Distributed Load Balancing Through Self Organisation of Cell Size in Cellular Systems

Ali Imran[1], Elias Yaacoub[1], Muhammad Ali Imran[2], and Rahim Tafazolli[2]
[1]QU Wireless Innovations Center (QUWIC), Doha, Qatar, 210531
[2]CCSR, University of Surrey, Guildford, United Kingdom. GU2 7XH.
email:{alii, eliasy}@quwic.com, {m.imran,r.tafazolli}@surrey.ac.uk

***Abstract*—Uneven traffic load among the cells increases call blocking rates in some cells and causes low resource utilisation in other cells and thus degrades user satisfaction and overall performance of the cellular system. Various centralised or semi centralised Load Balancing (LB) schemes have been proposed to cope with this time persistent problem, however, a fully distributed Self Organising (SO) LB solution is still needed for the future cellular networks. To this end, we present a novel distributed LB solution based on an analytical framework developed on the principles of nature inspired SO systems. A novel concept of super-cell is proposed to decompose the problem of "system-wide blocking minimization" into the local sub-problems in order to enable a SO distributed solution. Performance of the proposed solution is evaluated through system level simulations for both macro cell and femto cell based systems. Numerical results show that the proposed solution can reduce the blocking in the system close to an Ideal Central Control (ICC) based LB solution. The added advantage of the proposed solution is that it does not require heavy signalling overheads.**

I. INTRODUCTION

Poor resource utilisation efficiency often results from an unbalanced traffic load among different cells in Wireless Cellular Systems (WCS). This unbalanced load may result from natural variation in user traffic dynamics over the day and night, permanent shadowing and various socio economic and demographical factors. These factors, altogether make the geographic distribution of users as well as their *cell association* non uniform and ever changing. Non uniform cell association means different number of users get associated with different cells even if the cell sizes are the same.

Although these factors vary at much slower rate as compared to many other system parameters [1][1], but they render the user distribution ever changing and hard to predict. As a result, no specific fixed geographical user distribution can be assumed during planning and deployment phase of WCS unless we opt for a worst-case design strategy. The heterogenous traffic distribution resulting from such factors may cause congestion in one cell by increasing the *call blocking* probability while resulting in under utilization of resources in the other cells at the same time. This problem becomes more severe in heterogeneous WCS i.e. WCS that contain cells of different sizes e.g. Macro and Femto cells. Different propagation characteristics of Femto and Macro cells and their largely different cell sizes become an additional reason for uneven cell associations and unbalanced traffic load. As a result resource efficiency and Quality of Service (QoS) in WCS may remain suboptimal largely, if a dynamic LB mechanism is not in place with objective to minimise system-wide average blocking. Given the complexity and scale of the problem in the context of emerging WCS, this LB mechanism has to be self organising [2]. From, functional point of view self organisation of a system/algorithm/solution can be characterised by three key qualities of that system/algorithm/solution i.e. *agility*, *scalability* and *stability*. These characteristics are discussed in detail in our work in [1], here it would suffice to say that: agility means the fastness of a solution to adapt itself to the change in its environment; scalability means ability of solution to accommodate and remain operational if reasonable number of entities of the system leave or enter the system, and stability means the elasticity and reversible behaviour of the solution in response to all the dynamics it faces i.e. it should not be chaotic; e.g. one way to ensure stability would be to avoid single points of failure in the solution design. These characteristics of self organisation make self organisation very desirable element in future cellular networks, that are going to target ubiquitous coverage and service provision in cost effective manner by heterogenous deployment; i.e. planned deployment of macro cells overlaid by the impromptu deployment of Pico/Femto cells.

Since the need for LB to mitigate the effects of the spatio temporal variation in user distribution was realised immediately after the advent of commercialised WCS [3]; a large number of research works have proposed variety of useful LB strategies. However, the problem lies in the fact that most of these LB schemes are specific to the particular generations of WCS and only a few are applicable to the emerging WCS e.g. LTE and LTE-A due to their differences in the MAC and physical layer from the legacy WCS. Broadly speaking, all of these LB schemes can be classified in three general categories based on their main underlying approach towards LB. i.e. 1) *Resource Adaptation* based LB [4]–[6], 2) *Traffic Shaping* based LB [3], [7], [8]. 3) *Coverage Adaptation* based LB [9]–[11]. A detailed survey of these schemes can be found in [1].

In order to establish the novelty of our work, we discuss only the works in [9]–[11] that are most relevant to our work as they consider an OFDMA based system and propose algorithms for dynamic cell association or coverage

[1] A time scale classification of WCS dynamics can be found in [1]

978-1-4673-2569-1/12/$31.00 ©2012 IEEE

adaptation. Authors in [9] use coverage adaptation or *dynamic association*, as termed by authors therein, for joint objective of LB and interference avoidance through fractional frequency reuse. This work shows a significant gain in terms of designed utility metric as an indicator of system-wide performance. However, the underlying assumption of having access to network-wide feedback and channel estimation, for each user and Base Station (BS), at each scheduling instant and need for a omnipotent central control entity, makes this solution short of a practical level of scalability and agility required for self organization. Authors in [10] also proposed a similar algorithm for jointly solving the problems of cell association and channel assignment with the objective of LB. This solution is again purely centralised.

Recently a LB solution for OFDMA based WCS is presented in [11]. This solution is scalable as it is implementable in a fully distributed fashion. The basic idea is that each BS periodically broadcasts its average load and Mobile Station (MS) uses this information along with the signal quality in order to make the decision for cell association. This is contrary to the design of a legacy WCS where association decision is made only on the basis of received signal strength. This distributed algorithm has been shown to achieve the global optimal solution iteratively but with the two crucial assumptions 1) spatial loads are temporally stationary and 2) time scale at which BS broadcasts its load is much larger than the time scale of call holding. The impact of these assumption on the stability and agility and thus SO potential of the proposed LB solution might need further investigation.

In summary, though there are number of LB solutions in literature, generally they are not designed to be self organising in their operation and distributed in their implementation and thus may compromise on either scalability, stability of agility. To the best of our knowledge, a distributed and self organising solution for LB applicable to generic WCS as well as heterogenous OFDMA based WCS like LTE and LTE-A, is still missing. To this end, we present a novel LB solution for minimising the call blocking through SO of the cell coverage in a distributed manner. We follow the design principles of SO in natural systems to achieve desired characteristics of scalability, agility and stability.

The minimisation of average blocking as a function of the traffic load is formulated as an optimization problem. A solution is analytically derived and its complexity is further reduced to enable its decomposition. A novel concept of super-cell is proposed in order to decompose a system-wide solution into local sub-solutions that can be implemented independently at super-cell level. Finally a heuristic algorithm is used to implement the proposed solution and its performance is evaluated through system level simulations.

The rest of this paper is organised as follows. In Section II we present system model, assumptions and problem formulation. In Section III we apply principles of SO inspired by the SO systems in nature in order to achieve SO solution for our problem. Section IV presents the numerical results and Section V concludes this paper with the future work directions.

II. System Model and Problem Formulation

For analytical treatment of the problem of LB, we consider a generic WCS with N cells. These cells can be projected by omni or directional antenna of the macro BS or a Pico/Femto BS. For mathematical traceability we assume circuit switching model. Since, in the existing 2G and 3G WCS circuit switching is prevalent, this makes this model valid for existing WCS. Though, emerging WCS such as LTE and LET-A are packet switching based, there too, data services with strict QoS requirements are provided by allocating permanent resources throughout duration of call by some sort of tunnelling mechanism. Therefore, this model can represent several practical traffic scenarios in packet switched systems as well. Since the main objective of this investigation is to establish a LB mechanism, we assume a worst case scenario of a lossy system with no queuing in place. i.e. all calls that do not find a free channel in their respective cell, on arrival, are considered blocked.

The total traffic in the system is T_t (in Erlang) such that $T_t = \sum_{n=1}^{N} T_n$ where subscript n denotes association with n^{th} cell, and subscript t denotes total. The total number of available radio resource channels in the system are M_t such that $M_t = \sum_{n=1}^{N} M_n$, where M_n is the number of frequency channels in the n_{th} cell. The blocking in the n^{th} cell can be given by Erlang B formula [12]

$$B_n(M_n; T_n) = \frac{\frac{T_n^{M_n}}{M_n!}}{\sum_{m=0}^{M_n}\left(\frac{T_n^m}{m!}\right)} \tag{1}$$

The *expected blocking* in the whole system can be found as:

$$\bar{B}(\boldsymbol{M}_N; \boldsymbol{T}_N) = \frac{1}{T_t}\sum_{n=1}^{N}\left(\frac{\frac{T_n^{M_n}}{M_n!}}{\sum_{m=0}^{M_n}\left(\frac{T_n^m}{m!}\right)} \times T_n\right) \tag{2}$$

where $\boldsymbol{T_N} = [T_1, T_2, T_3, ...T_N]$ and $\boldsymbol{M_N} = [M_1, M_2, M_3, ...M_N]$ are vectors denoting the traffic and radio resources associated with N cells.

Given the system model our problem can be described as follows: *for given total traffic and radio resources in the system, system should self organise such that the user satisfaction is optimal for that total traffic and resources in the system*. User are unsatisfied due to either hard or soft blocking. By hard blocking we mean blocking due to unavailability of free channels, whereas soft blocking means although free channels are available but interference on those channels is too high to achieve the lowest required QoS and hence the attempted call is rejected. In our work [13] we focused on maximising spectral efficiency by reducing the interference and thus minimising soft blocking. Here we focus specifically on the hard blocking referred to simply as *blocking* onward. Hence, our problem is to minimise the system-wide average blocking for given total traffic and radio resources in the system. i.e.

$$\min_{\boldsymbol{T}_N} \bar{B}(\boldsymbol{M}_N; \boldsymbol{T}_N) \tag{3}$$

subject to: $\mathrm{T}_t = \sum_{n=1}^{N} T_n$ and $\mathrm{M}_t = \sum_{n=1}^{N} M_n$

The system-wide optimal solution of this optimisation problem can be anticipated to be complex and unscalable, as it requires achieving the right amount of traffic and thus right user association for each of the N cells in the system. This in turn will require system-wide cooperation that will be a compromise on agility and scalability. In next section we determine SO solution to this problem that is very less complex and avoids the need for the system-wide cooperation.

III. Designing A Bio Inspired SO Solution

In nature many systems can be observed to manifest self organization e.g. flock of common cranes, school of fish etc. A detailed discussion on design principles of self organisation can be found in our work in [1] as well as in [14]. Here it would suffice to say that, for a self organising solution, perfect system-wide objective need not necessarily be aimed for [14]. Rather, a simpler manifestation of the objective can be aimed for, given that, it can be divided into local sub problems that can be solved at local level independently or semi independently by the local entities of system. e.g. In flock of cranes each crane does not try to maximise group flight efficiency directly, rather, *flight efficiency maximisation* problem is manifested as a *maintaining V-formation* problem. Then each crane adapts certain flight attributes based on its local observation only, that results in emergence of *almost* V-formation, that in turn can achieve the original system-wide objective approximately, i.e. air drag minimisation for group flight efficiency is maximisation [15]. In following subsections we apply these design principles of SO in natural systems to achieve SO solution for our problem in (3).

A. Solving the System-Wide Problem

The solution of system-wide optimization problem (3) should return optimal traffic distribution among cells that yields minimum average blocking in the system. We present following theorem to provide solution to the problem in (3).

Theorem 1. *In a cellular system of N cells with traffic T_i and T_j and the resources M_i and M_j in i^{th} and j^{th} cell respectively, the average blocking will be minimum if*

$$\frac{T_i}{T_j} = \frac{(M_i + 1)}{(M_j + 1)} \quad ; \forall i \neq j, \ and \ i, j = 1, 2, 3...N \tag{4}$$

Proof: provided in Appendix A. ∎

Condition in (4) is the solution of (3) and as highlighted above is an unscalable solution because for practical implementation, it requires coordination among all cells. To achieve a SO solution (4) needs to be transformed into a decomposable manifestation.

B. Designing Simpler and Decomposable Approximation

As discussed above the next step of designing a SO solution will be to shape the solution into decomposable form. By building on the solution obtained in (4) we present following theorem to obtain a form of (4) that is decomposable into local sub-solutions.

Theorem 2. *In a cellular system of N cells with given radio resources allocated to each cell and total traffic T_t, the average blocking is minimum if traffic in each cell is such that:*

$$T_n = \frac{T_t - \sum_{\forall l \in \mathcal{N}/n} (M_l - M_n)}{N}, \quad \forall \quad n \in \mathcal{N} \tag{5}$$

where $\mathcal{N}$ denote set of cells i

Proof: provided in Appendix B ∎

Equation (5) now provides us a simple solution to calculate the optimal traffic each cell can have for minimum system-wide average blocking. The real advantage of solution in (5) compared to (4) is its ease of decomposability into localised solutions as explained in next subsection.

C. Decomposing Systemwide Solution into Local Sub-Solution

In order to enable the decomposition of system-wide solution into localised solution to achieve SO as suggested in [1], [14], we propose a novel concept of *super-cell*. A super-cell is fixed set of cells among which cooperation would require trivial overhead e.g. set of six BS cells and three Femto cells associated to same BS site as illustrated in figure 1. Or a set of adjacent cells linked with X2 interface. We propose to decompose the solution in (5) to local sub-solutions by exploiting the concept of super cell, then (5) will become

$$T_n = \frac{T_s - \sum_{l \in \mathcal{N}_s/n} (M_l - M_n)}{N_s}, \quad \forall \quad n \in \mathcal{N}_s \tag{6}$$

where subscript s denotes super-cell and thus $\mathcal{N}_s$ is set of cells in the super cell such that $|\mathcal{N}_s| = N_s$ and T_s is total traffic in the super cell.

Thus the basic idea of our proposed solution is that, LB is performed within each super-cell independently using (6). The actual actuators to implement (6) i.e. to achieve the matching between traffic and the resources can be any of the three approaches taken towards LB discussed in section I. We call this solution LB-BSOF i.e. LB based on Biomimmetic SO Framework as its basic idea of achieving global objective through local actions, is inspired from SO system in nature i.e. flock of common cranes as illustrated previously.

D. Practical Implications of LB-BSOF

It can be seen that the subproblem in (6) does not aim to optimise system-wide resource allocation. This feature has one major advantage i.e. this local solution can be solved independently within each super cell without requiring co-ordination with rest of the cells in the system. This brings, agility, scalability as well as stability into the solution making it self organising. The cost of this advantage is that the system-wide globally optimal load balancing is not aimed for through LB-BSOF; however, it is just like the case that in nature SO systems do aim for perfectly optimal objectives. For instance, due to reliance on only local observations cranes do not fly in perfect V-shape, yet even maintaining a near V-shape increases

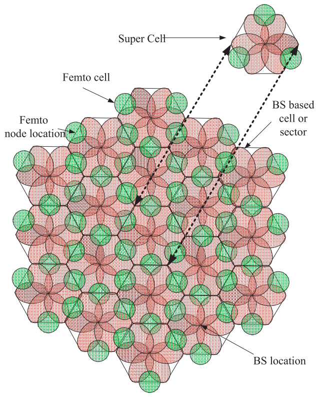

Fig. 1. Simulation model and super-cell concept illustration

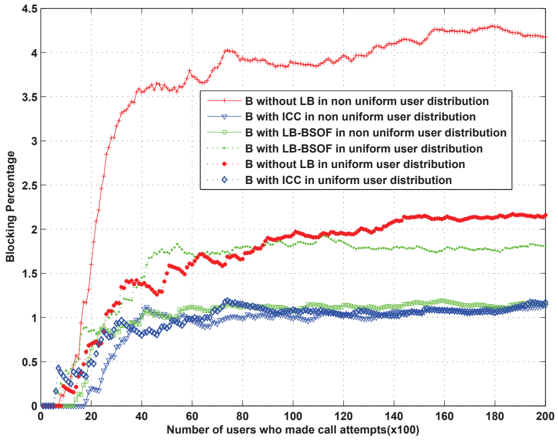

Fig. 2. system-wide average blocking in macro cells based scenario.

TABLE I
SYSTEM LEVEL SIMULATION PARAMETERS

Parameters	Values
System topology	19 BS× 6 sector, Freq. Reuse 1
Cell Radius, BS and user height	600,32 and 1.5 meters
BS and User antenna gain	20dB and 0 dB
Frequency	2 GHz
Pathloss model	3GPP Urban Macro
Bandwidth	5 MHz
Shadowing standard deviation	8 dB
User Scheduling	FIFO
Total user population	20000 users
Mean call holding time	120s (exponential)
User data rate	13 kbps (voice)
Call arrival process	Poison with adaptive mean
User Distribution	Uniform, Non Uniform

their group flight efficiency by 70% [15]. Similarly, as we will show in result section, our solution not only significantly reduces the blocking but it also brings it very close to the globally optimal solution in spite of its distributed nature.

IV. Performance Evaluation

In order to evaluate the performance of LB-BSOF, We assume a simple scenario where all cells have same amount of radio resources. Then (6) becomes $T_n = \frac{T_s}{N_s}, \quad \forall n \in \mathcal{N}_s$. We assume that each user produces same amount of traffic. Thus (6) becomes $K_n = \frac{K_s}{N_s}$ where K_n and K_s represents number of users within n^{th}cell and a super-cell respectively. Now we propose a simple heuristic algorithm to implement LB-BSOF based on coverage adaptation. i.e. *each super-cell independently adapts the coverage of its constituent cells by adapting their reference signal power determining cell association such that* (6) *is maintained*.

Performance of LB-BSOF is compared with scenario with no LB in place; and with an Ideal Central Control (ICC) based LB. ICC yields the *minimum* possible Blocking (B) for given radio resources and traffic in the system by perfect *system-wide* LB according to (5) unlike LB-BSOF that uses (6) for LB within each super-cell in the system independently. Note that ICC is not scalable and agile and thus lacks SO. However, ICC provides a useful benchmark as centralised LB scheme to be compared with our distributed SO solution i.e. LB-BSOF.

A. LB in Macro only WCS

Figure 2 shows system-wide average blocking B observed for a realistic scenarios of non uniform user distribution as well as ideal scenarios of uniform user distribution. Following key observations can be made here. Firstly, it can be seen in figure 2 that even in the hypothetical scenario of perfect uniform user distribution where cells are supposed to have same load, LB-BSOF reduces B noticeably. Secondly, By comparing the results for two scenarios it is worth noticing that for same traffic requirement and total number of users in the coverage area, the B with non uniform user distribution is much higher i.e. 5% compared to the just 2% observed in case of uniform user distribution. This shows the substantial impact user geographical distribution can have on system performance. Thirdly, it should be noticed that B with ICC stays same i.e. 1% in both cases of uniform and nonuniform user distribution. This is because ICC performs perfect LB among all cells in the system. By using hypothetical global control ICC adapts the cell sizes to take exactly same number of users in each cell throughout the system. Thus the B with ICC depends only on total number of users in system, their traffic demands, and the amount of resources available per cells. Since these parameters stay same in both scenarios hence B observed with ICC is same.

The B with LB-BSOF in terms of percentage of absolute minimum B, is shown for both non uniform and uniform user distribution, in figure 3. The gap in performance gain in two scenarios can be explained with help of figure 4 that compares user associations with and without LB-BSOF in both scenarios. A non uniform user distribution results in more diverse number of users per cell, compared to uniform user distribution, providing LB-BSOF more margin to adapt and thus converge to a local average user per super-cell that is closer to the global average user per cell.

B. LB with Femto/Pico cells

Figure 5 plots B for heterogeneous network containing Femto/Pico cells in addition to macro cells according to layout

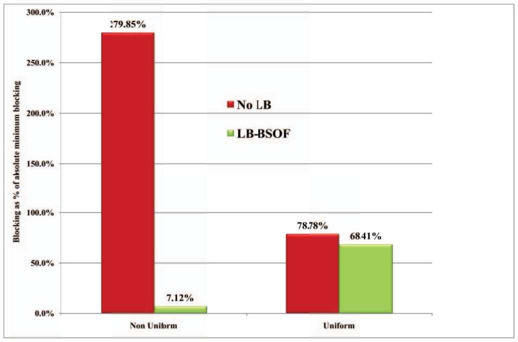

Fig. 3. B as percentage of minimum B i.e. achievable with ICC.

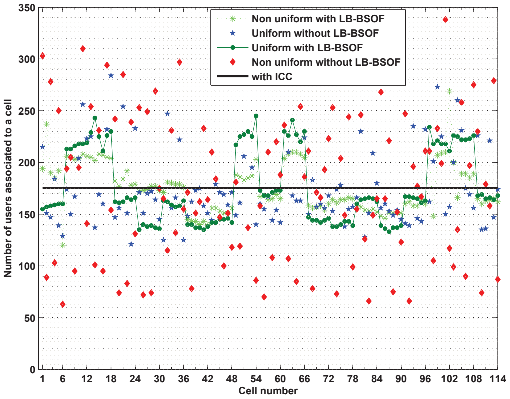

Fig. 4. User association per cell. Notice that each adjacent 6 cells make a super-cells and thus have even user association among them with LB-BSOF.

shown in the figure 1. It can be seen in figure 5 that without any LB, B on macro cells is very high, i.e.7% whereas B on Femto cells is very low i.e. 0.1%. This is mainly because of large difference in the size of Femto and macro cells. Very low transmission power and the low antenna gain of Femto cells means that very small number of users get associated with them. If same amount of radio resources are available in Femto cells, this results in large difference in the resource utilisation. It can be seen in figure 5 that LB-BSOF not only reduces the blocking in macro cell to very close to the absolute minimum, it also reduces the blocking in Femto cells to zero. This is because it not only performs inter macro cell LB, but also performs macro to Femto cell LB to to achieve better LB.

Impact of LB-BSOF on Interference : Since LB-BSOF adapts only the reference signal power that controls cell association (e.g. RSRP/cell ID carrying signal in LTE) and this adaptation is confined within super cell only, its impact on interference is observed to be negligible in simulations. However, ICC on other hand causes noticeable increase in interference. For space limitations these results are omitted.

V. Conclusion and Future Work

A novel framework of Load Balancing based on Biomimmetic Self Organisation Framework is presented that can be implemented in distributed and scalable way using concept of super-cell. Simulation results show the proposed solution can reduce blocking substantially (by 270% compared to no LB).

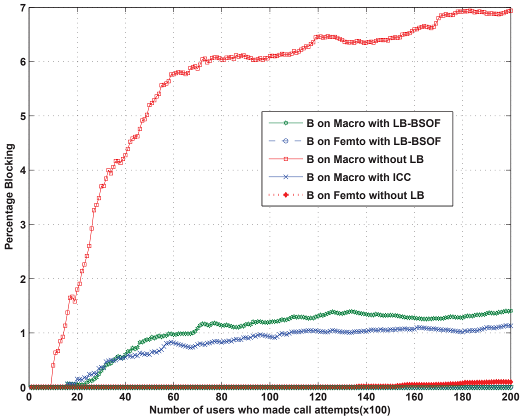

Fig. 5. Performance of LB-BSOF compared to no LB and ICC in Heterogenous WCS containing Femto cells

The key advantages of the proposed framework is its SO features because of being designed on principles of BSOF. It is fully scalable because of distributed implementation and very low complexity. It is perfectly suitable to cope with medium to large time scale (hours or longer) dynamics of WCS . However, it is not agile enough for short term dynamics because of the handovers it may trigger. In future we intend to extend this framework to other actuators in addition to coverage adaptations e.g. resource adaptation. Furthermore, impact of LB-BSOF on energy consumptions will also be investigated.

Appendix A

The blocking will be minimum or maximum if all partial derivatives of $\bar{B}$ with respect to traffic in each cell are zero.i.e.

$$\frac{\partial \bar{B}}{\partial T_n} = 0; \quad \forall \quad n = 1, 2, 3...N \tag{7}$$

solving for first cell, i.e. T_1

$$\frac{1}{T}\frac{\partial}{\partial T_1}\sum_{n=1}^{N}\left(\frac{\frac{T_n^{M_n}}{M_n!}}{\sum_{m=0}^{M_n}\left(\frac{T_n^m}{m!}\right)} \times T_n\right) = 0 \tag{8}$$

$$\frac{\partial}{\partial T_1}\left\{\frac{\frac{T_1^{M_1+1}}{M_1!}}{\sum_{m=0}^{M_1}\left(\frac{T_1^m}{m!}\right)} + \frac{\frac{T_2^{M_2+1}}{M_2!}}{\sum_{m=0}^{M_2}\left(\frac{T_2^m}{m!}\right)} + ... + \frac{\frac{T_N^{M_N+1}}{M_N!}}{\sum_{m=0}^{M_N}\left(\frac{T_N^m}{m!}\right)}\right\} = 0$$

By taking the derivative of each term we get:

$$\frac{\partial}{\partial T_1}\left\{\frac{\frac{T_1^{M_1+1}}{M!}}{\sum_{m=0}^{M}\left(\frac{T_1^m}{m!}\right)}\right\} = 0 \tag{9}$$

As M is supposed to be large number in OFDMA based systems i.e. large number of channels(sub-channels) are available per cell (this assumption is particulary true for OFDMA based systems since a much larger number channels are available per cell compared to legacy TDMA-FDMA systems). Therefore, for mathematical traceability, we can use the Taylor series approximation here. Equation (9) can then be written as:

$$\frac{\partial}{\partial T_1}\left\{\frac{\frac{T_1^{M_1+1}}{M_1!}}{e^{T_1}}\right\}=0$$

Taking the derivative:

$$\left(T_1^{M_1}\right)((M_1+1)-T_1)=0 \tag{10}$$

(10) implies that either

$$\left(T_1^{M_1}\right)=0 \tag{11}$$

or

$$((M_1+1)-T_1)=0 \tag{12}$$

Since (11) cannot be true for reasonable values of T_1 and M_1, hence the valid conditions for optimal blocking is (12):

$$T_1=M_1+1 \tag{13}$$

By second derivative test it can be shown that critical point represented by (12) is a minimum. Similarly by putting the partial derivatives of (2) with respect to traffic in other cells equal to zero, we get

$$T_n=M_n+1; \quad n=1,2,3...N \tag{14}$$

Theorem 1 can be obtained by mutually dividing (13)-(14).

Appendix B

From theorem 1 we know that the average blocking is minimum if:

$$T_n=M_n+1 \quad ,\forall n\in\mathcal{N} \tag{15}$$

we can write above set of equations as:

$$T_n=aM_n+b \quad ,\forall n\in\mathcal{N} \tag{16}$$

In order to solve this system of linear equations, to determine the optimal radio resources to be allocated or traffic to be offered to each cell, we use basic elimination method. By subtracting the equation for one cell from the other we proceed as follows:

$$T_i=a(M_i-M_j)+T_j \quad ; \quad i,j\in\mathcal{N} \tag{17}$$

As

$$T_1+T_2+T_3+...+T_N=T_t \tag{18}$$

Using (17) in (18) to solve for T_1 a

$$T_1+(a(M_2-M_1)+T_1)+(a(M_3-M_1)+T_1)+ \\ ...+(a(M_N-M_1)+M_1)=T_t \tag{19}$$

$$\frac{N\times T_1}{a}+(M_2-M_1)+(M_3-M_1)+...+(M_N-M_1)=\frac{T_t}{a} \tag{20}$$

$$\frac{N\times T_1}{a}+\sum_{\forall l\in\mathcal{N}/1}(M_l-M_1)=\frac{T_t}{a} \tag{21}$$

$$T_1=\frac{T_t-a\sum_{\forall l\in\mathcal{N}/1}(M_l-M_1)}{N}$$

Similarly solving for M_2, M_3 and so on we, can obtain the optimal traffic for each cell. The general formula to calculate optimal traffic in each cell for minimum system wide average blocking will be then given as:

$$T_n=\frac{T_t-a\sum_{\forall l\in\mathcal{N}/n}(M_l-M_n)}{N}, \quad \forall \quad n\in\mathcal{N}$$

■

Acknowledgment

This work was made possible by NPRP grant 4-353-2-130 from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] O. Aliu, A. Imran, M. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1 –26, 2012.

[2] 3GPP, "Self-organising networks; concepts and requirements," 3GPP TS 32.500 v10.1.0, Tech. Rep., 2010.

[3] T. Fujii and S. Nishioka, "Selective handover for traffic balance in mobile radio communications," *IEEE International Conference on Communications*, pp. 1840 –1846 vol.4, Jun. 1992.

[4] S. Das, S. Sen, and R. Jayaram, "A structured channel borrowing scheme for dynamic load balancing in cellular networks," *Proceedings of the 17th International Conference on Distributed Computing Systems, 1997*, pp. 116 –123, May 1997.

[5] K. Sungwook and P. Varshney, "Adaptive load balancing with preemption for multimedia cellular networks," in *IEEE Wireless Communications and Networking, 2003,(WCNC'03)*, vol. 3, March 2003, pp. 1680 –1684.

[6] S. Mitra and S. DasBit, "On location tracking and load balancing in cellular mobile environment-a probabilistic approach," in *Electrical and Computer Engineering, 2008. ICECE 2008. International Conference on*, 2008, pp. 121 –126.

[7] C.-Y. Liao, F. Yu, V. Leung, and C.-J. Chang, "A novel dynamic cell configuration scheme in next-generation situation-aware CDMA networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 1, pp. 16 – 25, 2006.

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

[9] K. Son, S. Chong, and G. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 7, pp. 3566 – 3576, July 2009.

[10] I. Koutsopoulos and L. Tassiulas, "Joint optimal access point selection and channel assignment in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 15, no. 3, pp. 521 –532, June 2007.

[11] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Alpha-optimal user association and cell load balancing in wireless networks," in *Proceedings IEEE INFOCOM, 2010*, March 2010, pp. 1 –5.

[12] A. Erlang, "Solutions to some problems in the theory of probabilities of significance in automatic telephone exchange," *Electroteknikeren*, vol. 13, p. 5, 1917.

[13] A. Imran, M. A. Imran, A.-u.-Q. , and R. Tafazolli, "Distributed spectral efficiency optimization at hotspots through self organisation of BS tilts," in *IEEE GLOBECOM Workshops*, Dec. 2011, pp. 570 –574.

[14] C. Prehofer and C. Bettstetter, "Self-organization in communication networks: principles and design paradigms," *Communications Magazine, IEEE*, vol. 43, no. 7, pp. 78 – 85, july 2005.

[15] P. B. S. Lissaman and C. A. Shollenberger, "Formation flight of birds," *Science*, vol. 168, no. 3934, pp. 1003–1005, 1970. [Online]. Available: http://www.sciencemag.org/cgi/content/abstract/168/3934/1003