-device (D2D) communications of the users into account. The results prove that the user transfer approaches with the DSM method with respect to the D2D communications can enhance the balancing results in some cases by 24.68% compared to the case without transfer. Additionally, the balance improvement ratio can reach 78.30%. Besides, the average inter-communications ratio between the access points can be reduced by 57.35%.

Index Terms—UDN, RoF, load balancing algorithm, user transfer algorithms, D2D communications, DSM method, end-to-end delay.

I. Introduction

The upcoming 5G networks are characterized by ultra-dense deployment of small cells. Ultra-dense networks (UDNs) are capable of providing the desired increase in capacity and data rates by improving the network coverage. The Radio over Fiber (RoF) system can be used as small cells in UDN networks to achieve higher data rate. The RoF technology combines the advantages of optical and broadband technology. This combination provides higher capacity, lower power consumption and easy installation [1]. However, the increased demand of mobile traffic results in heavily uneven load between the small cells. While the RoF system has attracted much attention, as it provides high data rate transmissions, it suffers from a load unbalance [2]. This is because of the frequent handovers of users, as the RoF coverage is relatively small. In studies that have applied the RoF system, the handovers were more frequent than in those discussing traditional cellular networks [1]. Therefore, a load balancing algorithm (LBA) becomes a necessity to redistribute the load among the access points (APs) of UDN networks in selective way and to transfer the extra users to the macrocells of the base stations (BSs) with respect to some constraints imposed on the users. Considering that, the system performance will be optimized, such as the throughput, user fairness, resource utilization and processing delays [3].

The design structure matrix (DSM) method provides a simple, compact, and visual representation of a complex system that supports innovative solutions to decomposition and integration problems used in the system engineering of products [4]. To the best of our knowledge, the DSM method has not been exploited yet in the previous studies that deal with load balancing with constraints imposed on some users and with transfer of the extra users to the macrocells.

The first load balancing algorithms within wireless networks were proposed by Balachandran and Aleo [5], [6]. Nonetheless, the proposed algorithms were very simple and only balance the load between two cells with an overlapping zone. A channel borrowing scheme has also been used to offload the overloaded cells by using an unused channel from the neighboring unloaded cells in [7], [8]. This method without a strict channel locking strategy may result in co-channel interference. Strategies based on cell breathing and power control have been presented by Hanly *et al* [9]. These can offload the overloaded cells by simultaneously reducing the power of the APs in the overloaded cells and increasing the power of the APs in the underloaded cells. However, these can cause a disconnection of some users located on the cell edges and increase the co-channel interference, and the AP can remain overloaded even after reducing the coverage area.

A new load balancing algorithm in UDN networks based on a stochastic differential game scheme and an RoF system was suggested in [1], [2]. This algorithm did not take into account the optimization issue of the overlapping zone selection, and was without any policy for transferring the extra users to the macrocells or any constraints imposed on the users.

On the other hand, the load balancing by transferring users has not been highlighted enough in the recent studies. Elgendi *et al* [10] have proposed new schemes to find the optimal number of sessions to be transferred

Manuscript received June 12, 2018; revised March 3, 2019.
doi:10.12720/jcm.14.4.249-260

from unlicensed long term evolution (U-LTE) networks to licensed long term evolution (L-LTE) or Wi-Fi networks. They have shown that it is possible to transfer the users from programmable BSs to Wi-Fi APs in order to achieve a win-win outcome for both networks. Nonetheless, they have focused on the speed of users and the distance between the user and the BS more than the data offloading. Besides, the proposed schemes have transferred a higher number of users. In this paper, we propose only transferring the necessary number of users in selective way based on data offloading. In addition, the DSM method with the (user) transfer approaches are adapted first to decrease the inter-communications between the APs of the small cells and second to redistribute the throughput among the APs with respect to end-to-end (E2E) delay of the users. When the load balance and the user transfer are achieved, the DSM method takes the device-to-device (D2D) communications of the users in consideration. The purpose of the transfer approaches is to offload the small cells of UDN networks by transferring the best candidate user (BC) to the macrocells. The transfer approaches are based on several load balancing approaches. The load balancing approaches are first based on selecting the best overlapping zone among several ones and then determining the BC to be handed over or transferred to the BS by selective way in order to reduce the number of the handed-over and transferred users.

The paper focuses on UDN hotspots where all the APs of the UDN network are considered to be always active. The user density can be 10 times larger than that of APs [11], and hence it can reach six users per each small cell. The first goal is to find an optimal policy for carefully selecting the part of load to be shifted from the overloaded APs. The second contribution is to employ the DSM method in reducing the inter-communications between the APs and in balancing the load at the same time. All that will be encountered with the aid of the transfer approaches, which transfer the extra users to the macrocells with intent to improve the load balancing within the small cells.

The rest of this paper is organized as follows. The system model is described in section II. The different algorithms are proposed in section III. The simulation model and the performance evaluation criteria are presented in section IV and section V. While section VI introduces the proposed approaches, section VII discusses the different approaches with respect to the D2D communications users. Then, section VIII explains how the DSM method can be exploited in reducing the inter-communications between the APs and balancing the load. Section IX describes the DSM algorithms. The results are discussed in section X. Finally, a conclusion and perspectives of this work are presented in section XI.

II. System Model

The proposed system consists of multiple macrocells, as shown in Fig. 1. We consider several APs of UDN small cells with overlapping zones. Each set of small cells constitutes a so-called RoF cluster. The small cells can be integrated with the Remote Radio Heads (RRHs), which are also connected to the central BS via high speed optical fiber or microwave links [12]. The APs of each cluster are controlled by a Virtual Base Station (VBS) through optical fiber. The VBS is considered as a router of the RoF system. The system of load balancing and user transfer can be either distributed in each VBS or centralized in each virtual BS controller (VBSC)/virtual mobile switching center (VMSC). Each small cell is modeled by a multi-processor queue. Due to the high density of small cells and in order to avoid the interference, some small cells are allowed to be inactive (idle mode) in the case of an interference occurring [13]. In this paper, the rates of the users (UEs) are limited by the core network. The proposed system model is assumed to accurately measure the user location from the user reference signals, and thus the location of each user is known [14]. The proposed model can support the standalone and network-assisted D2D communications.

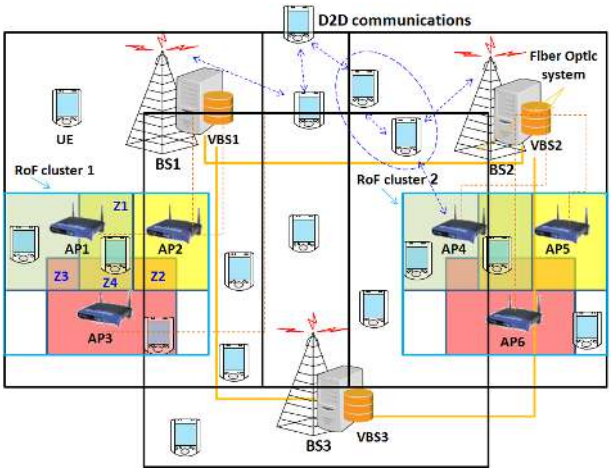


Fig. 1. Distributed system model

III. Load Balancing and User Transfer Algorithms

The purpose of this paper is to balance the load between the APs of the UDN network and to transfer the extra users to the macrocells with respect to the constraints imposed on some users. Consequently, our contribution consists of three parts. The first is the load balancing between the APs and the second part is to find a mechanism to migrate the extra users to the macrocells. Both parts are respectively explained in this section. In the third part, the DSM method, which reduces the inter-communications between the APs and balances the load is introduced in section VIII. The different proposed algorithms respect the constraints imposed on some users and can be implemented in a controller. During the busy hours, the following load balancing algorithm between the APs with one of the upcoming two (user) transfer algorithms is initialized. In other words, the LBA will be called by one of the two transfer algorithms before or after transferring the extra users.

A. Load Balancing Algorithm between the APs (LBA)

This section explains the LBA between the APs without any user transferring to the macrocells. The LBA **first** starts checking the user density (ρ) within each cluster and comparing the density of the cluster with the highest density to the density threshold ρ_{Th}. If ρ does not exceed the threshold, the algorithm is stopped and it waits for the next trigger. Otherwise, the algorithm sets the throughput of each user, its zone and the tolerance parameter α, which will be determined later. Then, the algorithm calculates the throughput of each AP ($T_{AP(i)}$) as a summation of throughputs of all users (j) connected to the serving AP(i), as given by

$$T_{AP(i)} = \sum_{j=1}^{m(i)} T_{user(i,j)} \tag{1}$$

where m(i) is the number of users connected to AP(i). Next, the algorithm calculates the average network load (ANL) of the whole cluster as follows:

$$ANL = (T_{AP1} + T_{AP2} + ...T_{AP(n)}) / n \tag{2}$$

where n is the maximum number of APs. Meanwhile, the LBA determines the state of each AP by using the transfer policy. This policy verifies which AP must exclude a user (overloaded AP) and which one must include this user (underloaded AP). For that, two thresholds (δ_1 and δ_2) are needed. The upper threshold and the lower threshold are given by

$$\delta_1 = ANL + \alpha \times ANL, \ \ \delta_2 = ANL - \alpha \times ANL \tag{3}$$

According to the transfer policy, an underloaded AP can accept new users and users handed over from an overloaded AP. While a balanced AP can only accept new users, an overloaded AP does not receive any new or handed-over users. Subsequently, the load balancing process will exclusively hand over the users from overloaded APs to underloaded APs.

With regard to the tolerance parameter α, the critical value of α is calculated before applying the LBA by setting the throughput of the most overloaded AP equal to δ_1 as follows:

$$\alpha_{critical} = (T_{APmost-overloaded} - ANL) / ANL \tag{4}$$

Then, the result is divided by 10 to obtain the required value of α. Note that increasing the value of α widens the balance zone and thus reduces the handovers. Therefore, this shortens the running time of the LBA and benefits the users with real time applications. In practice, the desired value of α can be empirically calculated so that the average value of α can be tuned based on the state and the location of the network.

In the **second step**, the algorithm checks if there is at least one overloaded AP within the cluster with the highest user density (cluster of first order). If not, the algorithm transits into the cluster of second or third order successively and rechecks the density condition. If this condition is not satisfied in these three clusters, the algorithm is stopped. Otherwise, the algorithm calculates Jain's fairness index (β) for each overlapping zone [15], which is determined as follows:

$$\beta = \frac{\left(\sum_{i=1}^{n} T_{AP(i)}\right)^2}{n \times \sum_{i=1}^{n} T_{AP(i)}^2} \tag{5}$$

where n is the number of small cells that overlap on the zone in question, i.e., each overlapping zone has its own β. When all the APs have exactly the same throughput, β is equal to one. Otherwise, β approaches 1/n, so β∈[1/n, 1]. The **third step** is to apply the selection policy in order to determine the BC user to be handed over as follows. First, the difference (delta Δ) between the selected overloaded AP and the ANL is calculated by

$$\Delta = T_{overloadedAP} - ANL \tag{6}$$

Of all the users located in the overlapping zone in question and connected to the chosen overloaded AP, the BC is the one for which the difference of the user throughput and delta has the smallest absolute value as follows:

$$BC_j = \left| T_{user\,j} - \Delta \right| \tag{7}$$

Note that some constrained users, e.g., the relay users of the D2D communications, are excluded from any handovers or transferring to the macrocells, as it will be explained later.

The **fourth step** is to determine the new β if the BC is handed over. This step is called the distribution policy. The aim of determining new β is to ensure that the expected handover will definitely improve the balance before doing the handover to avoid the ping-pong problem. Thus, the handover of the candidate user will be carried out if and only if $\beta_{new} > \beta_{old}$. If the latter condition is satisfied, the algorithm selects this candidate and the handover decision occurs. Otherwise, the algorithm transits into the next target zone. It is one of the overlapping zones, which changes or not according to the selected load balancing scheme, as it will be explained later. After that, the algorithm repeats the last policies with the new target zone. The **fifth step** is to check again if there is still an overloaded AP, and also if the improvement is still valid. If so, it evaluates the enhancement within the new target zone by updating all the values of β (βs) and so on. Otherwise, the algorithm is stopped and waits for the next trigger.

B. Transfer After Algorithm (TAA)

The TAA is one of the algorithms that take care of the users that should be transferred to the macrocells. This algorithm is composed of the following two stages. The first one is the balance stage that is achieved by the previous LBA. The second is the transfer stage, which is

carried out after the balance stage. Therefore, the TAA has the same first steps of the LBA; however, when there are no more balance improvements within the APs, the transfer stage with new selection and transfer policies are initialized. In the **first step** of the transfer stage, the algorithm checks if at least one of the APs is overloaded, i.e., its throughput exceeds the transfer threshold $T_{capacity}$, which is the maximum allowed capacity for each AP, as it will be determined later. If this condition is not satisfied, the algorithm is stopped and waits for the next trigger. Otherwise, the **second step** is to perform the new selection policy in order to determine the BC to be transferred by a vertical handover procedure as follows: First, the algorithm calculates the new delta as the difference between the most overloaded AP and $T_{capacity}$ as given by

$$\Delta_{new} = T_{most-overloaded AP} - T_{capacity} \quad (8)$$

Second, the best candidate value $BC_{(j)}$ is calculated, for each unconstrained user and connected to the chosen AP, as the difference between the user throughput and the new delta as follows:

$$BC_{(j)} = T_{user(j)} - \Delta_{new} \quad (9)$$

Of all the unconstrained users connected to the AP in question, the BC is the one for which the $BC_{(j)}$ has the smallest positive value. Otherwise, the BC is the one that is unconstrained and has the smallest negative value in case all the values of $BC_{(j)}$ are negative. The transfer is repeated until the AP throughput becomes less than or equal to $T_{capacity}$. In the **third step**, the algorithm determines the next most overloaded AP and then it repeats the second step. When all the APs have been checked and there is still not any more transfer possibility, the TAA is stopped.

C. Transfer Before Algorithm (TBA)

The TBA is similar to the TAA; however, the transfer stage is initialized as a first step for each AP that has a throughput exceeding $T_{capacity}$. When the throughputs of all the APs do not exceed $T_{capacity}$ any more or if there are no more available unconstrained users to be transferred, the balance stage starts by calling the LBA, which continues the load balancing task as usual. Note that δ_1, δ_2, $T_{AP(i)}$ and ANL required for the LBA are determined once the transfer stage is over.

D. Active Algorithm

In case there are no constraints imposed on the users, the active algorithm is able itself to balance the load between the APs and to transfer the extra users to the macrocells without any help from the LBA unlike TAA or TBA. Instead, in the case of constrained users, this algorithm needs to apply the ordinary LBA as a last step. Actually, the active algorithm has a specific policy and is triggered for each new user entering into the network. This algorithm takes care of individual users by taking into account the zone of the user, the throughput of the user and the APs. This algorithm is composed of the following steps. The **first step** is to set the throughput of the new user and its zone. If the throughput of each AP is zero, i.e., this user is the first user that enters into the cluster. In this case, the user is accepted by one of the APs based on the SNIR metric. Otherwise, the algorithm in the **second step** selects the best AP for this user. This step is considered the *first phase of balance*. The selected AP is the least loaded AP so that if the throughput of this unconstrained user is added to the throughput of this AP, the new AP throughput should not exceed $T_{capacity}$. If there is no AP satisfies this condition, the unconstrained user is transferred to one of the macrocells (BSs). In contrast, if the new user is a constrained user, this user will be accepted by the chosen AP even if this results in exceeding the $T_{capacity}$ limit. A constrained user is one of the following users: a D2D user, a relay user or a user that is communicating with another user located in the same cluster, which is a so-called DSM user. In the **third step**, the algorithm checks the density condition. The user density is calculated by considering the number of the users accepted in the UDN network and the number of the transferred users. If the user density within the chosen cluster is higher than or equal to ρ_{Th}, the algorithm applies the ordinary LBA as a last step to balance again the load between the APs because the constraints make the algorithm unable to balance the load without applying this *second phase of balance*. Otherwise, in case the user density condition is unsatisfied, the algorithm sets the throughput of the next new user and so on. In practice, the density condition is not necessary to be checked, as this algorithm is always on standby and triggers for each new user. This condition is only imposed in this study in order to compare the results of this algorithm to those in the previous algorithms with the same user density.

IV. Simulation Model

To simplify, we consider two macrocells with one overlapping zone. Multiple small cells are covered by each macrocell. Each set of three square overlapping small cells forms an RoF cluster. We consider a three intersecting cell model with four overlapping zones (Z1, Z2, Z3, Z4), i.e., with β1, β2, β3 and β4 (β), as shown in Fig. 2 (a). The load balancing and the user transfer processes are implemented at the cluster level. The tolerance parameter α is chosen to be 5%. The area of overlapping zones between each two small cells occupies about 25% of the total small cell area. The dimensions of each square small cell are 20x20 m^2 and the dimensions of each square macrocell are 0.5x0.5 km^2. The inter-sites distance is 15m. The user density ρ is on average equal to six users per small cell. Therefore, the density threshold ρ_{Th} is on average 18 users per cluster in addition to some users with D2D communications type, which establish connections with the relay users to reach the UDN network. Each user selects a specific throughput in the

range from 0 to 350 Mbps [16]. The average throughput of each user is around 175 Mbps. Accordingly, the throughput for six users is around 1Gbps, which is the maximum allowed capacity for each AP, $T_{capacity}$. The total number of small cells covered by each macrocell is 625 small cells. Consequently, the maximum allowed capacity of each BS is 625 Gbps. Subsequently, the number of clusters, which can be covered by each BS is about 208 clusters, and the density of small cells per square kilometer is 2500 small cell/km^2. This density is greater than the value of 10^3 small cell/km^2 used in the recent research studies, thus, the studied network is sufficiently overloaded. In this study, the potential communications of the users can be of D2D type or infrastructure type such as user-to-user via an AP.

V. Performance Evaluation Criteria

The considered criteria are Jain's index β and the standard deviation σ_T of the throughputs of APs, which is given by

$$\sigma_T = \sqrt{\frac{(T_{AP1} - ANL)^2 + (T_{AP2} - ANL)^2 + ... + (T_{APn} - ANL)^2}{n-1}} \quad (10)$$

The index β4 is not always available for any configuration of small cells, as it depends on existing of a zone common between all the overlapping cells. In contrast, σ_T is topology-independent and is a general parameter. σ_T has also a wider range than β, which is limited within [1/n, 1]. Actually, β and σ_T express the same state of balancing: the increment of the β value towards "1" leads to the decrement of the σ_T value towards "0". Another considered criterion is the standard deviation of all the values of σ_T, STDEV(σ_T). The latter represents an indicator about the change of traffic distribution among the APs during the different steps of the algorithm. Other criteria are the handover rate (HOR), the rate of the transferred users (TR), the rate of the replaced users (RR) in some cases and the balance improvement ratio (BIR), which is defined as the difference between the final value and the initial value of σ_T divided by the initial value as given by

$$BIR = \left| \frac{\sigma_{T\,final} - \sigma_{T\,initial}}{\sigma_{T\,initial}} \right| \quad (11)$$

The best algorithm is the one that minimizes the required signaling and maximizes the load balancing. For that, we consider the following criteria: the balance efficiency (BE), the transfer efficiency (TE) and the overall efficiency (OE), which are respectively given by

$$Balance\ \ Efficiency\ \ (BE) = \sigma_T \times (HOR + RR) \quad (12)$$

$$Transfer\ Efficiency\ \ (TE) = \sigma_T \times TR \quad (13)$$

$$Overall\ Efficiency\ \ (OE) = \sigma_T \times (HOR + RR + TR) \quad (14)$$

The reducing inter-communications ratio (RICR %) between the APs is also another criterion that is given by

$$RICR\ \% = \left| \frac{\tau_{final} - \tau_{initial}}{\tau_{initial}} \right| \quad (15)$$

VI. Proposed Approaches

In order to accomplish the load balancing within the small cells, the following approaches are proposed:

A. Common Zone (CZ) Approach

In this approach, the load is only balanced by the users located in Z4 depicted in Fig. 2 (a), which is the common zone (CZ) between the three overlapping cells, and it is always the target zone. This approach is quick and simple, as it does not require much processing. In contrast, it is not very convenient in the case of UDN networks, since the user density is relatively.

B. Worst Zone (WZ) Approach

The load balancing in this approach is executed in the worst zone (WZ), which is the target zone that has the smallest value of β(i). Accordingly, the CZ approach is less complicated than the WZ approach because the latter one must calculate the different β(i) to determine the WZ for each handover. Therefore, the CZ approach might shorten the running time of the algorithm.

C. Mixed Approach (MA)

A hybrid approach that combines the CZ approach and the WZ approach. It starts balancing the load in the CZ and then it transits into the WZ with or without returning to the CZ approach after each handover. Subsequently, the target zone alternates between the CZ and the WZ according to the selected policy. In this regard, we suggest five MA policies according to the transition into the zone of action as follows:

1) 2nd-AP policy

This policy tries to hand over all the available users in the CZ as long as there are users of first order (connected to the most overloaded AP) and second order (connected to the next most overloaded AP). Next, it transits into the WZ approach without returning to the CZ approach.

2) Early WZ policy

Using this policy, only one handover is executed for a user of first order in the CZ and then it transits early into the WZ.

3) Persist 1st-Users policy

This policy only hands over the users of first order in the CZ before transiting into the WZ approach. Subsequently, if there is only one user of first order in the CZ, the early WZ policy and the persist 1st-users policy will be identical. Once there are more than one overloaded AP and the handovers for the first order users in the CZ are over by using the persist 1st-users policy, does the algorithm come back to the CZ or not? What are the potential policies in this case? To answer to these questions, two additional policies are proposed:

4) Persist WZ policy

This policy only hands over one user of second order in the CZ, after handing over all the users of first order by

the persist 1st-users policy, then it permanently transits into the WZ approach.

5) Persist CZ policy

This policy is opposite to the persist WZ policy, i.e., after handing over all the users of the first order, it only hands over one user by the WZ approach and then it transits into the CZ approach. Thus, this policy is computationally the most complicated one. On the other hand, in order to transfer the extra users, the following transfer approaches are proposed:

D. Passive Approaches

We propose three transfer approaches. If the balance stage achieved by the LBA is carried out as a first step and then the transfer stage is performed as a next step, this approach is called the **transfer after (TA) approach**. The direction of transfer from the small cells to the macrocells is named an *up handover procedure*. The reverse direction of transfer, from BSs to APs, is carried out by *a down handover procedure* once a sufficient bandwidth is available in one of the small cells. Note that both types of the vertical handovers can be accomplished using diverse strategies of connections replacement that have been proposed by Salhani *et al* [17]. Alternatively, if the transfer process is achieved as a first step, this type of the approaches is called the **transfer before (TB) approach**. Both approaches are named passive, as they are only triggered when the density condition is satisfied.

E. Active Approach (AA)

The AA approach, opposite to the passive approaches, is always on standby and ready to be triggered each time a new user enters into any cluster. This approach depends on the throughput of the user and the APs and also on the zone of the new user. When an AP is selected to include the new user and the throughput of this AP will not exceed $T_{capacity}$ if it accepts this new user, this user is accepted by the AP in question. Otherwise, the algorithm transfers this unconstrained user to the macrocells. This process is repeated for each new user until the user density of the chosen cluster reaches ρ_{Th}.

VII. Proposed Approaches with D2D Communications

In this section, we discuss the results of the load balancing and user transfer approaches with respect to the D2D communications. Consider a cluster with three intersecting cell model and four overlapping zones (Z1, Z2, Z3 and Z4) covered by two macrocells, as shown in Fig. 2 (a). Four different values for Jain's index are hence assumed (β1, β2, β3 and β4) with a number of overlapping zones of n=2, 2, 2 and 3, respectively.

In this cluster, 18 users exist and each user is represented by its number j and its throughput T_{userj} such as j(T_{userj}). Z1 is the overlapping zone between the two cells of AP1 and AP2. In this example of applications, the following users are located in Z1: UE3, UE4 and UE9. Similarly, Z2 is the overlapping zone between the two cells of AP2 and AP3. UE1, UE2, UE6, UE8 and UE14 are located in Z2. UE11 and UE18 are located in Z3, which is the overlapping zone between the two cells of AP1 and AP3. UE5, UE7, UE10, UE12, UE13, UE15, UE16 and UE17 are located in Z4, which is the CZ between all the cells. With regard to the D2D communications, UE19 and UE20 are also considered. These users are so-called D2D users. These two users succeeded to be connected to AP3 by D2D communications type via two relay users UE1 and UE6. In this view, UE1 and UE6 are assumed able to convey the traffic of UE19 and UE20 to AP3.

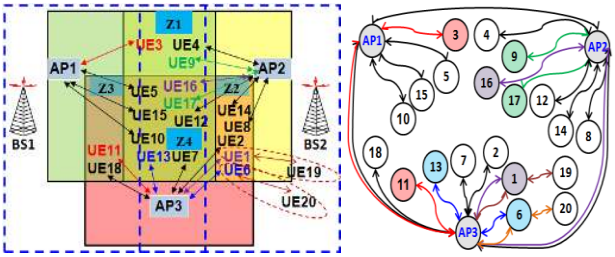
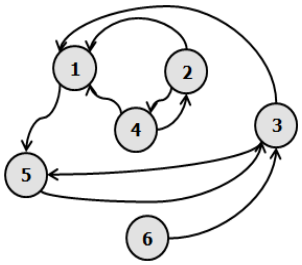


Fig. 2. A cluster (a) and its graph of the partition (b).

Once the algorithm is applied, the D2D constraints are respected so they prevent the relay and the D2D users from any handovers or transferring to the macrocells. Otherwise, two connections will be disconnected during the expected handover or transfer procedures: the connection of each relay user and the one of each D2D user. The disconnection could be risky for the D2D users with real time applications, which do not tolerate with the delay.

VIII. Use of the DSM Method



	1	2	3	4	5	6
1	X				1	
2	1	X		1		
3	1		X		1	
4	1	1		X		
5			1		X	
6			1			X

Fig. 3. The graph of the nodes (a) and adjacency matrix (b)

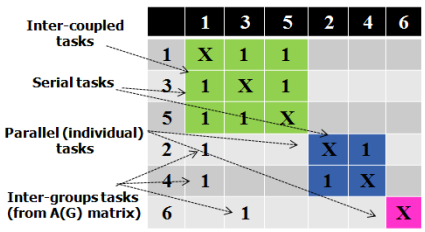
In the following, we employ the design structure matrix (DSM) method in decreasing the inter-communications between the APs in addition to balance the load among the small cells. This method deals with the partition of graphs in order to realize a cooperation between the nodes (terminals), and to organize the complex tasks in the projects with regard to the parallel, consecutive and coupled tasks. A simple example is considered prior to exploiting this method. Fig. 3 (a) shows a graph composed of six nodes, which communicate together to perform a predefined task. The intended aim is to distribute these nodes on two switches with respect to the type of tasks (serial, parallel). For that, the adjacency matrix A(G) is determined as pictured in

Fig. 3 (b). This matrix is concerned with the direct arcs among the nodes, i.e., the directional and the short communications between the current node and the neighboring nodes. Each arc that starts from a node and heads to another node is represented by "1", while the other cases are left empty or filled by "0".

Second, the attainability matrix R(G) is determined. The latter takes care of the direct and not direct connections between the nodes. Each arc starting from a node and reaching another - even after many hops- is represented by "1", and the other cases are left empty "0". Third, the parallel and in series tasks are deduced as follows. The coupled components matrix C(G) is obtained by

$$C(G) = R(G)\ \ AND\ R(G)^t \tag{16}$$

Each row from the R(G) matrix is multiplied by the corresponding column of this matrix and the results are put in the new row of the C(G) matrix. Next, the new groups (components) are determined after reordering these groups by the reorganized C(G) matrix, as depicted in Fig. 4. This matrix clarifies the relationships between the new groups, i.e., the parallel and serial tasks, whereas the inter-group tasks are deduced from the A(G) matrix.



	1	3	5	2	4	6
1	X	1	1			
3	1	X	1			
5	1	1	X			
2	1			X	1	
4	1			1	X	
6		1				X

Fig. 4. The reorganized C(G) matrix.

Therefore, the new groups become as follows: C1= (1, 3, 5), C2=(2, 4) and C3=(6). We notice that the nodes of group C1 are inter-coupled tasks. While the groups C2 and C3 are parallel tasks, the groups C2 and C3 are in series tasks with C1. Consider each switch has only four ports, the nodes can be partitioned on the two switches as follows: C1=(1, 3, 5) and C2=(2, 4, 6). This distribution can be developed using a refinement method in order to reduce the inter-communications between the switches as follows. The replacement gain for each node is introduced. It is the difference between the number of the connections of a node with the other groups and the number of the connections of this node with the nodes existing in its group. Refer to the obtained distribution, we refine it according to the replacement gain concept. Fig. 5 illustrates the gain matrix of each group (G1 and G2) in the two steps of the refinement.

	G1			G2		
	1	3	5	2	4	6
G1	2	3	3	1	1	1
G2	2	1	0	2	2	0
Gains	0	-2	-3	-1	-1	1

	G1			G2		
	3	5	6	1	2	4
G1	3	2	1	2	0	0
G2	1	1	0	2	3	3
Gains	-2	-1	-1	0	-3	-3

Fig. 5. The refinement steps of the DSM method.

The refinement process is based on checking and taking care of the nodes with positive gains (node 6 in group 1 with G=1). Accordingly, node 6 must be replaced by node 1, which has the biggest gain within group 2, G=0. In this context, the load index τ is introduced with intent to evaluate the replacement performance. The index τ is defined as the ratio of the number of inter-group connections (N_i) and the total number of interconnections of all the nodes (N_t) as follows:

$$\tau = \frac{N_i}{N_t} \tag{17}$$

The new distribution of the nodes becomes C1=(1, 2, 4) and C2=(3, 5, 6). The initial value τ_{inital} is 3/9 and the final value τ_{final} after the refinement becomes 2/9. Consequently, the inter-communications between the switches are reduced using the refinement concept. The refinement process is stopped once all the positive values of gains become negative or at least get zero. *The question is how the DSM method and the refinement process can be employed in balancing the load within the small cells of UDN networks with the transfer approaches?* As the LBA is triggered when the user density condition is satisfied, at that time, the users have already been connected to the APs and each AP has already been constituted a group of some connected users. Therefore, the required task is only how the refinement process can be applied. Actually, to use the DSM method, **either**, the user replacement stage is firstly applied and then the balance stage is carried out by one of the previous load balancing approaches and one of the transfer approaches. This policy is called the DSM_first (DSMf). **Or,** the constraints of the DSM method are respected by the LBA during the selection policy. This policy is called the DSM_included (DSMi). In both policies, the DSM constraints impose that the selection of a user to be replaced is only possible if the number of hops of the user's connection is kept constant, i.e., the load index τ remains constant, or rather this number of hops will be reduced from 3 to 2 hops. Thus, the DSM method reduces as far as possible the E2E delay between the DSM users.

IX. DSM Algorithms

In order to apply the DSM method, two algorithms are proposed with one of the transfer algorithms as follows. First, the two types of the DSM algorithms are explained without transferring the users to the macrocells. Second, the D2D&DSM constraints are considered with the transfer algorithms.

A. DSM Algorithms Without Transfer

The DSM_first algorithm (DSMf) first reduces the inter-communications between the APs and then, starts balancing the load using one of the previous load balancing approaches. Alternatively, in the DSM_included algorithm (DSMi), the replacement gain

of each user is taken into account during the steps of the LBA. Indeed, the DSMi algorithm is one of the load balancing algorithms described earlier; however, during the selection policy, the replacement gain is respected as follows. The selected user will not be the BC and thus handed over, if this handover will increase the replacement gain. Otherwise, the algorithm selects the user of second order at the cost of decreasing the quality of balance. The selected user is the one that has the highest value of the replacement gain.

On the other hand, if the DSMf or DSMi algorithm considers the D2D constraints, we call it D2D&DSMf or D2D&DSMi, respectively. The D2D&DSMf algorithm **first** checks if ρ of the cluster with the highest density exceeds ρ_{th}. If this condition is not satisfied, the algorithm is stopped and waits for the next trigger. Otherwise, the algorithm sets the throughput of each user, the user zone and α. Then, it calculates the following values: T_{AP1}, T_{AP2}, T_{AP3}, ANL, δ_1, and δ_2. In the **second step**, the algorithm checks if there is at least one overloaded AP within the chosen cluster. Otherwise, the algorithm transits into the next order cluster, which is the second order cluster or even to third order one from the user density perspective. If the user density condition for these three clusters is not satisfied, the algorithm is stopped. In case there is at least one overloaded AP, the gain matrix for each AP is computed. This matrix represents the replacement gains for each user connected to the AP in question. Next, the algorithm searches, in the gain matrix of the most loaded AP, for a user that has the highest positive gain and is connected to this AP. This means that this user is currently communicating with another user (its partner), which is connected to another AP. If there is no user that has a positive gain, the algorithm goes to the next most loaded AP. Conversely, if there are many users satisfying these conditions, the user with the highest throughput and positive gain is selected. In the **third step**, the algorithm checks the coverage condition: the AP of the candidate user and the AP of the partner should cover the two users. If so, this means that any one of them can be transferred (handed over) to the AP of the other one. Otherwise, the algorithm selects the user of the second order. This user has the next highest throughput, is connected to the most loaded AP and has the highest positive gain. In the **fourth step**, the algorithm checks the D2D constraints. If the selected user, which is connected to the most loaded AP, is a relay user, the partner should be transferred to the AP of the selected user. As a result, this partner is replaced by the BC. The BC is connected to the most loaded AP, has the highest throughput, is located in the same zone of the partner and is not a DSM user with a negative gain. In contrast, if the selected user is not a relay, it is transferred to the partner's AP. The selected user is hence replaced by the BC. The BC is thus the one that is connected to the partner AP, is located in the same zone of the selected user, has the lowest throughput and is not a DSM user with a negative gain. The **fifth step** is to check again if there are still other users that have a positive gain and are connected to the most loaded AP. If so, a new user is selected and the third step is repeated. Otherwise, the algorithm transits into the next most loaded AP and repeats the third step. When all the APs are checked and the replacement process is over, D2D&DSMf calls the LBA algorithm to evaluate any more improvement and the LBA continues its steps as usual. In the same manner, D2D&DSMi is one of the previous load balancing algorithms, which respects the D2D&DSM constraints during its algorithm steps. In that way, D2D&DSM algorithms will reduce the inter-communications of users in each cluster by making the gain of all the users negative, G=-2, and balancing the load at the same time. In that view, the LBA deals with the balance issue and the D2D&DSM algorithms take care of the number of hops of each user and its type, i.e., DSM or D2D user.

B. Constrained Transfer Algorithms

In this case, the previous transfer algorithms respect the D2D&DSM constraints. The D2D&DSMi algorithm is applied before the transfer in the case of TA approach or after the transfer in the case of TB approach. Regarding the D2D&DSMf algorithm, the replacement process is always achieved as a first step. Then, the transfer process is applied either as a first step in the case of TB approach or after the balance stage in the case of TA approach. In the case of AA approach, the transfer is applied with the first phase of balance for each unconstrained user that makes its AP exceed $T_{capacity}$. After that, the second phase of balance is achieved according to D2D&DSMi or D2D&DSMf. Note that the second phase of balance is needed to complete the balancing process owing to the constrained users.

X. Results Analysis

In this section, we discuss the results of the different approaches with the D2D&DSM constraints. The partition graph for the example of applications is depicted in Fig. 2 (b). The users are split into three groups, which represent the APs with the connected users. In addition to the D2D users, some DSM users are considered as follows: UE1 and UE6 are relay users. UE1 communicates with UE16 via AP3 and AP2. UE6 communicates with UE13 via AP3. UE3 communicates with UE11 via AP1 and AP3. UE9 communicates with UE17 via AP2.

We observed after applying the TA approach that the throughputs of the APs are finally located inside the desired balance range [δ_1, δ_2]. While β tends towards 1, σ_T goes to zero. Subsequently, the throughputs are redistributed better and thus the APs can accept new users. Besides, STDEV(σ_T) increases then decreases versus the handovers and the transferred users achieved during its algorithm steps. This decrease is due to the TA approach selecting the users with the lowest throughputs. In contrast, after applying the TB approach, we noticed that the final states of the APs become worse than those in the TA approach. Moreover, STDEV(σ_T) increases until

reaching the final value. This increase confirms that the handovers and the transfer processes are carried out for the users with the highest throughputs. Therefore, the throughputs of the APs deeply change in the TB approach. In addition, the number of the transferred users becomes higher than that in the TA approach; however, the handovers decrease. In fact, the TAA is based on first handing over the users and then transferring the extra users to the BSs. In that way, it guarantees the balance and then it transits into the transfer task. Furthermore, we observed in the case of TB approach that sometimes one AP is kept slightly overloaded. Indeed, as many users are transferred early on, no more BCs to extensively balance the APs.

Applying the AA approach, the balancing results become better than in the TB approach at the expense of more signaling load caused by the frequent triggering for each new user. Moreover, the value of STDEV(σ_T) is smaller than in the passive approaches. This value decreases smoothly due to the smaller change of the AP throughputs. Besides, all the considered criteria are more dependent on the throughput of the new user and the AP to which the new user will be connected. Additionally, the value of σ_T fluctuates much more than in the passive approaches.

In the following, the general results are discussed. We perceived that the **Balancing Results (σ_T)** of the transfer approaches with the D2D&DSM constraints become worse than those with only either the DSM or D2D constraints or without constraints. Indeed, once the constraints increase, the load balancing task becomes more and more difficult. The D2D&DSM algorithms will be forced to select the user of second order and even in the zone of second order. However, the transfer algorithms are able to reduce the impact of these constraints. The transfer approaches improve the balancing results (on average) by 29.77% and **24.68%** in the case of D2D&DSMi and D2D&DSMf, respectively compared to the case without transfer. The approach least affected by these two constraints is the AA approach. This is due to its specific policy and also, since it applies the LBA and the DSM method as a last step. Alternatively, the AA approach must be triggered for each new user. Consequently, the TA approach is more convincing and its balancing results are better than the TB.

By comparing the D2D&DSMi algorithm with the D2D&DSMf one, the average balancing results are similar. A small difference of 8.90% for D2D&DSMi is noticed. Comparing to the case with only D2D constraints, the balancing results of D2D&DSMi and D2D&DSMf become worse. As a result, once the constraints are numerous, the MA approach starts losing its capability and gradually converges to the WZ approach. In the case of TA approach with D2D&DSMi, the MA approach loses its capability and the WZ approach becomes better than the early WZ policy, which is the best policy in the MA approach, by 8.79%. In addition, in the case of TA approach with D2D&DSMf, the WZ approach is better than the early WZ policy only by 3.83%.

An important result is that the impact of DSM constraints on the balance with the transfer approaches is a little greater than the D2D constraints. Furthermore, β of the WZ approach is better than in the MA approach (i.e., the average value of the MA policies) in both cases: D2D&DSMi and D2D&DSMf.

The **handover rate (HOR)** is reduced using the transfer approaches with D2D&DSM compared to the case without transfer and the case without constraints. The HOR in the D2D&DSMi case is higher by 12.59% than the D2D&DSMf case; however, the balancing results are only better by 8.90% with D2D&DSMi. Besides, the TB and AA approaches are too sensitive to the constraints, as the balance stage (the second balance stage in case of the AA approach) is achieved after the user transfer. At that stage, these approaches have already been lost the BCs. In contrast, the TA approach achieves an HOR identical to the case without transfer. This confirms that in this approach the transfer process is independent of the balance, since the transfer stage carries out as a last step and the transfer process is just related to the APs with throughputs exceeding $T_{capacity}$.

Due to the D2D constraints, the **rate of the replaced users (RR)**, the RR is reduced by 10% compared to the DSMf case. The TA approach with D2D&DSMf replaces the same number of users as the TB approach and as the case without transfer, as the replacement in the passive approaches is independent of the transfer process. In contrast, the replacement is carried out after entering all the users in the case of AA approach. Consequently, the passive approaches replace users more than the AA approach in the case of D2D&DSMf or DSMf. Additionally, the RR in the case of the TA approach with D2D&DSMf is the same as the case without transfer. Consequently, the HOR and the RR in the TA approach are not affected by the transfer process.

With regard to the **rate of the transferred users (TR)**, the transfer approaches with the D2D&DSM constraints show a TR similar to the case without constraints. This means that the D2D&DSM constraints do not affect the transfer process. The transfer process is related to the APs, which exceed $T_{capacity}$ more than the constraints imposed on the users. Besides, the TB approach achieves a TR higher than the TA approach and the AA approach. Alternatively, the TA approach achieves the lowest TR. In fact, as the balancing results with this approach are better than the TB, thus there is no need to transfer many more users.

Presumably, UDN networks offer calls with less expensive cost and provide better QoS, particularly the latency, hence, the TA approach seems the suitable option although the constraints.

Comparing to the DSM case, the TR only increases by 1.88%. This increase is because the algorithms try to compensate the decreasing of the balancing results by transferring many more users. Moreover, D2D&DSMi

transfers users more than D2D&DSMf only by 3.01%. On the other hand, the TR is reduced by 12.22% in comparison to the D2D case. This means that the D2D constraints force the small cells to transfer users more than the DSM constraints. However, the DSM constraints affect the balancing results of the UDN network more than the D2D constraints. Another important result is that the criteria of the balance process (σ_T, β, HOR, RR and BE) are more affected by the constraints than the transfer process itself.

From the **signaling load** perspective and owing to the replaced users, the total ratio (HOR+RR+TR) is increased by 19.49% in comparison to HOR+TR with D2D&DSMi. This ratio is similar in the case of TB approach and the AA approach. However, the TA approach achieves the highest ratio.

The **Balance Improvement Ratio (BIR)** in the case of D2D&DSMi is better than D2D&DSMf only by 8.84%. Additionally, the BIR in the case without transfer is higher than that in D2D&DSMf and D2D&DSMi only by 7.13%. Moreover, with both algorithms, the TA approach leads to the best BIR. The BIR of the TA approach with the WZ approach reaches 83.07% and **78.30**% in the case of D2D&DSMi and D2D&DSMf, respectively. Besides, the value of the BIR with D2D&DSM is less than in the case of DSM or D2D constraints or without constraints. Nevertheless, when the TA approach with D2D&DSM is used, the BIR becomes better than in the case without transfer. In addition, a small enhancement is also observed in the case of TB approach. This means that the transfer process can improve a little the BIR using the passive approaches, even with these two constraints. Lastly, as the smallest BIR is noticed by the AA approach, which is only 50.98% against 76.71% in the case of the TA approach, this can be considered another reason to exclude the AA approach.

With regard to the **Balance Efficiency (BE)** with D2D&DSMi, it is better than with D2D&DSMf by 57.06%. This difference is mainly caused by the RR rate. Nevertheless, the transfer approaches enhance the BE compared to the case without transfer even with these two constraints. While the TB approach achieves the best BE with D2D&DSMi, the AA approach leads to the best BE with D2D&DSMf. This does not mean that the TB approach shows the best balancing results. These two approaches are better than the TA approach because of the lowest HOR.

On the other hand, the WZ approach seems the best choice specifically with the TA approach. In this context, we observed that the BE achieved by the WZ approach with the transfer approaches is better than the MA policies by 10.94%.

As the σ_T criterion is noticeably increased compared to the cases with DSM or D2D constraints. The **Transfer Efficiency (TE)** with D2D&DSM deteriorates. Besides, the TE with D2D&DSMi outperforms that with D2D&DSMf only by 3.82%. The main reason is that the balancing results with D2D&DSMi are better than with D2D&DSMf. The best TE is indicated using the AA approach, as it achieves the best balancing results and the lowest TR. Moreover, the TE of the TA approach is better than that in the TB approach by 20.69%, as the latter achieves the highest TR.

On the other hand, the **Overall Efficiency (OE)** with D2D&DSMi is better than with D2D&DSMf by 24.04%. Because of these two constraints together, the OE deteriorates compared to the cases of DSM and D2D. We noticed that the impact of the DSM constraints on the OE is a little greater than the D2D constraints. In addition, since the AA approach shows the smallest values of (HOR+RR+TR) and σ_T, thus, it achieves the best OE.

Comparing to the case without transfer, due to the high signaling, the OE is reduced by 69.86%. Only the AA approach is able to improve the OE compared to the case without transfer. Regarding the load balancing approaches, the WZ approach is a little better than the MA one. The WZ approach outperforms the MA and the CZ approach by 5.76% and 22.70%, respectively.

With respect to the **Reducing Inter-Communication Ratio (RICR)**, the replacement with the transfer process and D2D&DSMi is only affected in the case of the TB approach and the AA approach. This is because the replacement is achieved during the steps of the algorithm. In the case of D2D&DSMi, the RICR is decreased by 16.74% and 37.77% compared to the case of DSM and the case without transfer, respectively. Alternatively, the TA approach with D2D&DSMi is not affected and its RICR is the best. In fact, the transfer is left as a last step in the TA approach and it does not lose the BCs earlier. This can be considered as another reason to adopt the load migration mechanism based on the TA approach. Additionally, in the case of D2D&DSMf, the RICR is similar to the case without transfer and is identical in the case of TA approach. The RICR reaches 50.85% with D2D&DSMf against only 7.76% with D2D&DSMi. On the other hand, the RICR using the WZ approach with D2D&DSMf is better than the MA approach and the CZ approach by 4.29% and 14.09%, respectively. All the MA policies lead to identical results in this case. This means that the MA approach loses its capability because of these two constraints. Moreover, the RICR achieved by the WZ approach in the case of TA approach with D2D&DSMf reaches **57.35%** against 18.95% with D2D&DSMi. Furthermore, the RICR performed by the WZ approach with D2D&DSMi is better than the MA policies by 26.64%.

TABLE I. COMPARISON BETWEEN THE DIFFERENT STUDY CASES

IC model	without constraints	D2D	DSMi	DSMf	D2D&DSMi	D2D&DSMf
Balancing results	MA	MA	MA	WZ	WZ	MA
Signalling	WZ	WZ	WZ	WZ	MA	WZ
BE	WZ/MA	MA	WZ/MA	WZ	WZ	MA
OE						
RICR%			MA	MA/WZ	WZ	WZ
The best approach	MA	MA	MA	WZ	WZ	MA/WZ
	transfer	transfer+D2D	transfer+DSMi	transfer+DSMf	transfer+D2D&DSMi	transfer+D2D&DSMf
Balancing results	TA+MA	TA+MA	TA+WZ	TA+MA	TA+WZ	TB/TA+MA/WZ
Signalling	TB+WZ/MA	TB+MA/WZ	TB+WZ/MA	TB+MA/WZ	TB+WZ/MA	TB+WZ/MA
BE	TA+MA	TB+MA	TB+WZ/MA	TB+WZ/MA	TB+WZ/MA	TB+MA/WZ
OE	TA+MA	TA+MA/WZ	TA+WZ	TA+MA	TA/TB+WZ/MA	TB+WZ/MA
RICR%			TA+MA/WZ	TA+MA/WZ	TA+WZ/MA	TA+WZ
The best approach	TA+MA	TA+MA	TA+WZ	TA+MA	TA+WZ	TA+WZ

To summarize, Table I cites the best approach in all the study cases. We notice that while the MA approach would be applied in seven of twelve study cases, the WZ approach would be used in six study cases. This shows the importance of these two approaches with the TA approach once the transfer is considered.

XI. Conclusion

The importance of adopting a load migration mechanism with the D2D&DSM constraints is studied. Several load balancing approaches within the small cells of the UND network are proposed: a CZ approach, a WZ approach and a MA approach. The WZ approach and MA approach confirmed their efficiency in balancing the load. As a way to improve the balance by offloading the small cells, the transfer process even with the D2D&DSM constraints showed its efficiency. In this context, three transfer approaches are suggested: the before transfer (TB) approach, the after transfer (TA) approach and the active approach (AA). The TB approach leads to transferring many users to the macrocells and sometimes may keep an AP slightly overloaded. The AA approach requires much processing and signaling, and should be on standby for each new user entering into the network. The TA approach seems the suitable approach, as it achieves the best balancing results and transfers the smallest number of users.

The DSM method proved its robustness by reducing the inter-communications between the APs and also by balancing the load. In this context, it is not necessary to hand over a D2D user, a relay user or a user that is communicating with another user existing in the same cluster, since this handover/transfer will be risky for the users with real time applications and will increase the number of hops, i.e., the E2E delay. With the D2D&DSM constraints, it is better to adopt the D2D&DSMf algorithm in order to reduce the inter-communications between the APs as a first priority. In this view, two D2D&DSM algorithms can accompany the TA approach and the WZ approach: If the balance is more important than the E2E delay, the D2D&DSMi algorithm would be preferred with a balance improvement ratio of 83.07% and a reducing inter-communications ratio of 18.95%. In contrast, if the E2E delay has the priority, the D2D&DSMf algorithm would be a promoting solution with a BIR of 78.30% and a RICR of 57.35%.

The machine-type-communications with the DSM method can be discussed in the future work. An architecture of integrated UDN_LTE network may be also introduced with QoS classes mapping. A resource reservation mechanism can be also proposed to reduce the delay of the vertical handover procedures when the users are transferred to or from the small cells.

References

[1] H. Xu, Z. He, and X. Zhou, "Load balancing algorithm of ultra-dense networks: A stochastic differential game based scheme," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 7, pp. 2452-2467, 2015.

[2] H. Zhen and W. Jianping, "Non-cooperative differential game based load balancing algorithm in radio-over-fiber system," *IEEE*, vol. 11, no. 2, pp. 79-85, 2014.

[3] Y. Wang, X. Xu, and Y. Jin, "QoS constraint optimal load balancing for heterogeneous ultra-dense networks," *WPMC*, 2016, pp. 317-323.

[4] B. R. Tyson, "Applying the design structure matrix to system decomposition and integration problems: A review and new directions," *IEEE Transactions on Engineering Management*, vol. 48, pp. 292–306, August 2001.

[5] A. Balachandran, P. Bahl, and G. M. Voelker, "Hotspot congestion relief and service guarantees in public-area wireless networks," *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 1, pp. 59-59, 2002.

[6] V. Aleo, "Load distribution in IEEE 802.11 Cells," Master of Science Thesis, KTH, Royal Institute of Technology, March 2003.

[7] Y. T. Wang, "A fuzzy-based dynamic channel borrowing scheme for wireless cellular networks," in *Proc. 57th IEEE Conference on Vehicular Technology*, 2003, vol. 3, pp. 1517-1521.

[8] S. S. M. Patra, K. Roy, S. Banerjee, *et al*., "Improved genetic algorithm for channel allocation with channel borrowing in mobile computing," *IEEE Transactions on Mobile Computing*, vol. 5, no. 7, pp. 884-892, 2006.

[9] S. V. Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 13, pp. 1332-1340, 1995.

[10] I. Elgendi, K. S. Munasinghe, and A. Jamalipour, "Traffic offloading for 5G: L-LTE or Wi-Fi," in *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Atlanta, GA, 2017, pp. 748-753.

[11] M. Ding, D. L. Perez, and G. Mao, "A new capacity scaling law in ultra-dense networks," arXiv: 1704.00399V1 [cs.NI], 3 Apr 2017.

[12] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Network*, vol. 29, no. 2, pp. 6–14, Mar.-Apr. 2015.

[13] H. Claussen, I. Ashraf, and L. T. W. Ho, "Dynamic idle mode procedures for femtocells," *Bell Labs Tech. J.*, vol. 15, no. 2, pp. 95–116, .2010.

[14] M. Koivisto, A. Hakkarainen, M. Costa, P. Kela, K. Leppanen, and M. Valkama, "High-Efficiency device positioning and location-aware communications in dense 5G networks," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 188-195, 2017.

[15] M. Huang, S. Feng, and J. Chen, "A Practical Approach for Load Balancing in LTE Networks," *Journal of Communications*, vol. 9, no. 6, pp. 490-497, June 2014.

[16] P. Kela, "Continuous ultra-dense networks, a system level design for urban outdoor deployments," Book 1799-4942 (electronic), Aalto University Publication Series Doctoral Dissertations 86/2017.

[17] M. Salhani, R. Dhaou, and A. L. Beylot, "QoS mapping and connection admission control in the WiMAX / DVB-RCS Access Network," in *Proc. ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems and ACM Workshop on Performance Monitoring and Measurement of Heterogeneous Wireless and Wired Networks*, Tenerife, Canary Islands, Spain, October 2009, pp. 94-98.

Mohamad Salhani is an associate professor at the Department of Computer and Automation Engineering (CAE), Faculty of Mechanical and Electrical Engineering (FMEE), Damascus University since 2016. He received his B.S degree in Electrical Engineering. from the FMEE in 2000, M.Sc degree from National Polytechnic Institute of Lorain (INPL), France in 2005 and Ph.D degree from National Polytechnic Institute of Toulouse (INPT), France in 2008. He was an assistant professor at the CAE, FMEE, Damascus University in 2009. In 2016, he was a vice-dean for Administrative and Scientific Affairs at the Applied Faculty, Damascus University. He is currently a visiting professor at the Department of Communications and Networking, School of Electrical Engineering, Aalto University, Espoo, Finland. His research interests include 5G mobile communication systems, Ultra-dense networks (UDNs), Internet of Things and LoRa technology.

Markku Liinaharja received the D.Sc (Tech.) degree in communications engineering from Helsinki University of Technology in 2006. He is currently working as a research fellow at the Department of Communications and Networking within the School of Electrical Engineering, Aalto University, Espoo, Finland. His main research interests are in radio communications and error control coding.