Journal of Computer Science 2 (5): 466-472, 2006
ISSN 1549-3636
© 2006 Science Publications

Performance Evaluation of Implementing Calls Prioritization with Different Queuing Disciplines in Mobile Wireless Networks

Salman A. AlQahtani and Nasser-Eddine Rikli
Department of Computer Engineering, King Fahd University of Petroleum & Minerals,
Dhahran 31261, Saudi Arabia

Abstract: It is known that efficient call handling mechanisms can greatly improve cellular network performance. One way to improve system performance of cellular network is to use efficient handover schemes when users changes between cells. This paper focuses on the performance evaluation of originating and handover calls prioritization using different queue size and discipline. Both calls will be queued until they reach a certain threshold, or a channel becomes available. Higher priority will be given statically or dynamically to the handover calls based on different criteria. Tow different queuing policy known as minimum waiting time first out (MWFO) and first in first out (FIFO) are compared. In addition, the effect queue size on calls prioritization and system performance is studied. In this proposed scheme, we aim to decrease the probabilities of blocking and forced termination and increase the total carried traffic while improving the service quality.

Key words: Call prioritization, mobile wireless, performance evaluation, quality of service, queuing discipline.

INTRODUCTION

It is known that efficient call handling mechanisms can greatly improve cellular network performance. One way to improve system performance of cellular network is to use efficient handover schemes when users changes between cells. This paper focuses on the performance evaluation of originating and handover calls prioritization using different queue size and discipline. Both calls will be queued until they reach a certain threshold, or a channel becomes available. Higher priority will be given statically to the handover calls or dynamically based on some criteria. Tow different queuing policy known as minimum waiting time first out (MWFO) and first in first out (FIFO) are compared. In this scheme, we aim to decrease the probabilities of blocking and forced termination and increase the total carried traffic while improving the service quality. In [1-6] different calls prioritization schemes are studied, but the effect of queue size and queuing discipline on the call prioritization and system performance, which are the main focus of our study, were not studied.

The handover process usually consists of two phases. One is the handover initiation phase and another one is the handover execution phase. In the handover initiation phase, the QoS level is monitored in order to decide when to trigger the handover. In the handover execution phase, allocation of new resources to the handover call is performed. It should be noted that the focus of this paper is put on the handover execution phase, and we assume that the handover request detection and initiation procedures are perfect (i.e., all valid requests are detected and no invalid requests activate the handover procedure). Due to the increased traffic in cellular mobile networks, the availability of service within the supply area depends more and more on availability on free channels and thus on the proper traffic configuration of the system. From the traffic point of view, the QoS is determined by the probability of the two events, which occur due to the occupancy of all available channels. These events are the probability of call blocking and the probability of forced termination. There is trade off between these two performance measures and the configuration parameters. The performance parameter measures estimated by this study are the probability of call blocking, the probability of forced termination and the ratio of carried traffic to the total offered load.

System description: We consider a cellular mobile network where there are R cells in the network. Each base station has its zone where its radio waves can be received. The zone of each base station is called a cell. We divide the cell into overlapping areas and non-overlapping area. The area, which is covered by more than one cell, is called handover area. For simplicity, we assume that there are no areas where three or more cells overlap. Here, a service area is assumed to be of homogeneous topology (Fig. 1). We take out a cell from it, which is called marked cell (Fig. 2). There is a base station and channels in a cell. When a moving user holding a channel approaches from a neighboring cell toward the marked cell and the received signal strength goes below the handover threshold of the neighboring cell, a handover request is generated in the marked cell.

All the handover and the new calls were assumed to be independent of each other. We set a finite queue for

Corresponding Author: Salman AlQahtani, Department of Computer Engineering King Fahd University of Petroleum & Minerals, Mail Box 5065, Dhahran 31261, Saudi Arabia Tel.: +996-3-860-2110, Mobile:+966505225172

each call type. As shown in Fig. 2, in the base station there are two separate queues Q_h and Q_n with capacities K, and L for handover requests and new call requests, respectively. We also assume that the channel number needed for each call classes (bandwidth) are S_h and S_n for handover requests and new call requests, respectively. Queuing both new calls and handover streams is very likely to have merit of improving the perceived service quality. Firstly, fewer handovers will be terminated in the middle of a call due to queuing. Secondly, the total carried traffic will be increased, since the new calls are not blocked when no channels are available for them, but simply delayed for small period of time and will ultimately go through the system, contributing therefore to the carried traffic.

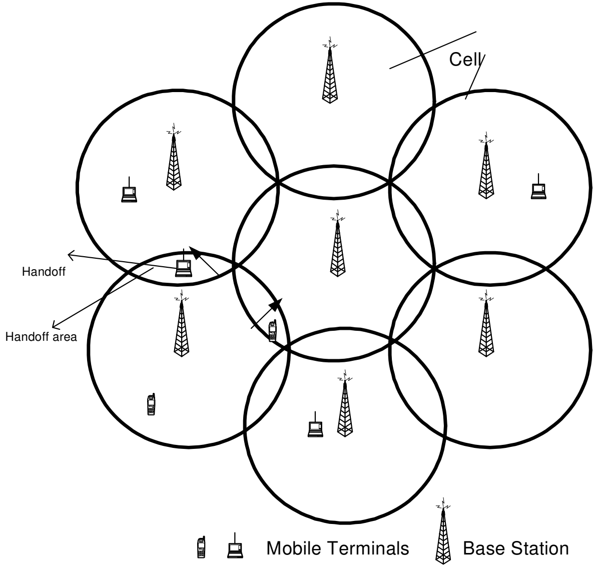

Fig. 1: Service area

System traffic model: When calls arrive to the system (Fg. 2) , they can be either: 1) Blocked, if on its arrival it finds its queue full 2) Dropped, if waiting time of the call in the queue exceeds the maximum time-out before getting the service (energy drops below the minimum), 3) Served by the system and completed within the cell if call holding time is less than cell residence time otherwise call is left for the adjacent cell.

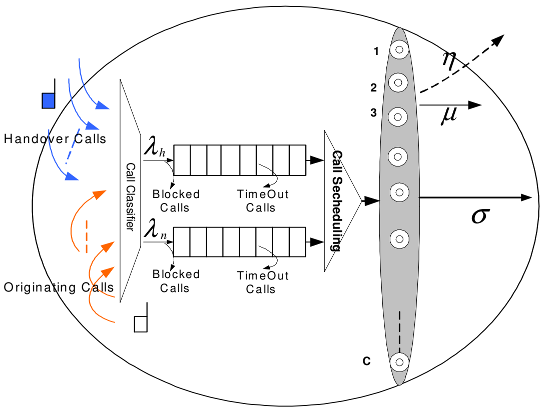

Fig. 2: System model

There are three characteristics of a traffic model: 1) Call arrival process,2) Call holding time, and 3) Cell Residence (dwell) time. A lot of research has been done to define the most suitable traffic model for wireless networks. For the call arrival process, we can safely assume that the call arrival is a Possion process. For the call holding time and the cell residence time, two main approaches can be found in the literatures. One way is to model the call holding time and the cell residence time as general independent identically distributed (iid) with nonlattice distribution using several distributions , namely: gamma, hyper-exponential, lognormal, hyper-Erlang , Weibul , or deterministic. This approach has been intensively studied by[8,9]. The second approach is based on the user's mobility, the shape and size of the cell, and exponential distribution to determine the distribution of the cell residence time and the channel holding time. This method has been extensively used by[1-9]. A lot of researches use the exponential assumptions to obtain analytical solutions for cellular systems.

Call arrival rates: New call requests and handover requests arrive at every cell according to Poisson processes with rates λ_h and λ_n for handover request and new call request respectively. All call generation processes are assumed mutually independent[1-9].

Call holding times (t_c): The call holding time of a call is the time duration between the beginning of a call and the completion of a call. It is also assumed that the call holding time is exponentially distributed with probability density function $f_c(t_c) = \mu e^{-\mu t}$ and has mean μ^{-1}. The call holding time can span several cells before the call is terminated. Call holding times are mutually independent irrespective of areas in which they are generated and base stations at which they are served. Call holding times are independent of call generation[1-9]. Each call class can have its own call holding time with mean μ_h^{-1} and μ_n^{-1} for handover request and new call request respectively.

Residence times (t_r): If a mobile is given a channel in a cell, and the mobile remains in the cell's coverage area for a period of time t_r, which is called the residence time. It is assumed to be exponentially distributed with probability density function $f_r(t_r) = \eta e^{-\eta t}$ and has mean η^{-1}. The relation between the call holding time and residence time are shown previously in Fig. 3. If the holding time of a call served at a base station ends before the residence time are over, then the call leaves the system (ends). If the residence time is over before the holding time ends, then the call try to move to a neighboring cell. Call residence times are independent of call generation process and call holding times[1-9].

Channel holding time (t_h): If the mobile is given a channel, this channel would be released either by the completion of the call in the cell or by a handover process to a neighboring cell. In this way, the channel occupancy time is the smaller of the call holding time and the cell residence time (because of the memoryless characteristic of the exponential distribution). Since the call holding time and cell residence time are assumed to be exponentially distributed with means μ^{-1} and η^{-1} respectively, and using the memoryless property of the exponential PDF, the channel holding time, which is equal to min (t_c,t_r), is exponentially distributed. Since the call holding time and cell residence time are exponentially distributed and are independent of each other as assumed before, then $f_h(t)=\mu e^{-\mu t} x \; e^{-\eta t}+\eta e^{-\eta t} x \; e^{-\mu t}=(\mu+\eta)e^{-(\mu+\eta)t}$.

For more detail[6]. Then the channel holding time being the minimum of the call holding time and cell residence time, is also exponentially distributed with a mean equal to σ^{-1} where $\sigma =$ $(\eta + \mu)$.

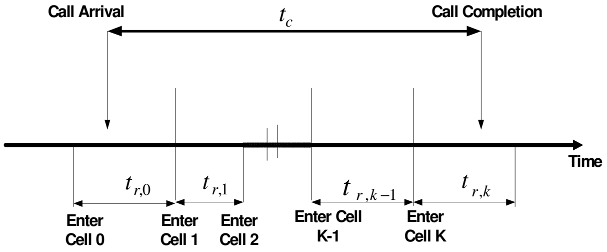

Fig. 3: Call holding time and cell residence time

Queuing time(t_q): Each handover call arriving to the cell and having to wait in the queue has a patience time (staying power time). If the virtual waiting time (i.e., the time that the handover call would have to wait until service) exceeds the patience time, the handover departs from the system and gets lost by impatience. The patience time of a handover call is assumed to be exponentially distributed with mean γ^{-1} . There are a lot of papers dealing with impatience phenomena[1-9]. Therefore, the time-out period determines the maximum time a call stays in the queue before departing from the system. We assume that the handover requests and new calls requests are subject to time-out with mean γ_h^{-1} and γ_n^{-1} respectively. For all calls, this time out period is assumed to be exponentially distributed. When a mobile is moving out of a cell (moves across the handover area), its RSS decreases until it becomes unacceptable for the communication. The average time handover requests will stay in queue (within the handover area), before it is forced to terminate, is t_q (this is also referred as degradation interval). This time is exponentially distributed with density function $f_q(t_q) = \gamma e^{-\gamma t}$ and mean γ^{-1}. Since for handover requests, queued handover requests are dropped as the mobile moves out of the handover area before the handover is completed, the time-out period will equals to t_q. For simplicity, we assume that each new call request has an average time-out identical to the queuing time of a handover request with different mean.

Performance measures: The performance parameter measures estimated by this simulation are the probability of call blocking, the probability of forced termination and the ratio of carried traffic to the total offered load. These parameters will be used in the presentation of numerical results and are defined as follows:

Blocking probability (P_b): is the probability that a new call attempt does not get service. The blocking probability of a new call is the sum of two terms (the blocking probability of new call and the time-out probability).

$$P_b = \frac{\text{No. of New Call blocked} + \text{No .of New Call Timed out}}{\text{No. of New Call Arrived}}$$

The probability of forced termination of handover (P_f): The probability that the mobile experiences an unsuccessful handover attempt. The blocking probability of a handover call is the sum of two terms (the blocking probability of handover and the time-out probability).

$$P_f = \frac{\text{No.of Handoff Call blocked} + \text{No of Handoff Call Timed out}}{\text{No. of Handoff Call Arrived}}$$

Grade of service (GoS): This is a cost function that penalizes the fact that handover forced termination probability is much more annoying than new call blocking and it is used as a reference of the grade of service offered by the system. The expression associated with this variable is:

$GoS = P_b + 10P_f$

Offered load: new calls and handover calls that arrived $\lambda = \lambda_h + \lambda_n$

Total carried traffic: is the amount of traffic admitted to the cellular network as opposed to the offered load . In general, the carried traffic is less than the offered load because of the blocking of calls and handover forced termination probability. The percentage of the offered load that is carried is certainly desired to be as high as possible. This percentage decreases with the increase of offered load and probability of call blocking and forced termination.

$$Total\,Carried\;Traffic = ((1-P_f)\lambda_h + (1-P_b)\lambda_n)$$

$$\text{OfferedTrafficUtilization} = \frac{\text{TotalCarriedTraffic}}{\text{OfferedTraffic}}$$

NUMERICAL RESULTS

Effects of queuing on the system performance: In this section, we test the effect of increasing the queue size of each call type on the system performance. Firstly, we will study the performance of two-queue system (queues are used for all call types) as shown in Fig. 3-8. Secondly, systems with one queue and two queues will be compared. In all cases, blocking probabilities, probability of forced termination and offered traffic utilization versus queue size will be compared. We consider the parameters shown in Table 1.

Table 1: Standard system parameters used to test queue usage

Call types	Arrival Rate (call/sec)	tc (sec)	tr (sec)	tq (sec)	Queue size
handover call	0.5	180	60	10	Variable
new calls	2	180	60	15	Variable
Number of channels =50 and handover requests are 25% of new calls.					

Figure 3 and 5 show the impact of increasing the queue size of all types of call simultaneously on the total system performance. As the queue size of each call increases, more traffic of each type enters the network. Therefore, the system utilization will increase as shown in Fig. 9. Figure 9 shows the over all system offered traffic utilization. The impact of queue size on the system utilization will not have any effect after a certain queue level in general (in this case level 17 (calls)) since no one of any call type will have more arrival traffic can enter the network (Fig. 3 and 5). The forced termination and blocking probabilities of each call type remains substantially constant when the queue size of each call type exceeds a constant value due to the limitation of queuing time of each call type in its queue. This is explained as follows. When the queue size of handover call increases, the carried traffic will increase. This continues until the queue size exceeds a size of 5 (calls). After that, the queuing delay will exceed the maximum waiting time of handover call and the carried traffic will not increase. The same idea will apply for new calls (17 calls). So that the total offered traffic utilization of the systems will increase as the queue size increases, until the queue size equals to 17 (calls). After that, any increase in any queue size will not have any impact on the system performance.

In general, allowing each call type to wait for service will increase the system performance and therefore decrease the blocking probability more than the system without using these queues. In addition, the queuing of new calls increases the system utilization at the expense of slight increase in the forced termination of handover calls.

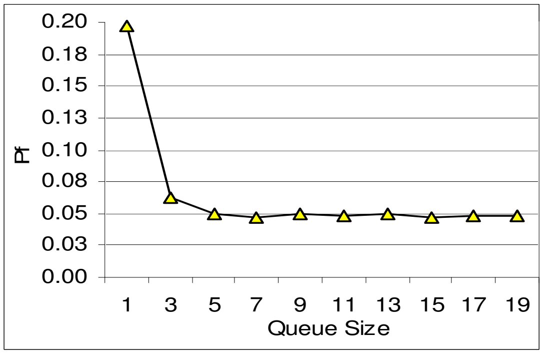

Fig. 3: P_f versus queue size

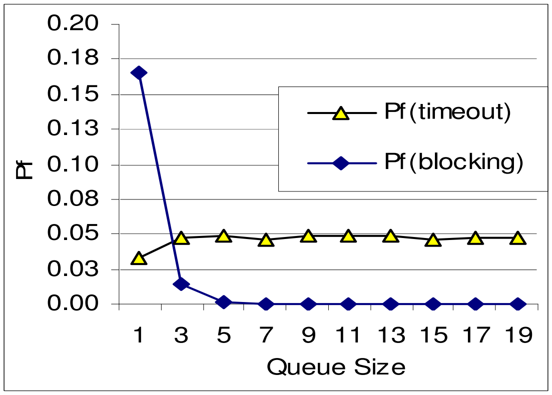

Fig. 4: Time out and blocking probability of handover requests

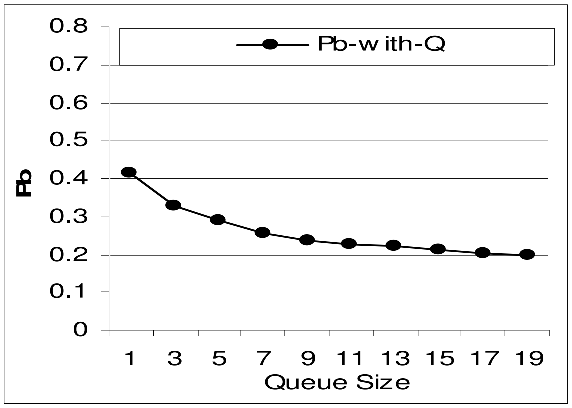

Fig. 5: P_b versus queue size

Figure 6 shows P_f as arrived request increases, with and without using queues. As we allow the handover request to be queued, P_f will improve. But this will has small effect on the new call P_b. As we allow handover request to be queued, the blocking probability will increase, but this can be solved by queuing the new calls themselves. Figure 8 shows the blocking probability of new calls versus the arrival rates using and without using the queues. Without allowing the new calls to be queued and allowing handover request to be queued the blocking probability of new calls will increase, but when we allow the new calls to be queued the blocking probability of new calls will increase.

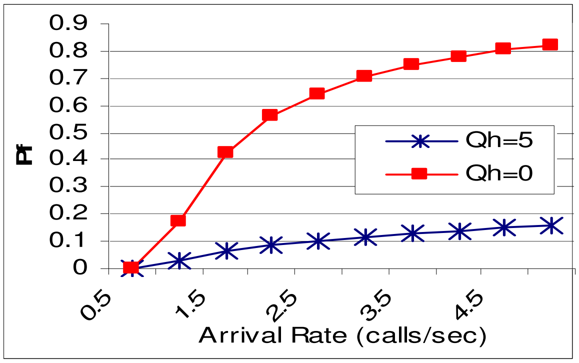

Fig. 6: P_{fh} with and without queuing versus arrived requests

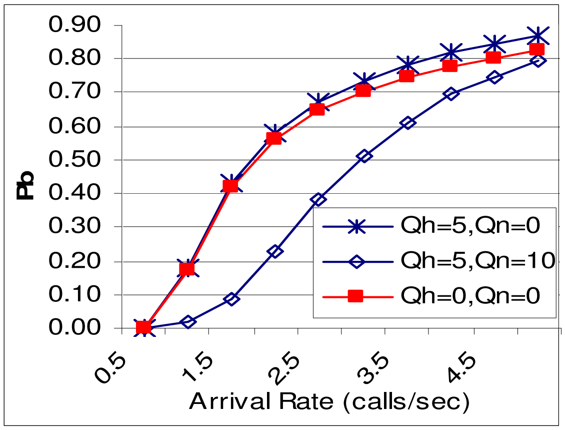

Fig. 7: P_b with and without queuing versus arrival rates

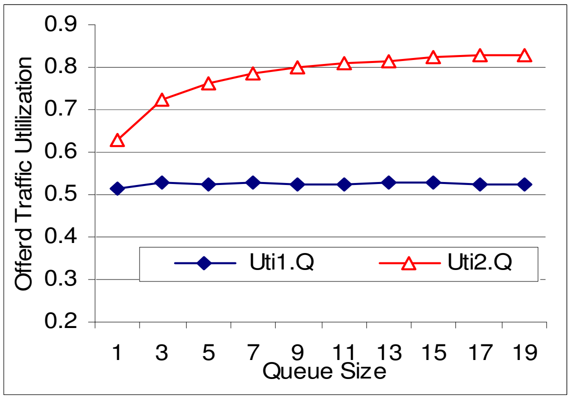

Fig. 8: Carried traffic utilization versus queue size

Figure 9 and 10 show P_f and P_b versus queue size at different handover request queuing times. As queuing time of handover increase the P_f will decrease, but this will be at the expense of small increase on the blocking probability of new calls as shown in Fig. 10.

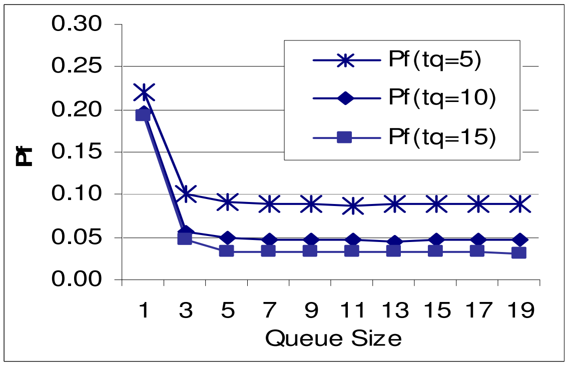

Fig. 9: P_f versus queue size at different queuing time

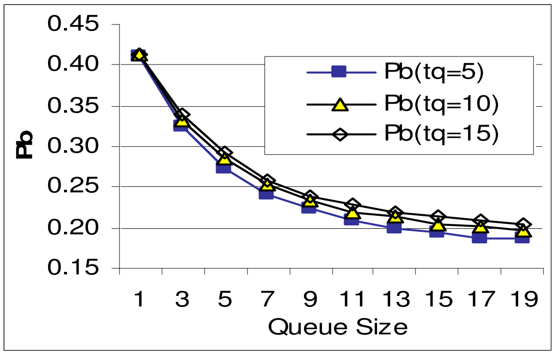

Fig. 10: P_b versus queue size at different handover request queuing time

System performance with different queuing policy: For the FIFO scheme the calls are inserted in their queue based on the call arrival times, then the call that arrive a first will be served first within its queue. For the Minimum Waiting time first out (MWFO), the calls are inserted based on the call dropping times, then the call in the head-of-line of the waiting queue (with minimum drop time) is served first within its waiting queue. In this section, we studied the performance of the system when FIFO and MWFO are used at two different cases. In the first case the handover calls have the same average speed (one mean of waiting time). The second case, the handovers call are divided in two parts, slow handover calls and fast handover calls and each part have a different mean of waiting time. In the first case, we use the parameters in Table 1 with K=5, L=10. Figure 11 shows the handover calls forced termination probability using FIFO and MBPS policy. From the Fig. 11, it indicates that the MWFO has is roughly the same as the FIFO and the difference between the two schemes is negligible.

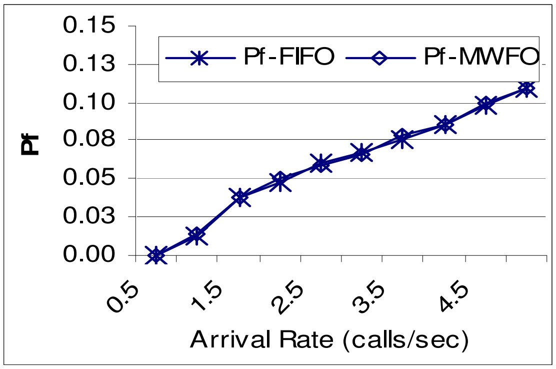

Fig. 11: P_{fh} versus arrival rate using FIFO and MWFO with one mean of waiting time

In the second case, we assume that the handover calls have different speeds. That is the mobile user can be slow or fast. In this case, we assume that the handover calls have two different mean of waiting time. The parameters used are the same as in previous case except that the second mean of waiting time is equal to halve of first waiting time. Figure 12 and 13 show forced termination probability of handover calls using FIFO and MWFO policy. From the Fig. 13, it indicates that the MWFO has a less forced termination probability than FIFO. Hence, the total carried traffic will increase as the handover calls forced termination probability decreases. As expected, as the difference between the call waiting time become bigger the difference between the two schemes become clear. In case 2, two mean of waiting time for each call type one of them is half of other the waiting time (the mobile user classified as slow and fast), and as Figures indicate, the system performance of the MWFO is better than FIFO scheme. Therefore, we can say that, the performance of MWFO will be better in case of large variation between the waiting times of handovers. Finally, in case1 and case2 the difference between MWFO and FIFO is negligible at high call arrival rates.

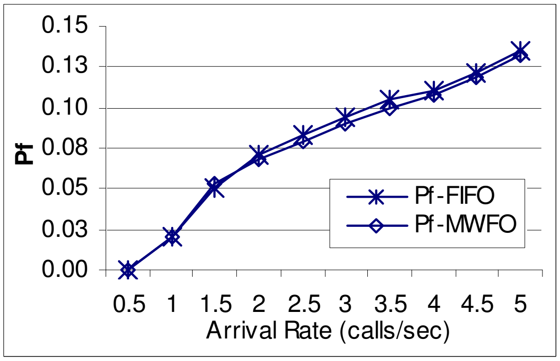

Fig. 12: P_f versus arrival rate using FIFO and MWFO with different mean of waiting times

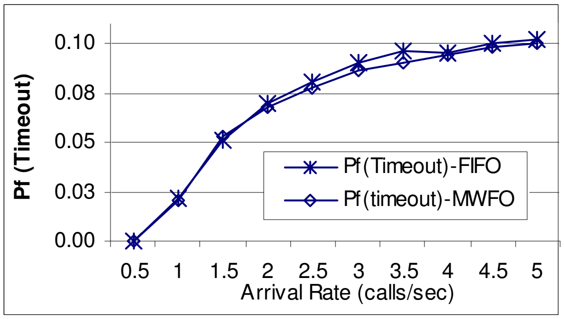

Fig. 13: Time out probability of handover versus arrival rate using FIFO and MWFO with different mean of waiting times

CONCLUSION

In this paper, a prioritized handover and admission control scheme in cellular mobile wireless networks has been studied. Finite buffers (queues) are used for both new calls and handover calls. We have focused on making blocking probability of new calls and the probabilities of forced termination of handover calls smaller than no queuing system by allowing new calls and handover calls to wait for services (using queue). According to the simulation results, as shown in all figures of the result section, we can make the blocking probabilities of each call type smaller than no queuing system and this increases the total carried traffics of the system. However, by making the blocking probabilities of handover calls small (as we increase the queue size of each handover calls), the time-out probabilities of handover calls become large. We think that making the blocking probability of handover calls small should be given priority over making the time-out probabilities of handover calls small so that the avoidance of blocking should be paid more attention than that of time-out. Then, the probability that handover calls are blocked as soon as they move to new cell becomes smaller. In addition, the idea of queuing policy was presented and the FIFO and MWFO policy were discussed and compared in two cases. The first case is when the handover calls have one mean of waiting time and the second case is when the handover calls have more than one mean of waiting time (varying waiting time). The effect of MWFO will be in the cell with varying waiting time. The results we gain from our simulation indicate that MWFO has better impact on the total system performance than FIFO policy in second case.

REFERENCES

1. Wang, J., Q.-A. Zeng and D.P. Agrawal, 2003. Performance analysis of a preemptive and priority reservation handoff scheme for integrated service-based wireless mobile networks. IEEE Trans. Mobile Computing, 2: 65- 75.

2. Miquel O. and B. Joan, 1999. Performance evaluation of variable reservation policies for hand-off prioritization in mobile networks. IEEE, July, pp: 60-67.
3. Huang, C., W. Kuang, C. Wang, Y. Jin and H. Chen, 2001. A rational channel assignment scheme for initial and Handover calls in mobile cellular systems. Computer Communication, 24: 308-318.
4. Wei, L., K. Makki and N. Pissinou, 2000. Waiting time of Handover calls for wireless mobile networks with dependent calls arrival processes and impatient calls. IEEE J. Selected Area in Communications, 2: 845-849.
5. EL-Sayed, E., Y. Yu-Dong and H. Harry, 2001. A learning approach for call admission control with prioritized handover in mobile multimedia networks. IEEE J. Selected Area in Communications, pp: 972-976.
6. Hong, D. and S.S. Rappaport., 1986. Traffic model and performance analysis for cellular mobile radiotelephone systems with prioritized and nonprioritized handover procedures. IEEE Trans. on Vehicular Technology, VT-35: 77-92.
7. Khan, F. and D. Zeghlance, 1997. Effect of cell residence time distribution on the performance of cellular mobile networks. 47th IEEE Vehicular Technology Conf., VTC'97, pp: 949-953, May 5-7, Arizona , USA.
8. Zonoozi, M.M. and P. Dassanayake, 1997. User mobility modeling and characterization of mobility patterns. IEEE JSAC, 15: 1239-1252.
9. Bolotin., V.A., 1995. Modeling call holding time distributions for CCS network design and performance analysis. IEEE J. Selected Areas in Comm., 44: 229-237.