Load Balancing Based on Clustering Methods for LTE Networks

Omar Altrad, *Member, IEEE*, Sami Muhaidat, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a general load-balancing algorithm to help congested cells handle traffic dynamically. The algorithm is based on clustering methods and can be applied to any wireless technology such as LTE, WiMAX and GSM. The algorithm can be automatically controlled and triggered when needed for any cell on the system. It can be implemented in a distributed or semi-distributed fashion. The triggering cycle for this algorithm is left for the operator to decide on; the underlying variations are slow so there is no need for fast self-optimizing network (SON) algorithms. We apply the load-balancing algorithm to an LTE network and different criteria are adopted to evaluate the algorithm's performance.**

***Index Terms*—About Load balancing, LTE, Handover, SON.**

I. Introduction

THIS load experienced by neighboring cells tends to vary depending on the time of day and centers of activity; this causes cells to be more or less congested. Different distributions of traffic occur in both space and time which leads to unbalanced loads in the cells and causes degradation in system performance. This temporary traffic concentration problem needs a dynamic mechanism to adapt for these changes, either by using more hardware resources or the careful design of an algorithm to treat these occurrences. The load-balancing algorithm aims to find the optimum handover offset (HO) value between the overloaded cell and a possible target cell. The use of load-balancing (LB), which belongs to the group of suggested SON functions for LTE network operations, is meant to deliver this extra gain in terms of network performance. In addition, the algorithm needs to adjust the network control parameters in such a way that overloaded cells can offload the excess traffic to low-loaded adjacent cells [1]

In [2] a method of balancing the load among cells which are operating at maximum capacity is described. However, this method has the disadvantage of handling the handover of the mobile station (MS) due to load balancing differently from the handover of the MS leaving the cell. Another approach [3] patented by Kojima, narrows the service area of the base station (BS) by reducing its output power. However, this patent does not discuss the manner in which the handing over of established calls takes place. Bodin [4] introduces the concept of adaptive handover boundaries and introduces a simple algorithm to solve this problem. However, this algorithm does not ensure the existence of a continuous overlapping area. In [5] a solution for the adaptive handover problem is considered based on the predictable pattern of traffic loads; however, this assumption becomes inefficient when a deviation between the current pattern and the analyzed historical traffic patterns occurs.

Most of the previous work on evaluating the performance of load-balancing algorithms for cellular networks emphasizes simulations [6], [7]. Other papers adopt the theoretical analysis approach, which involves using mathematical techniques such as queuing models and Markov chain models to model and study the performance of task scheduling algorithms [8], [9].

Our contribution in this paper is the following:

- We introduce a new load-balancing algorithm based on clustering methods, where the centroid of the cluster is the cell position;
- a mathematical formulation for the problem to analyze the algorithm is introduced;
- The triggering mechanism for the algorithm is the call blocking ratio (CBR), which is the real parameter reflecting the degradation of the system when overload occurs;
- a control function is introduced and implemented with the proposed message names for reducing signaling overhead between the cells.

The rest of the paper is organized as follows. A description of the algorithm and its behavior is discussed in section II. In section III, a mathematical analysis of the algorithm is presented. Section IV provides the simulation results and the paper is concluded in section V.

II. Algorithm Design and Description

The proposed algorithm is semi-distributed since it is invoked at each congested cell and controlled by a management entity. The input for this algorithm is the current load of each neighboring cell as well as the current handover margin which is shown in (11).

The proposed algorithm works as follows: First each mobile station i reports its measurements to its serving cell j in a periodic fashion. These measurements include the SNIR measurements of the neighboring cells as well as the serving

Manuscript received January 30, 2013.

O. Altrad and S. Muhaidat are with the School of Engineering Science, Simon Fraser University, Burnaby, BC, V5A1S6, Canada. Phone: 778-782-7376. Fax: 778-782-4951. E-mail: oaltrad@sfu.ca, muhaidat@ieee.org

cell. At any time and for any cell in the system if (12) is satisfied then the cell is considered to be congested, and a ***REQ_Load_Balance*** message is sent to the management entity.

This will respond by sending a ***Load_Balance_Res*** message to invoke the algorithm on the requested cell and at the same time this cell is added to the list of the update cell pool. We need this process before starting the algorithm in any cell in order to allow the management entity to exclude this cell after it has finished running the algorithm, so as to prevent an endless loop in the system. Also note that it is straightforward to modify the algorithm to be executed without the management entity, i.e., fully distributed. However, the drawback of this will be extra signaling overhead between cells as well as an increased ping-pong effect, since every congested cell will try to shift the overload to its neighboring cells when congestion occurs in more than one neighboring cell.

After that, the congested cell will request the estimated load of each neighboring cell in the list. This reflects the current load of the cells. Each cell with an estimated load less than the defined threshold load will be added to the candidate list to be assigned a portion of the load of other cells with the condition that this cell is a neighboring cell of the congested cell. Note that we have only one threshold defined to distinguish between congested and decongested neighbors. This threshold is enough to do the job, since the algorithm will recursively choose the next congested cell and exclude the cells that are already fully loaded. A mapping function is used to update the handover threshold of the overloaded cells as shown in (15).

The algorithm is simple and the only requirement is the measurement exchange of the estimated load. This extra signaling depends only on the size of the list of neighbors and the periodicity rate. It is not; however, appropriate to substantially decrease the size of the list of neighbors due to handover optimization issues. The following script is used to illustrate the proposed algorithm.

Proposed Algorithm

```
A mobile station reports its measurements to the serving cell j using (5).
The serving cell $j$ detects an overload using (12), and sends the
Req_Load_Balance message.
The management entity adds cell $j$ to the update balance pool and responds
by sending Load_Balance_Res.

FOR each cell j (congested cell), the clustering function is invoked to estimate
            the overloaded portion, with the constraint of (10)
    FOR each cell i neighbor to j where i = 1… k performs Update HM (j;
            i) using (15), so as to reduce the coverage  area which depends
            on the available resources of its neighbor.
        i = i + 1
    END
 Informs the management entity by sending Update_Balance_ Fin. The
management entity drops cell j from the balance pool if Update_Balance_Fin
is received.
    j = j + 1
END
```

III. Algorithm Analysis

Consider a cellular coverage area ***C*** consisting of n cells, where, $C = \{C_1,\ \mathrm{C}_2,\ ...,\ \mathrm{C}_n\}$, and a set of all mobile stations defined as, $M = \{M_1,\ M_2,\ ...,\ M_m\}$. Denote $M_{i,j}$, as the mobile station i connected to C_j, where $i = 1,...,m$, $j = 1,...,n$ and m□ n. Then the received power at the mobile station i from base station j is defined as

$$P_{r_{M_{i,j}}} = \frac{P_{t_j} G_{i,j}}{L_{i,j}} \tag{1}$$

Where P_{t_j} is the transmitted power of the cell j, $G_{i,j}$ is the gain between the mobile i and the cell j, and $L_{i,j} = l_\circ d_{i,j}^{\alpha} \gamma$. Where $l_\circ$ is constant depends on the frequency being used. $d_{i,j}$ is the distance between the mobile i and the cell j. α is the path loss exponent and γ represents the shadowing effect, which can be modeled as shown in [10]. The measured signal to interference and noise ratio at $M_{i,j}$ can be defined as

$$SINR_{M_{i,j}} = \frac{P_{r_{M_{i,j}}}}{I_{i,j} + N_\circ} \tag{2}$$

where $N_\circ$ is the thermal noise and $I_{i,j}$ is defined as

$$I_{i,j} = \sum_{p \neq j} X(j,p) \rho_p P_{r_{M_{i,j}}} \tag{3}$$

where $X(j,p)$ is defined as

$$X(j,p) = \begin{cases} 1, & when\, j,p \ use\ the\ same\ band \\ 0, & when\, j,p \ use\ different\ band \end{cases} \tag{4}$$

and ρ_p is defined as the load ratio of used resources.

We represent each $M_{i,j}$ as a point in space of a dimension equal to the length of the neighboring base stations list and the serving cell j, i.e.,

$$D_{M_{i,j}} = f\left(SINR_{i,1}, \ldots, SINR_{i,k}, SINR_{i,j}\right) \tag{5}$$

where $\{1, \ldots, k\}$, represents the length of the list of neighboring cells. Using this convention, we can represent the cell C_j as a point in space of the same dimension, i.e.,

$$D_{C_j} = f\left(SINR_1, \ldots, SINR_k, SINR_j\right) \tag{6}$$

We assume that the cell can measure the received power of its neighbors, as it will be the centroid of the clusters. Then the Euclidian distance between $M_{i,j}$ and C_j will be

$$X_{M_{i,j}} = \left| D_{M_{i,j}} - D_{C_j} \right| \tag{7}$$

The clustering algorithm must be applied recursively, so as to map the load in each congested cell to a number of clusters, and then use this mapping to adjust the handover margin with each neighboring cell. By default this prevents or delays the heading MSs from connecting to the congested base station or extending the connection of an MS in a light-loaded cell to a certain limit. To apply the clustering method, some preparation is needed and a number of requirements must be met for this process to be accomplished successfully. We can represent the overall load of the congested cell as

$$L_{C_j} = \sum_{K=1}^{2} S_K \tag{8}$$

where S_1 represents the first cluster size which is intended to be handed over to neighboring cells. S_2 represents the second cluster size which represents the acceptable load of C_j that can be handled, i.e., it can be constrained by

$$S_2 \leq L_{th} \tag{9}$$

where L_{th} is a predefined threshold for each cell, which represents the maximum allowable load on the cell. Note that this threshold does not mean we reserve resources in the cell, since it can simply be replaced by the maximum allowable load in that cell. However, we define it here as such for illustration purposes.

Estimating the size of one cluster will give us the size of the other cluster. To do so, we sort the mobile stations according to their Euclidian distances from the congested cell and keep adding the requested resources for each mobile till we reach the maximum load threshold that can be handled by the cell. This distance will indicate the crossover point between the two clusters. The size of cluster S_1 can be constrained by the sum of all available resources in the neighboring cells:

$$S_1 \leq \sum_{i=1}^{k} (1 - \widehat{L}_i) \tag{10}$$

Where $\widehat{L}_i$ is the estimated load of the neighboring cell i. Equation (10) is implemented to prevent an endless loop between cells and to reduce the ping-pong effect when the algorithm is executed.

Note that an entire cluster is not necessarily handed over to one neighboring cell only, since each mobile station in the cluster will be handed over to a preferred cell indicated by the best SNIR received, i.e., cluster S_1 can be considered to be divided into sets; each set will be connected to the neighboring cell with the best SNIR received. For more clarification of this point, see Fig.1, where we consider two neighboring cells for illustration purposes.

A. *Handover Triggering Condition*

The condition which triggers the handover from a serving cell to its neighboring cell can be dependent on many factors. Some examples of these are BER, SINR, and RSSI. In LTE networks, the hard handover algorithm or so called Power Budget Handover Algorithm is adopted. Two parameters are defined in the cell at the time of deployment to switch the mobile user from one cell to its neighbor, the handover margin (HM) and the time to trigger (TTT). These parameters are constant and implemented during the deployment phase. Different values are considered for each cell on the system. The received signal strength is called the reference signal received power (RSRP) which is used to evaluate whether the condition to trigger a handover has been met. This condition can be written as.

$$RSRP_T \geq RSRP_S + HM \tag{11}$$

where $RSRP_T, RSRP_S$ are the reference signal received power of the targeting cell and the serving cell, respectively. This condition must be satisfied for a period of time represented by the TTT.

B. *Detection of overloaded cells and handover adaptation*

Before invoking the algorithm, a triggering method should be used to detect the overloaded cells. We used CBR as the triggering method, i.e., when

$$CBR \geq B_{th} \tag{12}$$

where $CBR = \text{blocked calls / total accepted calls}$,and B_{th} is a predefined threshold kept for operator use which is determined by the quality of service (QoS) the operator promised to provide. In this paper we kept this threshold to 2% [11].

A mapping function to adjust the handover margin between each pair of cells should be used. The adjusted handover margin between the congested cell j and its neighbor i will be directly proportional to the estimated load in the neighboring i. The overloaded portion represented by the cluster S_1, defined in (10), can be divided into subsets; each subset will reflect the amount of load intended to be handed over to a neighboring cell. This subset will be constrained by the maximum allowable load that can be shifted to this neighboring cell, i.e.,

$$S_1 = \{s_{1,1}, s_{1,2}, \ldots, s_{1,k}\} \tag{13}$$

where $s_{1,i}$ is the subset of neighbor cell i. The size of the subset can be constrained by

$$s_{1,i} \leq 1 - \widehat{L}_i \tag{14}$$

The handover margin between the congested cell j and the neighboring cell i will then be adjusted as

$$HM\left(j,i\right) = HM_{def} + (HM_{def} - H_{max})s_{j,i} \qquad (15)$$

Note that the adjusting procedure requires only one step. This method will dramatically reduce the signaling overhead caused when compared to conventional adjustment methods. The adoption of the linear equation shown in (15) is because the handover margin is directly proportional to the overload portion. It thus follows that a linear function is sufficient for this procedure. Equation (15) is a smart way of adjusting the handover margin when the cells are congested, since this adjustment is pair-wise adjusted, i.e., each neighbor has its own adjusted handover margin with the congested cell. The algorithm we propose will handle this adjustment procedure.

IV. Simulation Results

For the performance evaluation of the proposed algorithm, a modified LTE model based on Opnet modular 16 simulation software is adopted as shown in Fig .2. A scenario that reflects the practical situations of MSs movement and environment is modeled.

Table 1 PARAMETERS VALUES FOR SIMULATIONS

Attribute	Value
Base Frequency	2 GHz
Network layout	7 BS site 3 sectors, 21 cells
Path loss Model	Okumura-Hata COST 231
Bandwidth 10 MHz	Bandwidth 10 MHz
Maximum transmission power	46 dBm
Physical Profile Type	OFDM
Mobility model Random	Waypoint Model
Thermal noise (N)	-114 dBm
Shadowing	zero mean and standard deviation 8 dB
Symbol Duration	100.8 microseconds
Number of subcarriers per RB	12
Sub-carrier Spacing	15 KHz
Packet Scheduler	Round Robin
Inter site distance	500 m
Threshold load L_{th}	0.85 of the maximum estimated load
Frame Duration	10 milliseconds

A. Scenario and Parameters

The major simulation parameters are shown in Table I, where we follow the reference settings for LTE [12]. Each BS site has three sectors; each sector represents a cell of hexagonal shape. A total of 21 cells are used in this study. The MSs are uniformly distributed in the lightly loaded cells. Each MS is constantly moving at a fixed speed and with an initial direction randomly chosen from 0 to 2π, where the MS is permitted to change its direction randomly so as to represent practical situations. Moreover, we have created a cluster of MSs having random movement which can be dropped at different times of the simulation into randomly chosen cells to represent buses and trains.

The Costa-231 HATA model for the urban environment is used for the path loss computation [13]. Shadow fading is modeled as a Gaussian log normal distribution with 0 mean and 8 dB variance [14]. A Round-Robin packet scheduler is used for fair transmission while the Hybrid Automatic Repeat Request (HARQ) technique is used for wireless transmission error recovery [15]. Assuming a constant bit rate of 256 kbps for each user and a bandwidth of 10 MHz, this will cause an approximately 38 users/cell. The algorithm described earlier will adjust the handover threshold for each congested cell with its neighboring cells, which requires only one step as discussed earlier.

Starting from the most congested cell, the adjusting procedure will follow the Euclidian distance mapping described earlier. 10 dB is the maximum handover adjustable margin. The default value is kept unchanged when there is no congestion in the cell, where the maximum means the cell is fully loaded. All cells of sites ***BS_2***, ***BS_4*** are considered as congested cells with an average 42 users/cell. All other cells belonging to sites ***BS_1, BS_3, BS_5, BS_6,*** and ***BS_8*** are lightly-loaded cells with an average of 12 users/cell.

B. Results

In the simulation scenario, we compare the performance of the proposed system when implementing the load-balancing algorithm and when not invoking the algorithm. The performance is measured in terms of CBR. Since this algorithm entails only one-step adaptation, the result was promising and the CBR was dropped to almost zero in all congested cells as shown in Fig. 3.

For example, before the algorithm was invoked, the CBR in cells 7, 8 and 9 was 12%, 9%, and 10% respectively, which contributed to a CBR average of 10% in the three congested cells. This congestion was reduced to almost 1% when the algorithm was applied. Thus a reduction in congestion of 90% was achieved.

Also notice that the average load of cell 14 exceeded 2, where this load represents the blocked and accepted calls as shown in Fig. 4. At the same time, the CBR of this cell was 10% as shown in Fig. 3. Right after invoking the algorithm as seen in Fig. 4, the load on cell 14 was reduced to the threshold 85% which we defined earlier. This reduction of load in the congested cell was carried by cell 3 and cell 5 where each cell carried a portion of the overload. Note that cell 15 and cell 13 did not handle any portion since they were already congested cells. Also notice that even though cell 11 is a lightly-loaded cell, it did not contribute to the process as we limited the list of neighbors to only three to reduce the simulation time.

Fig. 5 shows a comparison between the algorithm suggested by [4] and our proposed algorithm. The Bodin algorithm was implemented after some modification to fit LTE requirements. The algorithm proposed by [4] did not show the step size, or the way the load of the congested cell was estimated. Moreover, it did not explain the invoking procedure or the message exchange between cells. As a result, our proposed algorithm showed a reduction of more than 80% in CBR compared to the Bodin algorithm. This reduction was also caused by the fine adjustment of our proposed algorithm and the control procedure we adopted.

Moreover, in Fig 6, a fluctuation of the load in the congested cell is seen when applying the Bodin algorithm which explains the high CBR and the instability of the algorithm. Our proposed algorithm, on the other hand, shows consistency and a smooth control of the current load of the congested cell.

Finally, one of the most important features for any proposed load-balancing algorithm is fast adaptability to the dynamic changes in the load in the congested cells which results in a reduced CBR. Therefore, the number of satisfied users is increased. This feature is shown in Fig. 6 where the convergence of our proposed algorithm is achieved with less time compared to the Bodin algorithm.

V. Conclusion

In this paper, a load-balancing algorithm based on clustering methods is proposed. We applied this algorithm to LTE networks. Our results show a significant improvement compared to previous works. Using a pair-wise method to adjust the handover margin significantly improves the performance of the system compared to the conventional methods which use the cell-wise method. Our new method shows a reduction in the CBR exceeding 85% in some cells. Moreover, a total reduction of 75%   in CBR is achieved on the overall system. Our results show a distribution of the load of the congested cell to its neighbor in one step only, which significantly reduces the signaling overhead and wasting of resources in the lightly-loaded cells compared to conventional methods. Applying this algorithm to more practical scenarios and relaxing some of the assumptions made here is left for future work.

VI. References

[1] A. Lobinger, S. Stefanski, T. Jansen and a. I. Balan, "Load balancingin downlink LTE self-optimizing networks," in *Vehicular Technology Conference (VTC 2010-Spring)*, May 2010.

[2] Brody and C. George, "Load balancing for cellular radiotelephone system". US Patent 4 670 899, 2 January 1987.

[3] J. Kojima and K. Mizoe, "Radio mobile communication system wherein probability of loss of calls is reduced without a surplus of basestation equipments". US Patent 4 435 840, 6 March 1984.

[4] R. bodin, Spanga and A. Norefors, "Load sharing control for a mobile celluar radio system". US Patent 5 241 685, 1 August 1993.

[5] C. Chandra, T. Jeanes and W. Leung, "Determination of optimal handover boundaries in a cellular network based on traffic distribution analysis of mobile measurement reports," in *Vehicular Technology Conference*, 1997.

[6] T. Nihtila, J. Turkka and I. Viering, "Performance of LTE selfoptimizing networks uplink load balancing," in *Vehicular Technology Conference (VTC Spring)*, May 2011.

[7] J. Rodriguez, I. d. l. Bandera, P. Munoz and R. Barco, "Load balancing in a realistic urban scenario for lte networks," in *Vehicular Technology Conference*, May 2011.

[8] I. Viering, M. Dottling and A. Lobinger, "A mathematical perspective of self-optimizing wireless networks," in *International Conference on Communications (ICC)*, June 2009.

[9] S. Kourtis and R. Tafazolli, "Adaptive handover boundaries: a proposed scheme for enhanced system performance," in *Vehicular Technology Conference Proceedings*, 2000.

[10] O. Altrad, S. Muhaidat and M. Dianati, "A novel-dual trigger handover algorithm in wimax networks," in *International Wireless Communications and Mobile Computing Conference(IWCMC)*, April 2012.

[11] A. Technologies, "Agilent 3GPP Long Term Evolution: System Overview, Product Development, and Test Challenges.," Available online: http://cp.literature.agilent.com/litweb/pdf/5989-8139EN.pdf, 2009.

[12] 3GPP, "Physical Layer Aspects for evolved Universal Terrestrial Ra-dio Access (UTRA)," 3GPP.

[13] T. S. Rappaport, Wireless Communications: Principles and Practice, 3rd ed., Prentice Hall, 2003.

[14] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electronics Letters*, vol. 27, no. 23, pp. 2154-2146, 1991.

[15] J. Ikuno, C. Mehlfuhrer and M. Rupp, "A novel link error prediction model for OFDM systems with HARQ," in *International Conference on Communication (ICC)*, 2011.

Omar Altrad (M’09) was born in Irbid, Jordan in 1974. He received the B.Sc. in communication engineering from Mutah University, Jordan, 1996 and the M.sc. degree in electrical and computer engineering from New York Institute of technology, USA.

He worked as a field engineer, Director of staff at Royal Jordanian Air Force, (1996-2006). Currently, he is pursuing his Ph.D. in wireless communications at the school of engineering Science, Simon Fraser University, Canada.

First A. Author (M’05, SM’12) received his M.Sc. in Electrical Engineering from University of Wisconsin, Milwaukee, USA in 1999, and the Ph.D. degree in Electrical Engineering from University of Waterloo, Waterloo, Ontario, in 2006. From 1997 to 1999.

he worked as a Research and Teaching Assistant in the Signal Processing Group at the University of Wisconsin. From 2006 to 2008, he was a postdoctoral fellow in the Department of Electrical and Computer Engineering, University of Toronto. From 2008 till present he is assistance professor in Engineering school, Simon Fraser University.

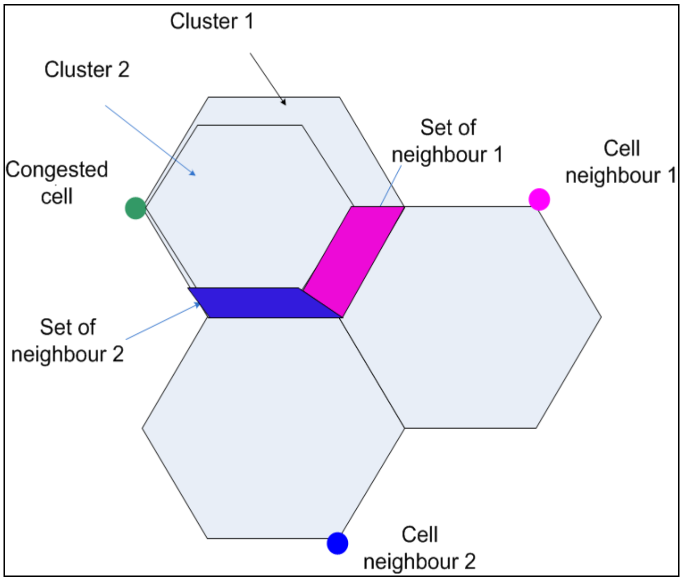

Fig. 1 Illustration of the subset for cluster S1

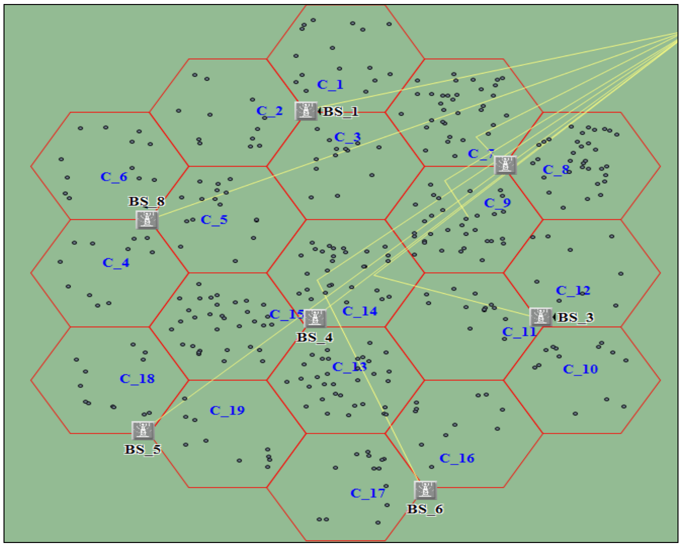

Fig. 2 Network Layout

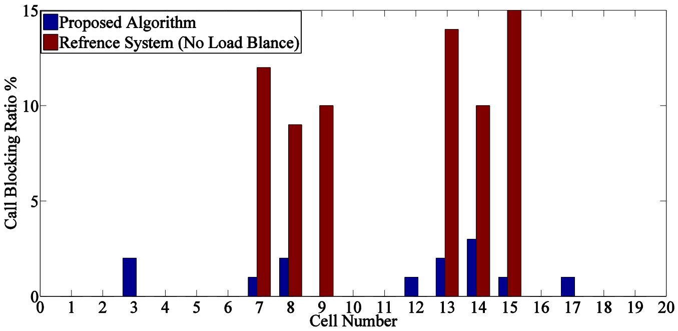

Fig. 3 Call blocking ratio for each cell in the network

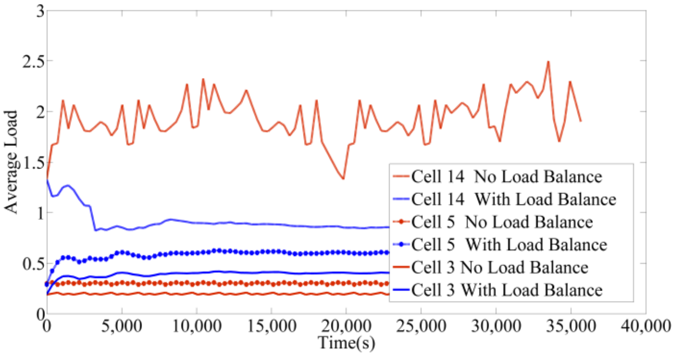

Fig. 4 Time average load in the congested cell 14 and its lightly-loaded

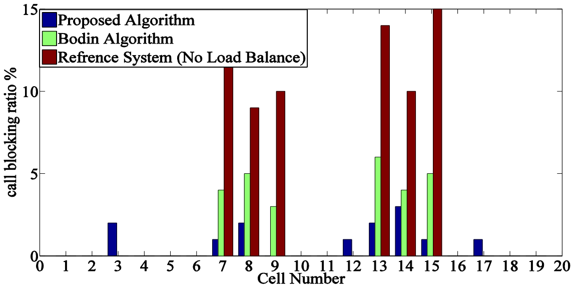

Fig. 5 Call blocking ratio of our proposed algorithm Compared to Bodin

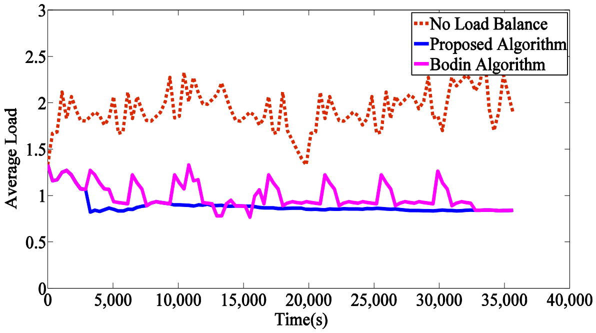

Fig. 6 Time average load of our proposed algorithm Compared to Bodin in