IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 15, 2019, accepted August 31, 2019, date of publication September 6, 2019, date of current version September 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939936

Dynamic Mobility Load Balancing for 5G Small-Cell Networks Based on Utility Functions

KHALED M. ADDALI[1], **(Member, IEEE), SUHIB YOUNIS BANI MELHEM**[2], **YASER KHAMAYSEH**[3], **ZHENJIANG ZHANG**[4], **AND MICHEL KADOCH**[1], **(Senior Member, IEEE)**

[1]Department of Electrical Engineering, École de technologie supérieure ÉTS, Université du Québec, Montreal, QC H3C 1K5, Canada
[2]Department of Electrical and Computer Science, York University, Toronto, ON M3J 1P3, Canada
[3]Department of Computer Science, Jordan University of Science and Technology, Irbid 3030, Jordan
[4]Key Laboratory of Communication and Information Systems, School of Software Engineering, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Khaled M. Addali (khaled.addali.1@ens.etsmtl.ca)

ABSTRACT Deployment of small cells was introduced to support high data rate services and expand macro cell coverage for the envisioned 5G networks. A small cell network, which has a smaller size, along with the user equipment (UE) mobility, frequently undergoes unbalanced load status. Consequently, the network performance is affected in terms of throughput, increasing handover failure rate, and possibly higher link failure rate. Hence, load balancing has become an important part of recent researches on small cell networks. Mobility Load Balancing (MLB) involves load transfer from an overloaded small cell to under-loaded neighbouring small cells for the more load-balanced network. This transfer is performed by adjusting the handover parameters of the UEs according to the load situations of the small cells in the vicinity. However, inaccurate adjustment of parameters may lead to inefficient usage of network resources or degrade the Quality of Service (QoS). In this paper, we introduce a Utility-based Mobility Load Balancing algorithm (UMLB) and a new term named load balancing efficiency factor (LBEF). The UMLB algorithm considers the operator utility and the user utility for the MLB-based handover process. While LBEF is proposed to order the overloaded cells properly for the MLB algorithm operation. The simulation results show that the UMLB minimizes standard deviation with a higher average-UE data rate when compared to existing load balancing algorithms. Therefore, a well-balanced network is achieved.

INDEX TERMS Small-cell network, mobility load balancing, measurement reporting, handover, cell individual offset, throughput.

I. INTRODUCTION

The increases in the use of smartphones and applications for information, and communications technologies are causing a rapid increase in the demand for mobile broadband services with higher data rates and higher QoS. According to the Cisco Visual Networking, the expected global mobile data traffic in 2022 is 77 exabytes, which is a seven-fold increase over 2017 [1]. As a result, mobile networks need serious steps to accommodate this massive traffic growth.

The small cell is viewed as a key part in the fifth-generation (5G) network to support the forecasted data demand and enhance the network capacity [2]. A small cell is a low power, cost-effective radio-access point with low service areas ranging from ten to several hundred meters [3]. Extending the service coverage within macro-cells was the prime objective behind designing small cells; however, they can be densely deployed to increase the capacity of the wireless network significantly [4]. Therefore, future networks may adopt the technology of small cells to support ever-increasing data demand.

The deployment of residential and non-residential small cells is growing rapidly [5]. This deployment can be planned or unplanned deployment according to the service operator's policy [6]. Unlike a macro network, the low cost of small cells encourages the subscribers to install their small cells without any network planning and site-specific system configuration settings. Hence, a significant number of small cells in the network will be randomly distributed.

Mobility of UEs in a small cell network with low service area cells may cause load-imbalance across the cells in the network. The performance of the network in terms of

The associate editor coordinating the review of this manuscript and approving it for publication was Sharief Oteafy.

capacity and handover success rate degrades as a result of such an unbalanced load. The shortage of resources in the overloaded small cells leads to poor QoS and increases the handover failure rate when UEs intend to enter those cells though they have lightly loaded neighbouring cells. Consequently, resources of the unloaded cells remain unutilized though some overloaded neighbouring cells cannot meet the QoS requirements. Thus, the network needs proper configuration and management mechanisms such that the QoS is improved.

System parameters are adjusted manually in the existing networks to reach high levels of operational performance. However, such manual tuning is becoming difficult with the fast evolution of networks. Self-organized network (SON) was introduced to configure, optimize, and heal itself automatically in LTE, and hence decrease the operational complexity [7]. SON algorithms are categorized into three classes: centralized, distributed, and hybrid. SON has several components such as mobility load balancing (MLB), frequent handover mitigation (FHM), mobility robustness optimization (MRO), and interference management (IM), that help small cells to deliver carrier-grade performance. MLB distributes the UEs load among small cells to enhance the QoS and to increase system capacity. MLB utilizes cells load information to optimize the cell boundaries to offload UEs. SON uses mobility/handover parameters for load balancing [8], [9].

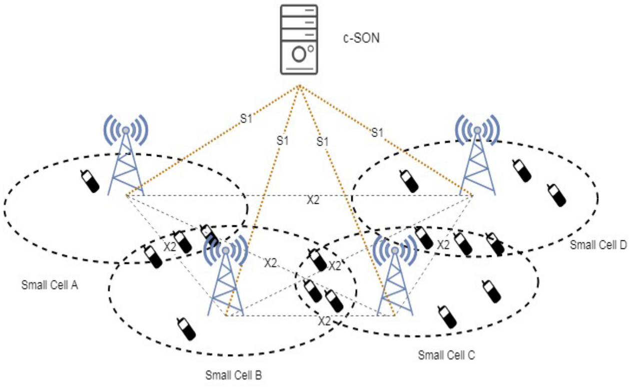


FIGURE 1. Radio access network architecture with a c-SON controller.

MLB distributes the load among the small cells by adjusting the mobility parameters (i.e., handover parameters) according to their load statuses. To shift the candidate UEs, the cell individual offsets (CIO) of the serving and neighbouring cells are adjusted by UEs based on the reported measurements. However, improper handover decisions and offloading sequence for overloaded cells in MLB might cause an inefficient usage of resources or degrade the service. For example, as shown in Figure 1, small cell A is under-loaded with a load of 50%, whereas small cells B, C, and D are overloaded and have load values of 70%, 75%, and 80%, respectively. If the fixed MLB algorithm is adopted, they sequentially take the highly-loaded cells from the list in the order of cell load and offload their excessive load to the lightly neighbouring cells. As a result, no cell can unload, but cell B, which comes third in the order and has an under-loaded neighbour cell A.

The main contribution of this paper is to introduce a Utility-based Mobility Load Balancing algorithm (UMLB) by considering both the operator utility and the user utility at the same time for each lightly loaded neighbouring cell. The main reason for choosing the utility function is that it expresses the satisfaction of any metric with a numerical value (e.g., cell load, reference signal received power (RSRP), or reference signal received quality (RSRQ)). Therefore, it is easier and less complex to compare these numerical values together to obtain the best decision during a handover process compared to the heuristic algorithms used in [10]. The algorithm starts by determining the edge-UEs of an overloaded cell that need to be offloaded to lightly loaded neighbouring cells. The offloading process takes place by handing over each candidate of edge-UEs of the overloaded cell to the best neighbouring cell by calculating the aggregated utility for each neighbouring cell. The aggregated utility is a function in the operator utility and the user utility. The operator utility is calculated for each potential handover based on the load of the neighbouring small cells.

On the other hand, the user utility calculation is based on the sigmoid function by considering different criteria (e.g., delay, data rate, etc.) for each edge-UE involving a handover process. Furthermore, we introduced a new term named load balancing efficiency factor (LBEF) that considers a load of neighbouring cells and the edge-UEs for each overloaded cell. This factor specifies the sequence of overloaded cells for the UMLB algorithm operation.

The remainder of the paper is organized as follows. Section II overviews the related work, and recent research has been done. Section III describes the system model, network architecture, and the assumptions considered. Section IV briefs handover parameters relative to load balancing, while Section V introduces the overload-status detection threshold and load estimation. The proposed work and algorithms are explained in Section VI, followed by simulation and results in Section VII. Section VIII concludes the paper.

II. RELATED WORK

Intra-LTE mobility parameters auto-adjustment based on the current load of the small cell network can enhance the system capacity compared to the static/non-optimized cell mobility parameters [11]. However, the UE QoS shall not be affected negatively with the forced load balancing.

Researchers in [12] was the first to demonstrate through simulation the effectiveness of simple load balancing algorithms in reducing the call blocking rate and increasing cell-edge throughput based on auto-adjustment of handover parameters.

In [13], based on RSRP measurements and load of neighbouring cells, overloaded cells group UEs according to the best neighbouring eNB, and for each handover offset value, it sorts the neighbouring eNBs in descending order concerning the number of possible handovers. A whole group will be handed over if their predicted load does not exceed the acceptable level at neighbouring eNB.

MLB algorithm proposed in [14] considered non-adjacent neighbourhood cells in the optimization area. The radio link condition of neighbouring cells is taken into consideration when offloading UEs from the source cells.

MLB optimization and handover parameter optimization (HPO) algorithms influence the handover decisions of the UEs. That interaction reduces the desired effects of each function. The coordination between MLB and HPO is investigated in [15]. The coordinator provides a solution that combines the strengths of the individual algorithms to better performance.

In [16], a multi-traffic load balance (MTLB) algorithm is presented to balance the traffic load and improve the network capacity with an appropriate handover procedure. A new cell selection is adapted to enhance the quality of service of UEs. Besides, the handover threshold and TTT (time to triggering) are adaptively adjusted to reduce the call drop rate with a more balanced-load network. Two conditions are accounted for the handover procedure: The signal strength condition and the RB condition. That helps in avoiding the wrong eNB or unnecessary handovers.

In [17], researchers presented an MLB algorithm and discussed its properties in terms of costs and gains. Their objective is to minimize the load standard deviation (LSD) to distribute the load evenly over the network. Simulation results show that the MLB algorithm can reduce the LSD significantly. However, the algorithm required more handovers than the no-MLB operation, and the number of RLF produced was higher as well.

Researches in [18], proposed an MLB algorithm based on cell reselection (CR) which works in accordance with the Mobility Robustness Optimization (MRO) function. In other words, when UEs are in the radio resource control (RRC) idle mode, the algorithm adjusts the CR parameters to make them camp on the lightly-loaded cell. Once UEs switch to the (RRC) connected mode, they will belong to the lightly-loaded cell selected in the idle mode.

Not only mobility of UEs can be utilized in load balancing, but also there are different types such as coverage and capacity optimization [19]. When a small cell is detected to be overloaded, the SON has a function that decreases the power and hence makes some edge-UEs offload to the lightly-loaded side of the network.

The mobility load balancing is performed by adjusting the mobility parameters; thus, the impacts of user mobility on the algorithm have been studied. Those impacts are evaluated and compared through computer simulations in [20]. Results show that MLB algorithms realize higher gain for pedestrian users with circle mobility and vehicle users in rectangle mobility.

In general, the previous work addressed the MLB problem by mainly considering the operator preference without considering the UEs preferences during the handover process. Moreover, the introduced algorithms in the literature did not follow a proper sequence in ordering the overloaded cells in the MLB problem. This paper exploits UMLB that integrates the UE utility and operator utility during load balancing in the network. Furthermore, it presents an LBEF term that is used to specify the sequence of overloaded cells for the MLB algorithm operation.

III. SYSTEM MODEL

This section defines the network model that will be investigated throughout the paper. Next, the most important system constraints are determined. Finally, we explain how the cell load is represented and calculated.

A. NETWORK MODEL

First, we introduce the network model, its parameters and assumptions. In this paper, we investigate a homogeneous network of several small cells indicated by the set $S = \{1, 2 \ldots S\}$, as depicted in Fig.1. The small cells belong to the same operator and operate in an open access mode. The centralized SON (c-SON) is adopted here. In c-SON, some optimization functions are executed at the Operation and Management system (OAM), while others are executed at eNBs.

The small cell network adopts two types of interfaces, as shown in Fig.1. Small cells connect to c-SON via the $S1$ interface [19]. However, they communicate among each other over the $X2$ interface. Small cells share handover-related information via $X2$ interface to execute handover of UEs from cell to another. Not only handover management is performed over $X2$ interface, but also load management such as resource status, and traffic load can be provided over the $X2$ interface [21]. The c-SON at the OAM periodically gathers information form small cells' c-SONs and uses them, if any overloaded cell is detected, to optimize and update the small cells' handover parameters to distribute the load over the network.

A Physical Resource Block (PRB) is the smallest unit of resources that can be allocated by a small cell eNB to a user in Long Term Evolution (LTE) $\{\text{PRB}\}$. One PRB occupies 180 kHz bandwidth in the frequency domain and 6 or 7 orthogonal frequency-division multiplexing (OFDM) symbols (i.e., one slot) in the time domain. The frequency part is composed of 12 consecutive sub-carriers of 15 kHz. Single sub-carrier with one symbol is defined as the resource element. Since LTE is capable of working with two types of duplex schemes frequency division duplexing (FDD) and time division duplexing (TDD), two different classes of radio frames are used. FDD uses Frame structure Type 1 (FST1), which is widely used type; however; TDD uses Frame structure Type 2 (FST2). Each radio frame length is 10 ms and composed of 20 slots each is 0.5 ms long. Every two sequential slots compose one subframe, which is the time unit for scheduling users, denoted as the transmission time interval (TTI) [22]. Each cell has some available PRBs that is set according to the system bandwidth. For our system, cells adopt 20 MHz bandwidth, which corresponds to 100 PRBs.

The network UEs are classified into two classes. The values of the requested QoS metrics and frequency and duration of connections differentiate between the classes. UEs that belong to the same class have the same demand for minimum data rate, the maximum delay, etc. Each class j UE can report measurements of neighbouring small cells if the RSRP of that neighbouring is greater than a predefined threshold. Those are the candidate UEs for handover, and the reported small cells are the qualified candidate small cells. However, the different classes UEs select the neighbouring cells for handover based on several criteria. The small cells' eNBs enter some network usage characteristics in their database and make them available to UEs. Examples of these characteristics could be the frequency and duration of the UEs' connections and (or) the QoS metrics (throughput, and delay).

B. SYSTEM MODEL CONSTRAINTS

This paper introduces a dynamic UMLB for LTE-A Small Cell Networks with c-SON. We assume that c-SON has a powerful capability in terms of memory to handle a set of small cells' operations. However, there are still a few constraints that severely affect the network operation if not taken into account.

1. The proposed architecture considers a centralized controller that balances the load across the small cells; consequently, the network may collapse if the controller fails.
2. Handover delay includes a pre-handover time that the UMLB takes in searching and examining the target cells. Thus, pre-handover time must be in a range that keeps the handover delay less than the allowed values [23]. Delays below 250 milliseconds (ms) are acceptable. Through simulation, we estimated the time needed for the proposed algorithm and found to be 21.404 ms.

Previous work introduced in [24] discussed possible solutions for these challenges.

C. CELL LOAD CALCULATION

To accurately measure the load of a network cell, a proper method is adopted. Several ways have been used to represent the load of the cells, such as the number of users served by a cell and the load of the transport network. In this paper, we adopt the resource block utilization ratio *RBUR*, which is the ratio of PRBs allocated by a small cell to that total available number of PRBs belong to that cell. The resource block utilization ratio directly limits the number of UEs that can be served by a specific data rate and delay constraints. The average *RBUR* of small cell s at time t over a time T is given by

$$\overline{RBUR_s}(t) = \frac{\sum_{\tau \in (t-T,t)} \sum_{j \in J} I_{s,j} * N_{s,j}(\tau)}{T * PRB} \tag{1}$$

where $I_{s,j}(\tau)$ is a binary indicator so that $I_{s,j}(\tau) = 1$ if user j is served by small cell$s$, $N_{s,j}(\tau)$ is the number of physical resource blocks assigned by small cell s to UE j at period t, and PRB is the total number of resource blocks available at cell s. Note that each UE is served by only one small cell. Furthermore, all small cells have the same limited number of PRBs. Hence, the total allocated number of PRBs by a small cell s at time t cannot exceed the maximum number of PRBs of the cell, $\sum_{j \in J} I_{s,j}(\tau) * N_{s,j}(\tau) PRB, \forall s$.

When the value of *RBUR* reaches 1, the cells' resources are depleted and, any UE coming into this cell will either be disconnected or served by a lower data rate. In this paper, we ignored any call admission control policy. Thus, when a new UE moves into an overloaded cell, it will be admitted by that cell, but the per-UE throughput in this cell will be affected. Hence, shifting UEs forcibly from the highly loaded cells to normal or lightly loaded cells is critical to mitigating the overload status.

D. LOAD BALANCING PROBLEM FORMULATION

Network performance determined by Key Performance Indicators (KPIs) indicates its QoS. Based on these KPIs, the c-SON identifies the optimum handover parameters for the edge-UEs and involving small cells to achieve a more stable network with the highest achievable QoS concerning load demand. Following is the KPI considered for the dynamic mobility load-balancing problem.

1) KEY PERFORMANCE INDICATOR FOR LOAD STANDARD DEVIATION

In this paper, a load standard deviation σ is monitored, which determines the level of load balancing in the network at a time and is evaluated by using the load distribution in all small cells. The load standard deviation is defined as

$$\sigma = \sqrt{\frac{\sum_{s \in S} \left(\overline{RBUR_s}(t) - RBUR_{Net}(t)\right)^2}{S}} \tag{2}$$

where $RBUR_{Net}(t)$ is the average $RBUR$ of a network of S small cells at time t during period T as well. For simplicity, we omit the time symbols throughout the paper.

The range of σ is in the interval [0, 1], with a lower value representing a highly balanced load distribution amongst all active small cells. Therefore, minimizing σ is one of the objectives of this work to achieve a highly balanced load in the small cell networks.

IV. MOBILITY CONTROL PARAMETERS

A network lets UEs report measurements of the signal quality of the serving and neighbouring cells either periodically or as event-driven reports. The signal qualities required to be measured by serving cells can be RSRP or RSRQ. However, LTE-A offers a set of event-driven measurement report mechanisms to minimize the signalling overhead in the network [25]. Those events are performed by UEs for both the serving cell and the neighbouring cells. Therefore, we adopted some of them in our work for network information gathering phase and handover execution procedure.

A. LTE EVENTS

LTE specified eight types of events a UE must report by Radio Resource Control (RRC) Connection Reconfiguration message. Events A1 to A6 are defined for intra-LTE mobility, while events B1 and B2 are tailored for inter-RAT mobility. Since we are dealing with intra-LTE mobility, we focus only on the A1 to A6 events.

If the criteria for a certain event have been satisfied for a predefined time, called time to trigger (TTT), the UEs perform intra-LTE event-triggered reporting. RSRP or RSRQ are the quantities used by the LTE network system for events measurement triggering and reporting. Some events are threshold-based events for which the network operator sets a predefined threshold. For instance, the A1 event is triggered by UE when the serving cell becomes better than a given threshold, whereas A2 event is triggered when the serving cell becomes worse than a predefined threshold. Event A4 is triggered by UE when a neighbouring cell becomes better than a predefined threshold. However, event A5 is triggered when the serving cell becomes worse than a given threshold, and a neighbouring cell becomes better than another given threshold.

On the contrary, events A3 and A6 are triggered using an offset value, which is a type of relative value regarding something such as the serving cell. Event A3 is triggered when a neighbouring cell becomes offset better than a primary serving cell, whereas event A6 is triggered when a neighbouring cell becomes offset better than a secondary serving cell. In our work, we will follow the standard in employing A3 event measurements for triggering and reporting handovers. Also, we adopt A4 event measurements to gather candidate edge-UEs and their corresponding neighbouring cells for the hastened handover.

B. A3 AND A4 EVENTS FOR LOAD SHIFTING AND EDGE-UEs FINDING

As stated above, event A3 is triggered based on the relative signal quality of the neighbouring cell. In other words, a UE triggers event A3 and report the measurement to its serving cell when a neighbouring cell shows a better signal quality than the serving cell by some offset in dB. Thus, event A3 has been commonly used for triggering handovers in wireless networks. The small cell eNB configures the UE to measure the signal quality (RSRP) of the serving cell and neighbouring cells and trigger a handover when an "entry condition" has been maintained for a duration of time larger than TTT. The entry condition to trigger and report the A3 event measurements are expressed as follows:

$$M_n + O_{fn} + O_{cn} - Hyst > M_s + O_{fs} + O_{cs} + Off \quad (3)$$

where M_n and M_s are the RSRPs of the neighbouring cell and the serving cell, respectively. The O_{fn} corresponds to the frequency-specific offset of the frequency of the neighbouring cell and the serving cell. However, O_{fs} corresponds to the frequency-specific offset of the frequency of the serving cell. O_{cn} is the cell-specific offset of the neighbouring cell, whereas O_{cs} is the cell-specific offset of the serving cell. $Hyst$ is a hysteresis term for cell s; and Off is the A3 event offset between the serving and neighbouring cells. Since we consider only intra-frequency handovers in this paper, we ignore the inter-frequency parameters, i.e., O_{fn} and O_{fs}.

By adjusting the values of the parameters, O_{cn}, O_{cs}, and Off of the above equation, it is possible to cause a particular UE currently served by cell s to hand-over to the neighbouring cell n. Therefore, we can deliberately perform early, or late handovers based on the load status of the serving and neighbouring cells.

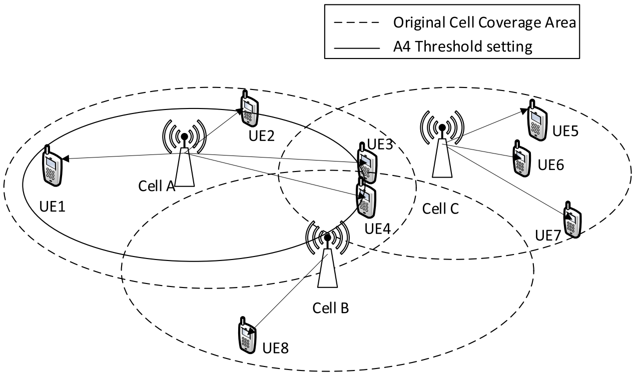


FIGURE 2. Original cell coverage and A4 event measurement reporting.

If we increase the value of O_{cn} of a peculiar neighbouring or decrease the value of O_{cs} of the serving cell, the cell range diminishes virtually, and hence the UEs will be handed over to the neighbouring cell. In contrast, reducing the value of O_{cn} or increasing the value of O_{cs} will increase the serving cell range which forces UEs to handover from the neighbouring to it; and hence the cell load increases. For example, in Fig.2 cell A is highly loaded with four UEs, and its neighbouring cells, B and C, both have less load with one UE and three UEs, respectively. At the overlap area of cells A, B, and C, there are two UEs, UE 3 and 4. Therefore, by increasing the O_{cn} of cell B or cell C, UEs 3 and 4 handover from cell A to either cell B or cell C, and the network might become more balanced. As mentioned before, changing O_{cs} of the serving cell might delay or speed the handover as well, but it may affect cell range of all the neighbouring cells. Thus, the parameter O_{cn} is more suitable since it shifts the load only to a neighbouring cell.

The c-SON can configure the handover parameters of event A3 to achieve a more balanced network. Still, the system needs information on the edge users' potential for early handovers. To that end, in our work, A4 event is used to gather information on edge users of the overloaded cells. The UE triggers A4 event: when the RSRP of a neighbouring cell becomes better than a provided threshold:

$$M_n + O_{fn} + O_{cn} - Hyst > Threshold \quad (4)$$

where *Threshold* is the event A4's threshold. If entry conditions of A4 event are satisfied by a UE, the UE can report measurements such as RSRP for the serving cell and neighbouring cell. The UE can report multiple neighbours, which means that such a UE has several $M_n s$, $O_{fn} s$, and $O_{cn} s$. The c-SON considers those neighbouring as neighbouring candidate cells. Small cells should set reasonable A4 event threshold to collect edge-UEs' information and the candidate neighbouring small cells, both of which are required for load balancing. In Fig. 2, let us assume that UEs 3 and four from cells C and B are outside of A4 event boundary of cell A and have reported RSRPs measurements as follows:

$$\left(M_B^{UE3}, M_B^{UE4}, M_C^{UE3}, M_C^{UE4}\right) + O_{fn} + O_{cn} - Hyst > Threshold_A \tag{5}$$

And in the meantime, the following conditions are true:

$$M_C^{UE3} > M_B^{UE3}$$
$$M_C^{UE4} > M_B^{UE4}$$

Based on measurement reports gathered from UEs three and four and load status of neighbouring cells B and C, cell A can hand UEs three and four over to cells B and/or C by increasing O_{cn} for cell B and C when it becomes overloaded.

Thus, we utilize A4 event measurement reports for collecting information on edge-UEs that can report multiple neighbouring cells, and then c-SON selects the best neighbouring cell based on a combination of criteria. The c-SON collects edge-UEs' information ε from small cells, which is $\varepsilon = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_S\}$, where ε_S is the set of edge-UEs of small cell s.

V. ADAPTIVE UTILIZATION THRESHOLD AND LOAD ESTIMATION

An adaptive threshold is needed to detect the overload state of the cells before starting the network balance by transferring suitable loads among cells.

A. AN ADAPTIVE UTILIZATION THRESHOLD FOR LOAD STATUS DETECTION

Two types of thresholds have been introduced to detect the status of the cells' load in the wireless networks: fixed and adaptive thresholds. The fixed threshold is not effective when applied to all scenarios as the load status of the mobile wireless network is dynamic and changes over space and time. Hence, an adaptive threshold has been introduced to adapt to the status of the network load. This method was proved to be better than the fixed threshold. Inspired by [27], the adaptive threshold is defined to be the maximum of the average network load or a pre-defined fixed threshold

$$TH_{Adp} = max\left(\overline{RBUR}_{NET}, TH_{INI}\right) \tag{6}$$

where TH_{INI} is an initial fixed threshold at which the MLB algorithm triggers.

B. CALCULATION OF USER's REQUIRED NUMBER OF PRBS AND AFTER-HANDOVER LOAD ESTIMATION

We are considering only the downlink of an LTE system. We assume that the interference of eNB is the only interference in the network. For a pair of cells and UE j in every time step, we assume that the Signal-to-Interference-plus-Noise ratio can be computed as

$$SINR_{s,j} = \frac{P_s \cdot L_{s,j}(d)}{N + \sum_{t \neq s} P_t \cdot L_{t,j}(d) \cdot RBUR_t} \tag{7}$$

where P_s is the transmit power for a cell s, L is the path loss mapping determined by the UE's location relative to a cell, N is the thermal noise per PRB, and $RBUR_t$ is the load of cell t. We assume that we can adopt the best modulation coding scheme (MCS) for a given SINR, which provides the highest data rate. That can be represented by Shannon formula as follows

$$r_{s,j} = B \cdot \log_2\left(1 + SINR_{s,j}\right) \tag{8}$$

where B is the bandwidth of one physical resource block (i.e., 180 kHz). Based on the data rate value demanded by an edge-UE, $R_{e_s^j}$, and the maximum achievable data rate using the assumed MCS at a given SINR $r_{s,j}$, the number of PRBs required by an edge-UE $e_s^j \epsilon \varepsilon_S$ to keep up with the throughput requirement is estimated by

$$N_{PRB}^{e_s^j} = \left\lceil \frac{R_{e_s^j}}{r_{s,e_s^j}} \right\rceil \tag{9}$$

where $\lceil\cdot\rceil$ is the ceil function. Note that the scheduler is the component that decides the number of PRBs allocated to UEs at a given time t. In our paper, we are considering the Channel-Aware Quality of service scheduler (CQA). In addition to the SINR, we are taking into consideration the UEs' RSRPs measurements of the serving and neighbouring cells. Number of RBs required serving edge-UEs before and after handover differs because they experience different RSRP values from two different small cells. To perform a good load balancing, and before triggering the handover procedure, the algorithm should determine the current load of the edge-UEs to be handed over and estimate their load at the neighbouring cells.

Therefore, for a given number of PRBs $N_{PRB}^{e_s^j}$, the average load to serve the UE e_s^j of small cell s is denoted by $\rho\left(s, e_s^j\right)$ and calculated as following:

$$\rho\left(s, e_s^j\right) = \frac{N_{PRB}^{e_s^j}}{PRB_s} \tag{10}$$

Then, we can estimate the after-handover generated load by edge-UEs, $\hat{\rho}\left(k, e_s^j\right)$, at a neighbouring small cell k by

$$\rho\left(k, e_s^j\right) \approx \rho\left(s, e_s^j\right) \cdot \frac{Ms_{\left(s, e_s^j\right)}}{Mn_{\left(k, e_s^j\right)}} \tag{11}$$

where $Ms_{\left(s,e_s^j\right)}$ and $Ms_{\left(k,e_s^j\right)}$ are the RSRPs of serving cell s and neighboring cell k measured by edge-UE $\mathrm{e}_\mathrm{s}^\mathrm{j}$, respectively.

VI. PROPOSED WORK

A. DATA GATHERING VIA NETWORKING MONITORING

The c-SON supports automatic information gathering by monitoring the network. It periodically collects various information from the network small cells of the studied model. If the load of a small cell s exceeds the computed adaptive threshold, the small cell is considered overloaded and should force a few UEs to handover to lightly load neighboring cells.

Not only the cell load status is monitored, but also the information on the UEs that are at the edges of the small cells is gathered. For that, the c-SON tunes the A4-event threshold for each small cell and collects information on the UEs that are moving close to the edge. The information gathered is relative to the UEs' serving cells and their neighbouring cells. The A4-event threshold is computed and adaptively adjusted based on the RSRPs reported by A3-events that are performed by a set of UEs to cell s for a predefined period, T. The RSRPs reported under A3 events by UEs in a serving cell s are averaged, $\overline{Ms}$ for that serving cell. Then, the A4 threshold of serving cell s is set to the average of $\overline{Ms}$ of the neighboring cells is defined as

$$TH_i = \frac{1}{||B_s||} \times \sum_{j\in B_s} \overline{Ms}_s \tag{12}$$

where $\overline{Ms}_s$ is the average RSRP reported by UEs at each serving small cell s; and $||B_s||$ is the neighbouring cells set of serving small cell s set that is reported by UEs based on A3 event measurement reporting during a time duration.

Next, the c-SON gathers measurement reports based on the A4-event threshold TH_i from edge-UEs of each serving cell. These UEs are the candidate UEs to be shifted to the neighbouring lightly loaded cells when their serving cells become severely loaded. The c-SON creates a database for each cell based on the A4 events-triggered measurements. The database contains information on reporting UEs: their Identities, their serving and neighbouring cells' RSRPs, SINRs, etc. Let $\varepsilon_s = \{e_s^1, e_s^2, \ldots e_s^{||\varepsilon_s||}\}$ indicates the set of edge-UEs that reported A4 measurements to serving cell s. They are listed in ascending order of the RSRP, Ms of the serving cell s. The c-SON creates another set that contains the neighboring cells reported by edge-UEs of the serving cell. Let us assume that each edge-UE in set ε_s can report multiple neighboring cells $\mathcal{T}_{e_s^j} = \{\mathcal{T}^1, \mathcal{T}^2, \ldots, \mathcal{T}^L\}$ where L is the number of the candidate neighboring cells for edge-UE e_s^j.

B. UMLB ALGORITHM

The UMLB algorithm is run periodically by c-SON. To that end, the c-SON hand-overs candidate edge-UEs from the highly loaded serving cells to the normal or under-loaded neighbouring cells based on the utility function. First, all the small cells report their load information, $RBUR$ to the c-SON. Next, reporting small cells are sorted in descending order of $RBUR$.

Afterward, the algorithm compares the max load, $RBUR_{max}$ in the list with the predefined initial threshold. If the $RBUR_{max}$ is greater that the initial static threshold, the network is in overload status, and it requires an immediate load-balancing.

Then, for the algorithm to be adaptive to the network load status, we set the adaptive threshold, TH_{Adapt}, using equation 6. The current load of each small cell $\overline{RBUR}$ is compared with the adaptive threshold TH_{Adapt} to detect the status of the load. If it is greater than the adaptive threshold, the cell is in an overload status, and accordingly, the c-SON algorithm performs load balancing.

The algorithm creates a new set O that contains all overloaded cells such that $\overline{RBUR}_o \geq TH_{Adapt}$ for $o \in O$ and $O \subset S$. Since the MLB algorithm mainly relies on the load status of the neighbouring cells and the UEs' positions of overloaded cells, the algorithm rearranges all overloaded cells in the set O according to the remaining capacity of the neighbouring cells and the estimated edge-UEs' load at the neighbouring cells. To that end, we introduce the load balancing efficiency factor (LBEF) for an overloaded cell, which is defined as follows:

$$\psi_o = \sum_{k\subset B_o} min(\sum_{j} \rho\,(k,j)\,, (1-\widehat{RBUR}_k)) \tag{13}$$

where B_o is the set of neighbouring cells reported by edge-UEs in cell o, $\rho\,(k,j)$ is the estimated after-handover edge-UEs load at the neighbouring cell, and $(1-\widehat{RBUR}_k)$ indicates the number of RBs remaining at neighbouring cell k. Subsequently, the c-SON rearranges the set O in decreasing order of ψ_o.

Then, the algorithm takes the overloaded cells one by one from the set and decreases its load to under-loaded neighbouring cells by handing-over the candidate edge-UEs. Each overloaded cell is computing the maximum load that can be moved to the target cells. That is to prevent the ultra-lightly loaded cells from becoming overloaded, and the serving cell from becoming underloaded. In other words, the UE is handed over if the maximum load moveable from the overloaded cell, $\tilde{\rho}_o$ is greater that the UE after-handover estimated load, $\rho\left(o, e_o^j\right)$. The moveable load is calculated by

$$\tilde{\rho}_o = \overline{RBUR}_{NET} - \frac{1}{2}\rho\left(o, e_o^j\right) \tag{14}$$

However, unlike the work that has been done in the literature, the handover process is based on the concepts of the utility function. In the following subsections, we model our system using the utility function.

The algorithm of selecting a neighbouring small cell eNB for early handover initiates with calculating the user utilities. We evaluate the UE utilities for each criterion for each neighbouring small cell candidate. The utility of a criterion is normalized to scale the interval [0, 1], i.e. $u\,(x) \in [0, 1]$, which indicates the UE satisfaction level from the criterion value offered by the small cell eNB. For instance, consider a scenario where a small cell eNB can serve a maximum of 18 Mbps and the requirement of a UE is in the range [5 Mbps, 30 Mbps]. With the help of the utility function, we can calculate the utility of the user for this data rate for this given eNB. In the second step, the c-SON evaluates the operator utility for the load criterion for each small cell eNB, which is a normalized function of a single criterion.

1) UEs' UTILITY CALCULATION

We let UEs report multiple neighbouring cells based on event A4 measurements. In other words, let us look at the scenario of Fig.2 again, in which UEs 3 and 4 report small cells B and C if both M_B^{UE3} and M_B^{UE4} are greater than $Threshold_A$ of event A4. That means each UE reports two neighbouring cells to the c-SON.

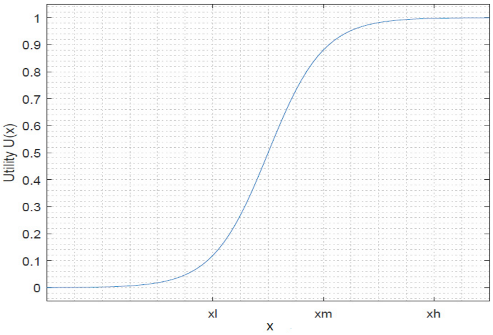


FIGURE 3. Utility calculation using sigmoid function [28].

We are considering a vector of n criteria, $X = \{x_1, x_2, \ldots, x_n\}$ with an associated vector of n weights $W = \{w_1, w_2, \ldots, w_n\}$ for the handover process. As stated above, the UEs concern about three criteria in this work: delay, data rate, and RSRP. The weights represent the UE's preference level for a criterion x_i. The utility of a class j UE for a small cell eNB s and a predefined criterion x_i, denoted as $u_j^s(x_i)$, is calculated using a Sigmoidal (S-shaped) function, as presented in [28]. The utility is used to quantify the UE's satisfaction for a given criterion. Several utility function forms were examined whether they satisfy the required properties: twice differentiability, increasing function, concavity and convexity conditions. As a result of the examination, it was proved that only the sigmoidal functions can satisfy the required conditions of a utility function. The sigmoid function is shown in Fig.3 and expressed by the equation

$$u(x) = \begin{cases} 0, & x_n < x_l \\ \dfrac{\left(\frac{x_n - x_l}{x_m - x_l}\right)^{\zeta}}{1 + \left(\frac{x_n - x_l}{x_m - x_l}\right)^{\zeta}}, & x_l \le x_n \le x_m \\ 1 - \dfrac{\left(\frac{x_h - x_n}{x_h - x_m}\right)^{\gamma}}{1 + \left(\frac{x_h - x_n}{x_h - x_m}\right)^{\gamma}}, & x_m \le x_n \le x_h \\ 1, & x_n > x_h \end{cases} \tag{15}$$

where $\zeta \ge max\left\{\frac{2(x_m - x_l)}{x_h - x_m}, 2\right\}$ and $\gamma = \frac{\zeta(x_h - x_m)}{x_m - x_l}$ are the parameters that determine the steepness of the utility curve, x_n is the value obtained for the criterion x, x_l is the minimum acceptable value for the criterion x, x_h is the maximum desired value for the criterion x, and x_m is a user-specific value that separates the satisfied from unsatisfied areas.

It is important to notice that equation (15) is defined for upward criteria for which the higher the values, the greater their utility (e.g., data rate, RSRP). However, we use $1 - u(x)$ for downward criteria for which the lower the metric value, the greater their utility (e.g., delay, load).

$$\mathfrak{U}_{e_o^j}^{\tau^l}(Z, W) = \prod_{i=1}^{n}\left[\mathrm{u}_{e_o^j}^{\tau^l}(x_i)\right]^{w_i}, \quad e_o^j \in J, \tau^l \in S,$$

$$\sum\nolimits_{\boldsymbol{i}} w_i = 1 \tag{16}$$

In this form, the interactions/dependence among the considered criteria is considered. Also, it can take into consideration the UE preference weights for different criteria.

2) OPERATOR UTILITY CALCULATION

In the second step, for every candidate edge-UE, e_o^j of an overloaded cell $o \in O$, we compute its load contribution in the current serving cell as $\rho\left(o, e_o^j\right)$ and estimate the after-handover load in each reported neighbouring cell as $\hat{\rho}\left(\tau^l, e_o^j\right)$. Hence, we can estimate the after-handover load of the neighbouring cell as

$$\widehat{RBUR}_{\tau^l} = \overline{RBUR}_{\tau^l} + \hat{\rho}\left(\tau^l, e_o^j\right) \tag{17}$$

where $\overline{RBUR}_{\tau^l}$ is the estimated current load of the neighbouring cell.

After computing the current edge-UE load contribution and estimating the after-handover neighbouring cell load, we use the following formula to compute the operator utility for each neighbouring cell eNB and each class j UE, which is expressed as follows

$$u_{e_o^j}^{\tau^l} = \begin{cases} 1 - \widehat{RBUR}_{\tau^l}, & \overline{RBUR}_o - \rho\left(o, e_o^j\right) > \widehat{RBUR}_{\tau^l} < TH_{Adp}^{\tau^l} \\ 0, & Otherwise \end{cases} \tag{18}$$

Note that this equation is for the downward criterion for which the lower the metric value, the greater the utility. The load cost values are normalized to prevent it from dominating the handover utility function. Here, conditions of the operator utility restrict the release of the load from the overloaded cell so that this cell does not become underutilized and the neighbouring cell does not get overloaded either, and hence the algorithm does not enter into an infinite loop of load balancing. When the operator utility is zero, it means that the small cell is overloaded, and it must be eliminated in the elementary operator utility.

3) AGGREGATE UTILITY AND HANDOVER

The overall utility for neighbouring small cell eNB τ^l and class j UE is obtained by aggregating the UE and operator utilities for this small cell eNB. To that end, we apply the multiplicative aggregation form again to calculate the neighbouring cell eNB utility as follows

$$\mathfrak{U}^{\tau^l}_{e^j_o} = \mathfrak{U}^{\tau^l w_u}_{e^j_o} \times u^{\tau^l w_o}_{e^j_o}, \tau^l \in S, e^j_o \in J, \quad w_u + w_o = 1 \tag{19}$$

where w_o and w_u represent the operator and user utility weights, respectively. Then, the best neighbouring cell eNB to handover the candidate edge-UE τ^l to is the one with the greatest utility value among all $\mathfrak{U}^{\tau^l}_{e^j_o}$, $\tau^l \in S$.

Hence, the algorithm updates the related cell individual offsets. Notice that the offsets are always set symmetrical to prevent ping-pong. Eventually, the algorithm updates the serving and neighbouring cells' load information as follows:

$$\widehat{RBUR}_o = \widehat{RBUR}_o - \rho\left(o, e^j_o\right) \tag{20}$$

$$\widehat{RBUR}_{\tau^l} = \widehat{RBUR}_{\tau^l} + \hat{\rho}\left(\tau^l, e^j_o\right) \tag{21}$$

Characteristics such as the behaviour of the UE, metrics for QoS are dynamically updated so that the UEs read the updated values and recompute the utilities again. The whole above process applies to the overload cells list. The process for utility-based mobility load balancing is depicted in Algorithm 1.

Algorithm 1 Utility-Based MLB Handover Algorithm

```
Input: cellist, selected_UE.
Output: a cell to receive the selected UE
$MaxAggragate \leftarrow MIN$
$allocated\_Cell \leftarrow None$
$TCells\ [\,] \leftarrow selected\_UE.getNeighboringCells$
foreach cell in $TCell[\,]$ do
    $Aggregate \leftarrow$
        $get.cell.UEUtilityget.cell.OperUtility$
    If ($Aggregate > MaxAggragate$) then
        $MaxAggragate \leftarrow Aggregate$
        $allocated\_cell \leftarrow cell$
    endif
end for
return $allocated\_cell$
```

4) ILLUSTRATIVE SCENARIO

Consider the scenario shown in Fig.2. Let us suppose that UE3 is a class 1 UE being served by cell A and report cells B and C. By using values of class 1 users in Table 2 and equation (15); we define the terms of equation (15) as follows: delay: $x_l = 0$, $x_m = 0.5$, $x_h = 0.75$, *the wight* $= 0.22$, data rate: $x_l = 128$, $x_m = 256$, $x_h = 512$, *the wight* $= 0.38$, and RSRP: $x_l = -144$, $x_m = -100$, $x_h = -55$, *the wight* $= 0.4$.

Then, we define x_n which might be the delay, data rate or RSRP values offered by cell B or cell C. For cell B: $x_n(\text{Delay}) = 0.4$, $x_n(\text{Data Rate}) = 768$, and $x_n\,(\text{RSRP}) = -110$.

For cell C: $x_n(\text{Delay}) = 0.3$, $x_n(\text{Data Rate}) = 200$, and $x_n\,(\text{RSRP}) = -55$.

Now substitute them in the first part of the equation (15). We start with the delay for cell B. Since $x_l \leq x_n \leq x_m$ $(0 \leq 0.4 \leq 0.5)$, we use the second part of the equation (15):

$$u^{\mathrm{B}}_3\,(Delay) = 1 - \frac{\left(\frac{x_n - x_l}{x_m - x_l}\right)^{\zeta}}{1 + \left(\frac{x_n - x_l}{x_m - x_l}\right)^{\zeta}}$$

We calculate ζ by $\zeta \geq max\left\{\frac{2(x_m - x_l)}{x_h - x_m}, 2\right\}$

$$\zeta \geq max\left\{\frac{2\,(0.5-0)}{0.75-0.5}, 2\right\} \geq max\,(4, 2) = 4$$

Thus,

$$u^{B}_3\,(Delay) = 1 - \frac{\left(\frac{0.4-0}{0.5-0}\right)^4}{1 + \left(\frac{0.4-0}{0.5-0}\right)^4} = 0.7094$$

Notice that we subtract it by one since delay criterion is a downward criterion.

Next, we do this again for data rate and because $x_n > x_h$, we use the last part of the equation (15) and hence

$$u^{\mathrm{B}}_3\,(Data\ Rate) = 1$$

Next, we compute the UE3 utility for RSRP criterion. We note that $x_l \leq x_n \leq x_m$ $(-144 \leq -110 \leq -100)$, and thus we use the second part of the equation (15)

$$u^{B}_3\,(RSRP) = \frac{\left(\frac{-110-(-144)}{-100-(-144)}\right)^{\zeta}}{1 + \left(\frac{-110-(-144)}{-100-(-144)}\right)^{\zeta}}$$

And ζ is calculated by

$$\zeta \geq max\left\{\frac{2\,(-100-(-144))}{-55-(-100)}, 2\right\}$$
$$= \max\,(1.95, 2) = 2$$

$$u^{\mathrm{B}}_3\,(RSRP) = \frac{\left(\frac{-110-(-144)}{-100-(-144)}\right)^2}{1 + \left(\frac{-110-(-144)}{-100-(-144)}\right)^2} = 0.3738$$

Hence, UE3 utility from cell B is combined using equation (16) as follows

$$U^{\mathrm{B}}_j(UE3\ utility\ comibation\ from\ cell\ B)$$
$$= u^{\mathrm{B}}_j\,(Delay)^{w_{Delay}} * u^{\mathrm{B}}_j\,(DataRate)^{w_{Datarate}}$$
$$* u^{\mathrm{B}}_j\,(RSRP)^{w_{RSRP}}$$
$$U^{\mathrm{B}}_3 = 0.7094^{0.22} * 1^{0.38} * 0.3738^{0.4} = 0.6255$$

For cell C, we do the same steps as we did for cell B. For the sake of brevity, we brought the final answers as follows

$$u_3^C\,(Delay) = 0.8852$$

$$u_3^C\,(Data\ Rate) = 0.24$$

Next, to calculate the RSRP UE3 utility, we use the third part of the equation (15):

$$u_3^C\,(RSRP) = 0.3970$$

Finally, we compute the UE3 combined utility from cell C as follows (equation (16)):

$$U_j^3(UE3\ utility\ comibation\ from\ cell\ C)$$

$$= u_j^C\,(Delay)^{W_{Delay}} * u_j^C\,(DataRate)^{W_{Datarate}}$$

$$* u_j^C\,(RSRP)^{W_{RSRP}}$$

$$U_3^C = 0.8852^{0.22} * 0.24^{0.38} * 0.3970^{0.4} = 0.3911$$

Now, let us calculate the operator utility. We assume that the cell A is 82% loaded, cell B current load is 70%, the current and estimated load for the UE3 at cell B is 10%, and the overload threshold is 81%. Then, we check the condition of equation (18) as follows

$$82 - 10 >?\ 80 + 10 <?\ 81$$

No, it's $72 < 90 > 81$. Thus

$$u_3^B\,(Operator) = 0$$

We do the same to calculate cell C operator utility. We assume that cell A is 82% loaded, cell C current load is 45%, the current and estimated load for the UE3 at cell B is 10% and 20%, respectively, and the overload threshold is 81%. Then, we check the condition of equation (18) as follows

$$82 - 10 >?\ 45 + 20 <?\ 81$$

Yes, the condition is satisfied: $72 > 65 < 81$

$$u_3^C\,(Operator) = \frac{1 - 65}{100} = 0.35$$

Finally, we compute the aggregate utility for UE3 using equation (16). w_u, w_o are user and operator utility weights, respectively. Let's suppose $w_u = w_o = 0.5$, thus cell B utility is given by

$$U_3^B\,(Aggregate\ utility) = 0.6255^{0.5} * 0^{0.5} = 0$$

And for cell C, the utility is given by:

$$U_3^C\,(Aggregate\ utility) = 0.3911^{0.5} * 0.35^{0.5} = 0.3699$$

We note that $U_3^C > U_3^B$, and hence UE3 is handed over to Cell C. We iterate these steps for all edge users and their reported cells.

TABLE 1. Simulation parameters.

Parameters	Values
System bandwidth	20 MHz
Transmission power	24 dBm
Number of small cells	10
Inter-site distance (ISD)	30m
Antenna mode	Isotropic
Number of UEs	80
Pathloss (NLOS)	$147.4 + 43.3\,log10(R)$
Fading	Standard deviation 4 dB, log-normal
Scheduler	CQA$_{ff}$
CIO_{min} and CIO_{max}	-6 dB, 6 dB
Hysteresis	2 dB
Δ	1 dB
Initial Threshold	0.75

TABLE 2. Criteria values requested by users.

Criterion	Requested values from user: min, mean, max/weight	
	Class 1 user	Class 2 user
Delay (ms)	0, 0.5, 0.75 / 0.22	0, 0.75, 0.9 / 0.35
Data rate (Mbps)	512,256,128/ 0.38	1000,512,256 /0.33
RSRP(dBm)×10^{-10}	-144,-100,-55/ 0.4	144, -90, -44 / 0.32

VII. SIMULATION

A. SIMULATION ENVIRONMENT

To study the proposed UMLB algorithm performance, we conducted a system-level simulation. A small cell network consisting of 10 small cells and 80 UEs were assumed in the simulation. UEs are split into two classes. Each small cell is assumed to use a bandwidth of 20 MHz. Hence, the number of total available resources is 100 PRBs. The transmission power is set to 24 dBm. The path loss was modelled as a non-line of sight propagation loss [29]. For allocating resources to the UEs, the channel and QoS aware (CQA) scheduler is used. We set the initial overload threshold to 0.75 for the proposed algorithm. The rest of the parameters are shown in Table 1. In the considered scenario, a full-buffer traffic model is used.

Regarding the initial UEs distribution over the network, 50% of UEs were static and non-uniformly distributed over the overlapping area of the small cells. For the sake of mobility, the remaining 50% of UEs were modelled with a circular way (CW) mobility model with a speed of 3.6 km/h and randomly distributed over the network coverage area.

B. CALCULATION OF USER AND OPERATOR UTILITIES

After gathering information from the cells' eNBs, detecting the overloaded cells and the candidate edge-UEs for handover, the UMLB algorithm calculates the UE utilities of the desired criteria. Table 2 gives the minimum, mean, and maximum values requested by the user for each criterion, as well as their corresponding weights. Then, the utility of each criterion is evaluated using values in Table 2 and

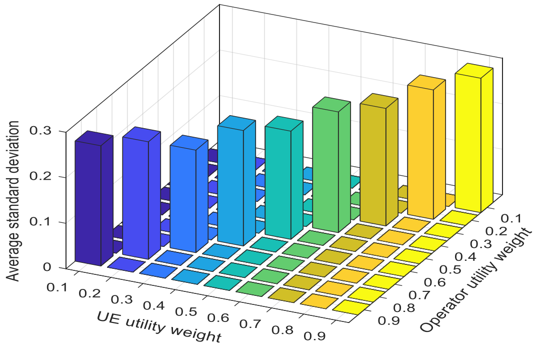


FIGURE 4. Standard devaition vs operater and UE utility weights.

Equation (15). Besides, the eNBs' offered values for data rate and delay are calculated based on eNB statistics that are logged over a sliding end of an interval for the users connected to it. while the RSRP values are calculated based on the propagation model mentioned in the previous section.

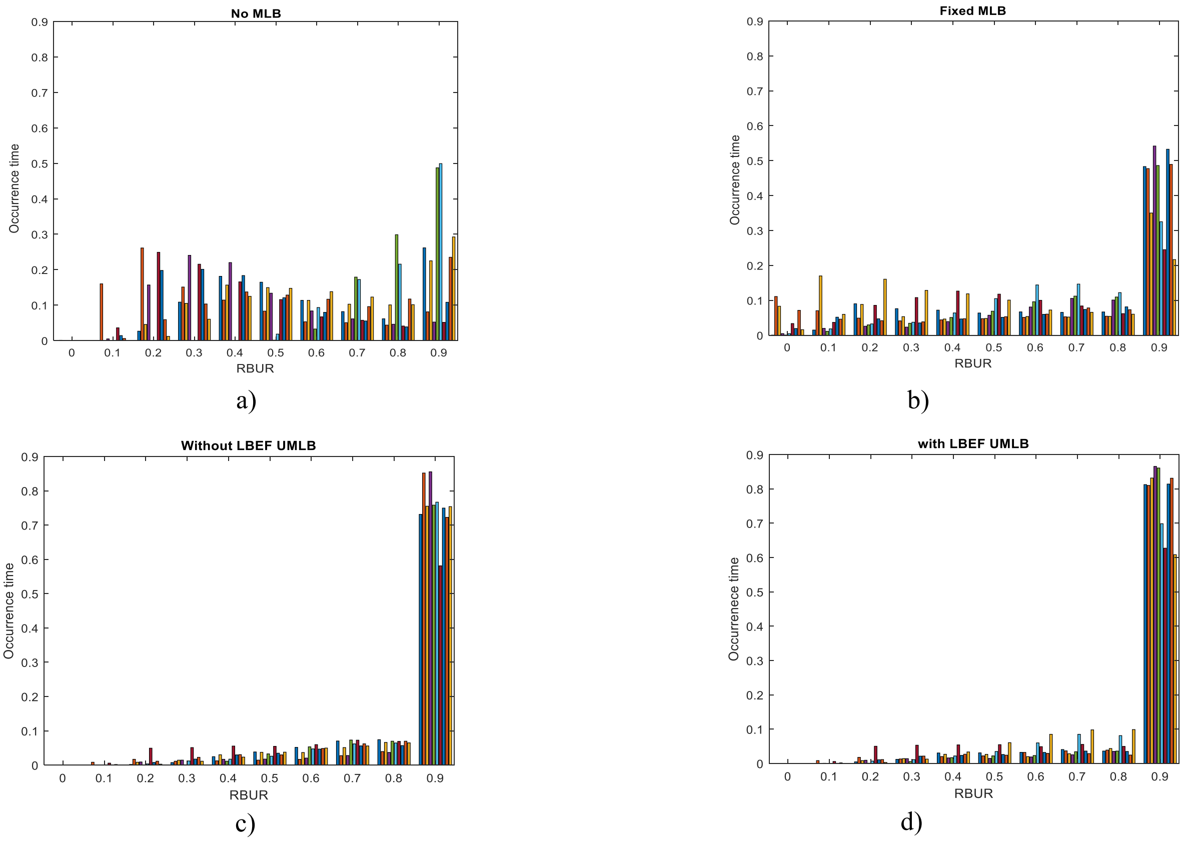
The c-SON uses instant cell RBUR values and their maximum capacity to calculate the operator utilities using Equation (18).

The weights of the aggregation attributed to the operator and UE utilities influence the standard deviation. We examined the potential combinations of the weights, as shown in Fig. 4. It is clear the standard deviation is at its minimum when the operator utility weight is 0.7, and the UE utility weight is 0.3, respectively.

C. PERFORMANCE EVALUATION METRICS

We investigated the performance of no-MLB, Fixed-MLB, W/O LBEF-UMLB, and LBEF-UMLB algorithms in terms of the standard deviation, the UE average data rate. The standard deviation is a metric used to measure the load distribution across the network. The effect of MLB algorithms on load distribution across the network was examined. Fig. 5 shows the RBUR for the scenarios that do not implement MLB as well as for the ones with MLB algorithms. Cells are represented by coloured bars ordered from left to right. Fig. 5a shows that cells 5 and 6 have a load greater than 0.9 for more than 50% of the operation time.

However, cells 4 and 7, which are neighbours of cells 5 and 6, respectively, have been underutilized at an RBUR less than 0.7 for 90% of the operation time. On the contrary, when the MLB algorithms are adopted, as shown in Fig. 5b, c and d, the highly loaded cells shifted some of their load to the lightly loaded cells. As a result, the load across small cells became more balanced. The figures show that RBUR of cells 4 and 7 became 0.9 for 80% and 55% of the time, respectively. Hence, for RBUR values greater than 0.9, the gap between the RBUR occurrence times for cells is reduced, which means the load became evenly



a)

b)

c)

d)

FIGURE 5. RBUR status of the network: (a) NO MLB algorithm (b) fixed MLB (c) W/O LBEF UMLB (d) with LBEF UMLB.

distributed among the cells. The proposed UMLB algorithm reports a higher RB utilization since the proposed load balancing mechanism considers multiple targets when handing over UEs. Furthermore, we can notice that the LBEF-UMLB algorithm introduces a slight enhancement to the UMLB for RBUR values greater than 0.9. That is due to the capability of the algorithm in offloading the proper overloaded cells first. The cell with a lightly-loaded neighbourhood has the priority to be offloaded. If this metric is ignored and the algorithm follows the classic sequence (starting with the maximum-loaded cell), some cells might not have the chance to shift some UEs, especially during the initial operation cycles.

Fig. 6 depicts the system performance for several MLB mechanisms in terms of load standard deviation across the small cell network. The proposed algorithms achieved smaller standard deviation compared with the other approaches. The proposed W/O LBEF-UMLB algorithm reduces the standard deviation by 75.86% and 74.07% for No-MLB and Fixed-MLB, respectively. The proposed LBEF-UMLB algorithm reduces the standard deviation by 77.58% and 75.92% for No-MLB and Fixed-MLB, respectively. Thus, the variance in load among small cells is lowered, and therefore, the system is more balanced.

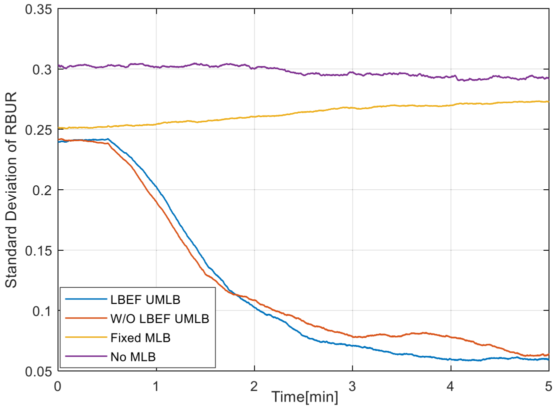


FIGURE 6. Standard deviation of RBUR among cells in the network.

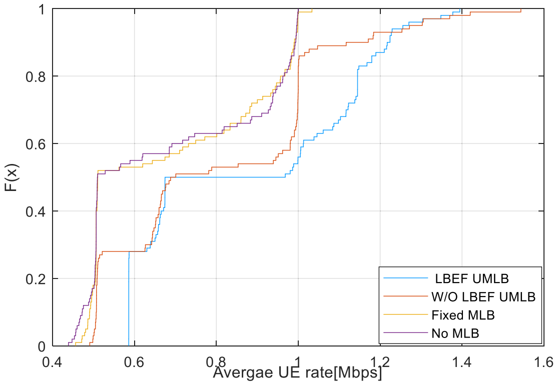


FIGURE 7. Average UE data rate [Mbps].

Moreover, we evaluated network performance in terms of UE average rate. Fig. 7 shows the average UE rate for multiple MLB approaches. Although there is always a trade-off between load balancing and throughput, the proposed algorithms increase the average UE data rate slightly. That is because shifted UEs are allocated the required RBs at the neighbouring cell. If the UE in that overloaded cell is not handed over, it will experience a limited throughput due to lack of RBs. The proposed LBEF-UMLB algorithms provide 40% of UEs with an average data rate of more than 1Mbps. On the other hand, approximately 1% of UEs can have an average data rate of 1Mbps when adopting No-MLB and Fixed-MLB algorithms.

Since the proposed algorithms consider the UEs' preferences during the handover process, the UE average delay is enhanced compared to the No-MLB and Fixed-MLB, as shown in Fig. 7. In this work, the delay is conceptualized as the difference between the achieved data rate and the required data rate.

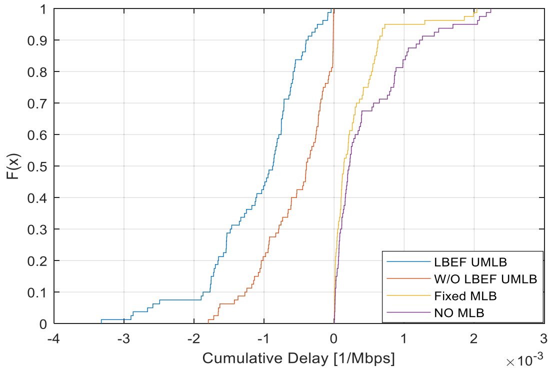


FIGURE 8. Average UE delay [1/Mbps].

Hence, it measures how much data is added to the transmission buffer. In other words, we measure delay as a fraction of offered load; this allows us to measure how many slots the UE has experienced a delay as well as calculate the size of the UE's buffer queue. Hence, the less the cell is loaded, the less the delay is. As a result, we showed the impact of the cell load status on the UE's average delay. It is apparent from Fig. 8 that the proposed algorithms exhibited a minimal delay in comparison to the other algorithms. The negative sign means that we are sending at a data rate higher than what is requested.

VIII. CONCLUSION

To conclude all this, the load-imbalance across small cells in the network due to its low service area and mobility of UEs is examined. In this paper, we introduced a UMLB algorithm and a new term named load balancing efficiency factor (LBEF). The UMLB balances the load across a small-cell network by considering the operator utility and the user utility for the handover process. The operator utility is calculated for each potential handover based on the load of the neighbouring small cells. Whereas, the user utility calculation is based on the sigmoid function by considering different criteria. Also, the LBEF considers a load of neighbouring cells and the edge-user equipment for each overloaded cell. This factor specifies the sequence of overloaded cells for the UMLB algorithm operation.

The simulation results show that the UMLB minimizes standard deviation with a higher average-UE data rate when compared to existing load balancing algorithms. Therefore, a well-balanced network is achieved. Future work is to study the impacts of UEs distribution and mobility patterns on the proposed UMLB algorithm.

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update 2017–2022," Cisco, San Jose, CA, USA, White Paper 1486680503328360, Feb. 2019.

[2] J. Hoadley and P. Maveddat, "Enabling small cell deployment with HetNet," *IEEE Wireless Commun.*, vol. 19, no. 2, pp. 4–5, Apr. 2012.

[3] "Small cells, what's the big idea?" Small Cell Forum, Bradenton, FL, USA, Tech. Rep. SCF030, Feb. 2014.

[4] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, Jun. 2011. doi: 10.1109/MWC.2011.5876496.

[5] Y. Q. Bian and D. Rao, "Small cells big opportunities," Huawei Technol., Shenzhen, China, Tech. Rep., Feb. 2014.

[6] *Qualcomm: Small Cells With UltraSON*. Accessed: Feb. 28, 2019. [Online]. Available: https://www.qualcomm.com/media/documents/files/small-cells-and-ultrason-presentation.pdf

[7] *Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication Management; Self-Organizing Networks (SON); Concepts and requirements*, document 3GPP TS 32.500, Version 13.0.0 Release 13, 2014.

[8] S. Feng and E. Seidel, "Self-organizing networks (SON) in 3GPP long term evolution," Nomor Res. GmbH, Munich, Germany, Tech. Rep., May 2008.

[9] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Overall Description Stage 2*, document TS36.300 v9.8.0, Sep. 2011. [Online]. Available: http://www.3gpp.org/

[10] R. Fedrizzi, L. Goratti, T. Rasheed, and S. Kandeepan, "A heuristic approach to mobility robustness in 4G LTE public safety networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2016, pp. 1–6.

[11] S. Feng and E. Seidel, "Self-organizing networks (SON) in 3GPP long term evolution," Nomor Res. GmbH, Munich, Germany, 2008.

[12] R. Kwan, R. Arnott, R. Paterson, R. Trivisonno, and M. Kubota, "On mobility load balancing for LTE systems," in *Proc. IEEE 72nd Veh. Technol. Conf.-Fall*, Sep. 2010, pp. 1–5.

[13] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Load balancing in downlink LTE self-optimizing networks," in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, May 2010, pp. 1–5.

[14] N. Zia and A. Mitschele-Thiel, "Self-organized neighborhood mobility load balancing for LTE networks," in *Proc. IFIP Wireless Days (WD)*, Valencia, Spain, 2013, pp. 1–6. doi: 10.1109/WD.2013.6686466.

[15] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Coordinating handover parameter optimization and load balancing in LTE self-optimizing networks," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC Spring)*, May 2011, pp. 1–5.

[16] Z. Huang, J. Liu, Q. Shen, J. Wu, and X. Gan, "A threshold-based multi-traffic load balance mechanism in LTE-A networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2015, pp. 1273–1278.

[17] P. Rajpoot and P. Dwivedi, "Optimized and load balanced clustering for wireless sensor networks to increase the lifetime of WSN using MADM approaches," *Wireless Netw.*, pp. 1–37, Aug. 2018.

[18] S. Oh, H. Kim, J. Na, Y. Kim, and S. Kwon, "Mobility load balancing enhancement for self-organizing network over LTE system," in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Cham, Switzerland: Springer, 2016, pp. 205–216.

[19] T. Yamamoto, T. Komine, and S. Konishi, "Mobility load balancing scheme based on cell reselection," in *Proc. ICWMC*, 2012, pp. 381–387.

[20] S. Oh, H. Kim, and Y. Kim, "User mobility impacts to mobility load balancing for self-organizing network over LTE system," in *Proc. 14th Int. Conf. Adv. Trends Radioelectron., Telecommun. Comput. Eng. (TCSET)*, Feb. 2018, pp. 1082–1086.

[21] *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S2 Application Protocol (S1AP)*, document TS 36.413, 3rd Generation Partnership Project, Aug. 2019.

[22] *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP)*, document TS 36.423, 3rd Generation Partnership Project, Aug. 2016.

[23] *Evolved Universal Terrestrial Radio Access (E-UTRA); FDD Repeater Radio Transmission and Reception*, document TS 36.106, 3rd Generation Partnership Project, Jul. 2012.

[24] K. Alexandris, N. Nikaein, R. Knopp, and C. Bonnet, "Analyzing X2 handover in LTE/LTE-A," in *Proc. 14th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2016, pp. 1–7.

[25] L. C. Schmelz, M. Amirijoo, A. Eisenblaetter, R. Litjens, M. Neuland, and J. Turk, "A coordination framework for self-organisation in LTE networks," in *Proc. 12th IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM) Workshops*, May 2011, pp. 193–200.

[26] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification*, document TS 36.331, 3rd Generation Partnership Project, Sep. 2018.

[27] M. M. Hasan, S. Kwon, and J.-H. Na, "Adaptive mobility load balancing algorithm for LTE small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2205–2217, Apr. 2018.

[28] Q.-T. Nguyen-Vuong, Y. Ghamri-Doudane, and N. Agoulmine, "On utility models for access network selection in wireless heterogeneous networks," in *Proc. IEEE Netw. Oper. Manage. Symp. (NOMS)*, Piscataway, NJ, USA, Apr. 2008, pp. 144–151.

[29] J. B. Andersen, T. S. Rappaport, and S. Yoshida, "Propagation measurements and models for wireless communications channels," *IEEE Commun. Mag.*, vol. 33, no. 1, pp. 42–49, Jan. 1995.

KHALED M. ADDALI (M'18) received the B.S. degree in electrical engineering from Mergheb University, Khoms, Libya, in 2004, and the M.S. degree in electrical engineering from Concordia University, Montreal, Canada, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering with the École de Technologie supérieure (ÉTS), Université du Québec.

From 2005 to 2008, he was an Operating Engineer with General Electricity Company of Libya (GECOL), Khoms. His research interests include resource and mobility management and the development of user association and load balancing techniques for 5G small-cell networks.

SUHIB YOUNIS BANI MELHEM received the bachelor's and M.Eng. degrees in computer engineering from the Jordan University of Science and Technology, Jordan, and the M.Eng. and Ph.D. degrees in electrical and computer engineering from Concordia University, Canada. He is currently a Postdoctoral Fellow with the Department of Electrical and Computer Science, York University, ON, Canada, collaboration with the National Research Council Canada, QC, Canada. His current research interests include cloud computing, resource management for virtual machine live migration, decision algorithms, cybersecurity for the Internet of Things, and load balancing for 5G small-cell networks.

YASER KHAMAYSEH received the B.Sc. degree in computer science from Yarmouk University, Jordan, in 1998, the M.Sc. degree from the University of New Brunswick, Canada, in 2001, and the Ph.D. degree in computer science from the University of Alberta, Edmonton, Canada, in 2007. In 2007, he joined the Faculty of Computer and Information Technology, Jordan University of Science and Technology, Jordan, as an Assistant Professor of computer science, where he is currently an Associate Professor. His research interests include wireless network optimization, resource management, and protocol design and analysis for future generation wireless communication networks and systems. He has served as a Technical Program Committee Member for various conferences.

ZHENJIANG ZHANG is currently a Professor in communication and information systems with Beijing Jiaotong University, where he is also a Subdecanal of the School of Software Engineering. He has authored several highly cited scientific articles, published in renowned journals with an impact factor in the communication and machine learning. He has edited many books and participated in many international academic activities. His current research interests include wireless sensor networks, data fusion, telecommunication, and information security.

MICHEL KADOCH (S'67–M'77–SM'04) received the B.Eng. degree from Sir George Williams University, in 1971, the M.Eng. degree from Carleton University, in 1974, the M.B.A. degree from McGill University, in 1983, and the Ph.D. degree from Concordia University, Montreal, QC, Canada, in 1991. He is currently a Full Professor with the École de technologie supérieure (ÉTS), Université du Québec, Canada. He is also an Adjunct Professor with Concordia University. His current research interests include cross-layer design and reliable multicast in wireless ad hoc and WiMAX networks. He has publications and patents in all these areas. He is serving as a Reviewer for a number of journals and conferences, as well as for NSERC grants.