Efficient Call Admission Control and Scheduling Technique for GPRS Using Genetic Algorithms

Shyamalie Thilakawardana and Rahim Tafazolli

Mobile Communications Research Group, Centre for Communications Systems Research (CCSR),

University of Surrey, Guildford, GU2 7XH, UK

S.Thilakawardana@eim.surrey.ac.uk

Abstract – **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. 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 also 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 comparison of this technique on a GPRS system is evaluated against data services and also a traffic mix comprising voice and data.

I. INTRODUCTION

Present communication networks are dominated by data traffic such as WWW and Email, which are bursty in nature. They possess different burst length distributions compared to traditional Exponential traffic models. For example in the case of WWW browsing with Pareto distributed burst lengths resulting infinite variance causes the undesirable queuing behaviour exhibiting self similarity at the aggregate level. Once these distributions are compared they exhibit a wide range of variation in first and second order characteristics such as mean and standard deviation. This affects the queuing performance characteristics opposed to the traditional traffic [1]. Thus more accurate and efficient performance can be obtained using dynamic resource allocation techniques.

The admission control and scheduling mechanisms implemented for Exponential models such as first in first out (FIFO) or best effort can no longer be used for efficient performance on these service classes. 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]. 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. Therefore the objective of this work is to design a CAC and scheduling algorithm to allocate resources in a fair and efficient manner among diverse set of services satisfying the QoS agreements.

Section II looks in to the GPRS dara communication architecture followed by QoS management of GPRS in Section III. Section IV presents the problem of efficient allocation of resources among diverse set of service classes as an optimization challenge. Encoding this problem into a genetic algorithm environment is discussed in Section V. Section VI discuss the performance evaluation of the proposed technique over a GPRS system is compared with state of the art algorithms. Focus is mainly on the down link behaviour. Finally results comparison and performance improvement in the proposed scheme are discussed.

II. GPRS DATA COMMUNICATION ARCHITECTURE

The General Packet Radio Service (GPRS) designated to support packet oriented data transmission is an extension of the Global System for Mobile Communications (GSM). Regarding the offered service, GPRS allows the subscriber to send and receive data in an end-to-end packet transfer mode, without using any network resources in circuit switched mode. This allows for autonomous operation of GPRS and best fits the bursty traffic characteristics. Radio communication between the mobile station (MS) and the GPRS network covers physical and data link layer functionality. The physical layer provides services for information transfer over a physical channel between the MS and the network. These functions include data unit framing, data coding, and the detection and correction of physical medium transmission errors [9].

The data link layer has been separated into two distinct sublayers. The radio link control/medium access control (RLC/MAC) mediates access to the shared medium between

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.

Authorized licensed use limited to: University of Surrey. Downloaded on May 17,2010 at 15:35:54 UTC from IEEE Xplore. Restrictions apply.

multitudes of MSs and the network. The packets, which are received from the network layer, are transmitted across the air interface using the logical link control (LLC) protocol. The LLC layer operates above the MAC layer. An LLC frame in the RLC/MAC layer is segmented into radio blocks, which are formatted into bursts on the physical layer. The size of the block depends on the applied coding scheme. Each radio block comprises 4 normal bursts in consecutive TDMA frames as shown in Fig. 1.

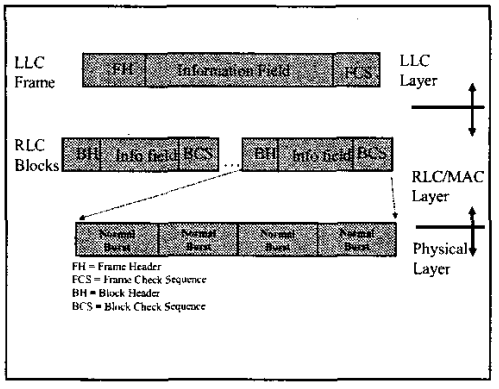

Figure 1. GPRS Radio Block Architecture

As a hybrid frequency division/time division system, GSM organizes radio transmissions by assigning carriers and time slots to logical channels. The frame duration is 4.615 ms, and each frame is divided into eight time slots. A cell that supports GPRS shall allocate one or more shared packet data channels, which are taken from the common pool of physical channels available to the cell and otherwise used for speech. A physical channel dedicated to packet data traffic is called a packet data channel. The need for efficient use of radio spectrum requires dynamic change of the mix of speech and data channels. It is also possible to interrupt a data transmission to one MS if a high priority service is to be sent to some other MS.

III. QOS MANAGEMENT IN GPRS SYSTEM

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. In four delay classes listed two types of SDU sizes are specified (i.e., 128 and 1024 octets) [Table 1]. By combining these attributes many possible QoS profiles are defined. To determine delay requirements for different packet lengths [4] came up with the set of equations derived with interpolation techniques.

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

IV. PROBLEM DEFINITION

The problem of resource allocation in agreement with QoS profiles of services can be seen as efficient distribution of n service classes among g number of resources. 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. Since an optimum allocation presents a combination of services among resources this kind of problems are called *combinatorial optimization* problem. In this work the QoS profiles of service classes are reflected in their *QoS index*.

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 call admission control (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 factors and at the same time capturing the traffic profile of service classes. The dynamic factors such as queue length, and static factors QoS profile, fairness among services needs to be considered in designing the scheduling algorithms for future services. Data traffic such as WWW browsing, which are bursty in nature exhibits self similarity behaviour at the aggregate level [1]. 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.

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

Authorized licensed use limited to: University of Surrey. Downloaded on May 17,2010 at 15:35:54 UTC from IEEE Xplore. Restrictions apply.

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].

A. 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. Fig. 2 shows the chromosome representation or the packing order of 8 service classes (S_1 ... S_8) among 6 resources (R_1 ... R_6). In the GA environment the number of resources in the system determines the length of a chromosome. In Fig.2 $Chromosome_1$ and $Chromosome_2$ represent two different ways of service class allocation among 6 resources. The chromosome length is 6 or the number of free resources waiting in the resource pool for service allocation. In $Chromosome_1$ multiple resources are given to S_5 and in $Chromosome_2$ that is given to S_1. If there are n services requesting a resource, then there are more than one ways 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 in GA reflects the criteria for the optimum resource allocation. When considering optimum allocation limiting factors such as QoS profile, fairness among service classes need to be reflected. The fitness function decides the survivability of the best chromosomes thus deriving optimum solution in the GA environment.

Apart from the above factors dynamic traffic profile is considered to understand the real time problem caused by traffic characteristics. Thus introduces the "*Refreshing Frame*" in the solution phase. 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.

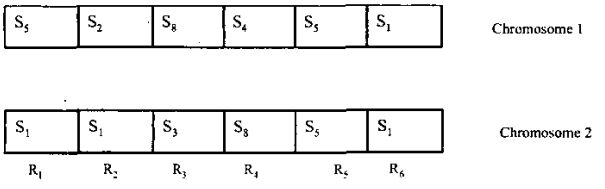

Figure 2. Chromosome Representation

V. FITNESS CRITERIA

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 of each service class. QoS parameters are graded according to their influence. For example in this work priority class has more weight than that of 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:

$$QoS_{index} = \frac{100}{p_1 * \sqrt{p_2}} \tag{1}$$

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

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

Next, fairness factor is considered in the scheduling mechanism. 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 (earliest deadline/delay first). The fitness decreases when the same service class request for more than one resources thus avoiding the exploitation. This is maintained in the fitness function evaluation with the introduction of 'resource frequency' (f_i).

Therefore by considering the above three factors in fitness evaluation the fitness of a service class (F_i) can be presented as;

$$F_i = K' \frac{Q_i * q_i}{\sqrt{f_i}} \tag{2}$$

Where K' is a proportionality constant

From (2) fitness of the chromosome structure (C_F) in Fig. 2 is the summation of service fitness included in the chromosome which can be represented as;

Authorized licensed use limited to: University of Surrey. Downloaded on May 17,2010 at 15:35:54 UTC from IEEE Xplore. Restrictions apply.

$$C_F = K \sum_{i=1}^{i=g} F_i \quad (3)$$

Where K is a constant and g is the number of resources in the resource pool (same as the chromosome length). The value of K is same for chromosomes of the same length but different when comparing chromosomes of varying length. In this research in down link scheduling algorithm a fixed chromosome length is used. If Service Class S_i is one of the service classes in the chromosome value of F_i is calculated from (2). If not value of $F_i = 0$ (i.e.; S_i does not contribute to the chromosome fitness C_F).

VI. Traffic Sources

Traffic consist of GPRS applications, which includes 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 for Email sessions is 40 % where as Railway and Mobitex traffic each presenting 20 % of the traffic mix. In this work focus is mainly on the down link performance.

Total 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]. The probability distribution function of Email connection sizes can be approximated by a truncated Cauchy (α = 0.8 and β = 1) distribution with a maximum message size of 10 Kbytes. The probability density function of the Cauchy distribution is as of (4).

$$f(x:\alpha,\beta) = \frac{\beta}{\pi(\beta + (x-\alpha)^2)} \quad (4)$$

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.

In a WWW browsing session a packet call corresponds to the downloading of a WWW document. After the document is entirely arrived to the terminal, the user is consuming certain amount of time for studying the information. This time interval is called the *reading time*. Hence a typical behaviour of a WWW browsing model is based on distributions described by session arrival process, number of packet calls per session, reading time between packet calls, number of bursts within a packet call, inter arrival time between bursts and the size of the burst. The modeling of WWW service application follows a Pareto burst size distribution (α = 1.1 and β = 81.5), with maximum burst size of 66666 bytes [5]. The probability density function of the Pareto distribution is as of (8).

$$f(x:\alpha,\beta) = \frac{\alpha.\beta^{\alpha}}{x^{\alpha+1}} \quad (5)$$

The average burst size of WWW browsing is 480 bytes.

Apart from the above applications uniformly and exponentially distributed packet sizes of the Mobitex and Rail data contribute towards the traffic mix [9].

VII. Performance Comparison

Comparisons are made with available techniques such as EDF and FIFO. Eight different service classes are considered with different QoS profiles. GPRS occupying a single carrier is considered. Each TDMA frame consists of eight time slots. Out of these eight slots one slot is allocated for signaling resulting for seven GPRS. QoS profile is based on the delay classes and precedence QoS parameters [Table 1]. The refreshing frame is selected as 200 frames where a GPRS frame is 18.46ms. Therefore in every ~ 4s time duration the optimum resource allocation is updated. Apart from the bursty data services a traffic mix comprising voice and data are also considered for comparison purposes. In this study voice calls are allocated to the resources as in GSM.

[Table 2] presents the service class classification of the traffic mix. Services are classified according to two QoS parameters namely delay and priority of a service. These parameters are taken in to consideration when calculating QoS index for the GA scheduling algorithm. Therefore [Table 2] categorize the 8 service classes.

Furthermore as a quantifiable performance measurement among different mechanisms, [4] introduced the comparison of performance between *average normalized delays*. 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 of the queuing system consists of variable packet sizes and different delay classes. If the QoS profiles are met satisfactorily this value is below 1.

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

VIII. Discussion of Results

Comparison of performances in terms of normalized delay against link utilization is analyzed.

Fig. 3 presents the normalized delay comparison for services of class I, denoting higher QoS profile classes, such as WWW Class I, Email Class I.

Authorized licensed use limited to: University of Surrey. Downloaded on May 17,2010 at 15:35:54 UTC from IEEE Xplore. Restrictions apply.

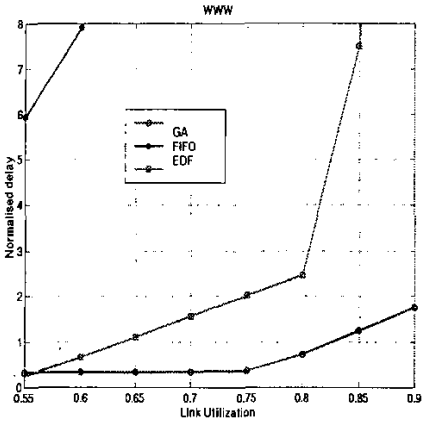

Figure 3. Performance Comparison for WWW Class I

Once the throughput is higher or the utilization of the link increases the normalized delay also get increased. Fig. 4 presents the performance comparison among the three different techniques for the services of class II. This follows Fig. 5 presenting that of services of class III. It is evident from these results for higher QoS service classes GA based scheduling mechanism outperforms the available algorithms such as EDF and FIFO.

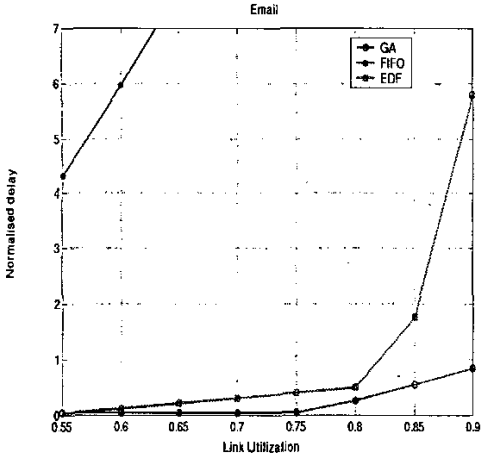

Figure 4. Performance Comparison for Email Class II

It is evident from the results GA based dynamic scheduling technique has better performance compared to EDF and FIFO. Finally performance is compared for a traffic mix of voice calls and data services. Voice calls having the highest priority get blocked when all the available channels are carried by voice calls. Also data channels will be pre-empted if there are no free channels available to carry the voice call. The dropping policy for data packets is two fold. If a data packet is pre-empted for voice the data packet will be dropped. Also if data packet exceeds the delay profile once waiting in the queue that packet will be dropped.

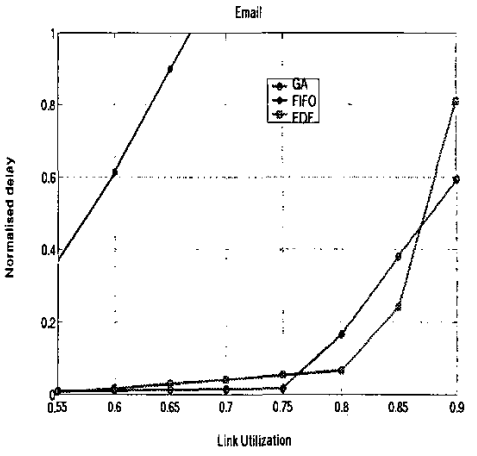

Figure 5. Performance Comparison Email Class III

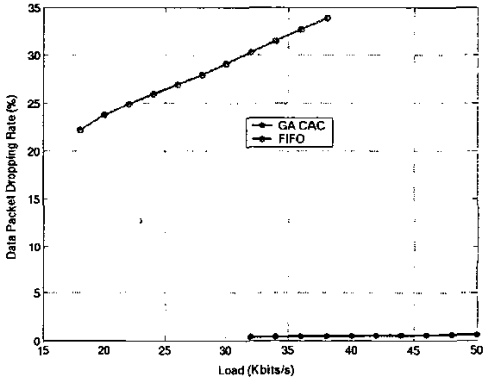

Figure 6. Performance Comparison of Data packet dropping rate (Data only)

From Fig. 6 to Fig. 7 the performance comparison between the admission control mechanisms is presented in terms of data packet dropping rate against the load. It is experimented with a voice load of 2.5E under 1% blocking probability. Fig. 6 compares the FIFO with the GA based technique only for data services. The better performance of the GA based mechanism is evident. Fig. 7 compares the same with a mix of data and 2.5E voice load.

Authorized licensed use limited to: University of Surrey. Downloaded on May 17,2010 at 15:35:54 UTC from IEEE Xplore. Restrictions apply.

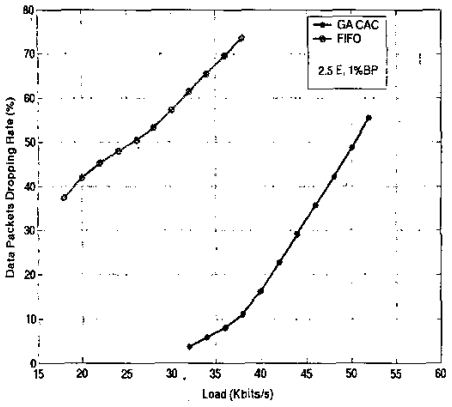

Figure 7. Performance Comparison Data packet Dropping rate (Data + Voice)

IX. CONCLUSIONS

From the simulation results it can be deduced the proposed GA based CAC and scheduling mechanism gives a reasonable efficiency for the data services irrespective of the service classes. Also at the same time this does not sacrifice the performances of the lower QoS profile services. The increase of performance compared to other available mechanisms is mainly due to the fitness calculation technique and the inclusion of refreshing frame concept considering dynamic nature of the traffic profile.

This novel admission control and scheduling criteria is based on 3 different parameters, embedded in the fitness function, addressing the multi dimensional nature of the problem. They are QoS index of the service class, dynamic queue length of each service class, and frequency of resources allocated for each service class. QoS index, which is a function of more than one QoS parameters, captures QoS profile of service classes more accurately and realistically. 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 frame concept. This novel concept captures the chaotic behaviour of traffic dynamically. The results demonstrate that the proposed algorithm achieves efficient resource allocation among diverse service classes for GPRS system. The proposed 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.

Authorized licensed use limited to: University of Surrey. Downloaded on May 17,2010 at 15:35:54 UTC from IEEE Xplore. Restrictions apply.