IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 5, 2018, accepted October 24, 2018, date of publication November 9, 2018, date of current version December 3, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2878489

Load-Dependent Handover Margin for Throughput Enhancement and Load Balancing in HetNets

MOHANAD ALHABO, (Student Member, IEEE), AND LI ZHANG, (Senior Member, IEEE)
School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, U.K.

Corresponding author: Mohanad Alhabo (mohanad.alhabo@gmail.com)

ABSTRACT Inbound handover (HO) or hand-in is done when the user equipment (UE) performs HO from a macrocell (MC) to a small cell (SC), while the outbound HO or handout is done when a UE hands over from SC to MC. On the other hand, the inter-SC HO is done when a UE performs HO between two SCs. The outbound HO is not as complex as the other two types, because the UE has only one HO target base station, i.e., the MC. Therefore, in this paper, we only consider the inbound and inter-SC HO types. The user may associate with a small cell for a very short time of stay (ToS), smaller than a short time of stay threshold, and this may cause frequently unnecessary HOs and result in service interruption causing degradation in the quality of service (QoS). In this paper, we propose a novel HO method for the purpose of load balancing and throughput improvement in heterogeneous networks (HetNets). The influence of interference from both MC and SC base stations is taken into account so as to offloaded the user from the congested cell and forced it to HO to the SC that gives a good data rate by choosing the best SC, which has the highest signal to interference plus noise ratio (SINR), from a reduced neighbor cell list (NCL). The NCL is optimized utilizing the SINR threshold and ToS. The proposed method utilizes a modified A3 HO initiation event taking into account the cell load and the interference. Results show that the proposed method can perform HO while maintaining the throughput to a good level. In addition, the proposed method has significantly reduced the inter-SC HOs and inbound HO and radio link failures compared to the existing methods. Under different network conditions, load factors, and call arrival rates, results show that the proposed method can give significantly better performance, thereby producing higher throughput for the user and the network.

INDEX TERMS Handover, heterogeneous networks, interference, small cells, neighborhood scanning, load balancing.

I. INTRODUCTION

The deployment of dense small cells in heterogeneous networks has a good impact on the performance of the network regarding capacity and coverage enhancement, in addition to minimizing the burden of traffic load on the MC by offloading the UEs to the SC. However, the dense deployment of SCs brought new issues into the HetNet, such as interference and mobility problems [1].

The robustness of mobility is an essential aspect in HetNets. The mobility states are either in Radio Resource Control (RRC) idle mode or in RRC-active mode. The RRC-idle mobility mode is related to the cell selection and reselection. While the mobility in the RRC-active mode involves the process of HO so as to maintain the ongoing call or data session. Therefore, the HO is a critical process that affects the services delivered to the user because it happens during the data transmission between the UE and the cell [1]. Generally, the HO happens in the cellular networks between different cells to provide a UE with a connection to the best cell in the network. In conventional homogeneous networks, the UE initiates the HO to the adjacent base station if its downlink power is higher than that of the source one by offset for a time period called as a time to trigger (TTT). Additionally, the TTT and the offset parameters are similar to every cells in the network [2]. Therefore, the conventional homogeneous HO methods are not suitable for HetNets.

Given the traditional HO scheme for HO to SCs as [3]

$$P^{r}_{m\rightarrow ue_k} < P^{th}_{min} \text{ and } P^{r}_{sc_i\rightarrow ue_k} > P^{r}_{m\rightarrow ue_k} + HM, \qquad (1)$$

where $P^{r}_{m\rightarrow ue_k}$, $P^{r}_{sc_i\rightarrow ue_k}$, represents the downlink received signal from the MC and SC respectively, P^{th}_{min} is the minimum required signal power threshold to guarantee QoS, and *HM* is the HO hysteresis margin.

Because of the huge difference in power transmission for the MC and SC, it is unlikely to achieve the above criteria. In particular, when the SC is positioned in the inner area of MC coverage, the UE will always be associated to the MC despite that $P^{r}_{sc_i \rightarrow ue_k}$ being strong enough. This may cause to high congestion in the MC because of the improper SC utilization and eventually leads to a lower network throughput. Therefore, it is necessary to consider other parameters for HO that account for different UE speeds, cell load balance, and interference levels introduced by a large deployment of SCs because the achievable data rate of the UE is largely affected by the interference level in the network as well as the distance between the cell and the UE in addition to the load of each cell.

For a UE to perform HO, the list of HO target cells is saved in each cell in a specific list known as a NCL, which includes all of the neighbouring cells for this base station. Upon the HO to this base station, the UE obtains the NCL. Then, the UE measures the signal quality of the cells stored in this NCL for the purpose of the next HO process [4]. With the dense deployment of SCs in the future 5G network, it is not efficient to consider a NCL containing a very large number of SCs. A shorter NCL means lower signal overhead, proper SC utilization, faster HO, and lower energy consumption.

The major contribution of this paper is to propose a novel HO method for the purpose of load balancing and throughput improvement in heterogeneous networks. Interference-based HO is taken into account to improve the throughput. This paper takes into account the minimization of the SC NCL by using the interference level, using SINR, and ToS as HO metrics, in order to improve the proper utilization of SCs and enhance the QoS by offloading the UE from the MC to the SCs. Considering the end user QoS, users will be offloaded from the congested cells and forced to HO to the SC that provides the higher data rate and has enough resources compared to the MC by applying our proposed HO triggering event that takes into considerations the interference and cell load. A modified A3 HO initiation event is proposed by taking into account the traffic load in the serving cell and an equivalent SINR received from a SC within the reduced NCL, which provides a sufficient data rate compared to the source MC. Reducing the interference will decrease the burden on the network and in turn will result in an energy efficient HO process for both user and base station. Results show that the proposed method gives high throughput for the users when compared to other work in the literature. The unnecessary HO and radio link failure are also reduced in our proposed method. Moreover, the throughput of the network is enhanced under different network conditions, such as load factor, different traffic data size, and different levels of noise. This paper is organized as follows. Section II presents the related works. The network system model is presented in section III, while section IV gives the proposed method process. In section V the performance of the proposed method and the results are analysed. Finally, section VI concludes the paper and outlines the future direction of this work.

II. RELATED WORKS

The user mobility in dense SCs HetNets is a big challenge since there are thousands of target cells [5]. Thus, minimizing the number of target HO cells is a good strategy to reduce the signalling overhead. In [6], the authors proposed a method that automatically manages the HM for outbound HO to MC. The method controls the HM based on the velocity of the UE such that for high-speed UEs the HM is reduced (preventing too late HO), and for the low speed UE the HM is increased (preventing too early HO). This method shows a reduction in late and early HOs in addition to the minimization in the radio link failures for different UE speeds. However, no strategy for managing the load between the MC and SC is taken into account, which may cause a high congestion in the MC, hence a high radio link failure rate is expected. In [7], the authors proposed a method to reduce the unnecessary HOs by minimizing the number of scanned SCs. The constructing of the SC NCL depends on the downlink received power and ToS, which neglects the SCs with a short time of stay. The HO is performed to the SC that gives the strongest downlink received power from the NCL. However, the cell load and interference scenario are not considered in this work, which may cause a throughput unaware HO and link failures. Authors in [8] present an algorithm for inbound HO to reduce the scanning of neighbouring SCs. The cell is considered in the list according to the HO probability to this cell and SINR at the UE side from its current serving SC. This process has highly reduced the scanning list, however, this work has not accounted for the problems of MC traffic offloading and SC utilization. In [9], the authors presented a method to minimize the target SC NCL and reduce the probability of unnecessary HO. The SC NCL is built utilizing the distance between the user and the SC in addition to the user angle of movement. The average walking speed is utilized to stop the high-speed UEs from performing HO to SCs. Results show a good performance in terms of SC NCL minimization, unnecessary HO reduction, and network throughput improvement. In [10], the authors presented a method to reduce the unnecessary HO and HO failure. A predicted ToS is utilized to remove SC, which may cause unnecessary HO or HO failure, from the NCL. The HO is performed to the SC, which gives the higher SINR and has enough capacity. Time threshold and the SINR are also utilized to obtain a trade-off between the HO failure and unnecessary HO in the network. Results show that this work has minimized the unnecessary HO and HO failure. In [4], a mechanism to reduce the scanning process is presented. This mechanism uses the estimated distance between the SC and the UE to perform the scanning by considering the previously visited SCs. However, this mechanism can only be applied to SC with a close subscriber group (CSG) and can not be used for an open subscriber group SCs in addition to the traffic offloading problem. A multiple HO criteria method is proposed in [11] to reduce the probability of HO and balance the load between the MC and the SCs. This method uses the estimated reference signal received power (RSRP) of the target cell, the transmit power of the UE, and the target cell capacity as HO decision making metrics. A HO method for load balancing is proposed in [24] where the influence of the predicted ToS and interference is utilized to achieve offloading from MC to SC. A HO margin based on source cell load is derived to perform the traffic offloading. Results reveal that this method has minimized the frequent unnecessary HO and failure probability in addition to improving the throughput. Authors in [12] proposed a dynamic cell association to increase the sum rate and considered a cell range expansion method for traffic load balancing in HetNet. The principles of SC range expansion is a good strategy to offload the traffic from MC to SC by increasing the transmit power of the SC, hence, more UEs associate with the SC and eventually the load balancing is accomplished. However, using this method of offloading has limited achievements because the biasing of SC power increases the interference and degrades the SINR received at the UE. Therefore, controlling the power biasing is a critical issue. In [13], the authors proposed a HO load balancing method for HetNet. The UEs are forced to perform the HO to the SCs when their speed is low and the capacity of the SC is available. However, these UEs are also permitted to connect to the MC temporarily if the capacity of the SC is not sufficient, to minimize the HO failure. On the other hand, fast moving UEs are connected to MC. However, this method is not efficient if deployed in a dense SC HetNet, which may result in a high number of SC in NCL, high number of unnecessary HOs, and signalling overhead.

This paper aims to reduce the congestion in the HetNet by forcing the HO to SCs, hence, balancing the load between the MC and SC tiers in addition to increasing the SCs utilization to enhance the throughput.

III. NETWORK SYSTEM MODEL

We first give the major symbols utilized in this paper in table 1.

System model in this work considers a two-tier HetNet, which consists of SCs overlaid under the coverage area of the MC, as illustrated in Fig.1. MC is deployed as a hexagonal with three sectors (120^o each). The SCs are randomly distributed following a uniform distribution. The MC and SCs are using the same frequency. The minimum distance constraint between the MC tier and SC tier is taken into account to reduce the influence of the interference and hence improve the anticipated capacity of the SCs. The minimum distances in meters are set as follows [1]: MC site to SC site is 75m and MC to UE is 35m. The UE mobility follows a Gauss mobility model which can be represented utilizing two parameters: direction, θ_k, and velocity, V_{ue}. The two parameters can be expressed as Gaussian distribution and are updated accordingly by the following [14]

$$V_{ue} = \mathcal{N}(v_m, v_{std}), \tag{2}$$

$$\theta_k = \mathcal{N}(\theta_m, 2\pi - \theta_m \tan(\frac{\sqrt{V_{ue}}}{2})\Delta t), \tag{3}$$

TABLE 1. Definition of abbreviations and symbols.

Symbol	Definition
$P^r_{m\rightarrow ue_k}$	downlink received signal from MC
$P^r_{sc_i\rightarrow ue_k}$	downlink received signal from SC i
V_{ue}	user velocity
θ_k	user direction of movement
$\mathcal{N}(\cdot,\cdot)$	Gaussian distribution with mean and standard deviation
$r_{i\rightarrow ue_k}$	maximum data rate from cell i
$\gamma^r_{i\rightarrow ue_k}$	the SINR received at user k from cell i
$ToS_{ue\rightarrow sc_i}$	user time of stay in SC i
HO^*_i	HO point
$\gamma^{req}_{sc_i\rightarrow ue_k}$	is the SC's i SINR equivalent to that of the MC's/SC's j that gives at least the same data rate as compared to the MC/SC's j
$\gamma^{pro}_{m\rightarrow sc_i}$	is the proposed interference-based load-dependent margin to control the HO point for inbound HO
$\gamma^{pro}_{sc_j\rightarrow sc_i}$	is the proposed interference-based load-dependent margin to control the HO point for inter-SC HO
n	set of all cells in the network
N_{sc}	total number of SCs in the network
N^*_{sc}	set of all SCs $\in N_{sc}$ with $\gamma^r_{sc_i\rightarrow ue_k} > \gamma_{th}$
N^{**}_{sc}	set of all SCs $\in N^*_{sc}$ with $ToS_{ue\rightarrow sc_i} > T_{th}$
$\mathbb{P}_{ue \text{ inside } sc_i}$	probability that the user is inside SC i coverage area
P_{out}	outage probability
T^{exe}_{ho}	time required to complete the HO process
L_{m_i}	load on MC sector i
L_{sc_j}	load on SC j
$RB^{ue}_{m_i}$	number of PRBs utilized by all UEs in sector i
$RB^{ue}_{sc_j}$	number of PRBs utilized by all UEs in SC j
RB_{tm}	total number of PRBs in MC
RB_{tsc_j}	total number of PRBs in SC j
L^{mr}_m	is the load-dependent parameter for inbound HO
$L^{mr}_{sc_j}$	is the load-dependent parameter for inter-SC HO

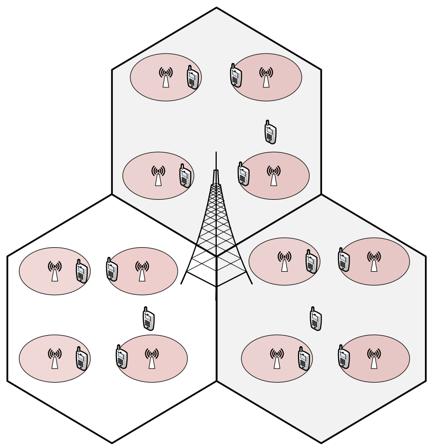

FIGURE 1. Network system model.

where v_m is the mean velocity of the UE, v_{std} is the standard deviation of the UE velocity, θ_m is the previous direction of the UE, Δt is the period between two updates of the mobility model, and $\mathcal{N}(x, y)$ is a Gaussian distribution with mean x and standard deviation y.

The data rate, $r_{i\rightarrow ue_k}$, is expressed as

$$r_{i\rightarrow ue_k} = \text{BW} \log_2(1 + \gamma^r_{i\rightarrow ue_k}), \tag{4}$$

where BW is the bandwidth and $\gamma^r_{i\rightarrow ue_k}$ is the SINR received from base station i at user k .

The SINR from SC i and MC can be expressed as

$$\gamma^r_{sc_i\to ue_k} = \frac{P^r_{sc_i\to ue_k}}{P^r_{m\to ue_k} + \sum_{j=1, j\neq i}^{N_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}, \tag{5}$$

$$\gamma^r_{m\to ue_k} = \frac{P^r_{m\to ue_k}}{\sum_{j=1}^{N_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}, \tag{6}$$

where $\gamma^r_{m\to ue_k}$ is the SINR received from MC at the user k, $\gamma^r_{sc_i\to ue_k}$ is the SINR received from SC i at the user k, σ^2 is the noise power, and finally N_{sc} is a set representing the total number of SCs in the HetNet.

The propagation model between the MC and the user is defined as in [15] by

$$\delta_{m\to ue_k} = 128.1 + 37.6\ \log_{10}(d_{m\to ue_k}) + \xi, \tag{7}$$

where $d_{m\to ue_k}$ is the distance between the UE and the MC in kilometres and ξ is a Gaussian distribution random variable with zero mean and 12 dB standard deviation [16].

For outdoor SC, the path loss is expressed as in [17] by

$$\delta_{sc_i\to ue_k} = 38 + 30\ \log_{10}(d_{sc_i\to ue_k}) + \xi, \tag{8}$$

where $d_{sc_i\to ue_k}$ is the distance between the user and SC i in metres. The UE time of stay can be defined utilizing the velocity, V_{ue}, and the estimated distance that the user will reside in the base station coverage area as shown in Fig. 2.

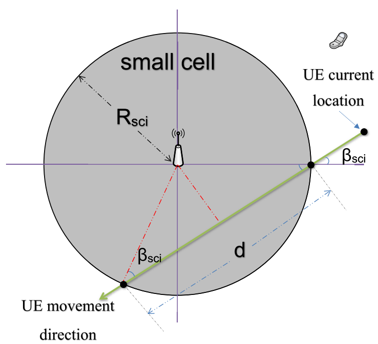

FIGURE 2. UE ToS measurement.

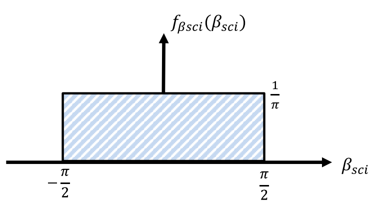

FIGURE 3. PDF of β_{sc_i}.

The angle β_{sc_i}, which is the UE angle of entry to the SC, and it can be expressed as a random variable, which is uniformly distributed and restricted to interval $[\frac{-\pi}{2}, \frac{\pi}{2}]$. This random variable has a constant density over the interval i.e., has a probability density function (PDF) $f_{\beta_{sc_i}}(\beta_{sc_i})$, as shown in Fig. 3 and equation (9).

$$f_{\beta_{sc_i}}(\beta_{sc_i}) = \begin{cases} \frac{1}{\mid \frac{-\pi}{2} - \frac{\pi}{2} \mid} & \text{if } \frac{-\pi}{2} \leq \beta_{sc_i} \leq \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Thus, we can express the mean ToS a user will stay in the SC as

$$\begin{aligned} \mathrm{E}\Big[ToS_{ue\to sc_i}\Big] &= \mathrm{E}\left[\frac{2R_{sc_i}\cos(\beta_{sc_i})}{V_{ue}}\right] \\ &= \int \frac{2R_{sc_i}\cos(\beta_{sc_i})}{V_{ue}} f_{\beta_{sc_i}}(\beta_{sc_i}) d\beta_{sc_i} \\ &= \int_{\frac{-\pi}{2}}^{\frac{\pi}{2}} \frac{2R_{sc_i}\cos(\beta_{sc_i})}{V_{ue}} \frac{1}{\pi} d\beta_{sc_i} \\ &= \frac{4R_{sc_i}}{\pi V_{ue}}, \end{aligned} \tag{10}$$

where R_{sc_i} is the SC radius.

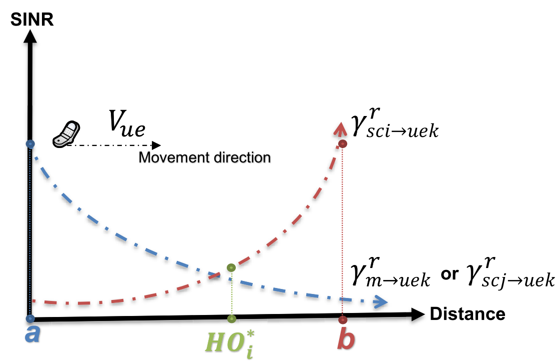

FIGURE 4. Handover point.

As given in Fig. 4, the blue curve represents the SINR of the serving cell (MC or SC j) and the red curve represents the SINR of the target cell (SC i). The HO takes place at point HO^*_i, which is the point at which the A3 HO event is satisfied. In other words, it is the point at which the HO should be performed.

The aim is to find $HO^*_i\ \forall i = 1, \ldots, N_{sc}$, to enhance the throughput and to obtain traffic load balancing between the MC and SCs by forcing the HO to SC to distribute the load. Where N_{sc} is the number of SCs.

In the following, we illustrate the analysis and calculations of the loads, the equivalent SINR required to perform the HO to SC and the proposed interference-based load-dependent margin.

A. LOAD CALCULATIONS AND RESOURCE ASSIGNMENT

The cell load factor is the amount of resource usage with respect to the available resources in the cell [18], i.e., a low load factor means that the base station has enough resources to serve the UE; on the other hand, a high load factor congests the base station and leads to poor network throughput [19].

1) **For Inbound HO**

For the i^{th} MC sector, the load L_{m_i} is expressed as the number of physical resource blocks (PRBs) being used by all UEs associated to the aforementioned sector divided by the total MC PRBs,

$$L_{m_i} = \frac{RB^{ue}_{m_i}}{RB_{tm}}, \tag{11}$$

where $RB^{ue}_{m_i}$ is the number of PRBs used by all UEs associated to the MC sector i and RB_{tm} is the total number of PRBs in the MC.

The number of PRBs used by all active UEs associated to the MC sector i, i.e., $RB^{ue}_{m_i}$, can be expressed as

$$RB^{ue}_{m_i} = \sum_{k=1}^{N^{sec}_{ue}} RB_{m_i,k}, \tag{12}$$

where N^{sec}_{ue} is the number of UEs in the sector and $RB_{m_i,k}$ is the number of PRBs used by UE k.

2) **For Inter-SC HO**

Whereas the load on the SC j is

$$L_{sc_j} = \frac{RB^{ue}_{sc_j}}{RB_{tsc_j}}, \tag{13}$$

where $RB^{ue}_{sc_j}$ is the number of PRBs used by all mobile users associated to SC j and RB_{tsc_j} is the total number of PRBs in SC j.

The number of PRBs used by all mobile users associated to SC j, $RB^{ue}_{sc_j}$, can be expressed as

$$RB^{ue}_{sc_j} = \sum_{k=1}^{N^{sc_j}_{ue}} RB_{sc_j,k}, \tag{14}$$

where $N^{sc_j}_{ue}$ is the number of active UEs residing in SC j and $RB_{sc_j,k}$ is the number of PRBs used by user k.

B. EQUIVALENT SINR ANALYSIS

From Fig. 4, HO^*_i is the point at which $\gamma^r_{sc_i\to ue_k} = \gamma^r_{m\to ue_k}$ for inbound HO and $\gamma^r_{sc_i\to ue_k} = \gamma^r_{sc_j\to ue_k}$ for inter-SC HO.

1) **For Inbound HO**

Using equations (5) and (6), we apply the condition ($\gamma^r_{sc_i\to ue_k} = \gamma^r_{m\to ue_k}$) to the two equations

$$\frac{P^r_{sc_i\to ue_k}}{P^r_{m\to ue_k} + \sum_{j=1,j\neq i}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2} = \frac{P^r_{m\to ue_k}}{\sum_{j=1}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}. \tag{15}$$

Reordering equation (15) and after some simplifications we get

$$P^r_{sc_i\to ue_k} = \frac{A}{\sum_{j=1}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}, \tag{16}$$

where A is

$$A = P^r_{m\to ue_k}\Big(P^r_{m\to ue_k} + \sum_{j=1,j\neq i}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2\Big). \tag{17}$$

The UE will initiate the HO to the SC with the highest data rate, i.e., at HO point HO^*_i. In other words, we can say that the HO is initiated when the downlink received power from the SC satisfying the criteria in (16). Substitute (16) in (5) to get the equivalent SINR $\gamma^{req}_{sc_i\to ue_k}$, for inbound HO from MC to SC i, that gives at least the same data rate as the current serving base station, that is

$$\gamma^{req}_{sc_i\to ue_k} = \frac{P^r_{sc_i\to ue_k}}{P^r_{m\to ue_k} + \sum_{j=1,j\neq i}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}$$

$$\therefore\ \gamma^{req}_{sc_i\to ue_k} = \frac{A/\sum_{j=1}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}{P^r_{m\to ue_k} + \sum_{j=1,j\neq i}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}, \tag{18}$$

2) **For Inter-SC HO**

Similarly, for a given SC i and SC j, we can derive an expression to find the equivalent SINR for the inter-SC HO from SC j to SC i, that is

$$P^r_{sc_i\to ue_k} = \frac{B}{\sum_{j=1}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}, \tag{19}$$

where

$$B = P^r_{sc_j\to ue_k}\Big(P^r_{m\to ue_k} + \sum_{j=1,j\neq i}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2\Big). \tag{20}$$

Substituting (19) in (5) and after some simplifications, we get the equivalent SINR $\gamma^{req}_{sc_i\to ue_k}$, for inter-SC HO from SC j to SC i, that provides at least the same data rate as the current serving base station, that is

$$\gamma^{req}_{sc_i\to ue_k} = \frac{B/\sum_{j=1}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}{P^r_{m\to ue_k} + \sum_{j=1,j\neq i}^{N^{**}_{sc}} P^r_{sc_j\to ue_k} + \sigma^2}. \tag{21}$$

Note that the summation of the interference term, in (15) to (21), takes into account only the SCs in set N^{**}_{sc} (as in section IV) as expressed in (32), which will, in turn, minimize the computation complexity because we have a few number of SCs in this set.

C. PROPOSED INTERFERENCE-BASED LOAD-DEPENDENT MARGIN

When the serving base station (MC or SC j) suffers from heavy traffic load and the target SC i has a light traffic load, the serving base station will undergo a high rate of radio link failure when a UE tries to perform HO to this serving base station. To maintain mobility load balancing in general, if the serving cell is congested then it increases the HO margin to initiate the HO early to another cell. However, this unplanned increase may result in link failure and ping-pong HO, and hence, poor QoS is delivered to the UE. Therefore, to maintain the mobility robustness, these parameters should be dynamically adjusted based on the actual cell load. For this reason, we aim to make the UE performs HO to SC i, which has a lower load. The proposed method will bias the HO point between the heavily loaded serving cell and the light-loaded target SC i.

The traditional A3 HO initiation event, which is based on a power-based margin, when the power of the neighbouring

SC i is offset greater than that of the serving MC for a period of TTT [1], that is

$$P^r_{sc_i \to ue_k} \geq P^r_{m \to ue_k} + HM_m - HM_{m,sc_i}, \tag{22}$$

where HM_m is the hysteresis parameter of MC and HM_{m,sc_i} is the SC i specific offset with respect to the MC (i.e., the hysteresis set by MC to HO to the SC i). Indeed the parameter HM_{m,sc_i} controls the HO point and can be optimized based on the load of the serving MC.

Inspired by (22), we proposed a modification to this condition to facilitate an interference-based load-dependent hysteresis HO margin. The SINR will be considered instead of the downlink received power and the power margin HM_{m,sc_i} will be replaced by the interference-based load-dependent margin, denoted $\gamma^{pro}_{m \to sc_i}$, namely the proposed interference-based load-dependent margin to control the HO point between MC and SC i.

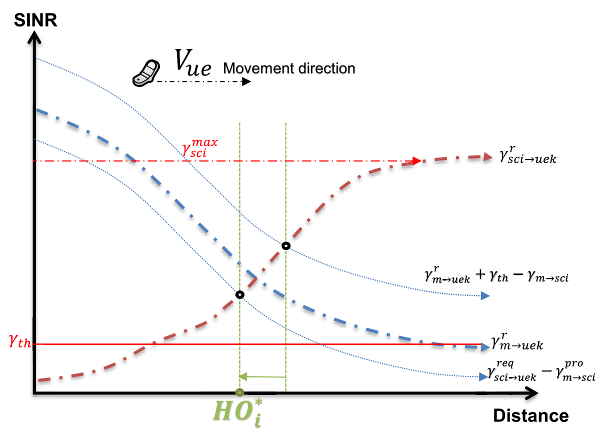

FIGURE 5. Handover point for inbound HO.

Since we are considering inbound and inter-SC HO, we need to find two margins. The first is $\gamma^{pro}_{m \to sc_i}$ for inbound HO from MC to SC i and the second is $\gamma^{pro}_{sc_j \to sc_i}$ for inter-SC HO from SC j to SC i.

1) **For Inbound HO**

We can rewrite equation (22) based on our proposal as shown in (23) for inbound HO from MC to SC i

$$\gamma^r_{sc_i \to ue_k} \geq \gamma^r_{m \to ue_k} - \gamma^{pro}_{m \to sc_i}, \tag{23}$$

For inbound HO, to balance the load, the HO point HO^*_i should be moved closer to the serving MC rather than being closer to the target SC i (the HO point HO^*_i will be changed based on the current load on the serving MC to perform offloading to SC). To change the HO point for a UE trying to make HO from MC to SC i, we must shift the HO^*_i point to the left as shown in Fig. 5, i.e., the HO point will be changed from the intersection point of the two curves $\gamma^r_{sc_i \to ue_k}$ and $(\gamma^r_{m \to ue_k} + \gamma_{th} - \gamma_{m \to sc_i})$ to the intersection point of the two curves $\gamma^r_{sc_i \to ue_k}$ and $(\gamma^{req}_{sc_i \to ue_k} - \gamma^{pro}_{m \to sc_i})$, note that $\gamma^{req}_{sc_i \to ue_k}$ is taken from (18). In other words, the congested MC adjusts the HO margin $\gamma^{pro}_{m \to sc_i}$ to allow the UE to perform early HO

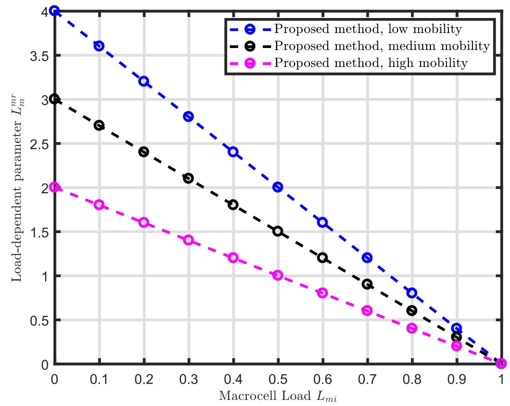

FIGURE 6. L^{mr}_m **vs.** L_{m_i}**.**

to SC i and preventing the UEs form performing HO to itself (i.e., to the MC) so as to avoid more congestion in the already congested MC. This is done by considering the load-dependent margin $\gamma^{pro}_{m \to sc_i}$. For Fig. 5, γ_{th} is the outage threshold and is set to 5 dB [20] and $\gamma^{max}_{sc_i}$ is the SINR from SC i when $\gamma^r_{m \to ue_k}$ is equal to γ_{th}. In order to keep the link failure to a lower level, the hysteresis can be assigned based on the UE speed [1]. Thus, the values of $\gamma_{m \to sc_i}$ and $\gamma_{sc_j \to sc_i}$ are adjusted to 4 dB for low speed UE ($V_{ue} \leq$ 20km/h), 3 dB for medium speed UE (20km/h $< V_{ue} \leq$ 50km/h) and 2 dB for high speed UE ($V_{ue} >$ 50km/h).

To incorporate the influence of the UE velocity on the proposed HO margin, we proposed to incorporate the margin $\gamma_{m \to sc_i}$ into (24) to find the load-dependent parameter, denoted as L^{mr}_m, which will be utilized to find the proposed HO margin

$$L^{mr}_m = (1 - L_{m_i}) \cdot \gamma_{m \to sc_i}, \tag{24}$$

where L^{mr}_m is the load-dependent parameter for inbound HO. Finally, the proposed interference-based load-dependent HO margin can be expressed as

$$\begin{aligned} \gamma^{pro}_{m \to sc_i} &= \gamma_{m \to sc_i} - L^{mr}_m \\ &= L_{m_i} \cdot \gamma_{m \to sc_i}. \end{aligned} \tag{25}$$

The parameter L^{mr}_m depends on L_{m_i}: the higher the value of L_{m_i} the smaller the value of L^{mr}_m (as shown in Fig. 6), the higher the proposed margin and eventually the closer the HO^*_i point to SC i.

When L^{mr}_m is low, this indicates that the serving MC is congested, therefore, the HO point is moved closer to the MC to speed up the HO initiation, which will balance the traffic load by offloading it from the heavily loaded MC to SC i.

In fact, the term $(1 - L_{m_i})$ is the key parameter to control the proposed interference-based load-dependent margin. As depicted in table 2 for low, medium and high MC loads, the term $(1 - L_{m_i})$ ensures that the load-dependent parameter L^{mr}_m is adjusted based on the current state of the load on the MC base station.

TABLE 2. L_m^{mr} and $\gamma^{pro}_{m\rightarrow sc_i}$ for different MC loads and different mobility states.

Mobility state		MC load		
		(10%)	(50%)	(90%)
Low	L_m^{mr}	3.6	2	0.4
	$\gamma^{pro}_{m\rightarrow sc_i}$ [dB]	0.4	2	3.6
Medium	L_m^{mr}	2.7	1.5	0.3
	$\gamma^{pro}_{m\rightarrow sc_i}$ [dB]	0.3	1.5	2.7
High	L_m^{mr}	1.8	1	0.2
	$\gamma^{pro}_{m\rightarrow sc_i}$ [dB]	0.2	1	1.8

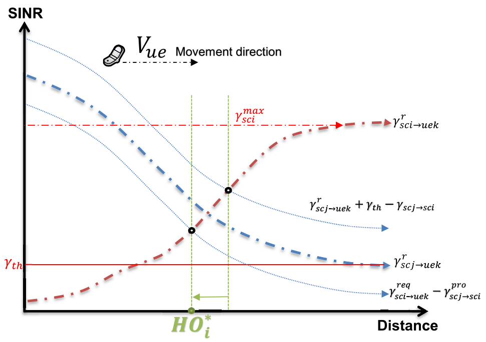

FIGURE 7. Handover point for inter-SC HO.

For instance, when the load is high at (90%) with the mobility state also high, then the parameter L_m^{mr} is 0.2, which will make the proposed margin high at 1.8 dB, hence, the HO is performed earlier so as to offload the congested traffic from MC to SC. On the other hand, when the load is low at (10%) with the mobility state also high, then the parameter L_m^{mr} is 1.8 and the proposed margin is low at 0.2 dB, hence, the HO is not performed early.

2) **For Inter-SC HO**

Also, we can rewrite equation (22) based on our proposal as shown in (26) for inter-SC HO from SC j to SC i

$$\gamma^r_{sc_i\rightarrow ue_k} \geq \gamma^r_{sc_j\rightarrow ue_k} - \gamma^{pro}_{sc_j\rightarrow sc_i}. \tag{26}$$

Similarly for the inter-SC HO, the HO point HO_i^* must be moved closer to the serving SC j rather than being closer to the target SC i. Thus, we move the HO point for a UE trying to perform HO from SC j to SC i to the left as shown in Fig. 7, i.e., the HO point will be changed from the intersection point of the two curves $\gamma^r_{sc_i\rightarrow ue_k}$ and $(\gamma^r_{sc_j\rightarrow ue_k} + \gamma_{th} - \gamma_{sc_j\rightarrow sc_i})$ to the intersection point of the two curves $\gamma^r_{sc_i\rightarrow ue_k}$ and $(\gamma^{req}_{sc_i\rightarrow ue_k} - \gamma^{pro}_{sc_j\rightarrow sc_i})$, where $\gamma^{max}_{sc_i}$ is the SINR from SC i when $\gamma^r_{sc_j\rightarrow ue_k}$ is equal to γ_{th}, note that $\gamma^{req}_{sc_i\rightarrow ue_k}$ is taken from (21). Then, the margin in this case is calculated as

$$L^{mr}_{sc_j} = (1 - L_{sc_j}) \cdot \gamma_{sc_j\rightarrow sc_i}, \tag{27}$$

$$\gamma^{pro}_{sc_j\rightarrow sc_i} = \gamma_{sc_j\rightarrow sc_i} - L^{mr}_{sc_j} = L_{sc_j} \cdot \gamma_{sc_j\rightarrow sc_i}, \tag{28}$$

where $L^{mr}_{sc_j}$ is the load-dependent parameter for inter-SC HO.

TABLE 3. TTT based on UE speed.

UE speed (km/h)	$V_{ue} \leq 20$	$20 < V_{ue} \leq 50$	$V_{ue} > 50$
TTT (ms)	1280	512	256

Now we have $\gamma^{req}_{sc_i\rightarrow ue_k}$, $\gamma^{pro}_{m\rightarrow sc_i}$ and $\gamma^{pro}_{sc_j\rightarrow sc_i}$, then we can rewrite equation (23) to represent our proposed modified A3 HO triggering event for inbound HO as

$$\gamma^r_{sc_i\rightarrow ue_k} \geq \gamma^{req}_{sc_i\rightarrow ue_k} - \gamma^{pro}_{m\rightarrow sc_i}. \tag{29}$$

While also rewriting equation (26) for inter-SC HO as

$$\gamma^r_{sc_i\rightarrow ue_k} \geq \gamma^{req}_{sc_i\rightarrow ue_k} - \gamma^{pro}_{sc_j\rightarrow sc_i}. \tag{30}$$

The above conditions in (29) and (30) should hold for a period of TTT based on the UE speed [21] as shown in table 3.

IV. PROPOSED METHOD PROCESS

Algorithm 1 shows the proposed method processes where $ToS_{ue\rightarrow sc_i}$ is the predicted time of stay of the user in the SC i coverage area, T_{th} is the time threshold for ToS and N^*_{sc} is a set that represents the total number of SCs with an SINR greater than the outage threshold.

Algorithm 1 Proposed Method

1: **Start**
2: UE **moves** to SC coverage area
3: **if** $\gamma^r_{sc_i\rightarrow ue_k} \leq \gamma_{th}$ **then**
4:     **Remove** this SC from the NCL N^*_{sc}
5: **end if**
6: **for** $i \leftarrow 1, N^*_{sc}$ **do**
7:     **Predict** $ToS_{ue\rightarrow sc_i}$
8:     **if** $\mathrm{E}\left[ToS_{ue\rightarrow sc_i}\right] > T_{th}$ **then**
9:         **Save** SC i in the new NCL N^{**}_{sc}
10:     **end if**
11: **end for**
12: **Convert** $\gamma^r_{m\rightarrow ue_k}$ or $\gamma^r_{sc_j\rightarrow ue_k}$ to its equivalent $\gamma^{req}_{sc_i\rightarrow ue_k}$
13: **Calculate** $\gamma^{pro}_{m\rightarrow sc_i}$ or $\gamma^{pro}_{sc_j\rightarrow sc_i}$
14: **Choose** the SC with the highest $\gamma^r_{sc_i\rightarrow ue_k}$ from N^{**}_{sc}
15: **if** $\gamma^r_{sc_i\rightarrow ue_k} \geq \gamma^{req}_{sc_i\rightarrow ue_k}$ - $\gamma^{pro}_{m\rightarrow sc_i}$ for TTT or $\gamma^r_{sc_i\rightarrow ue_k} \geq \gamma^{req}_{sc_i\rightarrow ue_k}$ - $\gamma^{pro}_{sc_j\rightarrow sc_i}$ for TTT **then**
16:     **if** $RB^{ue}_{sc_i} < 1$ **then**
17:         **Handover** the UE to sc_i
18:     **end if**
19: **end if**
20: **end**

First, the proposed method optimizes the NCL by minimizing its size. This is done by using γ_{th} and ToS metrics as illustrated in the pseudo code lines 3 through 11 and explained below.

The proposed method starts by removing the SCs that may cause a poor QoS, i.e., SCs with SINR below the outage threshold γ_{th}, resulting in a NCL N^*_{sc}, which is expressed as

$$N^*_{sc} = \{sc_i \in N_{sc} \mid \gamma^r_{sc_i\rightarrow ue_k} > \gamma_{th}\}. \tag{31}$$

Then, a NCL is built, denoted as N_{sc}^{**} set, containing all the SCs whose estimated mean UE ToS is larger than the time threshold T_{th}. Therefore, the new NCL can be written as

$$N_{sc}^{**} = \{sc_i \in N_{sc}^{*} \mid \mathrm{E}\big[ToS_{ue \to sc_i}\big] > T_{th}\}. \tag{32}$$

After the NCL reduction, the second phase of the method is applied to get the equivalent SINR and calculate the proposed margin.

The inbound HO is performed to SC i if its data rate is greater than that of the MC. This is accomplished by converting the SINR received from the MC, i.e., $\gamma^{r}_{m \to ue_k}$, to its equivalent SINR received from the SC, i.e., $\gamma^{r_{eq}}_{sc_i \to ue_k}$, which gives a higher data rate. Then, the resulting SINR is compared with the actual SINR received from the SC i and taking into account the proposed margin $\gamma^{pro}_{m \to sc_i}$ and TTT. On the other hand, if the UE's serving cell is SC j, the UE performs inter-SC HO to SC i if its data rate is higher than that of the serving SC j. This is accomplished by converting the SINR received from the serving SC j to its equivalent SINR received from the SC i, which gives a higher data rate. Then, the SINR from the last step is compared with the actual SINR received from the SC i and taking into account the proposed margin $\gamma^{pro}_{sc_j \to sc_i}$ and TTT metric. The HO is performed to SC i providing that its PRBs is good enough to supply resources to the UE. In line (16) in Algorithm 1, $RB^{ue}_{sc_i} < 1$, the value 1 indicates that the SC resources are all occupied by other UEs.

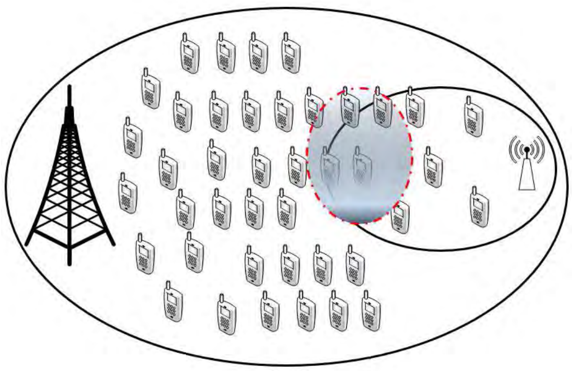

FIGURE 8. Inbound HO forcing to SC.

Fig. 8 simplifies the aim of the proposed interference-based load-dependent margins $\gamma^{pro}_{m \to sc_i}$ and $\gamma^{pro}_{sc_j \to sc_i}$. When the serving cell (MC or SC j) suffers from high load (congested cell), the margins $\gamma^{pro}_{m \to sc_i}$ and $\gamma^{pro}_{sc_j \to sc_i}$ will be adjusted to HO the UEs, located in the overlapped shaded region (i.e., in Fig. 8) between the serving and the target SC, to the target SC. This will attain the offloading purpose, increase the system throughout and eventually increase the proper utilization of SCs.

It is worth noting that the HO is only forced to the cell with SINR greater than the outage threshold, hence, the high throughput is expected with an acceptable HO signalling.

V. PERFORMANCE ANALYSIS AND RESULTS

The proposed method performance is compared against three competitive methods, namely the conventional method, the energy efficient and cell load balancing (ENCLB) method presented in [11], and the estimated time-of-stay-based cell selection (ETCS) method presented in [7]. System level Matlab simulations have been carried out for performance evaluations. We divide this section into three parts. The first part introduces the competitive methods. The second part presents the performance evaluation metrics. While the results and discussions are given in the last part.

A. COMPETITIVE METHODS

The three methods defined in this part are the conventional method, the ENCLB method given in [11] and the ETCS method presented in [7].

In the conventional method, the neighbourhood scanning is performed according to the downlink received power, to build a NCL. This indicates that there will be a significant time period needed to choose the target. Then, the HO is performed to the SC with the highest downlink received power as shown in (1) without taking the load balancing and interference into account, which means that the HO point HO_i^* for this method is downlink power dependent. This will result in a reduction in the UE throughput and wasting the battery power of the UE because of the frequent scanning measurement, especially in a dense SC networks. Thus, the HO target SC for the conventional method, denoted as sc^{t}_{conv}, is written as

$$sc^{t}_{conv} = \Big\{sc_i \in N_{sc} \mid \; P^{r}_{sc_i \to ue_k} \; > \; P^{r}_{m \to ue_k}\Big\}. \tag{33}$$

The ENCLB method in [11], forms the HO target cell list based on the predicted RSRP and the transmit power of the UE. The UE performs the HO to the SC, from the list, if its RSRP is offset greater than that of the serving cell and has enough capacity. Thus, the HO point HO_i^* is based on the power difference between the serving and the target cells with a fixed HO margin. The HO target SC for this method, denoted as sc^{t}_{enclb}, can be given as

$$sc^{t}_{enclb} = \Big\{sc_i \in N_{sc} \mid P^{r}_{sc_i \to ue_k} > P^{r}_{m \to ue_k} + HM_m \wedge \gamma^{up}_{ue \to sc_i} > \gamma^{up}_{th} \wedge RB^{ue}_{sc_i} < 1\Big\}. \tag{34}$$

where $\gamma^{up}_{ue \to sc_i}$ is the uplink SINR for the target SC i and γ^{up}_{th} is its threshold which is set to 3 dB.

The ETCS method in [7] builds the NCL according to the ToS and the downlink received power metrics. Then, the HO is performed to the cell with the highest power from the NCL. In addition, the cell load balance and interference scenario are not taken into account in [7] and the HO point HO_i^* is based on the power difference between the serving and the target cells. Thus, the NCL, sc^{t}_{etcs}, for ETCS can be expressed as

$$sc^{t}_{etcs} = \Big\{sc_i \in N_{sc} \mid \; (\mathrm{E}\big[ToS_{ue \to sc_i}\big] > T_{th}) \wedge P^{r}_{sc_i \to ue_k} > P^{r}_{m \to ue_k}\Big\}. \tag{35}$$

In contrast, our proposed method builds the NCL according to the interference constraints and ToS. Then, the HO is performed to the cell that supplies a better data rate with load balancing considerations, providing that the PRBs are sufficient enough, considering a modified A3 HO initiation event, this means that the HO point is interference based as shown in (29) and (30).

B. PERFORMANCE EVALUATION METRICS

Based on the density definition in [22], the SC density in a given coverage area can be found by utilizing the density metric, D_{sc}, as

$$D_{sc} = \frac{\mid N_{sc} \mid \pi R_{sc_i}^2}{\pi R_m^2}, \tag{36}$$

where R_{sc_i} and R_m are respectively the SC and MC radius. The denominator represents the MC coverage area. When the density metric of SC D_{sc} is equal to 1, this indicates that the deployment of the SCs occupies the complete zone of the MC coverage area. On the other hand, a higher than 1 value indicates that the SCs are covering the complete zone of MC and an overlapping is ensured among the SCs. We deployed 100 SCs, which means that $D_{sc} \approx 1.56$ and hence, the dense SCs scenario is obtained.

The probability that the UE is inside the coverage areas of SC i can be expressed as

$$\mathbb{P}_{ue \text{ inside } sc_i} = \mathbb{P}\Big[P^r_{sc_i \to ue_k} \geq P^{th}_{min}\Big]. \tag{37}$$

The probability of inbound HOs to the SCs is given as

$$\begin{aligned} P^{in}_{HO} = \mathbb{P}\Big[&P^r_{sc_i \to ue_k} \geq P^{th}_{min} \wedge \mathrm{E}\big[ToS_{ue \to sc_i}\big] > T_{th} \wedge \\ &\gamma^r_{sc_i \to ue_k} \geq \gamma^{req}_{sc_i \to ue_k} - \gamma^{pro}_{m \to sc_i} \text{ for } TTT \wedge \\ &RB^{ue}_{sc_i} < 1\Big]. \end{aligned} \tag{38}$$

In (38), the SINR $\gamma^{req}_{sc_i \to ue_k}$ is taken from (18).

Whereas the probability of inter-SC HOs, i.e., SC j to SC i, is expressed as

$$\begin{aligned} P^{inter}_{HO} = \mathbb{P}\Big[&P^r_{sc_i \to ue_k} \geq P^{th}_{min} \wedge \mathrm{E}\big[ToS_{ue \to sc_i}\big] > T_{th} \wedge \\ &\gamma^r_{sc_i \to ue_k} \geq \gamma^{req}_{sc_i \to ue_k} - \gamma^{pro}_{sc_j \to sc_i} \text{ for } TTT \wedge \\ &RB^{ue}_{sc_i} < 1\Big]. \end{aligned} \tag{39}$$

In (39), the SINR $\gamma^{req}_{sc_i \to ue_k}$ is taken from (21).

In fact, the performance of the network in terms of handover is expected to be enhanced with lower network load. Considering a constant network load, when we increase the number of SCs under the coverage area of the MC, the network load will be shared among the MC and the SCs. Thus, the load per cell is reduced resulting in a lower level of interference and hence reducing the radio link failure, which causes handover failure. The outage probability or the probability of transmission failure takes place either when the UE initiates HO procedures but an interruption stops the process before completion (before the HO execution time expires) due to the degraded SINR from the source and the target base stations, or when the SINR of the serving cell is degraded and the target SC has lack of resources. Thus, the outage probability can be written as

$$\begin{aligned} P_{out} = \mathbb{P}\Big[&\gamma^r_{sc_i \to ue_k} < \gamma_{th} \wedge \gamma^r_{m \to ue_k} < \gamma_{th} \text{ for } t < T^{exe}_{ho} \\ &\vee \\ &\gamma^r_{m \to ue_k} < \gamma_{th} \wedge RB^{ue}_{sc_i} = 1 \text{ for } t < T^{exe}_{ho}\Big], \end{aligned} \tag{40}$$

where T^{exe}_{ho} is the time needed to finish the HO process (including HO preparation time and HO execution time and is set to 1 second [4]).

C. RESULTS AND DISCUSSIONS

Initially, the UE is connected to the MC and receive $\gamma^r_{m \to ue_k}$, which gives $r_{m \to ue_k}$. The MC UE is moving from the MC towards the SC coverage area at a speed of V_{ue}. Due to its mobility, the UE approaches the vicinity of the SCs and follows the proposed method to perform HO to a SC $\in N^{**}_{sc}$ with the highest SINR i.e., the UE performs HO to SC i, which has an SINR equivalent to $\gamma^{req}_{sc_i \to ue_k}$, considering the proposed interference-based load-dependent margin, and also has the available resources in the target SC. The simulation parameters are listed in table 4 [7].

TABLE 4. Simulation parameters.

BW	10 MHz
Carrier Frequency (F_c)	2.5 GHz
MC Transmit power	43 dBm
MC Radius	800 m
SC Radius	100 m
Maximum SC Transmit power	23 dBm
Number of SC within the MC	100
Number of UEs within MC sector (N^{sec}_{ue})	40
Maximum number of UEs per SC	5
Minimum required signal for service continuity (P^{th}_{min})	-70 dBm
Outage threshold (γ_{th})	5 dB
Handover completion time (T^{exe}_{ho})	1 sec
Mean velocity of the UE (v_m)	{1,10,20,30, 40,50,60,70} km/h
Standard deviation for UE velocity (v_{std})	1 km/h
Period between two updates of the mobility model (Δt)	1 sec
Time threshold for ToS (T_{th})	5 sec
λ_i	[0.5:0.5:5]

To practically validate the impact of the proposed interference-based load-dependent margin $\gamma^{pro}_{m \to sc_i}$, it is tested against the MC load. Then we applied the HO condition in (29) by substituting the margin $\gamma^{pro}_{m \to sc_i}$. Here we assumed that $\gamma_{th} = 5$ dB [20], and the margin $\gamma_{m \to sc_i}$ is adjusted according to the mobility of the UE [1], i.e., low, medium, and high mobility as given in section III-C. Fig. 9 depicts the proposed HO margin against the load on the MC. As depicted in the figure, as the load on the MC increases the HO margin also increases linearly for all mobility states. Therefore, we expect an earlier HO to SC when the MC load increases since the HO

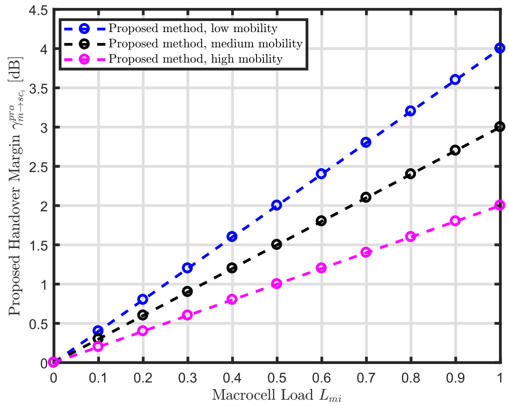

FIGURE 9. Proposed handover margin.

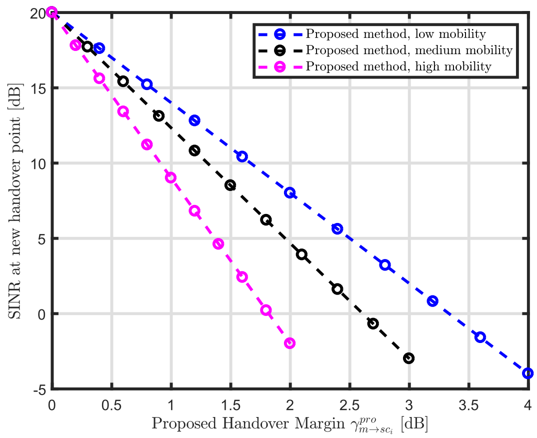

FIGURE 10. New handover point.

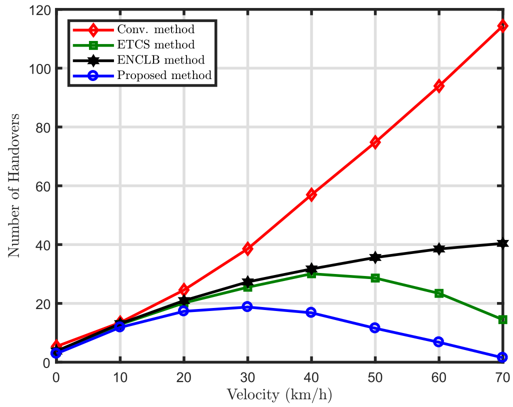

FIGURE 11. Total number of handovers.

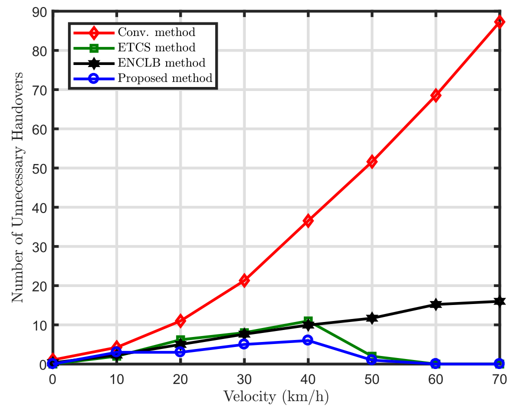

FIGURE 12. Number of unnecessary handovers.

condition subtracts the proposed HO margin from the SINR causing an early HO as the MC congests with a high load.

The proposed new HO point against the proposed HO margin is shown in Fig. 10. As the proposed HO margin $\gamma^{pro}_{m\to sc_i}$ increases (meaning the load on MC increases), the location of the new HO point decreases (i.e., the HO is initiated earlier) for all mobility states, which indicates that the new HO point from MC to SC is forced to be closer to the MC, i.e., the new HO point should be before the point at which the SINR of both the MC and the SC is identical. It can be noticed from Fig. 10 that the higher the proposed margin, the lower the new HO point for all mobility states. For example, when the proposed margin is 0.5 dB, then the new HO points for low, medium and high mobility states are respectively 17 dB, 16.2 dB and 14.5 dB. On the other hand, when the proposed margin is 1.5 dB, then the new HO points are 11 dB, 8.5 dB and 3.5 dB for low, medium and high mobility states respectively.

The total number of HOs is shown in Fig. 11. The conventional method has a higher noticeable increase in the number HOs including inbound and inter-SC HOs. Generally, for both of the proposed method and ETCS method, the number of HOs to SCs is highly reduced due to the reduction in the number of SCs in the NCL owing to the ToS condition. Our proposed method outperformed the three methods by minimizing the unnecessary HOs for different speeds because our method triggers the HO at a point when the data rate from the target SC is good enough with the consideration of the interference-based load-dependent modified A3 HO condition, unlike the other methods that depend on the downlink received power to trigger the HO to SC via the classical A3 HO event.

The HO is regarded as an unnecessary, if the user performs HO to a cell and then performs another HO to another cell before the expiry of the timer (5 seconds [7]). The proposed method minimizes the unnecessary inbound and inter-SC HOs as V_{ue} increases compared to the competitive methods because the final NCL only includes a few SCs (the proposed method removes the SCs that cause the short time of stay phenomena from the HO NCL) as the velocity increases, hence, the reduction in HO occurs, e.g., after 40km/h as depicted in Fig.12. It can also be observed from Fig.12 that a slight increase in the number of unnecessary handovers occurs in the proposed method, between 20km/h and 40km/h, due to the handover forcing to balance the load between cells. However, this increase is still below that of the other competitive methods.

Fig. 13 shows the outage probability. The proposed method gives a lower outage probability compared to the other three

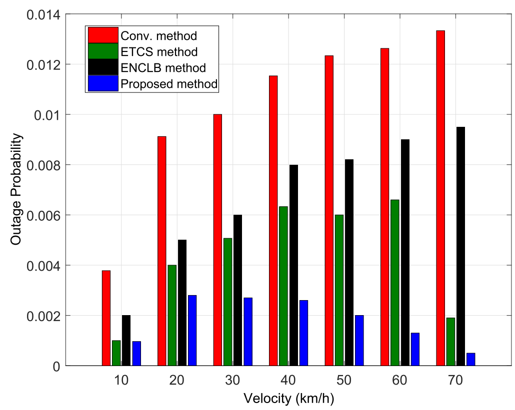

FIGURE 13. Outage probability.

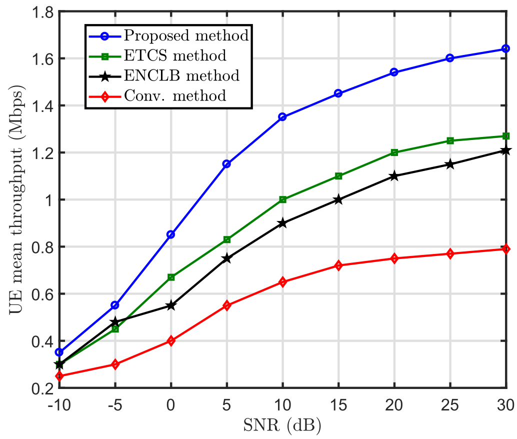

FIGURE 14. UE mean throughput vs. SNR.

methods because the proposed method only triggers the inbound and inter-SC HOs when there is a good data rate received from the target SC, which means that the HO is triggered with QoS consideration by taking into account the interference powers from the other adjacent base stations. The conventional and ENCLB methods have an instantaneous increase in the outage probability because of the fluctuated downlink received power due to the UE mobility in the HetNet, and the level of outage probability increases noticeably with the increase in UE velocity. The difference in the outage probability between the proposed and the ETCS method begins to be distinct at a velocity of 20km/h and it increases as the velocity increases because, in addition to the ToS criteria, the proposed method takes the interference from adjacent cells and the availability of PRBs into account when performing the HO to SC, resulting in QoS HO process. This reduction in the outage probability emphasizes that the proposed load-dependent margins, $\gamma^{pro}_{m\rightarrow sc_i}$ and $\gamma^{pro}_{sc_j\rightarrow sc_i}$, have properly managed the load distribution among cells in the network.

We also compared the performance of our proposed method with the other methods in terms of the maximum throughput a UE can obtain while moving through the network. Fig. 14 shows the UE's mean throughput with respect to different signal to noise ratio (SNR) values. The throughput is improved with the increase in the SNR accordant with common sense in all methods. The proposed method systematically gives the UE a high throughput compared to the other methods under different SNR values because the HO point for a UE trying to perform HO from an overloaded serving cell to a target SC is moved closer to the serving cell (i.e., the HO is triggered earlier), hence, the load is properly balanced between the two cells resulting in higher throughput.

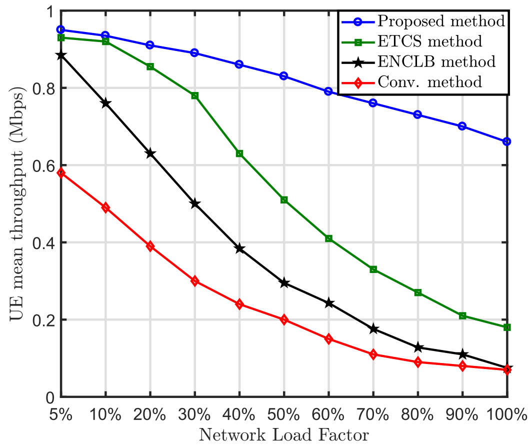

FIGURE 15. UE mean throughput vs. load factor.

For the range of MC load factors of 5% to 100%, Fig. 15 illustrates the UE mean throughput against load factor. The proposed method outperformed the other three methods. When the load increases, the cell becomes overloaded and its radio resources reduce, which may cause a drop in the throughput. As the load goes towards 1 (100% load), the interference will increase, which in turn will result in a reduced SINR leading to a lower UE mean throughput. From Fig. 15 we can see the sudden drop in the UE mean throughput for the ENCLB, the ETCS and the conventional methods since they initiate the HO to the target SC based on the downlink received power using A3 condition and also they do not apply the offloading policy, hence, higher dropping in calls is expected resulting in a lower throughput sudden decrease. On the contrary, the drop in the UE mean throughput for the proposed method is less than the other methods because the HO is performed when the modified interference-based load-dependent A3 condition occurs, where the users are offloaded to the SC by forcing the HO. Although there is a slight drop in the UE mean throughput for the proposed method as the load on the MC increases (due to the time needed for processing the HO from the serving cell to the SC), this drop is much slower than that of the other competitive methods.

Fig. 16 depicts the system throughput when the density of the UEs in the MC is varied. We assume that the density of the other cells in the network is fixed except for the MC, which is varied between 0 to 120 UEs. It is clear that the system throughput of the conventional method is always less than that of the other methods. Below 60 UEs in the network, the throughput of the conventional method continues

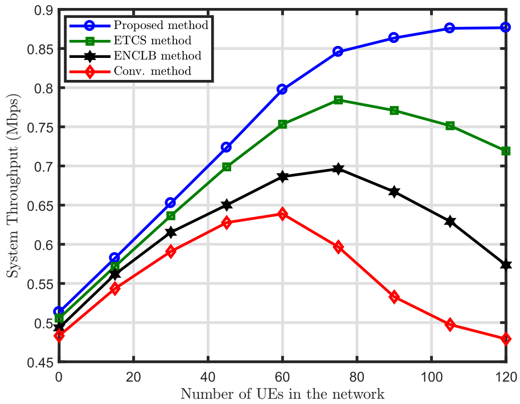

FIGURE 16. System throughput vs. number of UEs.

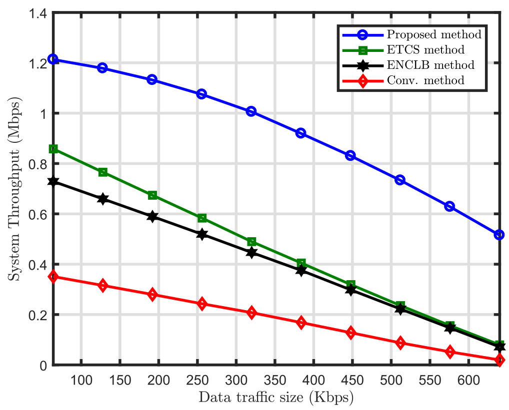

FIGURE 17. System throughput vs. traffic data size.

to increase because the capacity of the MC is still sufficient to deliver resources to the new incoming UEs, but a sudden drop in the throughput takes place after that owing to the fact that the MC will be heavily congested and its capacity will be limited. The same reason applies to the drop in the ENCLB method when the number of UEs exceeds 75. When the number of UEs is 60, we can observe that the proposed method has 15%, 12% and 2.5% enhancement in the throughput compared to the conventional, the ENCLB and the ETCS methods, respectively, and these percentages increase as the number of UEs increases. On the other hand, at 20 UEs and below, the performance of the proposed method is closer to that of the ETCS in terms of system throughput. However, above 20 UEs the proposed method's throughput is significantly higher than that of the ETCS due to the load-dependent margin incorporation which proves the proper distribution of the load between MC and SC tiers. Generally, for the proposed method, the average system throughput increases with the increase in the density of UEs. The reason behind this increase is that the HO point of UEs is moving closer to the overloaded cell, hence forcing the UEs to HO to a light load target SC. This means that the overloaded cell will not accept new HO requests, hence, reducing the load on this cell and eventually increasing its throughput.

When varying the traffic data size from 64KB to 640KB, the system throughput is depicted in Fig. 17. We assigned

TABLE 5. System throughput for different traffic data size.

Traffic data size [KB]	System throughput (Mbps)			
	Conv.	ETCS	ENCLB	Proposed
128	0.3175	0.765	0.658	1.18
256	0.2425	0.6	0.5416	1.1
384	0.17	0.3825	0.3652	0.91
512	0.09	0.255	0.2312	0.74
640	0.02	0.075	0.07	0.52

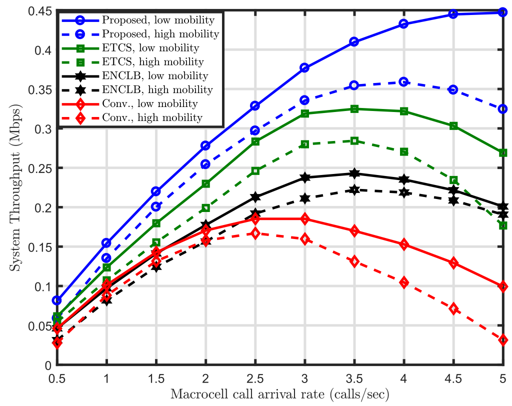

FIGURE 18. System throughput vs. call arrival rate.

120 UEs to the MC and no UEs to the other SCs. The proposed method always has the higher system throughput. As the traffic data size increases, the system throughput is sharply decreases for the conventional, the ENCLB and the ETCS methods because large data traffic size means that the UE will reside for a long time in the serving MC (i.e., the time that the user remains in the RRC-active mode will be longer), hence congestion is expected in the MC causing lower throughput. Unlike the proposed method, which offloads the traffic to the SC tier resulting in a slower decrease in the system throughput as the data traffic size increases. To clarify the differences in the total network throughput for the four methods, table 5 gives the achieved network throughput with respect to different traffic data sizes. For all values of traffic data size, the proposed method yields a higher network throughput when compared to the other methods. At a traffic data size of 256KB, the network throughput of the proposed method is 85% higher than that of the conventional method, 55% higher than that of the ENCLB method and 50% higher than that of the ETCS method. As the traffic data size increases, the obtained network throughput for the proposed method keeps above that of the other three methods. For example, at a 512KB traffic data size, our proposed method shows a better performance by producing 65%, 50% and 48% higher throughputs than that of the conventional, the ENCLB and the ETCS methods, respectively. This is due to the load balancing considerations in our proposed method.

Fig. 18 illustrates the system throughput against the MC new call arrival rate. The average MC call arrival rate is defined as a Poisson process with mean λ_i and the mean call duration is 60 seconds [23]. By varying the average MC new call arrival rate, the average system throughput is obtained for both low and high mobility states. Fig. 18 shows that the proposed method outperformed the other three methods by giving higher system throughput at different call arrival rates. From Fig. 18, for the low mobility scenario, we can notice that the throughput reaches its maximum level when the call arrival rate is 3 calls/sec for the ETCS method, when the call arrival rate is 2.5 calls/sec for conventional method and when the call arrival rate is 3.5 calls/sec for the ENCLB method. Then it starts to decrease as the call arrival rate increases because the network will congest with call sessions (no load balance strategy) and the interference will increase resulting in a lower throughput. In contrast, the proposed method balances the load between the congested cells and performs the offloading that leads to an increase in the system throughput. For the high mobility scenario, as the call arrival rate increases this will cause a reduction in the throughput. However, this has less impact on our proposed method due to the effect of the load-dependent margin that balances the traffic load between the MC and SCs.

VI. CONCLUSION AND FUTURE WORK

Inbound and inter-SC are considered as the most complicated HO scenarios in SC HetNets because there are high numbers of target SCs. Hence, the interference level should also be carefully considered in these HO scenarios. Therefore, in this paper, we proposed a novel HO method that takes into account the two scenarios. The effects of short ToS and interference are utilized to minimize the size of the final NCL so that the UE is forced to perform the HO to the cell that provides a good data rate and has enough capacity from a reduced NCL that contains a few and appropriate HO target SCs, in this way, traffic offloading from the MC to SC is achieved. We proposed a modified A3 HO initiation event taking into considerations the cell load and interference. Results reveal that the proposed method minimizes the outage probability and unnecessary HO compared to the existing methods. The proposed method also outperformed the competitive methods by giving greater throughput as the density of the users increased in the HetNet. Under different network conditions, including SNR, load factor, traffic data size and call arrival rate, we tested and compared the proposed method against the ETCS, the ENCLB and the conventional methods. Under all network conditions our proposed method outperformed the other three methods by providing a higher system throughput.

In the future, we aim to expand this work to focus on the reduction of NCL dynamically using a multi-tier small cells heterogeneous network.

REFERENCES

[1] X. Chu, D. López-Pérez, Y. Yang, and F. Gunnarsson, *Heterogeneous Cellular Networks: Theory, Simulation and Deployment*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[2] J. Acharya, L. Gao, and S. Gaur, *Heterogeneous Networks in LTE Advanced*. Hoboken, NJ, USA: Wiley, 2014.

[3] M. N. Halgamuge, H. L. Vu, K. Rarnamohanarao, and M. Zukerman, "Signal-based evaluation of handoff algorithms," *IEEE Commun. Lett.*, vol. 9, no. 9, pp. 790–792, Sep. 2005.

[4] *E-UTRA—Radio Resource Control (RRC), Protocol Specification*, document 36.331, 3GPP, 2013, vol. 290.

[5] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[6] Z. Li, Z. Xiao, D. Wang, and X. Li, "Hysteresis automatic configuration for outbound handover in heterogeneous small cell networks," in *Proc. 12th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, Aug. 2015, pp. 2127–2131.

[7] C.-H. Lee and J.-H. Kim, "Time-of-stay estimation-based cell selection scheme in multitier heterogeneous mobile networks," *IEEE Commun. Lett.*, vol. 19, no. 9, pp. 1596–1599, Sep. 2015.

[8] Z. Becvar, M. Vondra, and P. Mach, "Dynamic optimization of neighbor cell list for femtocells," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1–6.

[9] M. Alhabo and L. Zhang, "Unnecessary handover minimization in two-tier heterogeneous networks," in *Proc. 13th Annu. Conf. Wireless On-Demand Netw. Syst. Services (WONS)*, Feb. 2017, pp. 160–164.

[10] M. Alhabo, L. Zhang, and N. Nawaz, "A trade-off between unnecessary handover and handover failure for heterogeneous networks," in *Proc. Eur. Wireless*, May 2017, pp. 1–6.

[11] X. Shao, Z. Gao, W. Zhou, H. Liu, and Y. Wang, "A multi-criteria handover algorithm for UE energy efficiency and cell load balance in dense HetNets," in *Proc. 19th Int. Symp. Wireless Pers. Multimedia Commun. (WPMC)*, Nov. 2016, pp. 14–18.

[12] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 2457–2461.

[13] A. L. Yusof, M. A. Zainali, M. T. M. Nasir, and N. Ya'acob, "Handover adaptation for load balancing scheme in femtocell long term evolution (LTE) network," in *Proc. IEEE 5th Control Syst. Graduate Res. Colloq. (ICSGRC)*, Aug. 2014, pp. 242–246.

[14] J. Zhang and G. De la Roche, *Femtocells: Technologies and Deployment*. Hoboken, NJ, USA: Wiley, 2011.

[15] *HNB and HNB-Macro Propagation Models*, document 3GPP R4–071617, Qualcomm Europe, Oct. 2007.

[16] G. L. Stüber, *Principles of Mobile Communication*. Cham, Switzerland: Springer, 2011.

[17] *E-UTRA—Radio Frequency (RF) Requirements for LTE PICO Node B, Release*, document TR 36.931, 3GPP, 2012, vol. 9, p. V9.

[18] *E-UTRA—Mobility Enhancements in Heterogeneous Networks*, document TR 36.839, 3GPP, 2012.

[19] H. Holma and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE*. Hoboken, NJ, USA: Wiley, 2010.

[20] H. Kalbkhani, S. Yousefi, and M. G. Shayesteh, "Adaptive handover algorithm in heterogeneous femtocellular networks based on received signal strength and signal-to-interference-plus-noise ratio prediction," *IET Commun.*, vol. 8, no. 17, pp. 3061–3071, 2014.

[21] Y. Lee, B. Shin, J. Lim, and D. Hong, "Effects of time-to-trigger parameter on handover performance in SON-based LTE systems," in *Proc. 16th Asia–Pacific Conf. Commun. (APCC)*, Oct./Nov. 2010, pp. 492–496.

[22] N. Bulusu, D. Estrin, L. Girod, and J. Heidemann, "Scalable coordination for wireless sensor networks: Self-configuring localization systems," in *Proc. Int. Symp. Commun. Theory Appl. (ISCTA)*, Ambleside, U.K., 2001, pp. 1–6.

[23] H. H. Shuai, D. N. Yang, W. H. Cheng, and M. S. Chen, "MobiUP: An upsampling-based system architecture for high-quality video streaming on mobile devices," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1077–1091, Oct. 2011.

[24] M. Alhabo, L. Zhang, and O. Oguejiofor, "Inbound handover interference-based margin for load balancing in heterogeneous networks," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2017, pp. 146–151.

MOHANAD ALHABO received the B.Sc. degree in computer and information engineering from the Faculty of Electronics Engineering, University of Mosul, Mosul, Iraq, in 2007, and the M.Sc. degree in computer network administration and management engineering from the University of Portsmouth, Portsmouth, U.K., in 2009, and the Ph.D. degree with the School of Electronic and Electrical Engineering, University of Leeds, Leeds, U.K., in 2018. He was a Network Engineer for over four years. His research interest includes mobility management, handover and interference management for heterogeneous networks.

LI ZHANG received the Ph.D. degree in communications from the University of York, York, UK, in 2003. Since 2004, she has been with the School of Electronic and Electrical Engineering, University of Leeds, Leeds, U.K., where she became a Senior Lecturer in 2011. Her main research interests include communications and signal processing. She is serving as an active reviewer for a number of journals and a technical program committee member for a number of conferences. She has also been a member of the UK's prestigious Engineering and Physical Sciences Research Council Peer Review College since 2006. She received the Nuffield Award given to newly appointed lecturers in 2005 and the IEEE Communication Letter Exemplary Reviewer Award in 2010.