Delay Efficient Load Balancing Scheme for Component Carrier Selection in Carrier Aggregation in LTE-A

Aditi Gupta
Bharti School of Telecommunication Technology and Management
IIT Delhi, New Delhi 110016, India
gaditi80@gmail.com

Selvamuthu Dharmaraja
Bharti School of Telecommunication Technology and Management
IIT Delhi, New Delhi 110016, India
dharmar@maths.iitd.ac.in

Subrat Kar
Bharti School of Telecommunication Technology and Management
IIT Delhi, New Delhi 110016, India
subrat@ee.iitd.ac.in

ABSTRACT

In Long Term Evolution-Advanced (LTE-A), Carrier Aggregation (CA) allows user to simultaneously transmit and receive data on multiple carriers resulting in increased throughput. Carriers need to be selected to maximize load balancing and the spectral efficiency. In this paper, a load balancing CC selection scheme which can be optimized for the Quality of Service (QoS) required by users is proposed. Feedback fluid queue model has been used to analyze and optimize the performance of the proposed scheme.

Keywords

LTE-A, Carrier Aggregation, Component Carrier, Load Balancing, Disjoint Queue Scheduler, Feedback Fluid Queue

1. INTRODUCTION

The objectives of Long Term Evolution (LTE) are reduced latency, higher user data rate, improved system capacity and coverage [1]. Long Term Evolution-Advanced (LTE-A) targets for 1 Gbps downlink speed for each user. To achieve this, increase in bandwidth is required as, a single carrier of 20MHz in LTE is not sufficient. Carrier Aggregation (CA), a feature of LTE-A aggregates more than one carrier together to provide higher bandwidth. With CA, upto five carriers can be combined together to provide 100 MHz bandwidth. Each carrier is called Component Carrier (CC). The CC can be of two types, Primary Component Carrier (PCC) and Secondary Component Carrier (SCC). PCC is always connected to the user while SCCs can be activated or deactivated. A smart load balancing algorithm for CC selection can be utilized to maintain spectral efficiency and QoS requirements. In this paper, an adaptive load balancing algorithm for LTE-A CA based system is proposed. The proposed algorithm is capable of balancing the load across different carriers and also considering the service requirements of user.

The rest of the paper is organized as follows. In Section 2, research in CC selection methodologies in Radio Resource Management (RRM) framework and the motivation behind this research work are discussed. In Section 3, performance model is proposed and the fluid queue analysis is presented. The performance analysis of the proposed model is numerically illustrated in Section 4. Finally concluding remarks and future work are given in Section 5.

2. RELATED WORK

In RRM framework of LTE-A [2], first, PCC is selected for a user and then depending on traffic load and QoS requirements, SCC's are allotted. CC selection plays an important role in optimizing the system performance with CA. There have been different scheduling algorithms proposed in the literature for CA based systems. In [3], it is proposed to assign maximum CCs to the LTE-A user to achieve maximum efficiency. In [4], least load method is introduced in which the data is assigned to the CC that has the least amount of load. The users however, arrive randomly with different sizes of files for transmission and it is difficult to completely avoid the idle CCs. To overcome this issue, CC coupling schemes have been modeled in [5] [6]. In CC coupling, if any of the CC is in busy state the user can be switched to the other CC. There are two challenges. First, handling the CC switch delay and second, development of the efficient coupling methods for multiple CC.

Lesser scheduling delay is also a required factor, especially for real time data. A scheduler algorithm has been designed to meet the QoS level of real time traffic in [7]. The data arriving will first be classified into Real Time (RT) (e.g., live streaming) and Non Real Time (NRT) (e.g. emails) traffic by a classifier and divided in RT and NRT queue respectively. The algorithm proposed aims at optimizing the system overall throughput while maintaining the required QoS of the RT traffic by providing more RB's (Resource Block) to it. Each RB constitutes 12 sub-carriers providing a bandwidth of 180 KHz. It also corresponds to a sub-frame in time domain, with Transmission Time Interval (TTI) of 1 ms. To best of our knowledge, an efficient scheme for load balancing across different carriers has not been proposed. In real time networks, the scenario is quite dynamic and one carrier may be overloaded while the others idle. On switching the data there would be scheduling delay and overhead incurred which would add to the delay and bandwidth consumed.

VALUETOOLS 2016, October 25-28, Taormina, Italy

DOI 10.4108/eai.25-10-2016.2267676

Therefore, the scheduling algorithm should be devised such that there is always an even balance across the carriers to maximize the efficiency. Since LTE-A focuses on reducing the packet delay, it is important to keep the scheduling delay minimum. Keeping these requirements in mind, we have proposed a load balancing CC selection scheme that reduces the delay of the network.

3. PROPOSED SCHEME AND ITS PERFORMANCE MODEL

3.1 Load Balancing CC Selection Scheme

The first scheduler has the information of the CCs load and on the basis of the buffer content (the buffer content rises above a certain level), the data should be switched to other carriers. In this paper, we propose a model with two thresholds for CC selection. The two thresholds divide the buffer into three regimes, the rate of the data input to the carrier depending on the regime of the buffer. If the data is above the first threshold denoted by $T^{(1)}$, the rate of the input data should be decreased and routed to another carrier. If the input data rate increases further above the second threshold, denoted by $T^{(2)}$, the input rate turns down to 0, i.e., the data inflow to the carrier is stopped completely. To provide adaptable service according to the class of traffic, the parameters (input rate in different regimes, the two thresholds value) of the model should be adjusted. We have analysed performance model through the feedback fluid queue.

In fluid queue model, instead of individual customers, a continuous entity, fluid is considered. The fluid flows into the a fluid reservoir according to a background Markov process, and flows out dependent on the output rate of the server [8]. Depending on the state of the background process, the input rate to the buffer content changes. Fluid queues are particularly useful in telecommunication systems because the bursts of data is transmitted in smaller sized cells or data packets and can be considered as fluid.

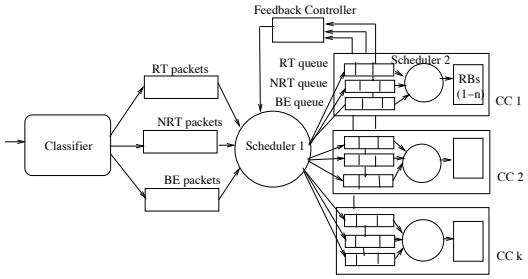

Figure 1: Schematic of the proposed performance model

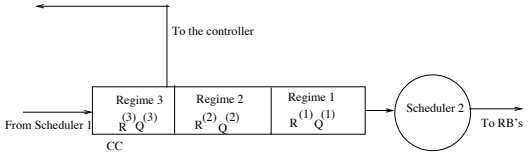

Figure 2: Buffer obtained for corresponding CCs

Feedback fluid queue is the type of fluid queue where the input rate to the buffer content depends on the background process and the behaviour of the background process depends on the content present in the buffer [9]. The meaning of feedback here is completely different from the traditional queueing systems. The feedback in fluid queue refers to control signals being sent to originating process depending on the buffer content unlike sending the data back in the conventional systems. In proposed model the input rate to the CC depends on the content of its buffer so feedback fluid queues are appropriate to model its behaviour.

3.2 Feedback Fluid Queueing Model

A feedback fluid queue with an infinite buffer content and a constant output rate c is considered to analyse the proposed performance model. In the proposed model there wouldn't be any overflows or losses, infinite buffer system can provide good approximations for the finite buffer. Figure 2 shows each CC's buffer corresponding to the users of a particular class of traffic. Let $W(t)$ be the content in the buffer for a class of traffic per CC at time t. The rate at which fluid enters the queue per unit time is dependent on the current state of a background irreducible continuous time Markov chain $\{X(t), t \geq 0\}$, defined on a state space $\{1, 2, \ldots, d\}$ for $d \in D$ where D is the total number of states in the background process.

In this performance model, the background process is the first scheduler that assigns the data to a particular CC. The background process or the first scheduler operates in two states, either it gives data to the CCs buffer (ON state) or it does not (OFF state). The $(i, j)th$ element of the scheduler generator is given by elements of matrix $Q_{ij}^{(k)}$($\tilde{Q_{ij}^{(k)}}$ at the threshold) such that the sum of row elements is 0. Q_{ij} for $j \neq i$ is the transition rate at which the background process jumps from states i to j. $R^{(k)}$ for a regime k is defined as the diagonal matrix with its ith diagonal element be given by $ir - c$, where r is the input rate to the buffer (r_h and 0 in regime 1, r_l and 0 in regime 2) and c is the constant output rate of the fluid outflow from the buffer. The output rate is constant, because some RBs are usually kept reserved for a class of traffic and continuous scheduling for those is carried out. While the output rate is constant, the value of r depends on two parameters. First, on the state of the system, whether it is ON or OFF. If it is ON then the rate depends on the second parameter, i.e., the regime of buffer. The two threshold model divides the buffer into three regimes. The input rate is high (in regime 1) as long as the lower threshold $T^{(1)}$ has not crossed. If that happens input rate is made low (in regime 2). If the higher threshold $T^{(2)}$ is reached the input flow is stopped completely (in regime 3) and the content is let to flow out until comes back to $T^{(2)}$ (in regime 2) wherein the input flow is started again.Thus, on the basis of the amount of fluid, content in the buffer at times t, the system is divided into three different regimes for our model.

For all the regimes, the subsets of D consisting of ON state, OFF state respectively are defined as:

$$D_{+}^{(k)} = \{i \in D | r_i^{(k)} > 0\}$$

$$D_{-}^{(k)} = \{i \in D | r_i^{(k)} < 0\}$$

where $k = 1, 2, 3$ depending on the regime of the buffer content. It is assumed that the input rate is not equal to the

output rate for the sake of calculations. The stability condition for the system is given by:

$$\sum_{i=1}^{D} \pi_i^{(k)} r_i^{(k)} < 0$$

where $\pi_i^{(k)}$ is the stationary distribution of the Markov process with generator $Q^{(k)}$.

Now, we find the expression for distribution of the buffer content, to calculate the performance measures like throughput, delay etc. Let $F^{(k)}(x)$ be the equilibrium distribution of the buffer where $k = 1, 2, 3$ depending on the regime of the buffer. i.e.,

$$F^{(k)}(x) = \lim_{t\to\infty} Prob\{W(t) \leq x\}, x \geq 0, k = 1, 2, 3.$$

The differential equations satisfied by the distribution can be expressed in matrix form as [10]:

$$\frac{d\mathbf{F}^{(k)}(x)}{dx} R^{(k)} = \mathbf{F}^{(k)}(x)Q^{(k)} + \mathbf{F}^{(k)}(T^{(k-1)})(\tilde{Q}^{(k-1)} - Q^{(k)}) +$$

$$\mathbf{F}^{(k)}(T^{(k-1)}-)(Q^{(k)} - \tilde{Q}^{(k-1)}) + \ldots + \mathbf{F}^{(1)}(T^{(1)}-)(-Q^{(1)} - \tilde{Q}^{(1)})$$

$$+\mathbf{F}^{(1)}(0)(\tilde{Q}^{(0)} - Q^{(1)}), k = 1, 2, 3. \quad (1)$$

The solution of the above differential equations is given by:

$$\begin{aligned}
\mathbf{F}_i^{(1)}(x) &= a^{(1)} exp[z_{(1)}^{(1)} x]\mathbf{v}_{(1)}^{(1)} + b^{(1)}\mathbf{v}_{(2)}^{(1)} + \mathbf{c}^{(1)} \text{ for regime 1} \\
\mathbf{F}_i^{(2)}(x) &= a^{(2)} exp[z_{(1)}^{(2)} x]\mathbf{v}_{(1)}^{(2)} + b^{(2)}\mathbf{v}_{(2)}^{(1)} + \mathbf{c}^{(2)} \text{ for regime 2} \\
\mathbf{F}_i^{(3)}(x) &= a^{(3)} exp[z_{(1)}^{(3)} x]\mathbf{v}_{(1)}^{(3)} + \mathbf{c}^{(3)} \text{ for regime 3}
\end{aligned}$$

$(z_i^{(k)}, v_i^{(k)})$ are the eigen value vector pair of $z_i^{(k)} v_i^{(k)} R^{(k)} = v_i^{(k)} Q^{(k)}$ and $a^{(k)}, c^{(k)}$ are the unknown coefficients for $k = 1, 2, 3$, $b^{(k)}$ for $k = 1, 2$ and $i = 1, 2$ depending on the regime of the buffer and on-off state of the system. We obtain total eleven unknown coefficients in the solution above whose values can be found by the following conditions:

1. $F_1^{(1)} = 0$. We get one equation from this condition.
2. $F_i(T^{(k)}-) = F_i(T^{(k+1)})$ for $i = 1, 2$ and $k = 1, 2$ This gives four equations from the continuity conditions at the thresholds $T^{(1)}$ and $T^{(2)}$.
3. $0 = \mathbf{c}^{(3)}Q^{(K)} + \mathbf{F}^{(K)}(T^{(K-1)})(\tilde{Q}^{(K-1)} - Q^{(K)}) + \mathbf{F}^{(K)}(T^{(K-1)}-)(Q^{(K)} - \tilde{Q}^{(K-1)}) + \ldots + + \mathbf{F}^{(1)}(T^{(1)}-)(-Q^{(1)} - \tilde{Q}^{(1)}) + \mathbf{F}^{(1)}(0)(\tilde{Q}^{(0)} - Q^{(1)})$ will give another equation.
4. $\sum_{i=1}^{3} c^{(i)} = 1$ is the normalization condition.
5. Substitution of the solution of the balance equations for $k = 1, 2$ in equation (1) gives four more equations.

The unique solution for the stationary distribution of the buffer content is obtained by which the performance measures have been calculated in the following section.

4. PERFORMANCE ANALYSIS

In this section, the performance measures are calculated and then some issues are discussed. Is the two threshold models better than the single threshold model, if so then how? What should be the relation between the thresholds to achieve the optimal performance of the model? How should the threshold and rates be varied according to the QoS required? To find the answers to above questions, we plot buffer content, throughput, mean delay with respect to the lower threshold $T^{(1)}$ by keeping different values of $T^{(2)}$. For illustration purpose, the generator matrices and rate vectors are given as follows:

$$Q^{(1)} = Q^{(2)} = Q^{(3)} = \tilde{Q}^{(0)} = \tilde{Q}^{(1)} = \tilde{Q}^{(2)} = \begin{pmatrix} -2 & 2 \\ 1 & -1 \end{pmatrix},$$

$$R^{(1)} = \begin{pmatrix} 15 \\ 0 \end{pmatrix}, R^{(2)} = \begin{pmatrix} 7 \\ 0 \end{pmatrix}.$$

With these values and different values of thresholds, the density distribution is found and then the performance measures are calculated.

4.1 Mean Buffer Content

Mean Buffer content is given by $\int_0^\infty (1 - F(x)dx$. It is plotted in Figure 3 and for increasing values of thresholds the buffer content increases as the accumulations in it increase.

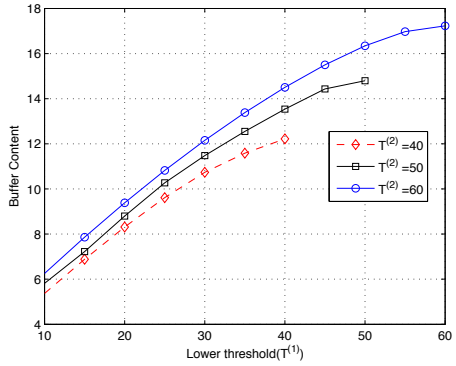

Figure 3: Mean buffer content Vs thresholds

4.2 Mean Throughput

Mean throughput is given by $c * (1 - F(0))$ where $F(0)$ is the density when the buffer content is 0. It is observed from Figure 4 that, for a given value of $T^{(2)}$, the throughput increases as the $T^{(1)}$ increases. Also, when $T^{(1)} = T^{(2)}$, which is the maximum limit for $T^{(1)}$, the throughput becomes maximum. This is also the case of a single threshold as both $T^{(1)} = T^{(2)}$ can be considered there. Hence, it is concluded on comparing with a single threshold model (in which the input is given at a single constant rate), proposed model will give lower throughput.

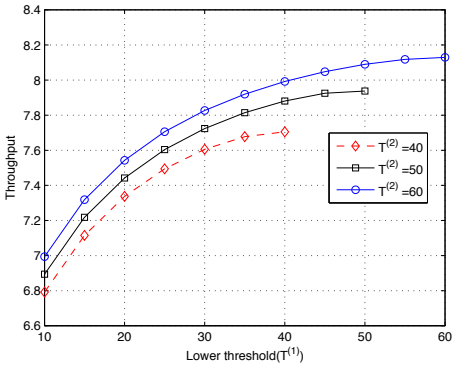

Figure 4: Throughput Vs thresholds

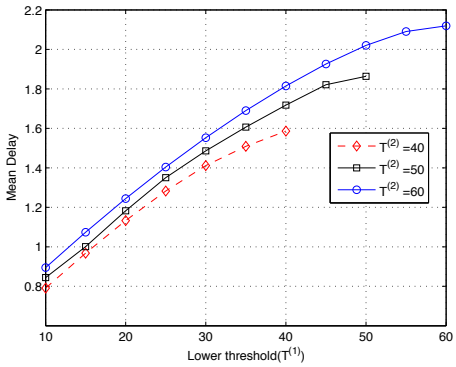

Figure 5: Delay Vs thresholds

4.3 Mean Delay

Mean Delay is given by average buffer content divided by throughput. In Figure 5, mean delay is plotted against the threshold. It is observed that as the two thresholds get closer, delay increase. Delay in a single threshold system would be thus higher than proposed model of two thresholds. Hence, it can be concluded that the proposed model results in lower delay and lower throughput as compared to model where the CC is assigned based on the single threshold. Apart from throughput and delay, number of feedback signals sent to the first scheduler also play role in adjusting the parameters for the incoming traffic. The feedback signals being sent result in causing overheads so it is an important factor while designing the system. Further, it is observed from the above results that, throughput increases when the difference between $T^{(1)}$ and $T^{(2)}$ increases though leading to more feedback signals being sent.

Keeping in mind the delay, throughput and amount of feedback involved in the model which can be designed. The real time data has the minimum delay requirement, thus for those packets, a lesser value higher threshold should be chosen. In other words, the switching of data to other users should be done even for low values of buffer content. The lesser the value of lower threshold is kept, lesser would be the delay. The bigger the difference between the thresholds, the lesser will be the feedback signals sent. The data sent as the best effort on the network such as emails, file sharing which do not require as least as delay possible, thus the other parameters such as throughput can be maximized for this class of data. Since, the data does not require minimal delay, the higher threshold can be kept more in this case. Thus, even a single carrier is sufficient for such data packets.

5. CONCLUSIONS AND FUTURE WORK

In this paper, a delay efficient load balancing scheme for CC selection in CA in LTE-A is proposed. Proposed scheme involving double threshold can be adapted according to different QoS requirement of user. Performance analysis of the proposed model is presented with fluid queue and the measures such as throughput and mean delay are compared with single threshold system. The proposed scheme is capable of balancing load across carriers while keeping the delay lesser than the single threshold model. The impact of this model and implementation on higher layers such as TCP can be studied as future work.

6. REFERENCES

[1] A. Hashimoto, H. Yoshino, and H. Atarashi. Roadmap of IMT-advanced development. *Microwave Magazine, IEEE, 9(4),* 80-88, 2008.

[2] K. I. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. G. U Garcia, and Y. Wang. Carrier aggregation for LTE-advanced: functionality and performance aspects. *Communications Magazine, IEEE, 49(6),* 89-95, 2011.

[3] Y. Wang, K. I. Pedersen , P.E. Mogensen, and T.B. Sorensen. Carrier load balancing methods with bursty traffic for LTE-Advanced systems. In *20th International Symposium on Personal, Indoor and Mobile Radio Communications,* 22-26. IEEE, 2008.

[4] L. Chen, W. Chen, X. Zhang, and D. Yang. Analysis and simulation for spectrum aggregation in LTE-advanced system. In *70th Vehicular Technology Conference Fall (VTC 2009-Fall),* 1-6. IEEE, 2009.

[5] L. Zhang, F. Liu, L. Huang, and W. Wang. Traffic load balance methods in the LTE-Advanced system with carrier aggregation. In *International Conference on Communications, Circuits and Systems (ICCCAS),* 63-67. IEEE, 2010.

[6] Y. Li, L. Zhange, X. Tan, and B. Cao. An advanced spectrum allocation algorithm for the across-cell D2D communication in LTE network with higher throughput. In *China Communications,* to appear, 2016.

[7] Y. L. Chung, L. J. Jang, and Z. Tsai. An efficient downlink packet scheduling algorithm in LTE-Advanced systems with Carrier Aggregation. In *Consumer communications and networking conference (CCNC),* 632-636. IEEE, 2011.

[8] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal, 61(8),* 1871-1894, 1982.

[9] W. R. W. Scheinhardt. *Markov-modulated and feedback fluid queues.* Universiteit Twente, 1998.

[10] M. Mandjes, D. Mitra, and Scheinhardt. Models of network access using feedback fluid queues. *Queueing Systems, 44(4),* 365-398, 2003.