A Practical Approach for Load Balancing in LTE Networks

Miaona Huang, Suili Feng, and Jun Chen
South China University of Technology, Guangzhou, 510641, China
Email: huang.mn@mail.scut.edu.cn; fengsl@scut.edu.cn; chenjun0789@163.com

Abstract —Load imbalance among multi-cell has tremendous impact on the network performance. The previous researches on load balancing place focus on maximizing the load balancing index of the network and cannot always guarantee the best performance of network's key performance indicators (KPIs). In this paper, taking both the network resource limitation and users' data rate demand into account, we aim to simultaneously optimize both the load balancing index and network average load for quality-of-service (QoS) requirements services, while maximizing the network utility for other services. Moreover, we proposed a practical algorithm with low complexity. Comparing with the previous methods, simulation results show that the proposed method can achieve better network performances, such as lower new call blocking rate and higher network resource utilization.

Index Terms—LTE, multi-cell, load balancing (LB), quality-of-service (QoS), multi-objective optimization problem

I. Introduction

In recent years, with the surging traffic demands, wireless communication network has been becoming more complex, resulting in higher operational costs. In order to reduce the substantial operational expenditure in network operational tasks, while optimizing network efficiency and service quality, the concept of self-organization in communication networks, which is referred to as Self-Organizing Networks (SON), has been introduced in Long Term Evolution (LTE) system to reduce manual operations by standardization bodies [1].

Load balancing (LB) which aims to balance the uneven traffic load among neighboring cells is one of important functionalities that belong to SON. In the wireless network, traffic load in different cells is frequently unequal, which has the characteristics of spatial and temporal distribution. It brings about higher call blocking rate and higher call dropping rate in hotspot cells. In contrast, a large part of resources in low-loaded cells stays in idle state, resulting in wasted resources and decreased network throughput. The network performance is seriously deteriorated by load imbalance among neighboring cells. LB plays an important role in improving the user-experience and network performance by redistributing the traffic load among neighbor cells. Therefore, the 3rd Generation Partnership Project (3GPP) documents define a framework of LB for possible researches [2]. Many solutions for LB have been proposed for LTE networks, which can be divided into two types: One is that the hot-spot cell borrows idle resources from neighboring low-loaded cells, such as channel borrowing [3], [4]. The other is that the over-loaded cells offload extra traffic to neighboring cells by cell breathing technique [5], [6] or by modifying the handover regions between neighboring cells [7]-[10].

A widely accepted and popular approach for LB, similar to the method we propose, is to achieve LB by formulating the problem as a convex problem. Different utility functions have been investigated in conventional papers [11]-[14], such as maximum and minimum fairness, proportional fairness etc. Based on the multi-objective optimization theory, Hao Wang proposed a scheme to achieve better load balancing index, which is solved by a practical solution framework [15], [16]. However, in order to maximize the load balancing index, the methods in [15], [16] can lead to unnecessary and blind handovers. Users with poor channel condition of the target cells may be handed over. After handover, these users would take up considerable resources to satisfy their data rate requirements. This leads to network resources inefficiently used. The resource left for the new arrival users is greatly reduced, resulting in higher new call blocking rate. Due to the poor signal strength from the target cells, it also has negative impacts on users experience by generating handover problems such as handover failure. As a result, load balancing index of the network may be optimal, but the key performance indicators (e.g. new call blocking rate) of the network are suboptimal. The efficiency of the algorithm in [15], [16] needs to be improved. So, the LB algorithm needs further research to achieve better effect.

To address the above issue, the user association is formulated as a multi-objective optimization problem, which jointly optimizes the load balancing index and the network average load for users with QoS requirements, while maximizing the network utility of others. Physical resource limitation and users'data rate demands are made as the constraints. Also, a distributed and practical algorithm with lower complexity is proposed. This scheme enables appropriate users to be handed over to the neighboring cells. Simulation results show that the proposed method can achieve better load balancing index and better performance in new call blocking rate and network resource utilization.

Manuscript received April 13, 2014; revised June 27, 2014.
This work was supported by the National Natural Science Foundation of China under Grant No.61340035;
Corresponding author email: fengsl@scut.edu.cn.
doi:10.12720/jcm.9.6.490-497

The rest of this paper proceeds as follows. In Section II, the network model is presented. The optimization objective functions for LB are shown in Section III. In Section IV, we formulate the LB problem as a multi-objective optimization problem. In Section V, a detailed solution algorithm is proposed. Simulation results are given in Section VI. The paper is brought to a conclusion in Section VII.

II. System Model

A. Network Model

Without loss of generality, a multi-cell network is considered in this paper, as shown in Fig.1, each cell is served by an eNodeB. Two kinds of users, Guaranteed Bit Rate (GBR) and Best Effort (BE) services, which represent the users with and without QoS requirements service, respectively, are considered in this work. For simplicity, suppose that in a multi-cell network, the frequency reuse factor is 1 and all cells have the same amount of time-frequency resources, denoted as S . Physical Resource Block (PRB) containing 12 adjacent OFDM subcarriers is the basic unit that can be assigned to users [17].

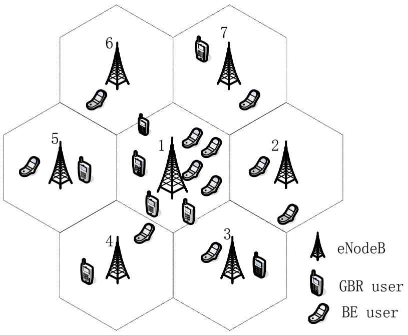

Fig. 1. System model

Let N denote the set of all cells. L represents the set of PRBs per cell. We use K_i , G_i and B_i to denote the total users, GBR and BE users in cell i , respectively. Obviously, $K_i = G_i \cup B_i$. A variable $I_{i,k}(t)$ is employed to indicate the affiliation of users at time t . $I_{i,k}(t)$ equals to 1 if the user $k \in K_i$ is associated to cell $i \in N$ at time t , otherwise, $I_{i,k}(t) = 0$. The symbol t denotes the time for LB and a LB cycle spans between any t and $t+1$, and it is much longer than a subframe (1ms).

B. Link Model

Each user is assumed to have the information of the instantaneous signal strength from all neighboring cells by pilot detection. The channel state information is sent back to its serving eNodeB by uplink transmission or periodical reports.

At subframe τ , the received signal-to-interference-and-noise-ratio (SINR) for user k from cell i at the *lth* PRB denoted as $SINR_{i,k,l}(\tau)$ can be expressed by

$$SINR_{i,k,l}(\tau) = \frac{g_{i,k,l}(\tau)p_{i,l}(\tau)}{N_0 + \sum_{j\in N, j\neq i} g_{i,k,l}(\tau)p_{j,l}(\tau)} \tag{1}$$

where $g_{i,k,l}(\tau)$ and $p_{i,l}(\tau)$ denote the channel gain of the *lth* PRB bwtween eNodeB i and user k and transmit power of the *lth* PRB of eNodeB i at subframe τ , respectively. $g_{j,k,l}(\tau)$ and $p_{j,l}(\tau)$ denote the channel gain of the \$l\$th PRB bwtween eNodeB j and user \$k\$ and transmit power of the *lth* PRB of eNodeB j at subframe τ , respectively. N_0 represents the power of Additive White Gaussian Noise per PRB.

The average bandwidth efficiency of user k in cell i at time t denoted as $\bar{e}_{i,k}(t)$ can be written as follows:

$$\bar{e}_{i,k}(t) = \frac{1}{|L|\cdot N_\tau}\left\{\sum_{\tau\in(t-1,t]}\sum_{l\in L}\log_2[1+SINR_{i,l,k}(t)]\right\} \tag{2}$$

where $|L|$ represents the number of the PRBs in a cell; N_τ denotes the number of the subframes in a LB cycle; e.g. if the LB cycle is 1second, a LB cycle contains 1000 subframes.

III. Objective Funcitons for LB

A. Load Balancing Index

Let $w_{i,k}(t)$ denote the allocated resources for GBR user k in cell i . $S_i^G(t)$ and $S_i^B(t)$ represent the resources that are occupied by GBR users and BE users in cell i at time t , respectively. As stated earlier, all cells have the same amount of time-frequency resources, denoted as S .

The load (or hard load) of cell i can be defined as the ratio of resources consumed by all GBR users and the total number of resources. Then the load of cell i , $\rho_i(t)$ can be calculated as follows:

$$\rho_i(t) = \frac{S_i^G(t)}{S} = \frac{\sum_{k\in G_i} I_{i,k}(t)w_{i,k}(t)}{S} \tag{3}$$

To measure the status of load balancing of the entire network, Jain's fairness index [18] is used, and represented as follows:

$$\varepsilon(t) = \frac{\left(\sum\rho_i(t)\right)^2}{|N|\sum\rho_i^2(t)} \tag{4}$$

where $|N|$ is the number of cells. The value of load balance index $\varepsilon(t)$ is between $\left[1/|N|\ ,1\right]$. Larger $\varepsilon(t)$ represents more balanced the load distribution among cells. So, the objective for GBR users is to maximize $\varepsilon(t)$ at each time t .

B. Network Average Load

$\rho(t)$ is used to represent the network average load at time t , which is

$$\rho(t)=\frac{1}{|N|}\sum_{i\in N}\rho_i(t) \tag{5}$$

Only considering $\varepsilon(t)$ when performing LB, users with extremely poor channel condition may be handed over to the neighboring cells and occupy considerable resources in target cells. In other words, aggressive handovers may happen. Then the network resources cannot be efficiently used and the new call blocking rates in target cells may increase. Thus, increasing the average load of the whole network as few as possible must be made as one of the objectives when performing LB. Thus, the objectives for GBR users when performing LB are to maximize $\varepsilon(t)$ and minimize $\rho(t)$ simultaneously at each time t.

C. *Network Utility for BE Users*

Since BE users do not have special QoS requirements, such as guaranteed bit rate, etc, throughput is a pertinent metric for performance evaluation. Let $R_{i,b}(t)$ denote the throughput of BE user $b\in B_i$ from cell i at time t and the calculation formula for $R_{i,b}(t)$ is shown in eq.(18). The utility function of throughput for BE user b can be represented as $U_b(I_{i,b}(t)R_{i,b}(t))$, and the utility function is usually concave function [19]. In order to ensure the fairness of allocating resources for BE users, the logarithmic function is adopted as the utility function, which is similar to proportional fairness, and achieves a desirable tradeoff between opportunism and fair allocation across BE users [20]. Then the total utility of all BE users in the network at time t can be expressed by

$$\psi(t)=\sum\nolimits_{i\in N}\sum\nolimits_{b\in B_i}\log_a(I_{i,b}(t)R_{i,b}(t)) \quad (a>1) \tag{6}$$

Then, the objective for BE users when performing LB is to maximize $\psi(t)$.

IV. Problem Formulation

In this section, the LB problem is formulated as a multi-objective optimization problem for the network. The objectives are to increase $\varepsilon(t)$ and decrease $\rho(t)$ as much as possible for GBR users while maximizing the total utility for BE users, so as to redistribute the traffic load among neighboring cells. Then at each time t, we try to minimize $\rho(t)$, maximize $\varepsilon(t)$ and $\psi(t)$ simultaneously and all of them are decided by the assignment between users and cells. The optimization problem with QoS and time-frequency resources constraints can be formulated as follows:

$$\max\ [\varepsilon(t),-\rho(t),\ \psi(t)]^T \tag{7}$$

$$s.t.\ \ \sum\nolimits_{k\in G_i} I_{i,k}(t)w_{i,k}(t)\le S\ ,\ \forall i\in N \tag{8}$$

$$\sum\nolimits_{i\in N} I_{i,k}(t)=1,\ \forall k\in K_i \tag{9}$$

$$\sum\nolimits_{i\in N} I_{i,k}(t)R_{i,k}(t)\ge D_k,\ \forall k\in G_i \tag{10}$$

where $[\bullet]^T$ denotes the transpose of a vector. D_k is the minimum data rate requirement of user k. The calculation formula for $R_{i,k}(t)$ is shown in eq. (17). The constraint (8) denotes that the resources taken up by all users in a cell should not exceed the total resources of that cell. (9) represents that a user can be connected to only one cell at any time. (10) explains that the minimum data rate requirement D_k of GBR users k should be rigorously satisfied in the current serving cell.

The objectives of maximizing $\varepsilon(t)$ and minimizing $\rho(t)$ are both for GBR users. Since both of them have the same magnitude, mathematically, the well-known linear weighted sum method can be utilized to construct them into a single aggregate objective function [21], [22]. So the optimization problem can be reformed as:

$$\max\ \ [\alpha\varepsilon(t)-(1-\alpha)\rho(t),\psi(t)]^T \tag{11}$$

$$s.t.\ \ \sum\nolimits_{k\in G_i} I_{i,k}(t)w_{i,k}(t)\le S\ ,\ \forall i\in N \tag{12}$$

$$\sum\nolimits_{i\in N} I_{i,k}(t)=1,\ \forall k\in K_i \tag{13}$$

$$\sum\nolimits_{i\in N} I_{i,k}(t)R_{i,k}(t)\ge D_k,\ \forall k\in G_i \tag{14}$$

$$\alpha\in[0,1] \tag{15}$$

Since the functions of $\alpha\varepsilon(t)-(1-\alpha)\rho(t)$ and $\psi(\mathrm{t})$ have different magnitude, they cannot be constructed into a single aggregate objective function. Note that $I_{i,k}(t)$ is 0-1 variable. Thus, the overall problem is an integer programming problem, which can be proved to be NP hard. As far as we known, there is no efficient algorithm except Exhaustive Search Method (ESM) to solve such a problem optimally. However, the computation complexity of ESM is enormous and intractable when N and the number of users in the network is large. Furthermore, it needs a central control unit to gather and process the information of all cells and users. However, there are no central controllers in LTE network. Hence, the implementation of LB algorithm should be distributed. In order to resolve the above multi-objective optimization problem, a distributed and practical algorithm is designed in what follows, which can be implemented in a distributed way with lower overhead.

V. Practical Algorithm

In this section, a distributed and practical algorithm is presented to solve the above multi-objective optimization problem. Since the system performance will be evaluated at each time t, we omit t in the following analysis for notational convenience.

A. *Pre-resource allocation*

Practically, users with higher QoS requirements should always be firstly and strictly guaranteed. Since the GBR users have higher QoS requirements than BE users, so we firstly allocate the resources to the GBR users to satisfy their QoS requirements. All the BE users have the same priority and better resource scheduling scheme should give attention to both efficiency and fair. That is to say, a good compromise should be traded off between maximizing the throughput and the fairness of resource allocation. For simplicity, in our practical algorithm, the residual resources are fairly allocated to the BE users [20].

The time-frequency resources needed for GBR user k in cell i is estimated as

$$w_{i,k} = \left\lceil \frac{D_k}{\bar{e}_{i,k}} \right\rceil \tag{16}$$

where D_k is the minimum data rate requirement of user k, and $\bar{e}_{i,k}$ is the average bandwidth efficiency of user k. $\lceil x \rceil$ is the minimum integer larger than x.

Then the achievable throughput of GBR user k in cell i is given by

$$R_{i,k} = w_{i,k}\bar{e}_{i,k} \tag{17}$$

The resources occupied by all GBR users in cell i is $S_i^G = \sum_{k\in G_i} I_{i,k} w_{i,k}$, and the residual resources for all BE users in cell i is $S_i^B = S - S_i^G$. Then, the resource allocated for each BE user can be written as $\frac{S_i^B}{B_i}$, and the achievable throughput of BE user b in cell i is

$$R_{i,b} = \frac{S_i^B}{B_i}\bar{e}_{i,b} \tag{18}$$

where B_i is the number of BE users served by cell i, and $\bar{e}_{i,b}$ is the average bandwidth efficiency of BE user b in cell i.

B. Handover Condition for GBR Users

The goal for GBR users is to maximize $\varepsilon(t)$ and minimize $\rho(t)$ simultaneously. Thus, the value of the aggregate objective function $\alpha\varepsilon(t)\text{-}(1\text{-}\alpha)\rho(t)$ for GBR users after handover should be larger than that before handover. Assume that GBR user k is switched from its original cell i to target cell j, the inequality $\alpha\varepsilon - (1-\alpha)\rho < \alpha\varepsilon' - (1-\alpha)\rho'$ should be met, where ε, ρ and ε', ρ' denote the load balancing index and network average load before and after handover, respectively. Combining with $\varepsilon(t)$ and $\rho(t)$ in (4) and (5), we can obtain

$$A_{i,j,k}^{GBR} = \frac{\alpha \sum_{n\in N} \rho_n^2 \left(\sum_{n\in N} \rho_n - \rho_{i,k} + \rho_{j,k} \right)^2}{\left(\alpha \left(\sum_{n\in N} \rho_n \right)^2 + (1-\alpha)\cdot\left(\rho_{j,k} - \rho_{i,k}\right)\cdot \sum_{n\in N} \rho_n^2 \right)\cdot p} > 1 \tag{19}$$

where $p = \sum_{n\in N} \rho_n^2 - 2\rho_i\rho_{i,k} + \rho_{i,k}^2 + 2\rho_j\rho_{j,k} + \rho_{j,k}^2$; $A_{i,j,k}^{GBR}$ denotes the LB gain of GBR user k switched from cell i to cell j; $\rho_{i,k} = w_{i,k}/S$ and $\rho_{j,k} = w_{j,k}/S$ are the load of user k in original cell i and target cell j, respectively. ρ_n denotes the load of cell n before handover.

Users whose LB gains are larger than 1 will have the chance to be handed over. To avoid oscillations of handover, many GBR users should not perform switching at the same time. Thus, the GBR user k^* with the largest LB gain will be selected to perform switching, i.e.

$$k^* = \arg \max_{k\in G_i, I_{i,k}=1} A_{i,j,k}^{GBR} > 1 \tag{20}$$

C. Handover Condition for BE Users

For BE user b in cell i, handover it to cell j should increase the total utility of $\psi(t)$. Let $\psi_{i,b}$ and $\psi_{j,b}$ represent the total utility before and after handover, respectively. Then it should be $\psi_{i,b} \prec \psi_{j,b}$. Together with $\psi(t)$ in (6), the $\psi_{i,b}$ and $\psi_{j,b}$ are given as:

$$\begin{aligned}\psi_{i,b} = \log_a \Bigg\{ & \left[\left(\prod_{m=1,2,\cdots,B_i} e_{i,m} \right) \cdot \left(w_{i,BE} \right)^{B_i} \right] \\ & \times \left[\left(\prod_{n=1,2,\cdots,B_j} e_{j,m} \right) \cdot \left(w_{j,BE} \right)^{B_j} \right] \Bigg\}\end{aligned} \tag{21}$$

$$\begin{aligned}\psi_{j,b} = \log_a \Bigg\{ & \left[\left(\prod_{m=1,2,\cdots,B_i, m\neq b} e_{i,m} \right) \cdot \left(w'_{i,BE} \right)^{B_i-1} \right] \\ & \times \left[\left(\prod_{n=1,2,\cdots,B_j} e_{j,m} \right) \cdot e_{j,b} \cdot \left(w'_{j,BE} \right)^{B_j+1} \right] \Bigg\}\end{aligned} \tag{22}$$

where m, n denote the BE users served by cell i and cell j, respectively. $w_{i,BE}$ ($w_{j,BE}$) and $w'_{i,BE}$ ($w'_{j,BE}$) represent the available resources for every BE user in cell i (j) before and after handover, respectively. Owing to the fair assignment of the residual resources for BE users in the cell, one can obtain

$$\begin{aligned} & w_{i,BE} = \frac{S_i^B}{B_i},\ w_{j,BE} = \frac{S_j^B}{B_j}, \\ & w'_{i,BE} = \frac{S_i^B}{B_i - 1},\ w'_{j,BE} = \frac{S_j^B}{B_j + 1} \end{aligned} \tag{23}$$

where S_i^B (S_j^B) represents the residual resources in the cell i (j) occupied by BE users.

$\psi_{i,b} < \psi_{j,b}$ together with (21) and (22) , then

$$e_{i,b} w_{i,BE} \left(\frac{B_i - 1}{B_i} \right)^{B_i - 1} < e_{j,b} w'_{j,BE} \left(\frac{B_j}{B_j + 1} \right)^{B_j} \tag{24}$$

Suppose that the number of BE users in the two cells are large enough, then, $\left((B_i - 1)/B_i\right)^{B_i - 1}$ and $\left(B_j/(B_j + 1)\right)^{B_j}$ can be approximated as e^{-1}.

Finally, we get

$$e_{i,b}w_{i,BE} < e_{j,b}w'_{j,BE} \tag{25}$$

(25) can be rewritten as:

$$\frac{e_{j,b}w'_{j,BE}}{e_{i,b}w_{i,BE}} > 1 \tag{26}$$

Similar to handover of GBR users, $A^{BE}_{i,j,b} = \left(e_{j,b}w'_{j,BE}\right)/\left(e_{i,b}w_{i,BE}\right)$ is defined as the LB gain of BE user b. Cell i only chooses the best BE user b^* that achieves the largest gain by changing its serving cell, i.e.

$$b^* = \arg \max_{k \in B_i, I_{i,k}=1} A^{BE}_{i,j,b} > 1 \tag{27}$$

D. Call Admission Control

For a new GBR user k, it will be admitted to access cell i only if there are enough resources to satisfy its QoS requirement, e.g.

$$1-\rho_i > \frac{w_{i,k}}{S} \tag{28}$$

For BE users, there is no constraint of resources for access.

E. Complexity Analysis

The algorithm executes in each eNodeB. The LB module in one cell (e.g. cell i) will be triggered when the load of the cell i exceeds a given threshold and chooses the eligible neighbor cell as the load balancing target cell (e.g. cell j). The overload cell i will pre-calculate the LB gains of users in cell i handed over to cell j. Suppose that there are K users in cell i, then the computational complexity of the algorithm is $O(K)$ (the complexity of calculating each user's LB gain is about $O(1)$). eNodeBs only need to exchange the load (or resource utilization) of the cell with each other periodically, hence its overhead is extremely low.

VI. Simulations

A. Simulation Setup

The network considered is shown in Fig.1. The distance between adjacent eNodeBs is 130 meters. The maximum transmission power of all eNodeBs is 38 dBm and the bandwidth of each eNodeB is 5 MHz, according to 3GPP in [23]. To make the simulation realistic, the simulation is carried out in dynamic environment. GBR and BE users arrive in any cell according to a Poisson process with rate λ_g and λ_b at uniformly distributed location and depart from the system after holding for a exponentially distributed period with mean 100 seconds. Assume that the minimum rate requirement of GBR users is 256 kbps.

Selection of LB period is a tradeoff between signaling overhead and performance gain of the algorithm (the shorter the period, the better the performance, but the heavier the overhead). In the following simulations, the LB cycle is 1 second and the simulation lasts for 1000 seconds. When each LB cycle begins, the network will detect the load of each cell. The MLB will be activated when the resource utilization of its cell reaches up to 85%. The neighbor cell whose load is lower than 60% will be selected to balance the load.

B. Simulation results

(1) Since the optimal value of α is difficult to derive, the proper value should be selected through simulation. Firstly, we evaluate the influence of α on network performance of the proposed LB algorithm (PLB) in a certain scenario(the arrival rates of both GBR and BE users in cell 1 are set as 0.8 users/second to make it the busy one in the whole network, while that of other cells are 0.4 users/second).

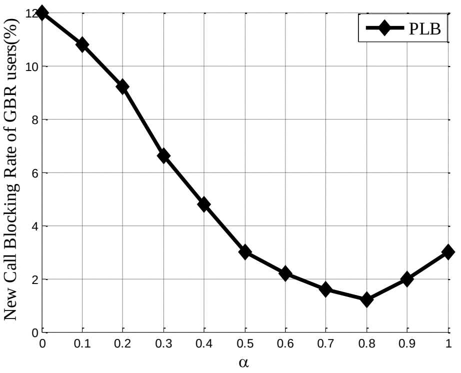

Fig. 2. New call blocking rate of GBR users with various α

The new call blocking rate of the networks varying with different α is shown in Fig. 2. It is found that the new call blocking rate of PLB is decreasing monotonously with α until $\alpha = 0.8$. It is reasonable that the larger α, the greater the weight of the load balancing index function, the more GBR users to be handed over for LB, and the lower the new call blocking rate. From Fig. 2 it can be seen that when $\alpha > 0.8$, the new call blocking rate of PLB is increasing monotonously with α. When $\alpha > 0.8$, the network average load is almost not considered for LB, and the result is that users with poor channel conditions may be handed over and occupy considerable resources in the target cells leading to higher new call blocking rate.

As shown in Fig. 3, the load balancing index varies with different α. It can be seen that the load balancing index increases monotonously with the increase of α. It is reasonable that the larger α, the greater the weight of the load balancing index function, the more GBR users to be handed over for LB, and the more balanced load distribution of the network.

From Fig. 2 and Fig. 3, we can come to conclusion that the larger load balancing index does not always bring lower new call blocking rate. It can be found that when $\alpha = 0.8$, the network can achieve better performance (the load balancing index is better, and the new call blocking rate is the lowest). Thus, in the following simulation, the value of α is selected as 0.8.

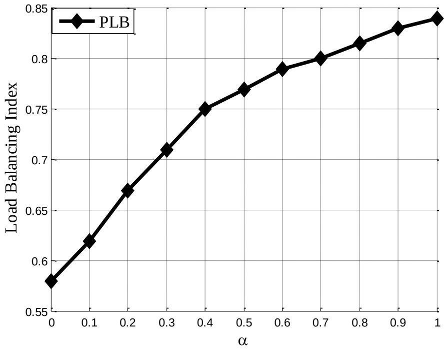

Fig. 3. Load balancing index with various α

(2) Simulations are made to evaluate the performance of the proposed algorithm varying with the load of the cell1, and take the algorithm in [15] for comparisons. To make the difference in the load situation of cells, cell 1 is set as the busy one with the alterable arrival rates from 0.4 users/second to 1.2 users/second stepped by 0.1users/second for both GBR and BE users, while the arrival rates of both GBR and BE users in other neighboring cells are set invariably to be 0.2 users/second ($\lambda_g = \lambda_b = 0.2$ users/second).

For expression convenience, in the following, NLB, OLB and PLB are used to represent no LB, the original LB method proposed in [15] and our proposed LB strategy, respectively.

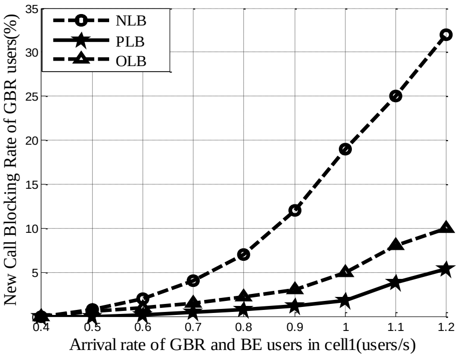

Fig. 4. New call blocking rate of GBR users with various arrival rates of cell 1

Fig. 4 shows that the new call blocking rates of NLB, PLB and OLB all increase monotonously as the arrival rate of users in cell 1. Since PLB and OLB handover users for LB so as to distribute the load of cell 1 among neighboring cells, the system could access more users and the new call blocking rates of them are less than that of NLB. Moreover, PLB achieves lower new call blocking rate than OLB. This is the performance gain of our PLB algorithm. In OLB, without considering the network average load, users with poor channel condition may be handed over and occupy much more resources in target cells than that in original cells, resulting in new call blocking rate increased.

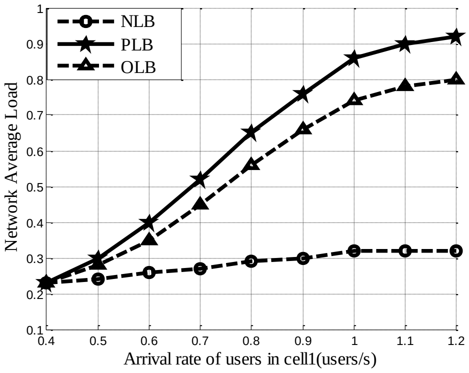

Fig. 5. Network average load with various arrival rates of cell 1

Fig. 5 shows the network average load of NLB, PLB and OLB. Actually network average load signifies the network resource utilization. Bigger arrival rate means more users to arrive and there will be more users served in the network. Thus, the network average load of all scenarios keeps increasing with the arrival rate of users in cell 1. From Fig. 5, we can see that the NLB has the lowest network average load of all the arrival rates of users. Since handover users for LB inevitably makes more users served in the network, it will bring about higher resource utilization of the network. So the network average load of PLB and OLB are larger than that of NLB. To minimize the network average load is one of the objectives, however, from Fig. 5, we could find that the network average load of the PLB is higher than the OLB. This is not conflict with the optimization objective. That is because the goal of minimizing the network average load is just for the handover decision moment, the users would be chose to handover only if it could minimize the network average load after handover, so more resources can be left for the new arrival GBR users. Therefore, the new call blocking rate of GBR users will be reduced and more GBR users can be served in the network, resulting in higher network average load.

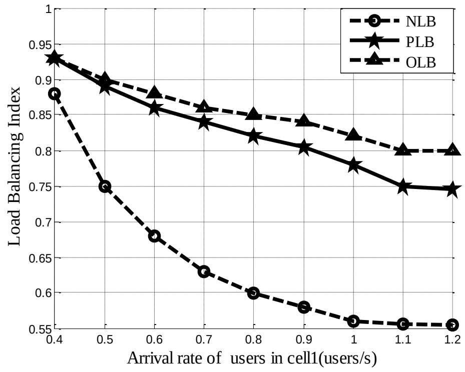

Fig. 6. Load balancing index with various arrival rates of cell 1

Fig. 6 shows that the load balancing index varies with alterable arrival rates of users in cell 1. We can find that the load balancing index of all scenarios decreases monotonously with the increase of the arrival rate of users in cell 1. Bigger arrival rate of users in cell 1 brings more unbalanced the load distribution of the network. We can find the load balancing index of NLB has the smallest under all arrival rates of users in cell 1. And the load balancing index of PLB is a little lower than that of OLB. The reason is that compared with OLB, PLB performs LB with considering the network average load. And the aggressive handovers can be hindered from switching, which could have positive effects on the load balancing index.

VII. Conclusion

In this paper, LB problem in LTE network with different QoS requirements is investigated. The LB is formulated as a multi-objective optimization problem. The objectives are to balance the load distribution among neighboring cells as much as possible and increase the network average load as few as possible for QoS requirements services, while maximizing the network utility of best effort services. Furthermore, a detailed solution algorithm is proposed. After that, the influence of α on network performance and the performance of the proposed scheme are evaluated. Simulation results show that the value of α has great influence on network performance. With specific α and different arrival rates, it also can be found that the proposed scheme can achieve better performance of the network's KPIs than conventional methods.

References

[1] H. Hu, J. Zhang, X. Zheng, Y. Yang, and P. Wu, "Self configuration and self-optimization for lte networks," *Communications Magazine, IEEE*, vol. 48, no. 2, pp. 94 –100, February 2010.

[2] 3GPP TS 36.300, E-utra and e-utran Overall Description, 2009.

[3] O. K. Tonguz and E. Yanmaz, "The mathematical theory of dynamic load balancing in cellular networks," *IEEE Trans on Mobile Computing*, vol. 7, 2008.

[4] S. K. Das, S. K. Sen, and R. Jayaram, "A dynamic load balancing strategy for channel assignment using selective borrowing in cellular mobile environment," *Wireless Netw.*, vol. 3, no. 5, pp. 333–347, Oct. 1997.

[5] K. Ali, H. Hassanein, and H. Mouftah, "Directional cell breathing based reactive congestion control in wcdma cellular networks," in *Proc. 12th IEEE Symposium on Computers and Communications*, July 2007, pp. 685–690.

[6] Y. Bejerano and S. J. Han, "Cell breathing techniques for load balancing in wireless lans, mobile computing," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 735–749, June 2009.

[7] T. Jansen, I. Balan, J. Turk, I. Moerman, and T. Ku andrner, "Handover parameter optimization in lte self-organizing networks," in *Proc. 72nd Vehicular Technology Conference Fall*, Sept. 2010, pp. 1–5.

[8] R. Kwan, R. Arnott, R. Paterson, R. Trivisonno, and M. Kubota, "On mobility load balancing for lte systems," in *Proc. 72nd Vehicular Technology Conference Fall*, Sept. 2010, pp. 1–5.

[9] A. Awada, B. Wegmann, I. Viering, and A. Klein, "A game-theoretic approach to load balancing in cellular radio networks," in *Proc. 21st International Symposium on Personal Indoor and Mobile Radio Communications*, Sept. 2010, pp. 1184–1189.

[10] R. Nasri and Z. Altman, "Handover adaptation for dynamic load balancing in 3gpp long term evolution systems," in *Proc. 5th International Conference on Advances in Mobile Computing*, 2007, pp. 145–154.

[11] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, 2006.

[12] Y. Bejerano, S. J. Han, and L. Li, "Fairness and load balancing in wireless LANs using association control," *IEEE/ACM Trans. Netw.*, vol. 15, no. 3, pp. 560–573, June 2007.

[13] K. Son, S. Chong, and G. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3566–3576, July 2009.

[14] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. Netw.*, no. 99, pp. 1–14, June 2011.

[15] H. Wang, L. Ding, and P. Wu, "QoS-aware load balancing in 3GPP long term evolution multi-cell networks," in *Proc. IEEE International Conference on Communications*, 2011, pp. 1-5.

[16] H. Wang, Z. H. Li, Z. W. Pan, X. H. You, and P. Wu, "QoS guaranteed dynamic load balancing algorithm in 3GPP LTE networks," *Science Chian Information Sciences*, vol. 42, no. 6, 2012.

[17] 3GPP TS 36.201 V9.1.0, Lte physical layer: General Description, 2010.

[18] D. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, no. 1, pp. 1-14, 1989.

[19] C. Mung, H. Steven, and A. R. Calderbank, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255-312. 2007.

[20] Q. Y. Ye, B. Y. Rong, and Y. D. Chen, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, 2013.

[21] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and Multidisciplinary Optimization*, vol. 26, no. 6, 2004.

[22] I. Das and J. E. Dennis, "A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems," *Structural Optimization*, vol. 14, 1997.

[23] 3GPP TR 25.814, Physical Layer Aspects for E-utra, 2006.

Miaona Huang was born in Guangdong Province, China, in 1983. She received her B.S. degree in Electrical information Science and Technology from Hanshan Normal University in 2006 and received her M.S. degree in Optics from South China Normal University in 2009. She is currently a Ph.D. candidate of the School of Electronic and Information Engineering at South China University of Technology. Her current research interests including the load balancing technology in multi-cell networks, handover control and call admission control in wireless networks.

Suili Feng was born in Guangdong Province, China, in 1955. He is Professor of the School of Electronic and Information Engineering at South China University of Technology. He received his B.S. degree in electrical engineering from South China Institute of Technology in 1982 and his Master and Ph.D. degree in electronic and communication system from South China University of Technology in 1989 and 1998, respectively. He was a research assistant in Hong Kong Polytechnic University during 1991–1992 and a visiting scholar in University of South Florida during 1998–1999. His research interests include wireless networks, computer networks and communication signal processing.

Jun Chen was born in Hubei Province, China, in 1981. He received his B.S. degree in Electronic and automation from Civil Aviation University of China in 2004 and received his M.S. degree in Circuit and System from South China University of Technology in 2010. He is currently a Ph.D. candidate of the School of Electronic and Information Engineering at South China University of Technology. His research interests include distributed antenna system, convex optimization and signal processing in wireless communications.