This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2011 proceedings

QoS-Aware Load Balancing in 3GPP Long Term Evolution Multi-Cell Networks

Hao Wang[1,2], Lianghui Ding[2], Ping Wu[2], Zhiwen Pan[1], Nan Liu[1], Xiaohu You[1]
[1]National Mobile Communication Research Laboratory, Southeast University, Nanjing, China
[2]Signals and Systems, Department of Engineering Sciences, Uppsala University, Uppsala, Sweden
hao_wang@seu.edu.cn, {lhding, ping.wu}@angstrom.uu.se, {pzw, nanliu, xhyu}@seu.edu.cn

***Abstract*—In this paper, we investigate load balancing problem in 3GPP Long Term Evolution (LTE) network. Since LTE network aims to serve heterogeneous users with different Quality of Service (QoS) requirements, the influence of load imbalance is quite different. For those users with minimum rate requirements, it may result in high block probability, while for others without minimum rate requirements, the throughput of boundary users may be degraded. In this paper, we take all the differences into account and formulate the problem as a multi-objective optimization problem. Then we analyze its complexity, and propose our solution framework, which includes QoS-guaranteed hybrid scheduling, QoS-aware handover for users with and without QoS requirements, and call admission control. Extensive simulations are conducted and the results show that the proposed framework leads to significantly better load balancing, and thus the decrease in call block probability of users with QoS requirements, and the increase in throughput of boundary best effort users.**

***Index Terms*—3GPP LTE, load balancing, Quality of Service (QoS)**

I. Introduction

3GPP LTE network has achieved high spectrum efficiency due to Multi-Input and Multi-Output (MIMO) antenna and Orthogonal Frequency Division Multiple (OFDM) access technology. However, the network performance is still influenced by load imbalance among neighboring cells. There has been a lot of researches to deal with the problem in cellular networks, such as "channel borrowing" [1] or "call transfer" [2] in GSM-like circuit-switched networks and dynamical association between users and cells according to different metrics in LTE-like packet-switched networks [3]–[5]. But most previous research only consider users without any QoS requirements.

In this paper, we deal with load balancing problem in 3GPP LTE multi-cell network from a new perspective, by distinguishing different QoS requirements. We first formulate this problem as a multi-objective optimization problem, then analyze its complexity, and propose a solution framework, which includes QoS-guaranteed hybrid scheduling, QoS-aware handover for users with and without QoS requirements, and call admission control. The hybrid scheduling is to strictly guarantee serving users' QoS requirements, while the network utility is proportionally maximized simultaneously. Two kinds of handovers are introduced to balance the load among cells and to increase the throughput of boundary users. And call admission control is to guarantee new access users' QoS requirements. Extensive simulations are then conducted to evaluate the performance of our algorithm.

The remainder of the paper is organized as follows. In Section II, we present the network model. In Section III, we formulate the problem as a multi-objective optimization problem, analyze its property and complexity, and then propose a solution framework in Section IV. Simulation results are given in Section V and the whole paper is concluded in Section VI.

This work is supported in part by VINNOVA Sweden (Grant 2008-00954); International Science and Technology Cooperation Program under grant 2008DFA12090; National Communications Research Laboratory Program (2010A02, 2011A02); National Special Key Program (2011ZX03003-002-02) and Huawei Corp. Ltd.

II. System Model

A. Network Model

A 3GPP LTE downlink multi-cell network serving users with heterogenous QoS requirements is considered here. In our model, there are two kinds of users with different QoS requirements, i.e., Constant Bit Rate (CBR) user with minimum rate requirement and Best Effort (BE) user with no QoS requirement. As shown in Figure 1, there are seven cells, each of which is controlled by a central eNodeB. Throughout this paper, cell and eNodeB are used interchangeably. Twelve adjacent OFDM subcarriers are grouped into one physical resource block (PRB), which is the smallest unit that can be allocated to each user in one subframe (1ms). There exists L PRBs in each cell.

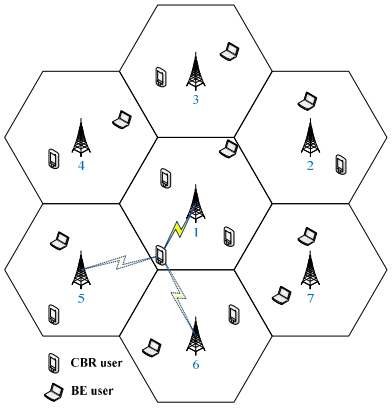

Fig. 1. Network model with heterogenous user.

$\mathbf{N}$, $\mathbf{K}$, $\mathbf{C}$ and $\mathbf{B}$ are used to denote the sets of cells, total users, CBR users and BE users, respectively. And it is obvious

that $\mathbf{K} = \mathbf{C} \cup \mathbf{B}$. An assignment indicator variable $I_{i,k}(t)$ is defined, which equals to 1 when user k is served by cell i at time t, and 0 otherwise. All *time* t mentioned here represents the time for load balancing. The span between t and $t+1$ is a load balancing cycle, which is much larger than a subframe.

B. Link Model

We assume that each user knows the instantaneous signal strength from its neighboring cells through pilot detection. And the channel status information is sent back to its serving cell within uplink data transmission or by periodical report.

The instantaneous Signal-to-Interference-and-Noise-Ratio (SINR) for user $k \in \mathbf{K}$ received on PRB l from cell $i \in \mathbf{N}$ at a subframe τ is

$$SINR_{i,l,k}(\tau) = \frac{g_{i,l,k}(\tau)p_{i,l}(\tau)}{N + \sum\limits_{j \in \mathbf{N}, j \neq i} g_{j,l,k}(\tau)p_{j,l}(\tau)} \quad (1)$$

where N is the power of Additive White Gaussian Noise (AWGN) on a PRB, $g_{i,l,k}(\tau)$ and $p_{i,l}(\tau)$ represent the instantaneous channel gain between eNodeB i and user k and the transmit power of eNodeB i on PRB l at subframe τ, respectively, and thus $g_{i,l,k}(\tau)p_{i,l}(\tau)$ is the signal strength received by user k from cell i on PRB l at τ.

Then, the average bandwidth efficiency $e_{i,k}(t)$ of user k from cell i during the time period $[t-1, t)$, is computed in the following manner

$$e_{i,k}(t) = \frac{1}{L}\sum_{l=1}^{L}\int_{t-1}^{t} log_2(1 + SINR_{i,l,k}(\tau))d\tau \text{[bps/Hz]} \quad (2)$$

For user k, the allocated resources depend on its QoS requirement and channel condition. Let $w_{i,k}(t)$ denote the time-frequency resources allocated to user k by eNodeB i at time t, then its achievable rate at time t is $R_{i,k}(t) = w_{i,k}(t)e_{i,k}(t)$, assuming that adaptive modulation and coding is used to achieve the Shannon rate limit.

C. Load Balance Index for CBR Users

$s_i(t)$, $s_i^c(t)$ and $s_i^b(t)$ are used to represent the total resources, that occupied by CBR users and that occupied by BE users at time t, respectively. Then the load of cell i is

$$\rho_i(t) = \frac{s_i^c(t)}{s_i(t)} = \frac{\sum_{k \in \mathbf{C}} I_{i,k}(t)w_{i,k}(t)}{s_i(t)} \quad (3)$$

In a multi-cell network, all cells often have the same amount of time-frequency resources. Thus we use s instead of $s_i(t)$ for simplicity. To measure the status of load balance of the entire network, Jain's fairness index [6] is used as follows

$$\xi(t) = \frac{(\sum \rho_i(t))^2}{|\mathbf{N}|\sum(\rho_i(t))^2} \quad (4)$$

where $|\mathbf{N}|$ is the number of cells. The value of load balance index is between $[\frac{1}{|\mathbf{N}|}, 1]$. A large ξ means a more balanced load distribution among cells. The objective of load balancing for CBR users is to maximize $\xi(t)$ at each time t.

D. Network Utility for BE Users

Let $R_{i,m}(t)$ denote the throughput of BE user m from cell i at time t, and $U_m(R_{i,m}(t))$ as the utility function of user m, then the total utility of BE users in the network at time t is

$$\Psi(t) = \sum_{i \in \mathbf{N}}\sum_{m \in \mathbf{B}} U_m(I_{i,m}(t)R_{i,m}(t)) \quad (5)$$

The objective of load balancing for BE users is to maximize $\Psi(t)$ at each time t.

III. Problem Formulation and Decomposition

In this section, we present the optimization problem of the above network. At each time t, we try to maximize $\xi(t)$ and $\Psi(t)$ simultaneously. Since both of them are determined by the assignment between cells and users, the problem is thus equivalent to the following multi-objective optimization problem with QoS and resource constraints.

$$\max\ [\xi(t),\ \Psi(t)]^T \quad (6)$$

$$s.t.\quad \sum_{k \in \mathbf{K}} I_{i,k}(t)w_{i,k}(t) \leq s,\ \forall i \in \mathbf{N}, \quad (7)$$

$$\sum_{i \in \mathbf{N}} I_{i,k}(t) = 1,\ \forall k \in \mathbf{K}, \quad (8)$$

$$\sum_{i \in \mathbf{N}} I_{i,k}(t)R_{i,k}(t) \geq \theta_k,\ \forall k \in \mathbf{C}, \quad (9)$$

Constraints in (7) present that the occupied resources of a cell by all users in it could not exceed the total resource limit. Constraints in (8) tell that one user can only be served by one cell at a certain time t. Constraints in (9) explain that the minimum rate requirement θ_k of any CBR user k has to be satisfied strictly.

For a multi-objective optimization problem, the most intuitive approach is to sum all the linear weighted objectives. Since the two objectives have different order of magnitude, it is hard to design the weights and evaluate their influence on network performance. And for recent popular approaches which base on particle swarm optimization or simulated annealing [7], [8], a central controlling unit is necessary to collect the QoS and SINR information of all users. Besides, it is time consuming to get the set of Pareto optimal points.

Since 3GPP LTE network has a flat network structure without a centralized controlling unit, the handover decisions have to be made by each eNodeB individually in a prompt response to the varying network conditions. Besides, the overhead of user status information exchange for decision making at each eNodeB should be minimized. Thus the load balance solution should be distributed, realtime and low-overhead. In the following, we will give a heuristic but practical algorithm which satisfies all the above requirements and its performance is justified through extensive simulation.

IV. Practical Algorithm

In this section, we propose a framework to solve the above multi-objective optimization problem as follows. For convenience, symbol t is omitted at each load balancing cycle in the following analysis.

A. QoS-Guaranteed Hybrid Scheduling

In practice, users with higher QoS requirements often have to be guaranteed strictly and firstly. Thus, in our network model, we first allocate resources to satisfy the rate requirements of CBR users, and then schedule residual resources for BE users to maximize the network utility.

Since multi-user diversity among users with QoS requirements is still an open issue, we do not consider it here and assume the amount of resources allocated to a user with QoS requirement is only determined by its rate demand and the average bandwidth efficiency. For a CBR user k in cell i, the time-frequency resource allocation is written as

$$w_{i,k} = \lceil \frac{\theta_k}{e_{i,k}} \rceil \tag{10}$$

where θ_k is the rate requirement of user k, and $e_{i,k}$ is the average bandwidth efficiency of user k. $\lceil x \rceil$ is the minimum integer larger than x. Then the resources occupied by all CBR users in cell i is $s_i^c = \sum_{k \in \mathbf{C}} I_{i,k} w_{i,k}$, and residual resources for all BE users in cell i is $s_i^b = s - s_i^c$.

For BE users, the well-known proportional fair scheduling scheme is utilized in which all users have the same log utility function $U(\cdot) = log(\cdot)$. Then the achievable throughput of BE user m in cell i is

$$R_{i,m} = s_i^b e_{i,m} \frac{G(Y_i^b)}{Y_i^b} \tag{11}$$

where Y_i^b is the number of BE users served by cell i; $G(\cdot)$ is the multi-user diversity gain as that in [9].

B. Handover Condition for CBR Users

For CBR user k in cell i, switching it to cell j should increase load balance index ξ. Let $\xi_{i,k}$ and $\xi_{j,k}$ represent the load balance index before and after the switching (handover), then there should be $\xi_{i,k} < \xi_{j,k}$. Assuming the numerator of $\xi_{i,k}$ and $\xi_{j,k}$ are the same, that is reasonable because boundary users which consume almost equal resource in source and target cells are preferred to perform load balancing handover, then $\xi_{i,k} < \xi_{j,k}$ together with (4) yields

$$\begin{aligned} \rho_i^2 + \rho_j^2 &> (\rho_i - \frac{w_{i,k}}{s})^2 + (\rho_j + \frac{w_{j,k}}{s})^2 \\ \Rightarrow \frac{w_{i,k}(2s_i^c - w_{i,k})}{w_{j,k}(2s_j^c + w_{j,k})} &> 1 \end{aligned} \tag{12}$$

We define $\psi_{i,j,k}^c = w_{i,k}(2s_i^c - w_{i,k})/w_{j,k}(2s_j^c + w_{j,k})$ as the CBR user load balancing gain for switching CBR user k from cell i to j. It should be larger than 1, or the handover is worthless. If many CBR users change their serving cells at the same time, this may result in oscillations of handover, thus cell i only chooses the best CBR user k^* that achieves the largest benefit by changing its serving cell, where

$$k^* = \arg \max_{k \in \mathbf{C},\ I_{i,k}=1} \psi_{i,j,k}^c \tag{13}$$

A threshold $\psi^{cbr} > 1$ is introduced to control the number of handover for load balancing of CBR users and reduce possible ping-pong effect. Only if $\psi_{i,j,k^*}^c > \psi^{cbr}$, user k^* is switched from cells i to j.

C. Handover Condition for BE Users

For BE user m in cell i, changing its serving cell from i to j will not significantly affect the total throughput in the two cells if the number of BE users in the two cells are large enough and the load balancing gain of BE user only depends on the throughput increment of itself. The proof is a slight extension of [4], and is omitted here for space limitation. And the achievable throughput of user m in cell i or j is

$$R_{i(j),m} = s_{i(j)}^b e_{i(j),m} \frac{G(Y_{i(j)}^b)}{Y_{i(j)}^b} \tag{14}$$

Similar to handover of CBR users, we also define $\psi_{i,j,m}^b = R_{j,m}/R_{i,m}$ as the load balancing gain of BE user m. Cell i only chooses the best BE user m^* that achieves the largest gain by changing its serving cell, i.e.,

$$m^* = \arg \max_{m \in \mathbf{B},\ I_{i,m}=1} \psi_{i,j,m}^b \tag{15}$$

We also introduce a threshold $\psi^{be} > 1$ to control the number of handover for load balancing of BE users and reduce possible ping-pong effect.

D. Call Admission Control

For a new CBR user k, it will be admitted to access cell i only if there are enough time-frequency resources to satisfy its QoS requirement, that is:

$$s - s_i^c > w_{i,k} \tag{16}$$

And for all BE users, there is no constraint for access.

V. Simulations

In this section, we first evaluate the influence of ψ^{cbr} on the performance in terms of block probability of CBR users, the 5th percentile throughput of BE users and total throughput of BE users in a certain scenario with fixed arrival rates. Then we give the performance variance according to the load of the busy cell. Since the performance gain of BE users' load balancing is similar to that in [4], we do not give the evaluation results on ψ^{be}, and just select it as 1.5 according to our simulation results.

A. Simulation Setup

The network considered here is composed of 7 hexagonal micro cells with heterogenous users as shown in Figure 1. The distance between neighboring eNodeBs is 130 meters. The maximum transmission power of all eNodeBs is 38 dBm and the bandwidth is 10 MHz, which are consistent with the simulation scenario recommended by 3GPP in [10]. Wrap-around technique is used here to avoid border effects. CBR and BE users arrive in any cell i according to a poisson process with rate λ_i^c and λ_i^b at uniformly distributed locations and depart from the system after holding for a exponentially distributed period with mean 100 seconds. We assume that the rate demands of all CBR users are 250 kbps. To differentiate the load of neighboring cells, cell 1 is set as the busy one with same alterable arrival rates for both CBR and BE users, while

that of both CBR and BE users in other neighboring cells are assumed to be 0.2 ($\lambda^c = \lambda^b = 0.2$ user/second).

Selection of load balancing cycle is a tradeoff between signaling overhead and the performance gain of the algorithm (the shorter the period, the better the performance, but the heavier the overhead). Since the marginal utility of the performance gain decreases very fast as the scale-down of the load balancing cycle, 1 second is taken as an example.

B. Simulation Results

For expression convenience, in the following, *N/A*, *CBR LB* and *CBR+BE LB* are used to represent no load balancing, load balancing only among CBR users and load balancing among both CBR and BE users, respectively.

1) Influence of ψ^{cbr} with fixed user arrival rates: In this subsection, we evaluate the influence of handover threshold of CBR users on the performance of our algorithm. The arrival rates of both CBR and BE users in cell 1 are set as 0.6 user/sec to make it the busy one in the whole network.

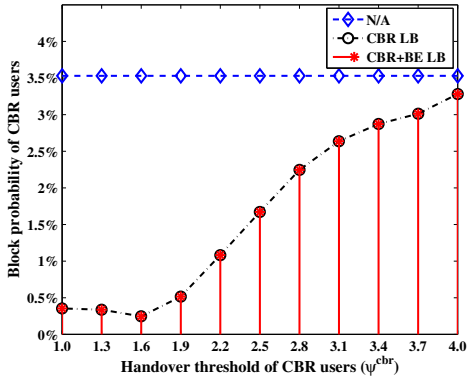

Fig. 2. Call block probability of CBR users with various handover thresholds of CBR users (ψ^{cbr}).

The variance of block probability of CBR users with different ψ^{cbr} is shown in Figure 2. We can find that the block probabilities of both *CBR LB* and *CBR+BE LB* increase monotonously along with the load balancing threshold ψ^{cbr} of CBR users. Since the value of ψ^{cbr} determines the proportion of CBR users to do handover for load balancing, it is reasonable that the larger the ψ^{cbr}, the fewer the CBR users to do handover for load balancing, and the higher the block probability. It also can be found that the curves of *CBR+BE LB* and *CBR LB* are overlapped with each other, which indicates that *CBR+BE LB* has no gain over *CBR LB* on block probability of CBR users. This result verifies that the priority of CBR users is higher than that of BE users due to our QoS-guaranteed hybrid scheduling scheme.

The 5th percentile throughput of BE users in cell 1 is shown in Figure 3. When ψ^{cbr} is low, more CBR users are allowed to do load balancing handover, and more resources are left for BE users, thus the 5th percentile throughput of BE users is also high. With the increasing of ψ^{cbr}, the number of CBR users allowed to do handover for load balancing becomes less, hence the 5th percentile throughput of BE users also decreases.

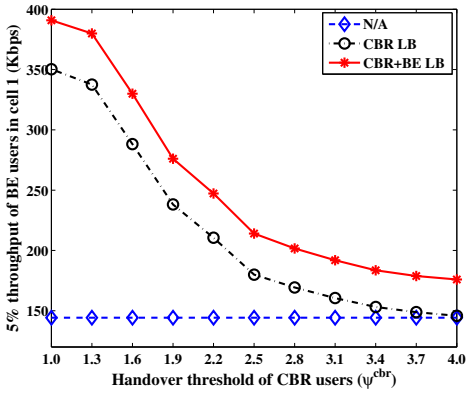

Fig. 3. 5th percentile throughput of BE users in Cell 1 with various handover thresholds of CBR users (ψ^{cbr}).

Furthermore, we can find that 5th percentile throughput with *CBR+BE LB* is larger than that with *CBR LB*, which shows that load balancing of BE users yields the throughput gain of boundary BE users by 10% to 20%.

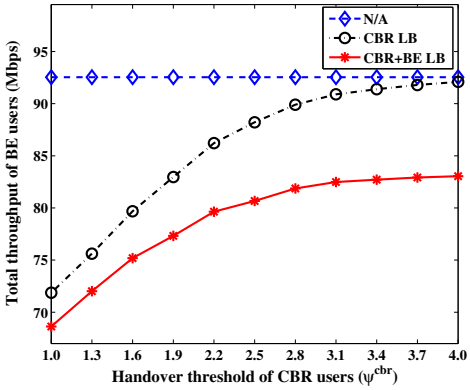

Fig. 4. Total throughput of BE users in the whole network with various handover thresholds of CBR users (ψ^{cbr}).

Total throughput of BE users in the whole network is shown in Figure 4. When ψ^{cbr} is small, there are more CBR users to do handover for load balancing which brings a more balanced load distribution. Thus the network could serve more CBR users so as to leave less resource for BE users. Along with the increase of ψ^{cbr}, the gap between the throughput with *CBR LB* and *N/A* becomes smaller. Here, the throughput with *CBR+BE LB* is smaller than that with *CBR LB*, which shows that the cost of throughput gain of boundary users in Figure 3 is the $4.5\% \sim 9.8\%$ deterioration of total throughput. Note that the results are reasonable, because handover of BE users from a busy cell to a relatively idle one often increases its throughput with the cost of lower spectrum efficiency. This phenomenon is consistent with the results presented in [4] without QoS consideration.

2) Performance variance with different arrival rates: In this subsection, we evaluate the performance of our algorithm with fixed handover thresholds and different arrival rates of users in Cell 1. Both ψ^{cbr} are ψ^{be} are set to be 1.5 here.

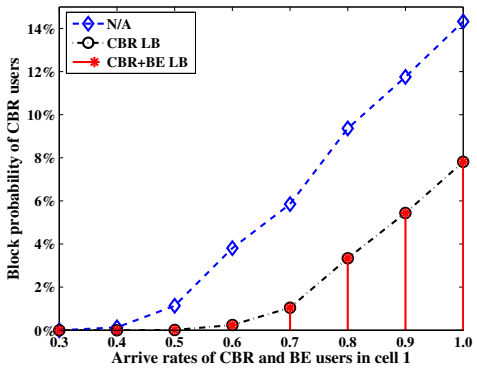

Fig. 5. Block probability of CBR users with various arrival rates of Cell 1.

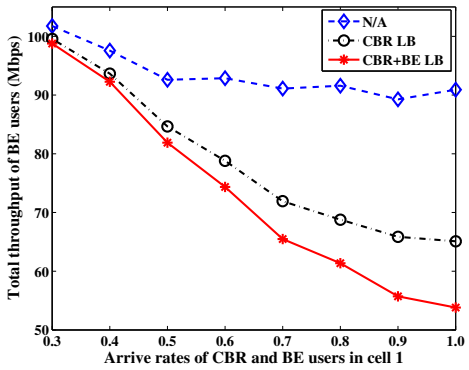

Fig. 7. Total throughput of BE users with various arrival rates of Cell 1.

The block probability of CBR users is shown in Figure 5. The block probabilities in all scenarios increase as the arrival rates. Through utilizing our load balancing algorithm, the block probability of CBR users is decreased by about 72% in average, and up to 100% in some scenarios. As explained in Figure 2, *CBR+BE LB* has no gain over *CBR LB* on block probability.

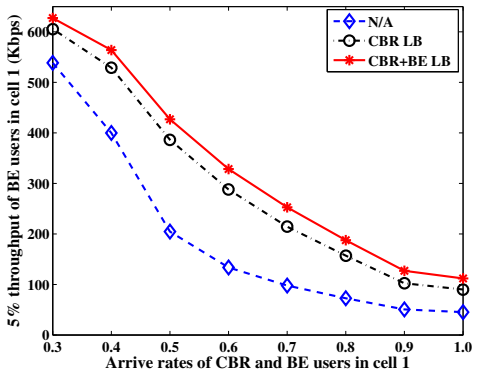

Fig. 6. 5th percentile throughput of BE users in Cell 1 with various arrival rates of Cell 1.

The 5th percentile throughput of BE users in Cell 1 in all scenarios is shown in Figure 6, which decreases as the increase of arrival rates. The average 5th percentile throughput with *CBR LB* and *CBR+BE LB* is larger than that with no load balancing by 53.8% and 70.2%, respectively. Furthermore, the average 5th percentile throughput with *CBR+BE LB* is larger than that with *CBR LB* by about 9.6%.

The total throughput of BE users under different arrival rates is shown in Figure 7. As the increase of arrival rates, the total throughput decreases due to more resources are occupied by more CBR users and less resources are left for BE users. The gap between *CBR LB* and *N/A* also increases because a higher arrival rate of CBR users bring a larger probability for CBR users to do handover for load balancing, thus less resources are left for BE users. The average total throughput with *CBR LB* is 15.9% less than that with no load balancing. And the average 7.1% total throughput deterioration in *CBR+BE LB* comparing with that in *CBR LB* is the cost of throughput gain of boundary users in Figure 6.

VI. Conclusion

In this paper, we have dealt with load balancing problem in LTE network with different QoS requirements taken into account. We first formulated it as a multi-objective optimization problem. Then we analyzed the complexity of the problem and proposed a heuristic but practical framework to solve it in a distributed manner, which includes QoS-guaranteed hybrid scheduling, QoS-aware handover of users with different QoS requirements, and call admission control. After that, we evaluated the influence of handover thresholds on network performance, and the performance variance according to different arrival rates. Simulation results show that the handover threshold of CBR users has a significant influence on network performance. With specific handover thresholds and different arrival rates, we also found that the load balancing framework proposed in this paper can decrease the block probability of CBR users and increase the throughput of boundary BE users in a busy cell with only a bit degradation of total throughput.

References

[1] H. Jiang, and S. S. Rappaport, "CBWL: A new channel assignment and sharing method for cellular communication systems," *IEEE Trans. Veh. Technol.*, vol. 43, no. 4, pp. 313–322, May 1994.

[2] S. K. Das, and S. K. Sen, and R. Jayaram, "A novel load balancing scheme for the tele-traffic hot spot problem in cellular networks," *Wirel. Netw.*, vol. 4, no. 4, pp. 325–340, Jul. 1998.

[3] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handover/cell-site selection, and scheduling in multi-cell packet data systems," *Wirel. Netw.*, vol. 14, no. 1, pp. 103–120, 2008.

[4] 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, Jul. 2009.

[5] H. Wang, L. Ding, P. Wu, Z. Pan, N. Liu, X. You, "Dynamic load balancing and throughput optimization in 3gpp lte networks," in *Proceeding of IWCMC'10*, June 2010, pp. 939–943.

[6] R. Jain, D. Chiu and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared systems," Digital Equipment Corp., DEC-TR-301, Tech. Rep., 1984.

[7] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[8] Carlos A. Coello Coello, Gregorio Toscano Pulido, and Maximino Salazar Lechuga, "Handling multiple objectives with particle swarm optimization," *IEEE Trans. Evol. Comput.*, pp. 256–279, Jun. 2004.

[9] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, Jul. 2004.

[10] 3GPP TR 25.814 V7.1.0 (2006-09), "Physical layer aspects for eutra."