Carrier Load Balancing Methods with Bursty Traffic for LTE-Advanced Systems

Wang, Yuanye; Pedersen, Klaus; Mogensen, Preben; Sørensen, Troels Bundgaard

Published in:
Personal, Indoor and Mobile Radio Communications Symposium 2009

DOI (link to publication from Publisher):
10.1109/PIMRC.2009.5450152

Publication date:
2009

Document Version
Accepted author manuscript, peer reviewed version

Carrier Load Balancing Methods with Bursty Traffic for LTE-Advanced Systems

Yuanye Wang*, Klaus I. Pedersen†, Preben E. Mogensen*†, and Troels B. Sørensen*

*Aalborg University, †Nokia Siemens Networks – DK-9220 Aalborg East – Denmark

Email: ywa@es.aau.dk

***Abstract*—in this paper we focus on LTE-Advanced performance under bursty traffic conditions, and devote our effort to the different methods for balancing the load across multiple carriers. These carriers are the component carriers (CCs) that belong to the same carrier frequency. They are bonded together in order to fulfill the requirement of wide spectrum in LTE-Advanced. We first derive the analytical model for a bursty birth-death traffic model with fixed payload size for OFDMA with different frequency domain packet schedulers. Applying this model for a multi-carrier system, we compute the performance for different system setups. The obtained analytical results are verified using extensive system level simulations. Based on the analytical and simulation results, it is suggested to assign the users on all CCs if a cell is not heavily loaded. Otherwise, assign each user with only one CC using the load balancing method of Round Robin is preferable, in the sense that it maintains good performance with low uplink overhead.**

I. INTRODUCTION

The Long Term Evolution (LTE) systems are currently deployed with a maximum bandwidth of 20MHz. It has significant improvements over the previous generation systems e.g. High Speed Packet Access (HSPA). It can provide a peak data rate over 100Mbps in the downlink and 50Mbps in the uplink [1]. To further improve the system performance and meet the International Mobile Telecommunications – Advanced (IMT-Advanced) requirements specified by International Telecommunications Union – Radio Communication Sector (ITU-R) [2], the 3rd Generation Partnership Project (3GPP) has set the new study item of LTE-Advanced [3]. With LTE-Advanced, a wider bandwidth than the 20MHz of the current LTE Rel'8 systems will be supported, going up to 100MHz [4].

The wide bandwidth required by LTE-Advanced is obtained via carrier aggregation (CA) of individual component carriers (CCs), where each CC follows the LTE Rel'8 numerology. It has been decided to use independent Link Adaptation (LA) and Hybrid Automatic Repeat request (HARQ) per CC in coherence with the LTE Rel'8 assumptions, which basically allows backward compatibility so LTE Rel'8 and LTE-Advanced terminals can co-exist [5].

With the availability of multiple CCs, it is possible to schedule one user on all CCs, leading to carrier aggregated mode at the user side; or to operate in a Rel'8 manner with separate CCs, where one user is scheduled on a single CC. The performance comparison between the two operation modes is addressed in this paper. For the latter case, different methods for balancing the load across the multiple CCs are also considered. An intelligent load balancing method can avoid the case where a large number of users are competing for resources in one CC, while resources in the other CCs are wasted due to lack of transmissions, and thereby increases the trunking efficiency of the system. For the case with mixed LTE-Advanced and Rel'8 terminals, both carrier-aggregated mode and separate CC mode should be supported, one can refer to [6] for the performance and fairness issues in such a scenario.

The support of dual carrier transmission is already standardized in [7] for the deployment of High Speed Downlink Packet Access (HSDPA) systems and many studies on the carrier load balancing already exist in literature [8]-[12]. However, these investigations all assume Code Division Multiplexing Access (CDMA) systems, which do not have the flexibility of time/frequency domain user multiplexing. LTE-Advanced systems use Orthogonal Frequency Division Multiplexing Access (OFDMA) in the downlink transmission, which allows time/frequency domain multiplexing of users. While the previous studies in [8]-[11] are carried out with voice transmission and mainly rely on simulations and heuristic algorithms, we instead consider a best effort traffic model with fixed payload size and provide additional insight by deriving a simple theoretical prediction model.

The rest of this paper is organized as follows: Section II provides the modeling of bursty traffic using birth-death process, with OFDMA and fixed payload size; Section III presents the load balancing methods under investigation; In Section IV we describe the simulation methodology and assumptions; Section V includes results from extensive system simulations as well as comparison against the theoretical estimations; Section VI concludes the paper.

II. BURSTY TRAFFIC MODELING

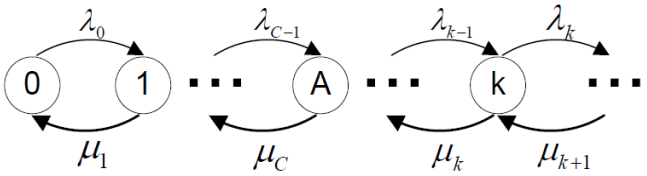

Fig. 1. The birth-death process in queueing modeling. λ_k is the arrival rate going from state k to state $k+1$ (a 'birth'); μ_k is the departure rate of going back to state $k-1$ (a 'death').

The arrival, or departure, of users in a network is usually modeled as the birth-death process, as shown in Fig. 1. The birth-death process is a special case of a continuous-time Markov process, where the states represent the current number of active users and the transitions are between neighboring states. The 'birth' is the transition towards increasing the number of active user by 1; and 'death' is the transition

towards decreasing the number of active user by 1 [13]. This is a typical *M/M/S/A* process, where the arrival of users follows the Poisson distribution, and the service time follows a negative exponential distribution (*M* for Markov). *S* is the state of the system, defined in our case by the number of active users, and *A* is the corresponding maximum number of users in the system. Although the negative exponential distribution of service time is usually assumed for voice calls [14], it can also roughly represent the time for users to download a fixed payload buffer due to channel quality variations. The latter is the case considered here, and the aforementioned assumptions are later verified via simulation results, showing a good match.

In our application of the model, in which Fig.1 is applied to model the active number of users in a macro-cell, we will use the following notation:

- S_k System state with *k* users being served
- λ_k Arrival rate in state *k*, in users per second
- μ_k Service rate in state *k*, in users per second (the average service time is μ_k^{-1})
- F Fixed payload size for the user packet transmission, in Mbits
- A Maximum number of users each cell can serve
- C Average cell throughput in Mbps, assuming a cell is fully loaded
- L The load indicator which tells how the cell is utilized: $L = \lambda F / C$, where λF is the offered load in Mbps

The arrival, λ_k, and departure, μ_k, in Fig. 1 now applies to the arrival and departure of users in the cell. In every state of the system, S_k, the *k* active users will transmit fixed payload size packets of size *F*. Hence, the departure rate is now determined by the service rate, that is, how fast we can serve the fixed payload *F*.

The admission control in LTE (-Advanced) is assumed to limit the number of users per cell to maximum *A* users. This leads to the following arrival rate:

$$\lambda_k = \begin{cases} \lambda & 0 \le k < A \\ 0 & k \ge A \end{cases} \tag{1}$$

For the study of bursty traffic with multiple carriers, we first present the modeling with single carrier and finite buffer transmission, under the access scheme of OFDMA. In an OFDMA system with multiple users, a channel aware packet scheduler can exploit the frequency and user domain diversity to improve the system performance compared with a channel blind Round Robin (RR) scheduler. This is termed as the Frequency Domain Packet Scheduling (FDPS) gain. FDPS gain is found to follow a logarithmic function of the active number of users [15]. The actual relation depends on the available transmission bandwidth, the scheduling frequency resolution, the channel conditions, and the distribution of users within each cell. For our modeling purposes, we represent the FDPS gain for a LTE system in [15] with the sample approximation as:

$$G(k) = \begin{cases} 1 & k = 1 \\ 0.11 * \ln(k) + 1.10 & 1 < k \le 13 \\ 1.38 & k > 13 \end{cases} \tag{2}$$

where *k* is the number of users for the CC and the average user throughput with k users per cell is $\frac{C}{k}\frac{G(k)}{G(\infty)}$ Mbps. The service rate is thereby:

$$\mu_k = \begin{cases} \frac{C}{F}\frac{G(k)}{G(\infty)} & 0 \le k < A \\ 0 & k \ge A \end{cases} \tag{3}$$

This is different from voice transmission with fixed user throughput, in which the service rate is proportional to the number of users [8]-[11].

The probability of the system being at state S_k is [8]:

$$P_k = \begin{cases} P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = P_0 L^k \prod_{i=1}^{k} G(\infty)/G(i) & 1 \le k \le A \\ 0 & k > A \end{cases} \tag{4}$$

where P_0 is the probability for being in state S_0, and

$$\sum_{k=0}^{A} P_k = 1 \tag{5}$$

Inserting (4) into (5), we obtain:

$$P_0 = \left(1 + \sum_{k=1}^{A} L^k \prod_{i=1}^{k} G(\infty)/G(i) \right)^{-1} \tag{6}$$

With the probability of each state derived above, we can see that if the offered load is no larger than the cell capacity ($L \le 1$), the equilibrium can always be achieved. However, in the case when the arrival rate is larger than what the cell can accommodate ($L > 1$), some users cannot get enough resources and they will remain in the system; in this case, the number of users will increase over time to *A*, and thereafter remain at that level.

When the equilibrium is reached, either because of the low offered load or because of the admission control restriction, the average number of users in a system can be expressed as:

$$\widetilde{K} = \sum_{k=0}^{A} k P_k \tag{7}$$

The average user throughput with arrival rate λ, payload size F, cell capacity C and maximum number of users A, using a frequency domain PF scheduler is calculated as:

$$TP_{user}(\lambda, F, C, A) = \frac{\sum_{k=1}^{A} \left(P_k \frac{C}{k} \frac{G(k)}{G(\infty)} \right)}{1 - P_0} \tag{8}$$

The average cell throughput is equal to the offered load as long as it stays below the threshold C, which is the throughput when a cell is fully loaded, hence:

$$TP_{cell}(\lambda, F, C, A) = \min\{C, \lambda F\} \tag{9}$$

For RR packet scheduler, the maximum cell capacity is $C / G(\infty)$, thus with $G(k) = 1$ for any positive value of k, we get also the performance with frequency domain RR.

III. Load Balancing Methods Under Investigation

As mentioned in the introduction, in a system with multiple carriers, it is possible to assign one or all CCs to each user, resulting in the Rel'8 or the aggregated operation modes. Here we consider the case when all CCs are assigned to each user as a special method of carrier load balancing, where the load is fully balanced across CCs. For the Rel'8 operation mode, different methods for balancing the load are also investigated. These load balancing methods are discussed below.

A. All CCs assigned to each user:

Although the LTE Rel'8 terminals support the transmission on only 1 CC at a time, the LTE-Advanced terminals can potentially be assigned simultaneously on all CCs. In this case, the load across CCs is automatically balanced thereby the best performance is achieved. The same bursty model as with single carrier can directly be applied for this case.

B. RR balancing with 1 CC per user:

The RR balancing [8] is also referred to as Combined Carrier Channel Assignment in [9]. The basic principle is to assign the newly arrived user to the carrier that has the least number of users. Thus, it tries to distribute evenly the load to all carriers. However, there might be small variation for the cell load in different CCs, because the number of users does not necessarily make an exact even over the CCs or because one or more users may leave the system at random.

Assume the average channel quality and bandwidth is the same for all CCs, they are expected to have the same performance. We therefore focus on the performance in one CC, and then multiply it by a factor of N to get the overall cell throughput, N being the number of aggregated carriers. The users are assigned with only one CC, their throughput equals the per CC user throughput.

C. Mobile Hashing (MH) with 1 CC per user:

The MH balancing [8], or the Independent Carrier Channel Assignment [9] method, relies on the output from the terminal's hashing algorithm. The output hash values are uniformly distributed among a finite set, which maps directly on the CC indices. Thereby, it provides balanced load across CCs in the long term. However, at each instant, the load across CCs is not balanced and the system will suffer from reduced trunking efficiency.

With MH, if some CCs are heavily loaded, one CC with low load may refuse to accept new arrivals because of the maximum user number limit. Thereby the user arrival is not independent across the CCs. Due to the difficulty in modeling the correlated user arrival, we rely solely on simulations to quantify the performance for the MH method.

IV. Simulation Methodology and Assumptions

The performance of the algorithms is evaluated in a quasi static downlink multi-cell system level simulator that follows the LTE specifications defined in [16], including detailed implementations of Layer-3 carrier load balancing, Layer-2 PS, HARQ and LA functionalities. The simulation scenario is Macro-cell case #1 as defined in [17]. The simulation parameters are summarized in Table I. The link to system mapping is based on the exponential effective metric model [18].

Note that, we aggregate 4 CCs, each of 10MHz to form a wide bandwidth of 40MHz. In case an even wider bandwidth is needed, more CCs can be aggregated together, or the bandwidth per CC can be extended. Only LTE-Advanced users are considered, which means, all users have the ability to be scheduled on multiple CCs. Simulation campaigns are conducted with one long simulation run (up to 200 seconds). This can offer sufficient statistics because users are created and terminated dynamically during the simulation.

The following measures are used in our study as performance indicators:

- Average cell throughput: Average throughput per cell, i.e. equals the summation of the user throughput in each cell.
- Average user throughput: Average throughput over the simulated users.
- Coverage: This is the 5th percentile worst user throughput over the simulated users.
- Service rate: The average number of users who finish their transmissions in a cell during one second.

TABLE I
System Simulation Settings

Parameter	Setting / description
Test scenario	3GPP Macro-cell case #1 (19 sites, 3 cells per site)
Carrier frequency	2 GHz
Aggregation configuration	4 CCs, with 10MHz per CC
Number of PRBs per CC	50 (12 subcarriers per PRB)
Sub-frame duration	1 ms (11 OFDM data symbols plus 3 control symbols)
Modulation and coding schemes	QPSK (1/5 to 3/4) 16-QAM (2/5 to 5/6) 64-QAM (3/5 to 9/10)
User receiver	2-Rx Interference Rejection Combining
HARQ modeling	Ideal chase combining
Max. number of retransmissions	4
Ack/nack & CQI feedback delay	6 ms
CQI frequency domain resolution	1 CQI per 3 PRBs
CQI reporting error	Log normal with 1dB std.
CQI reporting resolution	2 dB
Time domain PS	Round Robin
Frequency domain PS	Proportional fair
1st transmission BLER target	10%
Traffic type	Fixed payload with Poisson arrival
Payload size	2 Mbits
Admission control constraint	Maximum 50 users per cell

V. Analytical and Simulation Results

In section II we provided the model for the bursty traffic in a single-CC OFDMA system. Extension of this model to a multi-CC system with different load balancing methods was presented in Section III. Based on the latter analysis, we can analytically estimate the performance with different system settings, e.g. payload size, arrival rate. This accuracy of the derived expressions is verified via simulations. Detailed performance comparison based on extensive simulations is also made.

A. Comparison between estimation and simulation results

Based on simulation, we know the average cell throughput

in a 4x10MHz LTE-Advanced system with finite buffer is on the order of 49 Mbps for the considered environment when there is full load in all cells. Taking this value as input for the bursty traffic model (*C=49* Mbps in (8) and (9)), we obtain the performance from the analytical study.

Fig. 2 shows the average cell and user throughput for the aggregated mode and the Rel'8 mode of assigning CCs to the users. In terms of average user throughput, assigning all CCs to each user can achieve significant gain over 1 CC to each user when arrival rate is low. If the arrival rate is high, both methods provide the same user throughput. In terms of cell throughput, the performance is the same even at low arrival rates. This can be understood from (9), which shows the cell throughput is proportional to the arrival rate and upper bounded by the cell capacity, but is independent of the load balancing method. From Fig. 2 we can also see that there is a reasonable good match between the theoretical and simulated results.

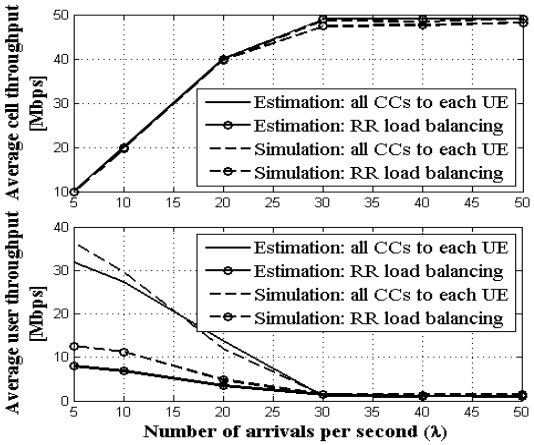

Fig. 2. Performance comparison between estimated and simulated results, with different load balancing methods.

B. Detailed comparison between different load balancing methods

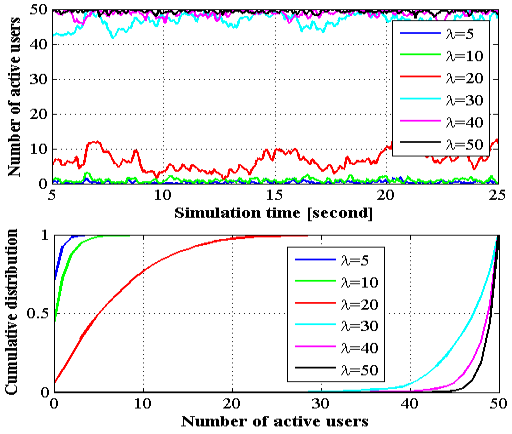

Fig. 3. Time trace and CDF distribution of active users per cell with different arrival rates. Load balancing method is assigning all CCs to each user.

Fig. 3 shows the time trace as well as the Cumulative Distribution Function (CDF) curves for the number of active users in a cell with different arrival rates. From this figure we can see that the average number of users will first increase with arrival rate, and then saturate for $\lambda > 25$/s. The reason is when the arrival rate is low, the offered load is no larger than the cell capacity. Thereby the system can reach the equilibrium without needing the maximum user limitation. At high load, the users will accumulate to approach 50 users per cell, which is the limitation put by admission control. Note that here we assume the load balancing method of assigning all CCs to each user, and the results are collected after a warm-up period of 5 seconds, which is used to stabilize the number of users in a system.

The user throughput for the three methods is summarized in Fig. 4. At low load it is observed that assigning all CCs to each user improves the performance by 150~190% relative to assigning only single CC per user with RR balancing, and 220% compared with MH balancing. However, the gain decreases with arrival rate. Indeed, when $\lambda > 25$/s, MH balancing achieves even higher user throughput than the other two methods.

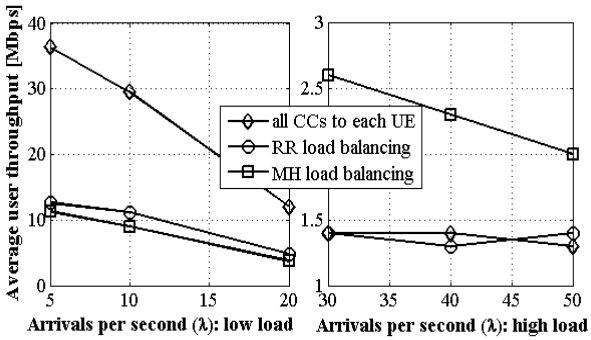

Fig. 4. Average user throughput versus arrival rate, with different load balancing methods.

The reason for the superior performance of MH balancing at high load is as follows: Even if the total number of users is upper bounded in both cases, some CCs may still be in a low load situation because of the unbalanced instantaneous CC load. Low load results in high user throughputs.

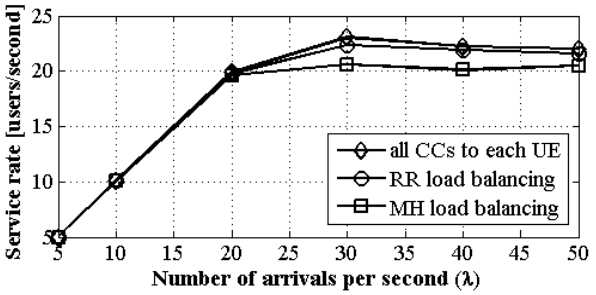

Fig. 5. Simulation results for service rate versus arrival rate, with different load balancing methods.

The service rate, which tells how many users can be served in a given time, is a better performance metric than the average user throughput with bursty traffic. Fig. 5 shows the service rate for different load balancing methods. When the offered load is low, all users are served and the service rate equals the arrival rate. When the arrival rate is higher than what the cell can accommodate, the balancing method of assigning all CCs to each user offers a service rate which is 2% higher than RR balancing, and 7~12% higher than MH balancing.

When looking at the coverage performance, the gain by assigning all CCs to each user is even obvious as compared with the other two algorithms. From Fig. 6 we can see the gain is 100~200% and 230~300% over RR and MH based balancing at low to medium load. With high cell load, it is 3~6% and 26~53%, respectively.

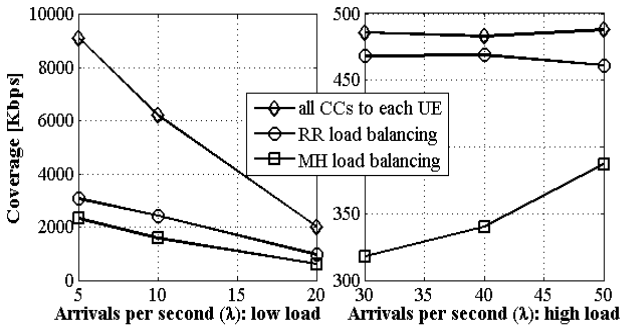

Fig. 6. Simulation results for coverage throughput versus arrival rate, with different load balancing methods.

C. Comparison between finite buffer and full buffer mode

We have seen before the cell throughput for bursty traffic increases with arrival rate, but it is bounded by the capacity *C*. Fig. 7 shows that *C* is 22% lower than the achievable throughput with full buffer and 20 users per cell. The reason is, with finite buffer mode, the users with good channel quality can finish data transmission faster than those with poor channel quality. As a result, most of the users left in the system are with poor channel quality. Note that 20 users are enough to offer maximum FDPS gain and the highest average cell throughput.

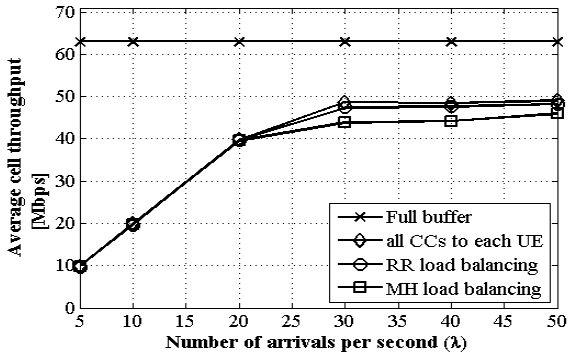

Fig. 7. Simulation results for average cell throughput. Finite buffer with different load balancing methods compared with full buffer mode.

VI. Conclusion

In this paper we have studied the performance for different load balancing methods with bursty traffic in a multi-carrier LTE-Advanced system. The bursty traffic model with fixed payload size. With this model, we can analytically estimate the performance for different load balancing methods. The accuracy of the theoretical estimates is further verified via comparison against results from extensive system level simulations.

Based on the results, we find that assigning all CCs to each user maximizes the trunking efficiency and the FDPS gain. Thereby we achieve much better performance than the other cases with one CC per user: When the cell load is low to medium, assigning all CCs to each user achieves 100~300% higher user throughputs / coverage than the other two methods; however, when the cell load is high, the different load balancing methods achieves similar performance. It is therefore suggested to assign all CCs to each user when the arrival rate of incoming traffic is low to medium, while at high arrival rate, RR load balancing with single CC per user should be used. Using single CC per user results in savings of the required uplink feedback overhead and user power consumption. More detailed performance and algorithm studies are coming as the standardization of LTE-Advanced continues, and more assumptions are being fixed.

Acknowledgment

The authors are grateful to Daniela Laselva, Jens Steiner and Mads Brix of Nokia-Siemens-Networks for their valuable suggestions and help in carrying out this study.

References

[1] H. Holma and A. Toskala, "LTE for UMTS, OFDMA and SC-FDMA Based Radio Access," John Wiley & Sons, 2009.

[2] Recommendation ITU-R M.1645, "Framework and overall objectives of the future development of IMT 2000 and systems beyond IMT 2000," June, 2003.

[3] S. Parkvall, E. Dahlman, A. Furuskar, Y. Jading, M. Olsson, S. Wanstedt, and K. Zangi, "LTE-Advanced – Evolving LTE towards IMT-Advanced," in Proc. IEEE VTC, pp. 1-5, Sept. 2008.

[4] 3GPP TR 36.913 v8.0.0, "Requirements for further advancements for E-UTRA (LTE-Advanced)," June, 2008.

[5] 3GPP TR 36.814 v1.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA Physical layer aspects," Feb. 2009.

[6] Y. Wang, K. I. Pedersen, P. E. Mogensen, and T. B. Sørensen, "Resource Allocation Considerations for Multi-Carrier LTE-Advanced Systems Operating in Backward Compatible Mode," in Proc. IEEE PIMRC, Sept. 2009.

[7] 3GPP TR 25.825 v1.0.0, "Dual-cell High Speed Downlink Packet Access (HSDPA) operations," June, 2008.

[8] T. W. Wong, and V. K. Prabhu, "Capacity Growth for CDMA System: Multiple Sectors and Multiple Carrier Deployment," in Proc. IEEE VTC, vol. 2, pp. 1608-1611, May, 1998.

[9] B. Song, J. C. Kim, and S. Oh, "Performance Analysis of Channel Assignment Methods for Multiple Carrier CDMA Cellular Systems," in Proc. IEEE VTC, vol. 1, pp. 10-14, May, 1999.

[10] A. He, "Performance Comparison of Load Balancing Methods in Multiple Carrier CDMA Systems," in Proc. IEEE PIMRC, vol. 1, pp. 113-118, 2000.

[11] N. D. Tripathi, and S. Sharma, "Dynamic Load Balancing in a CDMA System with Multiple Carriers," in Proc. IEEE VTC, vol. 2, pp. 1010-1014, Sept. 2001.

[12] K. Johansson, J. Bergman, D. Gerstenberger, M. Blomgren, and A. Wallén, "Multi-Carrier HSPA Evolution," in Proc. IEEE VTC, Apr. 2009.

[13] K. Leonard, "Queueing Systems Volume 1: Theory," John Wiley & Sons, 1975.

[14] D. Hong, and S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," IEEE Trans. Vehicular Technology, vol. 35, pp. 77-92, Aug. 1986.

[15] A. Pokhariyal, T. E. Kolding, and P. E. Mogensen, "Performance of downlink frequency domain packet scheduling for the UTRAN long term evolution," in Proc. IEEE PIMRC, pp. 1-5, Sept. 2006.

[16] 3GPP TS 36.300 v8.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description," Mar. 2007.

[17] 3GPP TS 25.814 v7.1.0, "Physical layer aspects for evolved universal terrestrial radio access," Sept. 2006.

[18] K. Brueninghaus, D. Astely, T. Salzer, et al., "Link performance models for system level simulations of broadband radio access systems," in Proc. IEEE PIMRC, vol. 4, pp.2306 - 2311, Sept. 2005.