Dynamic Load Balancing and Throughput Optimization in 3GPP LTE Networks*

Hao Wang
National Comm. Research Lab., Southeast University
hao.wang@angstrom.uu.se

Lianghui Ding
Signals and Systems
Uppsala University
lhding@angstrom.uu.se

Ping Wu
Signals and Systems
Uppsala University
ping.wu@angstrom.uu.se

Zhiwen Pan
National Comm. Research Lab., Southeast University
pzw@seu.edu.cn

Nan Liu
National Comm. Research Lab., Southeast University
nanliu@seu.edu.cn

Xiaohu You
National Comm. Research Lab., Southeast University
xhyu@seu.edu.cn

ABSTRACT

Load imbalance that deteriorates the system performance is a severe problem existing in 3GPP LTE networks. To deal with this problem, we propose in this paper a load balancing framework, which aims at balancing the load in the entire network, while keeping the network throughput as high as possible. In this framework, the objective is formulated as a network-wide utility function balancing network throughput and load distribution, and then it is transformed to an integer optimization problem under resource allocation constraints. After that, the complexity of the problem is analyzed, network structure constraints are presented, and a practical suboptimal algorithm, called Heaviest-First Load Balancing (HFLB), is proposed. Extensive simulation is made and the results show that using the HFLB algorithm the network can get significantly better load balancing while maintaining the same network throughput at the price of a bit more handovers compared with the traditional signal strength-based handover algorithm.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*wireless communication*

General Terms

Algorithms, Performance

Keywords

load balancing, 3GPP LTE, handover, HFLB

*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.

IWCMC '10, June 28-July 2, 2010, Caen, France.

1. INTRODUCTION

Inter-cell optimization in GSM and UMTS networks is usually delivered at the stage of network planning, and often done manually. Therefore, when the networks' environments are changed, it will not be optimal. In this case, it is necessary to conduct inter-cell optimization for the networks dynamically and adaptively according to their environments, especially when the loads of cells are not uniformly distributed, namely, not balanced, and vary with time. This is one of the important self-optimization issues in self-organizing networks for 3GPP LTE, and has received much attention until now [1]. When the loads among cells are not balanced, the block probabilities of heavily loaded cells may be higher, while their neighboring cells may have resources not fully utilized. In this case load balancing can be conducted to alleviate and even avoid this problem. There has been a lot of research done on load balancing, which can be classified into two categories: block probability-triggered load balancing [2–4], and utility based load balancing [5–7].

Although many algorithms have been proposed, the assumptions made do not fully satisfy the requirements of load balancing or the designing strategies for practical systems. In the first category, the overhead is low because the load balancing is triggered only when the block probability is larger than a certain threshold. However, the block probability is not minimized, since load balancing can be done before block happening to reduce it. For the second category, i.e., utility-based load balancing schemes, the performance is better because the load balance and throughput are considered in both cell selection and handover phases. However, their overheads are heavy, because the load of each cell has to be exchanged instantaneously.

In this paper, we deal with load imbalance in LTE networks, and propose a load balancing algorithm intended to balance the load among cells and keep network throughput with reasonable overhead but still in the line of the designing strategies of practical cellular system. We first formulate the problem as an optimization problem, which considers the tradeoff between network throughput and load balancing and employs user-cell matching indicator as parameters. Then we analyze its complexity and propose a suboptimal algorithm, called Heaviest-First Load Balancing (HFLB) here. In the algorithm, new users access the cell

with the maximum signal strength and load balancing is triggered when the load of the busiest cell in the network exceeds a threshold in each load balancing cycle. Appropriate users are switched out under the metric considering throughput and load jointly. Since only the heaviest loaded cell is chosen to do load balancing, the overhead of HFLB is low. Simulation results show that the HFLB algorithm can significantly decrease the load balance index (i.e., enhance load balancing) while keeping network throughput as high as possible at a small price of only a bit more handovers.

The rest of this paper is organized as follows. In Section 2, we present the network model. In Section 3, we formulate the optimization problem. After that we propose a suboptimal algorithm in Section 4, and show simulation results in Section 5. Finally the paper is concluded in Section 6.

2. SYSTEM MODEL

2.1 Network Model

We consider an OFDMA based LTE downlink cellular network with partial frequency reuse (PFR) as shown in Figure 1. There are seven cells numbered with 1,$\cdots$,7, respectively. Each of them is controlled by an eNodeB. All the cells use the same frequency band in their center areas, and different frequency bands at their edges. Twelve adjacent OFDM subcarriers are grouped into a physical resource block (PRB), which is the smallest unit that can be allocated to each user. Throughout this paper, cell and eNodeB are used interchangeably, and the following assumptions are made:

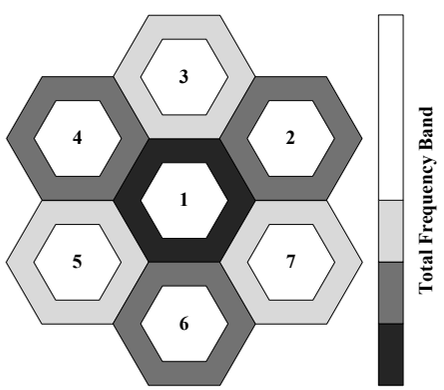

Figure 1: Network model with partial frequency reuse.

1) Each user knows instantaneous signal strength from all the cells through pilot detection. And all users send them back to their serving eNodeBs periodically.

2) Each eNodeB equally allocates one PRB to each user attaching to it, and fading differences among PRBs are not considered.

3) Neighboring eNodeBs exchange their load status information periodically through X2 interface between them [8].

4) All *time slot* mentioned in this paper represents a load balancing cycle, which is much larger than a frame length.

2.2 Load Balance Index

Let $b_i(t)$ and $b_i^u(t)$ denote the total number of PRBs and the number of used PRBs in cell i at time slot t, respectively. Then the load of cell i at time slot t is: $\rho_i(t) = b_i^u(t)/b_i(t)$. In LTE networks, all the cells are often allocated the same number of PRBs so that we use b instead of $b_i(t)$.

For performance analysis in our model, we define a load balance index measuring the degree of load balancing of the entire network, as follows:

$$\xi(t) = \sum_{i \in \mathbf{N}} (\rho_i(t) - \overline{\rho}(t))^2 \quad (1)$$

where $\mathbf{N}$ is the set of cells (or eNodeBs) in the network, and $\overline{\rho}(t) = \sum_{i \in \mathbf{N}} \rho_i(t)/|\mathbf{N}|$ is the average load of the network at time slot t, where $|\mathbf{N}|$ is the number of cells in the network. The load balance index is 0 when load is completely balanced among cells. The bigger the value of $\xi(t)$, the severer the unbalanced load distribution among cells. The target of load balancing is to minimize $\xi(t)$.

2.3 Throughput of User and Network

Since the utilization of PFR can reduce or even avoid inter-cell interference for most of the users, we only consider signal to noise ratio (SNR) for simplicity. The set of users is denoted by $\mathbf{K}$. Then the SNR of the signal received by user $k \in \mathbf{K}$ from eNodeB $i \in \mathbf{N}$ at time slot t can be written as: $SNR_{i,k}(t) = S_{i,k}(t)/N_0$, where $S_{i,k}(t)$ represents the power of received signal by user k from eNodeB i on the allocated PRB at time slot t, and N_0 is average power of the additive white Gaussian noise (AWGN) on the same PRB.

Given $SNR_{i,k}(t)$, the achievable Shannon rate at time slot t for user k from cell i is:

$$r_{i,k}(t) = W_{i,k} log_2(1 + SNR_{i,k}(t)) \quad (2)$$

where $W_{i,k}$ is the bandwidth of the PRB allocated by cell i. Considering that adaptive coding and modulation is used in LTE networks, we will use Shannon rate in equation (2) as the throughput of a user. The throughput of the entire network $R(t)$ is the sum of all the users' throughput at time slot t.

3. PROBLEM FORMULATION

We first present a network-wide utility function in the multi-cell LTE network as shown above. Our object is to make use of enforced handover to balance the load between different cells and keep the network throughput as high as possible at the same time. Given $|\mathbf{N}|$ eNodeBs and $|\mathbf{K}|$ mobile users, we try to find an optimal assignment between users and cells. The corresponding utility function is defined as:

$$U(\alpha, \beta)(t) = \alpha R(t) - \beta \xi(t) \quad (3)$$

where $\alpha \geq 0$ and $\beta \geq 0$ are weighting coefficients on network throughput and load balance index, respectively. Different values of α and β in solving the joint optimization problem in equation (3) can appropriately be selected according to the tradeoff between network throughput and load balancing. Influence of different α and β will be shown and explained in Section 5.

Since $R(t)$ and $\xi(t)$ are both determined by the assignment between users and cells, the problem is to find the optimal assignment that maximizes $U(t)$ for the current time slot t.

Define an assignment indicator variable $I_{i,k}(t)$, which is equal to 1 when eNodeB i assigns a PRB to user k at time

slot t, or to 0 otherwise. Then the load definition of cell i can be rewritten as: $\rho_i(t) = \sum_{k \in \mathbf{K}} I_{i,k}(t)/b$.

Denoting the assignment by the matrix $I(t) = (I_{i,k}(t) : i \in \mathbf{N}, k \in \mathbf{K}, \forall t \geq 0)$, the problem is thus equivalent to the following utility maximization problem with $I(t)$:

$$\max_{I(t)} \; U(I, \alpha, \beta)(t) = \alpha R(I(t)) - \beta \xi(I(t)) \tag{4}$$

$$s.\ t. \; \sum_{k \in \mathbf{K}} I_{i,k}(t) \leq b, \; \forall i \in \mathbf{N}, \tag{5}$$

$$\sum_{i \in \mathbf{N}} I_{i,k}(t) = 1, \; \forall k \in \mathbf{K}, \tag{6}$$

$$\sum_{i \in \mathbf{N}} I_{i,k}(t) r_{i,k}(t) \geq \theta, \; \forall k \in \mathbf{K}, \tag{7}$$

where $R(I(t)) = \sum_{k \in \mathbf{K}} \sum_{i \in \mathbf{N}} r_{i,k}(t) I_{i,k}(t)$ is the network throughput at time slot t. θ is the minimal throughput threshold of each user. The constraint in equation (5) presents that all cells have the same capacity limitation, and the number of users served by one eNodeB can't exceed the number of its total PRBs. Constraint in equation (6) tells that one user can only be served by one eNodeB at a specific time slot t. Constraint in equation (7) explains that a user can only be served by the eNodeB which can afford it a throughput large than θ.

Then we get the joint optimization utility function in equation (4), which aims at maximizing network throughput and minimizing unfairness of the load distribution among cells simultaneously with constraints in equations (5-7).

4. COMPLEXITY ANALYSIS AND SUBOPTIMAL ALGORITHM

In this section, we analyze the complexity of the problem and propose a practical suboptimal algorithm. Since handover would affect network throughput and load balance index simultaneously, to the best of our knowledge, there is no effective algorithm available until now to solve such an optimal problem that has two coupled parts in equation (4). If we use exhaustive search method, it requires a central controller and the computation complexity is huge. Besides, load of each cell and throughput of each user should be sent to the controller, which accordingly leads to a large overhead.

Unlike RNC in UMTS, LTE does not have a central controlling unit, thus the handover decisions have to be made by each eNodeB using limited cooperation with others. Besides, the overhead mentioned above is excessive. So we try to design a suboptimal algorithm which could be executed in a distributed manner with a low overhead.

Firstly, we begin with decomposing the network-wide utility function in equation (4) and yields the following relation:

$$\begin{aligned} \max_{I} \; U(I, \alpha, \beta) &= \alpha R(I) - \beta \xi(I) \\ &= \sum_{i \in \mathbf{N}} [\alpha \sum_{k \in \mathbf{K}} r_{i,k} I_{i,k} - \beta (\rho_i - \overline{\rho})^2] \end{aligned} \tag{8}$$

Since we will evaluate the system performance at each time slot, we omit the symbol t in equation (8) for convenience. Then the original utility function in equation (4) can be expressed by the sum of individual utility functions of all the cells in the system. The individual utility function of cell i can be expressed as:

$$u(i) = \alpha \sum_{k \in \mathbf{K}} r_{i,k} I_{i,k} - \beta (\rho_i - \overline{\rho})^2 \tag{9}$$

Assuming that cell i handovers a user k to a target cell j for load balancing, the following condition should be satisfied:

$$u(j)' + u(i)' > u(j) + u(i) \tag{10}$$

where $u(j)'$ and $u(i)'$ are the updated values of individual utility functions after handover for cell j and i, respectively. We use $\Delta u = u' - u$ to express the variation of individual utility function of each cell. Then $\Delta u(i)$ and $\Delta u(j)$ are:

$$\Delta u(i) = -\alpha \cdot r_{i,k} + \frac{\beta}{b}(2\rho_i - 2\overline{\rho} - \frac{1}{b}) \tag{11}$$

$$\Delta u(j) = \alpha \cdot r_{j,k} - \frac{\beta}{b}(2\rho_j - 2\overline{\rho} + \frac{1}{b}) \tag{12}$$

Then the condition in inequality (10) becomes:

$$\alpha(r_{j,k} - r_{i,k}) + \frac{2\beta}{b}(\rho_i - \rho_j - \frac{1}{b}) > 0 \tag{13}$$

The gain of user k which is switched from cell i to j is $\Delta u(j) + \Delta u(i)$ which we define as $g^k_{i,j}$, given as follows:

$$g^k_{i,j} = \alpha(r_{j,k} - r_{i,k}) + \frac{2\beta}{b}(\rho_i - \rho_j - \frac{1}{b}) \tag{14}$$

There may be a lot of users satisfying inequality (13) in cell i. For the one cell case, we have proved that switching out the user with the currently largest gain one by one may lead to the optimal solution. Due to space limitation, the theorem is given with the proof omitted.

THEOREM 1: For load balancing handover in a cell, the total gain in equation (8) is maximized if the user with the currently largest gain is switched out one by one until no user satisfies inequality (13).

Since information exchange for updating loads of cells after each handover of the entire network may lead to a heavy overhead and there is no need for all cells to do load balancing, thus in each load balancing cycle we only choose the heaviest loaded one whose load exceeds the threshold σ to perform load balancing according to the proposed algorithm, Heaviest-First Load Balancing (HFLB), as follows:

Algorithm 1: Heaviest-First Load Balancing Algorithm.

At the mth load balancing cycle:

- ***Step 1:*** All eNodeBs receive load status from its neighboring cells. And cell i is the heaviest loaded one.
- ***Step 2:*** If the load of cell i exceeds threshold σ, go to step 3. Else, go to step 5.
- ***Step 3:*** In cell i, find user k and target cell j with the largest $g^k_{i,j}$. If it satisfies inequality (13), switch user k to cell j and update other users' gain in cell i, then go to step 4. Else, go to step 5.
- ***Step 4:*** If load status of cell i still exceeds σ, go to step 3. Else, go to step 5.
- ***Step 5:*** Stop.

5. SIMULATIONS

5.1 Simulation Setup

The network considered here is composed of 7 hexagonal micro cells as shown in Figure 1, where the distance between neighboring eNodeBs is 130m. All eNodeBs have the same maximum transmission power 38 dBmw and 10 MHz bandwidth [9]. To avoid border effects, wrap-around technique is used. Users arrive in any cell i according to a poisson process with rate λ_i at uniformly distributed locations and depart from the system after a holding time that is exponentially distributed with mean $1/\mu = 100$sec. During their lifetime, we assume that users keep moving with a fixed speed of 3km/h and their direction is randomly selected at the beginning of user access. 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 only large scale path loss is considered in channel modeling for handover, load balancing cycle is set as 1 second. Load balancing threshold is set $\sigma = 90\%$ and θ is set to be such a value that could ensure enough throughput for handover of moving users in extreme conditions.

To verify the robustness and generality of the proposed HFLB, simulations are performed in both static and mobile scenarios with different arrival rate configurations. In the static scenario, all the users satisfying constraint in inequality (7) could be switched out for load balancing. While in the mobile scenario, ordinary handover is considered due to user movement. To avoid ping-pang effect, only such a user whose signal strength is below the threshold for ordinary signal strength based handover is selected for load balancing. All the simulation cases are listed in table 1. In each case, we repeat the simulation to get the average performance with each pair of weighting coefficients α and β. For expression convenience, we define "β/α" as balance to throughput weight ratio and use it as abscissa.

Table 1: Six Simulation Cases

Case	User speed	Arrival rate of cells 1,$\cdots$,7
Case 1-1	0 km/h	1.3 1.1 1.0 0.8 0.6 0.4 0.2
Case 1-2	3 km/h	
Case 2-1	0 km/h	1.2 1.2 0.8 0.6 0.5 0.4 0.3
Case 2-2	3 km/h	
Case 3-1	0 km/h	1.3 0.8 0.7 0.6 0.5 0.4 0.3
Case 3-2	3 km/h	

5.2 Simulation Results

Simulations are made in terms of load balance index, network throughput and handover ratio to examine the performance of the proposed algorithm, HFLB. Because of space limitation, more results on block and drop probabilities are omitted here.

5.2.1 *Load Balance Index*

Load balance index varying with "β/α" in static and mobile scenarios are shown in Figures 2 and 3, respectively. In both scenarios, new users only choose the cell with the strongest signal strength to access and handover users have a higher priority than new users.

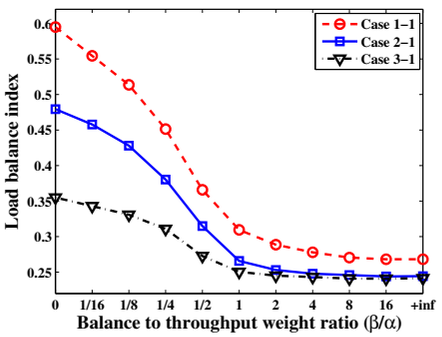

Figure 2: Load balance index in static scenario.

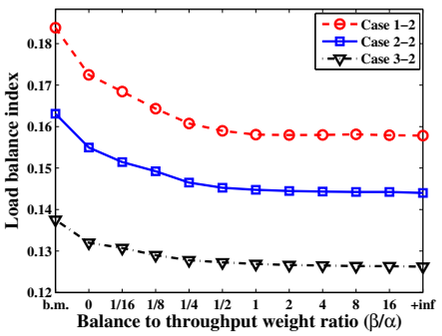

Figure 3: Load balance index in mobile scenario.

In the static scenario, when "β/α" is 0, there is no handover for load balancing. With "β/α" increasing, the load balance index is decreasing monotonously. It verifies that if we put more weight on parameter β, which is related to load balancing, the load of the entire network becomes more balanced. Due to the constraint on minimum throughput threshold θ in inequality (7), not all users in the heaviest loaded cell can be switched out for load balancing. Thus, when "β/α" is larger than 1, the load balance index decreases slower and becomes saturated as "β/α" becomes large.

In the mobile scenario, besides load balancing handovers, there are also handovers based on signal strength due to user movement. We use the traditional handover algorithm which only considers signal strength as the benchmark (on the leftmost of Figure 3 and labeled with "b.m." on the horizontal axis). The variation of load balance index is similar to that in the static scenario.

Comparing Figure 3 with Figure 2, we can find that the load balance index in the mobile scenario is lower than that in the static scenario for the same "β/α" and arrival rate configuration. This is because random movement of all users could lead to inherent load balancing capability.

5.2.2 *Network Throughput*

The results of network throughput in the static and mobile scenarios are shown in Figures 4 and 5, respectively. In the static scenario, the network throughput increases first and then decreases slightly. This is because the entire network could accommodate more users with more weight put on load balancing parameter β, which thus yields more throughput. When "β/α" is larger than 1, more users are switched

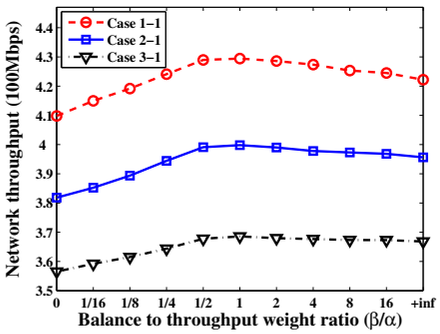

Figure 4: Network throughput in static scenario.

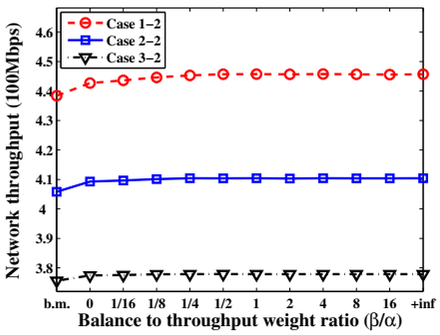

Figure 5: Network throughput in mobile scenario.

out to the cells with lower throughput for load balancing, which causes network throughput decreasing slightly. In the mobile scenario, since the number of users for load balancing is limited by user movement, the network throughput only changes a little with varying "β/α".

5.2.3 *Handover Ratio*

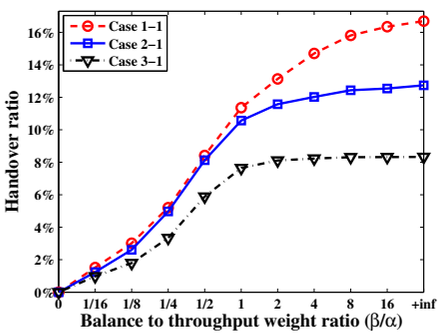

Figure 6: Handover ratio in static scenario.

Handover ratio, defined as the ratio between the number of handover and total successful calls, are shown in Figures 6 and 7. We can find that the gain in load balance index and network throughput are not free. The cost for the improvement of load balancing is a bit more handovers.

Observing all the three metrics in different cases may reveal that choosing "β/α" as 1 and 1/2 for the static and mobile scenarios, respectively, would give the best tradeoff between the gain and the cost.

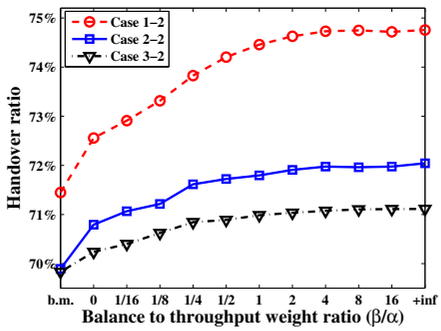

Figure 7: Handover ratio in mobile scenario.

6. CONCLUSION

In this paper, we have dealt with the optimization issue on dynamic load balancing and network throughput in 3GPP LTE networks. We formulate the issue as a joint integer optimization problem and propose a suboptimal method to solve it in a distributed manner. The algorithm for the method is developed. The generality and validity of the algorithm has been evaluated in various cases. Simulation results have showed that the algorithm could improve load balance index significantly, while the network throughput maintains high in both static and mobile scenario, and the cost for the improvement is only a bit more handovers. The tradeoff between improvement and cost is investigated and suggested.

7. REFERENCES

[1] NEC Corporation, "Self Organizing Networks - NEC's proposals for next generation radio network management," White Paper, Feb. 2009.

[2] O. K. Tonguz, and E. Yanmaz, "The mathematical theory of dynamic load balancing in cellular networks," *IEEE Trans. on Mobile Computing*, vol. 7, no. 12, pp. 1504–1518, Dec 2008.

[3] B. Eklundh, "Channel utilization and blocking probability in a cellular mobile telephone system with directed retry," *IEEE Trans. Comm.*, vol. 34, no. 3, pp. 329–337, Apr. 1986.

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

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

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

[7] 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.

[8] 3GPP TR 36.420 V8.1.0 (2009-05), "X2 general aspects and principle."

[9] 3GPP TR 25.814 V7.1.0 (2006-09), "Physical layer aspects for eutra."