Use of Genetic Algorithms in Efficient Scheduling for Multi Service Classes

Shyamalie Thilakawardana and Rahim Tafazolli
Centre for Communications Systems Research (CCSR),
University of Surrey,
Guildford, GU27XH, UK

Abstract - Increasing demand and limited bandwidth available for mobile communication services require efficient use of radio resources among diverse services. In future wireless packet networks, it is anticipated that a wide variety of data applications, ranging from WWW browsing to Email, and real time services like packetized voice and videoconference will be supported with varying levels of QoS. Therefore there is a need for packet and service scheduling schemes that effectively provide QoS guarantees and at the same time are simple to implement. This paper describes a novel dynamic admission control and scheduling technique based on genetic algorithms focusing on static and dynamic parameters of service classes [1]. Performance of this technique is evaluated against data services and also a traffic mix comprising voice and data. Better performance of this technique over the state of the art algorithms is illustrated on an example GPRS system. Extension of this is developed for dynamic handling of handover and new call admission control mechanisms [Appendix A]. The results conclude that this adaptive dynamic handover scheme outperforms the available mechanisms.

Keywords: Genetic Algorithm, Quality of Service, General Packet Radio Service.

I. INTRODUCTION

Present communication networks are dominated by data traffic such as WWW and Email, which are bursty in nature. They possess different characteristics compared to traditional Exponential traffic models, moreover exhibiting self similarity at the aggregate level. This affects the queuing performance characteristics opposed to the traditional traffic [1]. Therefore admission control and scheduling of these services can no longer be determined by mechanisms such as first in first out (FIFO) or best effort implemented for Exponential models. Recent admission control techniques are concerned only on one QoS parameter. These mechanisms involve Weighted Fair Queuing (WFQ), Start Time Fair Queuing (STFQ), Worst case Fair Weighted Fair Queuing (WF^2Q), Earliest Deadline/Delay First (EDF), Weighted Round Robin (WRR), and other techniques derived or related to these mechanisms [2], [3]. Comparison of FIFO with other two mechanisms namely static priority scheduling (SPS) and earliest deadline first (EDF) for GPRS service classes illustrates that EDF is more suitable for bursty services [4]. Therefore the objective of this work is to design a call admission control (CAC) and scheduling algorithm to allocate resources in a fair and efficient manner among diverse set of services satisfying the QoS agreements.

First section of this paper looks in to the problem of efficient allocation of resources among diverse set of service classes as an optimization challenge. This is encoded into a genetic algorithm environment. In the next section available admission control and scheduling techniques such as more complex, but with better performance EDF and more simple and poor performance FIFO are discussed. The performance evaluation of the proposed technique with the available techniques on an example GPRS system is discussed in the last section. Focus is mainly on the down link behaviour. [Appendix A] describes the extension of this mechanism into a dynamic handover technique. Results are compared for this adaptive dynamic handover scheme with the available mechanisms.

II. PROBLEM DEFINITION

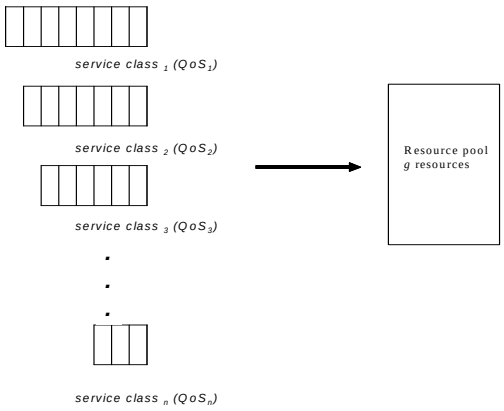
The problem of resource allocation in agreement with QoS profiles of services can be seen as an efficient distribution of n service classes among g number of resources ($n>>g$). Each service is graded according to their QoS parameters. It is needed to find an optimum way of allocating n number of services among a resource pool of g resources [Figure 1]. Since an optimum allocation within a time frame presents a combination of services among resources this kind of problems are called *combinatorial optimization* problem.

The scheduling mechanism determines the serving of each service class queue in order to stay within the agreed QoS range. When meeting this QoS range, which QoS categories needed to be accepted, which are to be rejected are determined by the CAC. CAC decides whether to accept, reject or delay a call. Therefore the CAC and scheduling algorithm must look at a wider view on dynamic as well as static characteristics of the system and at the same time capturing the traffic profile of service classes. The dynamic issues such as queue length, and static factors such as QoS profile, fairness among services needs to be considered in designing the scheduling algorithms for future services. For example data traffic such as

The work reported in this paper has been part of the Networks & Services Work Area of the Core II Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com, whose funding support is gratefully acknowledged by the authors. More detailed information and software tools of this research are available to Industrial Members of Mobile VCE.

[1] S. Thilakawardana & Rahim Tafazolli, "Method and system for determining optimum resource allocation in a network", UK patent 0302215.9 March 2003.

WWW browsing, which are bursty in nature exhibits self similarity behaviour at the aggregate level [5]. Hence to avoid undesirable features arising due to bursty characteristics, it is needed to watch the traffic profile, which is dynamic in nature. This needs to become a real time solution supporting dynamic nature of traffic.



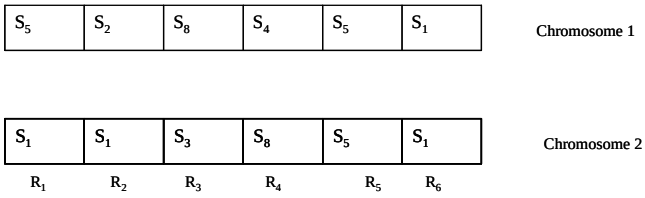
[Figure 1] Problem of efficient resource allocation

Since this problem needs a dynamic real time solution reasonably fast efficient algorithms are required. Given a hard optimisation problem it is often possible to find an optimum solution facing minimum space and time complexity. For small search spaces, classical exhaustive search algorithms can be applied, but for larger search spaces special AI techniques must be applied. Genetic Algorithms (GAs) are among such techniques. They are stochastic algorithms whose search methods model natural phenomena. This natural evolution is based on operations like selection criteria, cross over, mutation etc.. [6].

III. MAPPING THE PROBLEM TO GA ENVIRONMENT

Allocation of *g* resources among *n* service classes in a fair and efficient manner can be represented as a chromosome in a GA environment. [Figure 2] shows the chromosome representation or the packing order of 8 service classes among 6 resources. In a GA environment length of a chromosome determines the number of resources in the system. $Chromosome_1$ and $Chromosome_2$ represent two different ways of service class allocation among 6 resources. Chromosome length is 6 or the number of free resources waiting in the resource pool for service allocation. $Chromosome_1$ gives multiple resources to service $class_5$ and $Chromosome_2$ gives to service $class_1$. If there are *n* services requesting a resource, then there is more than one way of allocating these services among the resource pool. Each feasible solution represents a unique chromosome in the search space. Optimum service allocation is determined using the fitness criteria and using standard GA operations. Fitness function reflects the criteria for the optimum resource allocation. In the optimum allocation limiting factors such as QoS profile, fairness among service classes need to be reflected in the fitness criteria. The fitness function decides the survivability of the best chromosomes thus deriving optimum solution in the GA environment.

The real time problem caused by traffic characteristics is alleviated using dynamic traffic profile in the solution. Thus "refreshing frame concept" is introduced. Each solution is valid for only one refresh frame duration. After each refreshing frame resources must be reallocated according to the new optimum solution. Refreshing frames act as a dynamic way of looking and estimating real time traffic characteristics when allocating resources among multi service classes.



[Figure 2] Chromosome Representation

A. *Fitness Criteria*

The optimum solution selection is based on the fitness calculation of each chromosome. Therefore fitness function plays an important role in GA optimization procedure. The fitness function consists of three parameters, namely 'QoS index (Q_i)' of the service class, 'dynamic queue length (q_i)' of each service class and 'frequency of resources (f_i)' allocated for each service class. Following section describes the influence of each factor on the fitness function.

The QoS index of the service class depends on QoS parameters such as delay and priority. This index reflects the interaction between the QoS parameters. QoS parameters are graded according to their influence. For example priority having more weight than delay classes. QoS index ranges from 1 to 100, from the highest QoS service with QoS index 100 to the lowest at QoS index 1. There is a non-linear relationship among QoS indexes of different service classes. QoS parameter influence is inversely proportional to the QoS index. The weighting of QoS parameters to QoS index decreases according to the *square root law*. For example consider a QoS profile of the service class is defined with QoS parameters p_1 and p_2. When determining the QoS index of this service class the weight of highest QoS parameter (p_1) is inversely proportional to the QoS index with weight 1. The next QoS parameter (p_2) is inversely proportional to the QoS index with a weight of $\sqrt{p_2}$. Therefore the *QoS index* of a service with QoS profile defined in p_1 and p_2 can be represented as:

$$QoS_{index} = \frac{100}{p_1 * \sqrt{p_2}} \qquad \text{(Equation 1)}$$

Where p_1 parameter has more influence than p_2 on QoS profile of this service.

The factor considered next in the fitness calculation is dynamic queue length (q_i) of each service class queues. This factor reflects the call arrival rate distributions, call duration distributions and average service rate distributions of each queue. In fitness calculation the queue length of each service class is measured at the beginning of the refreshing frame.

Fairness is considered as the final factor. The main reason of introducing fairness to the fitness criteria is to avoid exploitation of resources by one service class. This is a major weakness of the available scheduling schemes such as EDF. The fitness decreases when the same service class request for more than one resource thus avoiding the exploitation. This is maintained in the fitness function evaluation with the introduction of 'resource frequency (f_i)'. Considering the above three factors the fitness of a service class (F_i) can be presented as;

$$F_i \quad \alpha \quad Q_i \qquad \text{(Equation 2)}$$

$$F_i \quad \alpha \quad q_i \qquad \text{(Equation 3)}$$

$$F_i \quad \alpha \quad \frac{1}{\sqrt{f_i}} \qquad \text{(Equation 4)}$$

From the above equations;

$$F_i \quad = K' \frac{Q_i * q_i}{\sqrt{f_i}} \qquad \text{(Equation 5)}$$

Where K' is a proportionality constant. Therefore fitness of the chromosome structure (C_F) in [Figure 2] is the summation of service fitness included in the chromosome which can be represented as;

$$C_F \quad = K \sum_{i=1}^{i=g} F_i \qquad \text{(Equation 6)}$$

Where K is a constant and g is the number of resources in the resource pool (same as the chromosome length). If Service Class S_i is one of the service classes in the chromosome value of F_i is calculated from [Equation 5]. If not value of $F_i = 0$ (i.e.; S_i does not contribute to the chromosome fitness C_F).

IV. PERFORMANCE COMPARISON

Performance of the above GA based dynamic CAC and scheduling mechanism is investigated applied to an example GPRS system. The success of the deployment of GPRS will be significantly influenced by the introduction of efficient and variable QoS management and supporting mechanisms. Although QoS profiles for a number of GPRS service classes has been specified by ETSI, implementation issues plays a major role in achieving that. This includes QoS management in the areas of traffic scheduling, traffic shaping and call admission control techniques.

QoS in GPRS is defined as the collective effect of service performances, which determines the degree of satisfaction of a user of the service. QoS enables the differentiation between provided services. The QoS attributes used in [7] and [8] are very similar apart from the difference related only to the throughput QoS attributes. In [7] five QoS attributes are defined. These are the precedence, delay class, reliability class, mean throughput and peak throughput class. There are four delay classes in the GPRS QoS profile; delay classes 1,2 and 3 offer predictive services and require QoS management, while class 4 provides a best effort service. Two types of delay profiles are specified as QoS parameters. One of them is the mean delay and the other one is the maximum delay in 95% of all transfers [Table 1]. In four delay classes listed two types of SDU sizes are specified (i.e., 128 and 1024 octets). By combining these attributes many possible QoS profiles are defined.

Parameter	Values				
Precedence	*High, Normal, Low*				
Reliability	*Packet loss probability: e.g., 10^{-9}, 10^{-4}, 10^{-2}*				
Delay 128 bytes	*Class*	*1*	*2*	*3*	*4*
	Mean(s)	*< 0.5*	*< 5*	*< 50*	*Best Effort*
	95%(s)	*< 1.5*	*< 25*	*< 250*	*Best Effort*
Delay 1024 bytes	*Mean(s)*	*< 2*	*< 15*	*< 75*	*Best Effort*
	95%(s)	*< 7*	*< 75*	*< 375*	*Best Effort*
Maximum bit rate	*8 kb/s – 2 Mb/s* [1]				
Mean bit rate	*0.22 b/s – 111 kb/s*				
[1] Current GPRS limit 160 kb/s					

[Table 1] GPRS QoS Profile

A. *Traffic Sources*

Traffic sources consist of GPRS applications, including email, railway traffic, mobitex and web browsing representing different probability distributions in burst sizes opposed to traditional Exponential models [9]. WWW data contributes 20 % of the total traffic mix and that of Email sessions is 40 % where as Railway and Mobitex traffic each presenting 20 % of the aggregate traffic. Focus is on the down link performance.

The Email sessions are presented by the FUNET model, which is based on statistics collected on Email usage from the Finnish University and Research Network [9]. WWW session is a characteristic application of hierarchical call architecture. Browsing *session* consists of sequence of *packet calls* and during a packet call several *packets* may be generated constituting a bursty sequence of packets. It is very important to take this phenomenon in to account in the traffic model. This burstyness during the packet call is a characteristic feature of packet transmission in the network. The modeling of WWW browsing is according to [5]. Apart from the above applications uniformly and exponentially distributed packet sizes contributes towards the mixed traffic with the inclusion of the Mobitex and Rail data [9]. Comparisons are made with available techniques such as EDF and FIFO. One of the drawbacks in EDF is higher complexity of this technique leads to implementation difficulties in practical situations. The EDF mechanism needs to sort the packet queue using at least *O(logN)* insertion operation for each arrived packet. This affects its

application due to implementation difficulty. Also at the same time with a mix of bursty and non bursty services EDF allows resource exploitation of burtsy or high QoS services [4].

Eight different service classes are defined with different QoS profiles [Table 2]. GPRS cell environment occupying a single carrier is considered. Each TDMA frame consists of eight time slots. Out of these eight slots one is allocated for signaling resulting seven GPRS. QoS profile is based on the delay classes and precedence QoS parameters [Table 2]. The refreshing frame duration is selected as 200 frames where a GPRS frame is 18.46ms. Therefore in every ~ 4s resource allocation is updated. Apart from the bursty data a traffic mix comprising voice and data are also considered. In this case voice calls are allocated to the resources as in GSM.

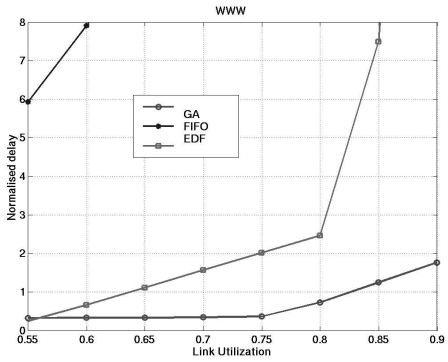
Service Type	(Priority, Delay)	QoS Index	% Mix
WWW Class 1	(1,1)	100	5
WWW Class 2	(1,2)	70	5
WWW Class 3	(1,3)	3	10
Email Class 1	(2,1)	50	10
Email Class 2	(2,2)	18	10
Email Class 3	(2,3)	2	20
Rail Data	(3,3)	1	20
Mobitex Data	(3,3)	1	20

[Table 2] Service Class categorization

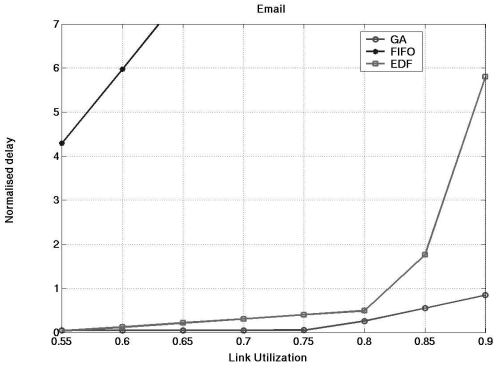
Furthermore as a quantifiable performance measurement among different mechanisms [4] introduced the comparison of performance between average normalized delay. Average normalized delay is defined as the ratio between *experienced mean delay* and the *imposed delay* for the service class with the agreement of the QoS profile. Using this measurement of *normalized delay* it is more convenient to evaluate the delay performance with variable packet sizes and different delay classes. If the QoS profiles are met satisfactorily this is below 1.

V. DISCUSSION OF RESULTS

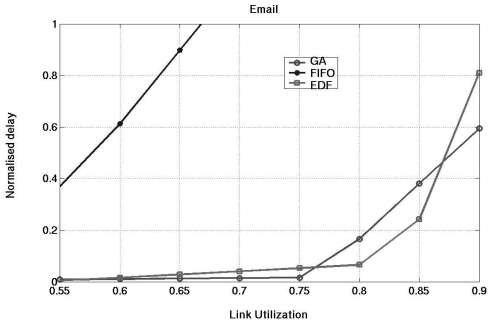
Comparison of performances in terms of normalized delay against link utilization (i.e. normalized load) is analyzed. [Figure 3] presents the normalized delay comparison for services of class I, denoting higher QoS profile classes, such as WWW class I, Email Class I. Once the throughput is higher or link utilization increases the normalized delay get increased. [Figure 4] presents the performance comparison among three different techniques for the services of class II. This follows [Figure 5] presenting that of services of class III.



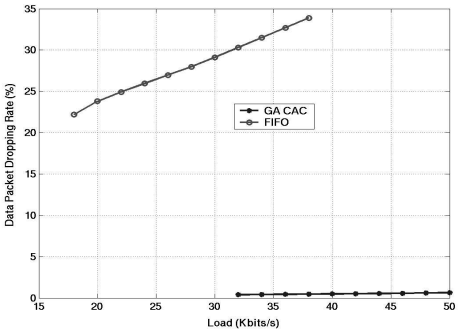
[Figure 3] Performance comparison www class I



[Figure 4] Performance comparison email class II



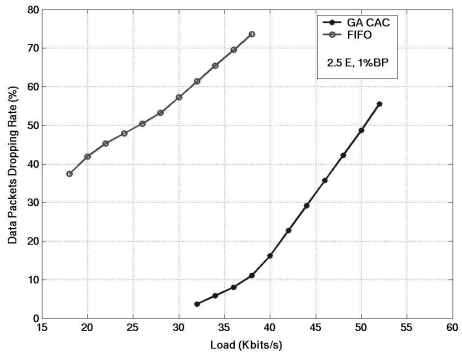
[Figure 5] Performance comparison email class III



[Figure 6] Packet dropping rate - data services

It is evident from these results GA based scheduling algorithm performs better for multi service classes maintaining QoS agreements. Finally performance is compared for a mix of voice and data. Voice calls having the highest priority get blocked only when all available slots are carried by voice. GPRS slots will be pre-empted if there are no slots available to carry voice. The dropping policy for data is two fold. If data is pre-empted for voice the packet will be dropped. Also if a

packet exceeds the delay profile once waiting in the queue it is dropped.



[Figure 7] Packet dropping rate - data and voice

[Figure 6] and [Figure 7] show the performance comparison between the admission control mechanisms in terms
of packet dropping rate against the load. It is experimented under a voice load of 2.5E with 1% blocking probability. [Figure 6] compares FIFO with the GA based technique only for data and [Figure 7] compares the performance with a voice load of 2.5E.

CONCLUSIONS

The results show the proposed GA based CAC and scheduling mechanism gives a reasonable efficiency for the data services irrespective of the service classes. Also this does not sacrifice the performances of the lower QoS services. The better performance is mainly due to the fitness calculation technique and the inclusion of refreshing frames concept considering dynamic nature of the traffic profile. The fitness function, which calculates the suitability for getting the resource, is designed to allocate resources among services in a fair manner. Moreover dynamic resource allocation among the services is achieved more practically and realistically with the introduction of refreshing frames concept. This novel concept GA based call admission control and scheduling algorithm gives a better control on resource allocation compared to the existing methods.

REFERENCES

[1] W. Leland, et al., "On the self-similar nature of Ethernet traffic", IEEE/ACM, Transaction on Networking, 2(1), pp. 1-15, Feb. 1994

[2] P. Goyal, H. M. Vin and H. Cheng, "Start Time Fair Queuing: A Scheduling Algorithm for Integrated Services Packet Switching Networks", Proceedings of SIGCOMM'96, 1996.

[3] L. Zhang, "Virtual Clock: A New Traffic Control Algorithm for Packet Switching Networks", Proceedings of ACM SIGCOMM'90, pp. 19-29, September 1990.

[4] Q. Pang, A. Bigloo, V. C. Leung and C. Scholefield, "Service Scheduling for GPRS Service Classes", Proceedings of WCNC 99, New Orleans, LA, September 1999.

[5] S. Thilakawardana and Rahim Tafazolli, "Effect of Service Modelling on Medium Access Control Performance", PIMRC 2001, San Diego, USA, September 2001

[6] J. H. Holland, "Adaptation in Natural and Artificial Systems", University of Michigan Press, Ann Arbor, MI, 1975.

[7] ETSI GSM 03.04 Standard, "Digital Cellular Telecommunications System (Phase 2+); Overall Description of the GPRS Radio Interface", ETSI.

[8] 3rd Generation Partnership Project, "General Packet Radio Service (Release 1999); Service Description Stage I", 3G TS 22.060, March 2000.

[9] G. Brasche, B. Walke, "Concepts, services, and protocols of the new GSM phase 2+ general packet radio service," IEEE Communications Magazine, vol. 35, no. 8, pp. 94-104, August 1997.

[10] S. Tekiny, "A Measurement Based Prioritisation Scheme for Handovers in Mobile Cellular Networks", IEEE Journal on Selected Areas in Communications, vol. 10, no. 8, October 1992.

[11] R. Guerin, "Queuing Blocking System with Two Arrival Streams and Guard Channels", IEEE Transactions on Communications, vol. 36, pp.153-163, 1988.

APPENDIX A: DYNAMIC HANDOVER SCHEME USING GENETIC ALGORITHMS FOR MOBILE CELLULAR NETWORKS

A.1 Introduction

Objective is to improve the perceived QoS of cellular service by minimizing both probability of forced termination of ongoing calls due to handover failures and degradation in the spectrum utilization. Both new and handover calls are queued until a channel is available. The waiting time in the queue depends on the handover margin and cell residence time. Once a call needed to be handed over an expiry time is generated. The proposed handover technique based on GA is compared with available schemes [10].

A.2 Proposed Mechanism

The probability of handover failure, reflected by the probability of forced termination of calls, is a major criterion in performance evaluation of cellular systems. In non-prioritized call traffic handling schemes, handover requests are treated in the same manner as originating calls thus of handover failure probability equals the probability of call blocking. Proposed scheme considers queuing of handover and new calls at the admission and maximum queuing delay depends on the handover algorithm. For power based handover algorithms maximum queuing delay depends on the *degradation level* [10]. For distance based handover algorithms maximum delay in the queue is proportional to the *handover threshold* and frequency of *handover retry*. The serving of handover and new call queues are based on a dynamic resource allocation technique based on GA resulting as an extension of the dynamic GA based scheduling and admission control algorithm.

A.3 Mapping the Problem to the GA and Methodology

The length of the chromosome is same as the number of resources (channels in this case) available for assignment of new or handover calls. Fitness function reflects the *Grade of Service (GoS)* of new and handover calls. The concept of *GoS* comes from the *Erlang B*. *Erlang B* is generally used for dimensioning cellular networks. In mobile cellular networks besides blocking probability, call dropping probability affects the *GoS*, which is due to the mobility of users. This is reflected in the new definition of *GoS*, which is more specific to mobile communications. Therefore the *GoS* in mobile cellular networks can be defined as:

$$GoS = Pb + w * Pd \qquad \text{(Equation A.1)}$$

Where P_b is the blocking probability from *Erlang B*, P_d is the call dropping probability and w is the weighting factor. In practical situations w takes a value of 10. Reducing the blocking probability requires a good system plan and sufficient number of radio channels. Therefore the fitness is a measurement of *GoS*. Lower the overall *GoS* the system is better resulting from less call blocking and call dropping probabilities. Therefore the fitness function needs to be minimized for the case of optimum solutions. Apart from *GoS* the dynamic nature of the solution is maintained using dynamic queue length (d), and the fairness among new calls and handover calls is maintained with the use of allocated channels (f) for each call type. Therefore the fitness of each chromosome (C_F) can be calculated with the use of the fitness function as in [Equation A.2]. N is the number of channels/resources available and (d_{type}/Q_{type}) is the cost ratio or the cost of serving each queue. W_{type} is the weighting factor where for new calls this value is 1 and for handover calls this is w. Frequency of allocated resources for each type of calls (i.e., new calls or handover calls) is given as f_{type}.

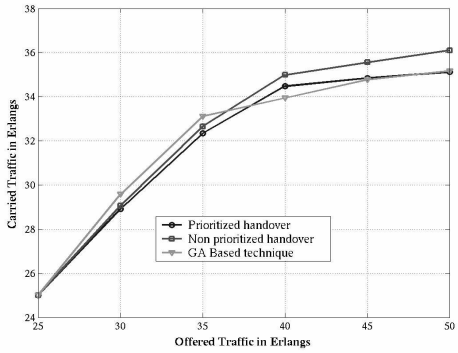
$$C_F = \sum_{i=1}^{i=N} F_i \quad where \quad F_i = \left(\frac{d_{type}}{Q_{type}} \right) * \frac{w_{type}}{\sqrt{f_{type}}}$$

(Equation A.2)

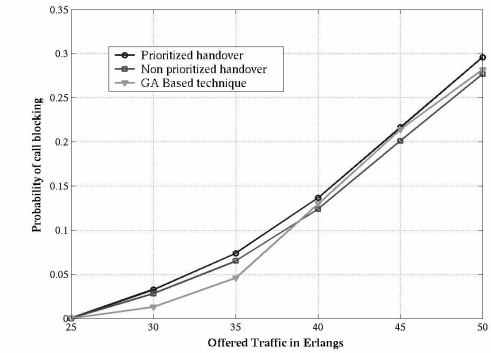
A.4 Performance Evaluation

The performance evaluation and comparison of the proposed GA based dynamic techniques with the non-prioritized call handling policy and FIFO queuing of handover requests presented in this section. The performance parameters measured are: probability of call blocking, probability of forced termination and the ratio of carried traffic to total offered traffic. Calls initiated within the cell are assumed to arrive at a *Poisson* rate, which is varied to obtain different traffic loads. Handover request arrivals also follow a *Poisson distribution* whose rate is input to the simulation. The fraction of the total traffic due to handovers is kept fixed while the total offered traffic is varied. The simulation has been run for the case where handovers account for 50% of the total traffic. Results presented are obtained assuming a fixed channel assignment strategy with the cell having a set of 50 channels. If all channels are occupied, new call arrivals and handover requests are queued until the next allocation of channels in the subsequent refreshing frame. In each refreshing frame depending on the number of available channels the allocation of new and handover calls are determined. The allocation is based on the fitness function described in [Equation A.2].

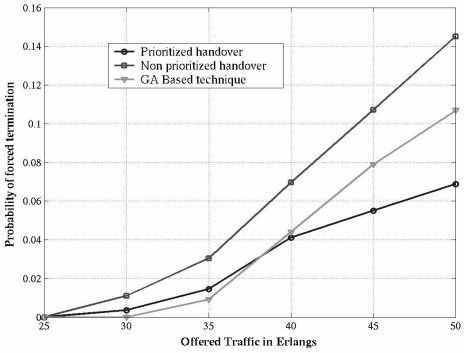
For the simulations refreshing frame duration is taken as 20 ms, (i.e.; in every 20 ms the available channels will be allocated accordingly). Channel occupancy times or channel holding times are drawn from an *Exponential* distribution for the simplicity of the model. Considering the memory less property of the *Exponential* distribution the only effect of handovers on the channel holding time distribution is that the mean duration value is less than that of the call duration distribution [11]. For new calls the average channel holding time is assumed as 100s whereas for handover calls it was taken as 60s. The maximum delay any type of call can endure in the queue was considered as a uniformly distributed random variable with minimum of 25ms to the maximum of 100ms. Q_{type} ratio of handover calls to new calls is 1000 to 1. Also the weighting factor w is 10 for handover calls and 1 for new calls. The following are the comparisons between the proposed scheme, prioritized and non-prioritized schemes [Figure A.1] to [Figure A.3].



[Figure A.1] Carried Vs Offered Traffic



[Figure A.2] Probability of Call Blocking



[Figure A.3] Probability of Termination

A.5 Discussion of Results

It can be seen from the results dynamic handover criteria based on genetic algorithms is more adaptable than the available handover techniques in terms of performance measures such as: probability of call blocking, probability of forced termination, and the ratio of carried traffic to total offered load. It is evident from results that the proposed GA based scheme

manages to tune to the optimal allocation between probability of call blocking and probability of forced termination. This technique measure the perfect balance between the fairness of allocation of channels among new call and hand over calls and at the same time maintaining the minimum call blocking and call dropping rates. Therefore the system benefit from having higher carried to offered traffic ratio and lower call blocking and dropping probabilities simultaneously.