Dynamic Load Balancing in 3GPP LTE Multi-Cell Networks with Heterogenous Services

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, Dept. of Engineering Sciences, Uppsala University, Uppsala, Sweden
{hao.wang, lhding, ping.wu}@angstrom.uu.se, {pzw, nanliu, xhyu}@seu.edu.cn

***Abstract*—Load balancing among multi-cells in 3GPP Long Term Evolution (LTE) networks with heterogeneous services is investigated. It is formulated as a multi-objective optimization problem, the objectives of which are load balancing index of services with QoS requirements and network utility of other services. The constraints are physical resource limits and QoS demands. Then the property and complexity of the problem are analyzed, and sequential optimization method is employed to solve it. After that, a practical algorithm for load balancing is developed which includes QoS-guaranteed hybrid scheduling, handover of users with and without QoS requirements, and call admission control. Simulation is made extensively and the results show that the proposed load balancing algorithm can significantly enhance the performance of LTE networks with heterogeneous services, decreasing call block probability of users with QoS requirements, and increasing throughput of boundary users with only a bit degradation of total throughput.**

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

I. Introduction

3GPP LTE is a promising candidate for next generation wireless networks. But like GSM and WCDMA, it still has the problem of load unbalance. Much research has been done to deal with the load unbalance problem in LTE-liked packet-switched network [1]–[5]. Most of them use proportional fairness (PF) as the scheduling metric among competing users, and do not consider QoS requirements. However, the networks in reality have different QoS requirements. Hence people have proposed weighted PF scheduling schemes to include the influence of QoS requirements [6], [7], wherein different types of services are differentiated with weights. It should be noted that the weighting method cannot strictly guarantee users' QoS requirements.

This paper is concerned with the dynamic load balancing problem in 3GPP LTE multi-cell networks with heterogenous QoS requirements, and organized as follows. In Section II, we present the network model. In Section III, we formulate the problem to be a multi-objective optimization problem, and then analyze its property and complexity and propose a solution framework in Section IV, which includes QoS guaranteed hybrid scheduling, handover of both users with and without QoS requirements, and call admission control. Simulation results are given in Section V and the whole paper is concluded in Section VI.

This work is supported by VINNOVA (Grant 200800954), Sweden; International Science and Technology Cooperation Program (Grant 2008DFA12090) and National Communication Research Laboratory Program (2009A02), China.

II. System Model

A. Network Model

A 3GPP LTE downlink multi-cell network serving users with heterogenous QoS requirements is considered here. Specifically, two kinds of QoS requirements, Constant Bit Rate (CBR) and Best Effort (BE) services, are taken into account. Other QoS requirements can, however, be incorporated easily. In the following, users with CBR and BE services are called simply CBR and BE users. The scenario considered here is shown in Fig. 1, where there are seven cells, each of which is associated with an eNodeB. Twelve adjacent OFDM subcarriers are grouped into a physical resource block (PRB), which is the smallest unit that can be allocated to a user in one subframe [8]. The sets of cells, total users, CBR users and BE users are assumed to be $\mathbf{N}$, $\mathbf{K}$, $\mathbf{C}$ and $\mathbf{B}$, respectively. It is easily to see $\mathbf{K} = \mathbf{C} \cup \mathbf{B}$. An assignment indicator variable is denoted $I_{i,k}(t)$, which equals 1 when user k is served by cell i at time t, and 0 otherwise. Time t used throughout this paper represents a time for load balancing and all variables changed at time t will take effect in the next load balancing cycle, which is the span between time t and $t+1$ and is much large than a subframe (1ms).

B. Link Model

For link model, we assume that each user knows the instantaneous signal strengths from its neighboring cells through pilot detection. Channel status information is sent back to its serving eNodeB through data transfer or by periodical report.

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

$$SINR_{i,k}(\tau) = \frac{g_{i,k}(\tau)p_i(\tau)}{N + \sum\limits_{j \in \mathbf{N}, j \neq i} g_{j,k}(\tau)p_j(\tau)} \qquad (1)$$

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

Since load balancing is periodically done on a lager time scale than a subframe, we use $E[SINR_{i,k}(t)]$ to represent the

expectation of instantaneous SINR between time $[t-1, t)$, thus the average bandwidth efficiency $e_{i,k}(t)$ of user k from cell i at time t is computed in the following manner

$$e_{i,k}(t) = log_2(1 + E[SINR_{i,k}(t)]) \text{ [bps/Hz]} \quad (2)$$

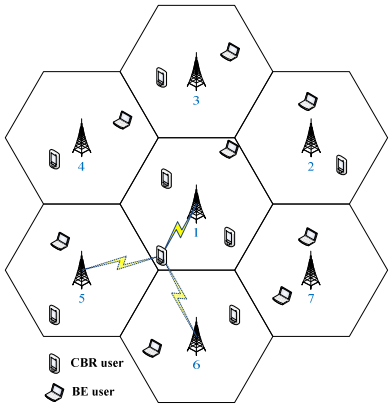

Fig. 1. Network model with heterogenous user.

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

C. Load Balance Index of CBR Users

We use $s_i(t)$ to represent the total resources, and $s_i^c(t)$ and $s_i^b(t)$ to represent the resources occupied by CBR users and BE users at time t, respectively. Then the load of cell i at time t 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 the 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, we use Jain's fairness index [9] 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 in the network, and the load balance index takes the value in the interval $[\frac{1}{|\mathbf{N}|}, 1]$. A larger ξ means a more balanced load distribution among the cells. The objective of load balancing for CBR users is to maximize $\xi(t)$ at each time t.

D. Network utility of BE Users

Let $R_{i,m}(t)$ denote the throughput of BE user m from cell i at time t, and $U_m(R_{i,m}(t))$ the utility function of user m. The network utility of all BE users at time t can be written as

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

Load balancing for BE users is aimed to maximize $\Psi(t)$ at each time t.

III. Problem Formulation and Decomposition

The purpose of load balancing, as above mentioned, is to maximize both load balance index $\xi(t)$ for CBR users and utility function $\Psi(t)$ for BE users. And load balancing is realized through enforced handover.

Then it can be formulated as the following multi-objective optimization problem with QoS and resource constraints

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

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

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

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

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

To deal with a multi-objective optimization problem, one of the feasible approaches is to construct a single Aggregate Objective Function (AOF), e.g., a linear weighted sum of the objectives. Since the objective functions may have different dimensions, it is still hard to design the weights and evaluate their influence on network performance. In practice, users with higher QoS requirements are often guaranteed first. For example, the CBR users in the present problem have higher QoS requirements than the BE users then $\xi(t)$ should be optimized first. Thus, we propose to use a sequential optimization method to deal with the above multi-objective optimization problem, i.e., optimizing the two objective functions one after the other according to the priority of QoS requirements.

Since both $\xi(t)$ and $\Psi(t)$ are determined by $I_{i,k}(t)$ $(i \in N, k \in K)$, to the best of our knowledge, there is no effective algorithm available until now to solve such a problem. If we use exhaustive search method, it requires a central controller and the computation complexity will be huge. Besides, resource occupation of each CBR user and throughput of each BE user to all cells should be sent to the controller, which accordingly leads to a large overhead.

Unlike UMTS that has radio network controller (RNC), 3GPP LTE network has a flat network structure without a central controller. Each eNodeB in the network makes handover decisions independently and promptly in response to varying network conditions. Besides, the overhead of user status information exchange for decision making at each eNodeB should be minimized.

In this case, we will design a heuristic and practical real-time algorithm which could be executed in a distributed manner with low overhead, and could solve the multi-objective problem in the sequential manner.

IV. Practical Algorithm

To solve the above multi-objective optimization problem, a framework is proposed that consists of three aspects: QoS-guaranteed hybrid scheduling, QoS-aware handover and call admission control. For convenience, we omit symbol t in the following analysis.

A. QoS-Guaranteed Hybrid Scheduling

Because CBR users have higher QoS requirements than BE ones, we first allocate resources according to the rate requirements of CBR users, and then schedule residual resources for BE users to maximize the network utility. For CBR user k in cell i, the appropriate time-frequency resource allocation is

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

where θ_k is the rate requirement of user k, and $e_{i,k}(t)$ is the average bandwidth efficiency of user k in the current load balancing cycle. $\lceil x \rceil$ represents the minimum integer larger than x. The resource allocation depending on average bandwidth efficiency is conservative because we could use opportunistic scheduling among all CBR users to achieve less resource occupation for each CBR user.

The resources occupied by CBR users s_i^c, and the residual resources for BE users s_i^b in cell i are given, respectively, by

$$s_i^c = \sum_{k \in \mathbf{C}} I_{i,k} w_{i,k} \tag{11}$$

$$s_i^b = s - s_i^c \tag{12}$$

For BE users, the proportional fair scheduling is used in which all users have the same log utility function $U(\cdot) = log(\cdot)$. Following the procedure analogous in [10], the achievable throughput for BE user m in cell i is

$$R_{i,m} = \frac{s_i^b}{Y_i^b} e_{i,m} G(Y_i^b) \tag{13}$$

where Y_i^b is the number of BE users served by cell i; $G(y) = \sum_{z=1}^{y} \frac{1}{z}$ represents the multi-user diversity gain depending only on the number of BE users [10].

B. QoS-Aware Handover

For CBR user k in cell i, switching to cell j should increase load balance index ξ. Letting $\xi_{i,k}$ and $\xi_{j,k}$ to represent the load balance index before and after the switching (handover), then there should exist $\xi_{i,k} < \xi_{j,k}$. 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 for load balancing handover, then $\xi_{i,k} < \xi_{j,k}$ together with (4) yields

$$\rho_i^2 + \rho_j^2 > (\rho_i - \frac{w_{i,k}}{s})^2 + (\rho_j + \frac{w_{j,k}}{s})^2$$

$$\Rightarrow \frac{w_{i,k}(2s_i^c - w_{i,k})}{w_{j,k}(2s_j^c + w_{j,k})} > 1 \tag{14}$$

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. If many CBR users change their serving cells at the same time, this may result in oscillation of handover. In this case cell i chooses only the best CBR user k^* that achieves the largest benefit by changing its serving cell, where

$$k^* = \arg \max_{k \in \mathbf{C},\ I_{i,k}=1} \psi_{i,j,k}^c \tag{15}$$

For BE user m in cell i, switching it to cell j should increase the network utility ψ defined in Eq. (5). The increment of ψ only depends on the utility increment of user m if the number of BE users in the two cells is large enough. The proof of this is quite similar to [4], and omitted here due to space limitation. For handover the following condition should be satisfied

$$log(R_{j,m}) > log(R_{i,m})$$

$$\Rightarrow \frac{R_{j,m}}{R_{i,m}} = \frac{\frac{s_j^b}{Y_j^b+1} e_{j,m} G(Y_j^b)}{\frac{s_i^b}{Y_i^b} e_{i,m} G(Y_i^b)} > 1 \tag{16}$$

Similarly, we define $\psi_{i,j,m}^b = R_{j,m}/R_{i,m}$ as the load balancing gain of BE users. Cell i only chooses the best BE user m^* that achieves the largest gain because of changing its serving cell, where

$$m^* = \arg \max_{m \in \mathbf{B},\ I_{i,m}=1} \psi_{i,j,m}^b \tag{17}$$

C. Call Admission Control

For a new CBR user k, it will be admitted to access cell i only if there is enough time-frequency resource available to satisfy its QoS demand, that is

$$s^{c_{max}} - s_i^c > w_{i,k} \tag{18}$$

where $s^{c_{max}}$ is the maximum of time-frequency resource that could be allocated to all the CBR users in a cell during one load balancing cycle.

For new BE users, there is no constraint on access.

V. Simulations

Simulations are made to evaluate the performance of the proposed algorithm in terms of load balance index ξ, block probability of CBR users, network utility ψ, 5th percentile throughput of BE users in the busiest cell and total throughput of BE users. The 5th percentile throughput of BE users is defined as the average of the lowest 5% throughput of BE users and usually regarded as a representative performance metric of boundary users.

A. Simulation Setup

The network considered is composed of 7 hexagonal micro cells with heterogenous users as shown in Fig. 1. The distance between neighboring eNodeBs is 130m. The maximum transmission power of all eNodeBs is 38 dBmw and the bandwidth is 10 MHz, which are consistent with the simulation scenario recommended by 3GPP in [11]. To avoid border effects, wrap-around technique is used.

In order to provide practical simulation results, we have investigated our algorithm in a dynamic setting. 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 an exponentially distributed period with mean $1/\mu = 100$ seconds. The average numbers of CBR and BE users in each cell depend on both the arrival rates and the holding time. We assume that rate demands of all CBR users are uniformly chosen from (64, 128, 256) Kbps. To differentiate the load of neighboring cells, we let Cell 1 be the busiest one with the same alterable arrival rates for CBR and BE users, while those of both CBR and BE users in other neighboring cells are assumed to be 0.2 ($\lambda^c = \lambda^b = 0.2$ user/second). Then we set $s^{c_{max}}$ to be 80% of the total time-frequency resource in one load balancing cycle.

Selection of load balancing cycle needs to consider tradeoff between signaling overhead and the performance gain of the algorithm (the shorter the period, the better the performance, while the heavier the overhead). However, the marginal utility of the performance gain decreases very fast as the scale-down of the load balancing cycle according to our simulation. Thus, a period of 1 second is used in the following simulations.

B. Simulation Results

In the simulations, we consider three cases: (1) no load balancing, (2) load balancing only among CBR users and (3) load balancing among both CBR and BE users, which are labeled with *N/A*, *CBR LB* and *CBR+BE LB*, respectively.

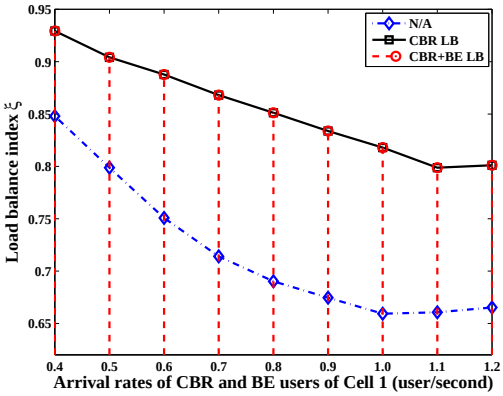

Fig. 2. Load balance index ξ with various arrival rates of Cell 1.

1) Load balance index ξ: The variance of load balance index ξ with different arrival rates is shown in Fig. 2. We can find that the load balance index ξ in all three cases decrease monotonously as the arrival rates increase. In other words, the larger the arrival rates, the more unbalanced the load distribution among cells, and the lower the load balance index ξ. That is reasonable since the value of arrival rates determines the degree of load unbalance. In addition, Fig. 2 shows that the load balance index ξ in *CBR+BE LB* and *CBR LB* is large than that in *N/A* by about 19.4% on average. This demonstrates that the proposed load balancing algorithm yields significant gain of performance. It also can be seen that the curves of *CBR+BE LB* and *CBR LB* are overlapped with each other, which indicates that *CBR+BE LB* has no advantage over *CBR LB* on load balance index ξ and load balance index ξ is only associated with load balancing among CBR users.

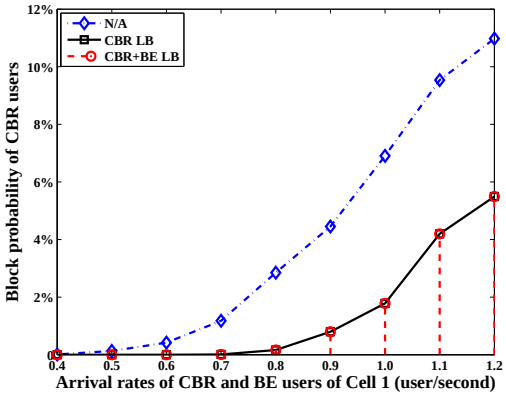

Fig. 3. Block probability of CBR users with various arrival rates of Cell 1.

2) Block probability of CBR users: The block probability of CBR users is shown in Fig. 3 and it increases with the arrival rates in all three cases. As shown in the figure utilizing the proposed load balancing algorithm leads to the decrease of the block probability of CBR users by about 71.3% on average, and up to 100% in some cases. Similar to the results in Fig. 2, *CBR+BE LB* has no advantage over *CBR LB* on block probability of CBR users.

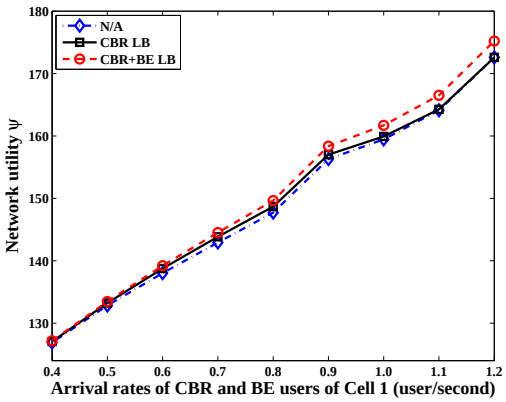

Fig. 4. Network utility ψ with various arrival rates of Cell 1.

3) Network utility ψ: The variance of network utility ψ with different arrival rates is shown in Fig. 4. It increases monotonously with the arrival rates in all the three cases. That tells that the larger the arrival rates, the more the BE users in the network, and the large the network utility ψ. That is because the value of arrival rates determines the number of BE users in the network. We can find in the figure that the network utility ψ for *CBR LB* is a bit larger than for *N/A*, which indicates that load balancing only among CBR users is good for network utility of BE users. That is reasonable since resource released in the original busy cell could bring a large utility gain for all BE users in the same cell than utility loss in the target idle cell which has less BE user and more residual resource. And network utility ψ with *CBR+BE LB*

is the largest in all of the three cases, which shows that the increment of network utility ψ mainly depends on the load balancing handover of BE users.

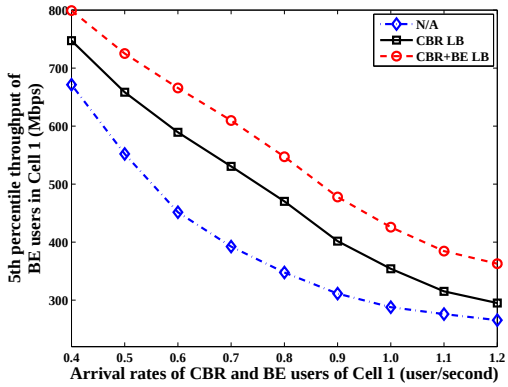

Fig. 5. 5th percentile throughput of BE users in cell 1 with various arrival rates of Cell 1.

4) 5th percentile throughput of BE users in cell 1: The 5th percentile throughput of BE users in Cell 1 is shown in Fig. 5. When arrival rates of Cell 1 are low, there are less CBR users, and more resources are left for BE users, thus the 5th percentile throughput is also high. With the increasing arrival rates, the number of CBR users becomes large and less resources are left for BE users, hence the 5th percentile throughput of BE users decreases. The average 5th percentile throughput in *CBR LB* and *CBR+BE LB* is larger than that in *N/A* by about 23.2% and 43.1% in average, respectively. Furthermore, the average 5th percentile throughput in *CBR+BE LB* is larger than that in *CBR LB* by 7.0% to 23.0%, which shows that the load balancing of BE users yields the throughput gain of boundary BE users.

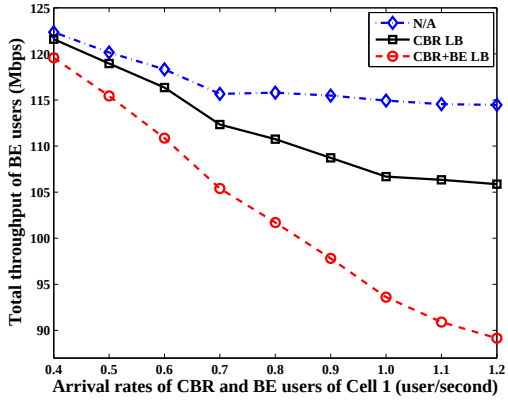

Fig. 6. Total throughput of BE users with various arrival rates of Cell 1.

5) Total throughput of BE users: The total throughput of BE users with different arrival rates is shown in Fig. 6. 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 the throughput with and without load balancing also increases because a higher arrival rates of CBR users bring a larger probability for them to do handover for load balancing, thus less resources are left for BE users. The average total throughput in *CBR LB* is 4.3% less than that with no load balancing. And the average 8.0% total throughput deterioration in *CBR+BE LB* compare with that in *CBR LB* is the cost of throughput gain of boundary users in Fig. 5.

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.

VI. Conclusion

Load balancing for LTE networks has been investigated in terms of services with different QoS requirements. The load balancing for heterogeneous QoSs was formulated as a multi-objective optimization problem. Then the property and complexity of the problem was analyzed, and a heuristic but practical algorithm proposed, which includes QoS-guaranteed hybrid scheduling, handover of users with different QoS requirements, and call admission control. The optimization problem was solved sequentially. A practical algorithm was developed. After that the performance variance according to different arrival rates was looked into via extensive simulation. The simulation results show that the load balancing framework proposed in this paper can significantly enhance the performance of LTE networks with heterogeneous quality of services, specifically decreasing the block probability of CBR users and increasing the throughput of boundary BE users in a busy cell with only a bit degradation of total throughput.

References

[1] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *IEEE Proc. INFOCOM*, 2003.

[2] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handovercell-site selection, and scheduling in multi-cell packet data systems," in *IEEE Proc. MobiCom*, Philadelphia, Pennsylvania, USA, Oct 2004, pp. 302–314.

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

[4] K. Son, S. Chong, and G. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Trans. on Wireless Communications*, 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," accepted by IWCMC 2010.

[6] S. Shakkottai and A. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in hdr," in *In Proceedings of ITC-17*, Sep. 2001, pp. 793–804.

[7] M. Lundevall, B. Olin, J. Olsson, N. Wiberg, S. Wanstedt, J. Eriksson, and F. Eng, "Streaming applications over hsdpa in mixed service scenarios," in *IEEE Proc. VTC*, vol. 2, Sep. 2004, pp. 841–845.

[8] 3GPP TS 36.201 V9.1.0 (2010-03), "LTE Physical Layer: General Description."

[9] 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.

[10] 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.

[11] 3GPP TR 25.814 V7.1.0 (2006-09), "Physical layer aspects for eutra."