Radio Resource Management in GPRS with Quality of Service

Josep Bada, Ferran Casadevall, Javier Jiménez and Eduardo Medrano
Universitat Politècnica de Catalunya (UPC),
Dept. of Signal Theory and Communications, Campus Nord
Barcelona, Spain
Tel: +34 93 401 6524, email: ferranc@tsc.upc.es

ABSTRACT

In wireless data networks, Radio Resource Management (RRM) strategies are an important issue for supporting Quality of Service (QoS). This paper evaluates the performance of two innovative scheduling disciplines that allow the subscribers to transmit data with guaranteed QoS in a packet switched radio network based on the General Packet Radio Service (GPRS). In particular, the analysed strategies are: Modified Advance-Time (MAT) and Virtual Clock (VC). Both scheduling algorithms allow the GPRS system to achieve a high usage of the available radio resources while guaranteeing the quality of service. Realistic traffic models for Web browsing, e-mail, fleet management and Short Message Service are assumed.

From the obtained results, it could be concluded that both disciplines are suitable to guarantee the quality of service in GPRS network even for high load traffic conditions.

I. INTRODUCTION

In the last years, data transmission and mobile telephony technologies have emerged as the major driving force behind the new developments in the area of the telecommunications networks. GSM (Global System for Mobile Communications) has gone beyond the imagination of the most optimistic forecasters. At the same time, the continued growth of Internet increases the need for efficient wireless data communications systems.

For bursty traffic, such as the generated by Internet, it is well known that packet switched access provides better use of the radio resources in comparison to the circuit switched (e.g. GSM). The reason is that radio resources are used on demand while many users can share a single radio channel, due to statistical multiplexing capabilities of the packet networks.

GPRS, standardized by European Telecommunications Standards Institute (ETSI), is an on-demand based bearer data service, sharing the same GSM air interface, that offers quality of service (QoS) to the users by dynamic allocation of different amounts of resources depending on the user requirements (throughput, delay, reliability, priority, etc.). This system allows to a subscriber to send and receive data in an end-to-end packet transfer mode through the GSM infrastructure. The GPRS architecture provides a platform able to support a wide range of data services while requires few modifications in the air interface. To introduce GPRS in the existing GSM infrastructure, additional network elements are added to the GSM architecture, namely, serving GPRS support node (SGSN) and gateway GPRS support node (GGSN) [1]. The physical layer is not affected but a specific RLC/MAC [2] (Radio Link Control/Medium Access Control) protocol is introduced. Four coding schemes for error correction (CS-1, CS-2, CS-3 and CS-4) have been defined for adapting the transmission rate of the data blocks to the different radio propagation conditions.

In order to meet the different requirements of a wide variety of user applications, a number of Quality of Service profiles have been specified in [3]. In GPRS Release 99 there are five parameters to define the QoS profile, namely, precedence, reliability, rate (mean, maximum) and delay.

The Packet Data Protocol (PDP) context specified in GPRS Release 99 is defined as the information sets, allocated in the MS, SGSN and GGSN nodes, that are used to specify the tight relationship between an application of a mobile subscriber, a PDP type and one QoS profile. Several PDP contexts with different QoS parameters can share the same PDP address (Secondary PDP Context). To define a QoS contract between the mobile station (MS) and the network, PDP contexts containing QoS profiles are negotiated between the MS and the SGSN node. With the introduction of QoS concept, it is possible to use the network resources in a more efficient way. That is, the application data flows are managed according to their actual needs.

Even though, from the application viewpoint the QoS is end to end defined, in this paper only QoS issues related to the radio network segment are considered, because they represent the bottleneck of a typical GPRS data connection. In Release'99, the QoS profiles are characterized by the five parameters listed in Table 1.

<table>
<tr><th>Parameter</th><th colspan="6">Values</th></tr>
<tr><td>Precedence</td><td colspan="6">Hight,normal, low</td></tr>
<tr><td>Reliability</td><td colspan="6">Packet loss probability; duplicate, out of sequence</td></tr>
<tr><td rowspan="5">Delay</td><td>Size SDU</td><td>Class</td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td rowspan="2">128 (bytes)</td><td>Means)</td><td>< 0,5</td><td>< 5</td><td>< 50</td><td>Best Effort</td></tr>
<tr><td>95%(s)</td><td>< 1,5</td><td>< 25</td><td>< 250</td><td>Best Effort</td></tr>
<tr><td rowspan="2">1024 (bytes)</td><td>Mean(s)</td><td>< 2</td><td>< 15</td><td>< 75</td><td>Best Effort</td></tr>
<tr><td>95%(s)</td><td>< 7</td><td>< 75</td><td>< 375</td><td>Best Effort</td></tr>
<tr><td rowspan="2">Rate</td><td>Mean</td><td colspan="5" rowspan="2">Depend</td></tr>
<tr><td>Maximum</td></tr>
</table>

Table 1. -The QoS Profile

The paper is organized as follows:

- Section two describes several radio resource strategies suitable to be applied to the GPRS radio interface.

- Section three shows the performance results of the analysed scheduling disciplines. The results are shown in terms of throughput and delay (mean and percentile 95%) taking into account the network load conditions of the radio interface. E-mail, fleet management, WWW and SMS traffic flows are assumed in the analysis.
- Finally, the conclusions of the study are highlighted in section four.

II. SCHEDULING DISCIPLINES

The use of efficient algorithms for Radio Resource Management (RRM), which are not standardized because their specific realization is implementation dependent, becomes mandatory to guarantee the required QoS profiles and to optimise the system capacity in a GPRS multi-service environment.

However, simple scheduling strategies, like FIFO (First Input First Output) or Round Robin, do not guarantee the QoS requirements when the network load increases. For example, when a FIFO strategy is assumed, the Figure 1 shows that to cope with the QoS delay requirements (normalized delay lower than 1) the link utilization[1] must be lower than 30%. In this paper, the link utilization is defined as the degree of occupation of the available resources for an hour, (i.e. if all resources are occupied for an hour then the link utilization is 1 or 100%)

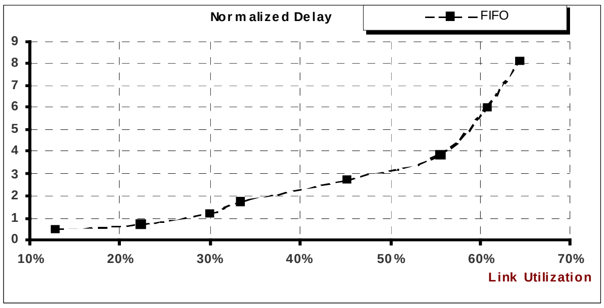

Figure 1. Normalized Delay FIFO

Therefore, the investigation in advanced scheduling techniques that allow guaranteeing the QoS parameters becomes mandatory. The work presented in this paper is based in the comparative performance evaluation of two service disciplines, namely, MAT (Modified Advanced Time) [4] and VC (Virtual Clock) [5] and [6], able to provide the requested QoS values. The analysed strategies were firstly studied in the context of the ATM service and they have been properly adapted to the GPRS service.

II.A. Modified Advanced-Time

The scheduling strategy denoted by MAT (Modified Advanced Time) is an adaptation of a previous proposal [4] so-called ATS (Advance-Time Scheduling). For each data flow, this scheduling strategy establishes the decision to transmit a given data block on the use of a index (Iu (t)) that reflects if the flow has over used or under used the assigned radio resources. The major advantages of the proposed scheduling strategy are:

- The available bandwidth is distributed between the different flows in proportion to their negotiated QoS.
- Its implementation simplicity

The expression (1) shows how the utilization index is updated taking into account its value in the previous scheduling instant. The expression defines the procedure for updating the utilization index for served and non-served flows as well as the initial conditions, $I_u(t_0)$. The assumed time granularity is 20ms, the time needed for transmitting a radio block. In the expression, "l_k" is the length, in bits, of the transmitted radio block; "d_{QoS}" is the negotiated bit rate value in bps, and "t_h" is a fixed value, denoted by historic time, which takes into account the memory of the algorithm.

A value of the utilization index less than 1 means that

$$I_u(t_{k+1})\big|_{flow\,j} = \begin{cases} 1 & for\, t_k = t_0 \\ \dfrac{t_k}{t_h+t_k}\cdot I_u(t_k)\big|_{flow\,j} & flow\,inactive\,t_k \\ \dfrac{t_k}{t_h+t_k}\cdot I_u(t_k)\big|_{flow\,j} + \dfrac{l_k}{(t_h+t_k)\cdot d_{QoS}} & flow\,active\,t_k \end{cases} \qquad (1)$$

the flow throughput is below the negotiated one. A bandwidth over-consumption happens when the value of the utilization index is greater than 1. From the expression, notice that the importance of the historic part (first term) increases as the time elapses, when compared to the current part (second term for the active flow). For instance, consider the case of a data flow that remains inactive for a long period of time. For this data flow, the utilization index value approaches zero. When new incoming packets appear in this data flow, then the very low value of its utilization index enables the flow to almost monopolize the transmission link. In this way, the flow recovers the unused portion of the agreed bandwidth. However, it is also clear that the historic part should have a limited weight in the expression in order to reduce the old traffic influence and to take the scheduling decisions also considering the recent traffic evolution. Usually t_h is assumed to be a fixed value, which is one or two orders of magnitude greater than the scheduling time

Then, each operative data flow in the system is queued and each queue has an associated index ($I_u(t)$). This index grows when the flow is active and decreases when it is inactive. At instant t_k, the flow to be served is the one that, having a radio block to be transmitted, has the lowest value of the utilization index.

Figure 2 shows an example of transmission scheduling based on the MAT service discipline. Six flows (three e-mail and three SMS sessions) with different QoS throughputs are assumed in the system. The figure shows the service order according to the index utilization of each flow. Notice that utilization index evolves around 1.

[1] The link utilization is directly related to the network load

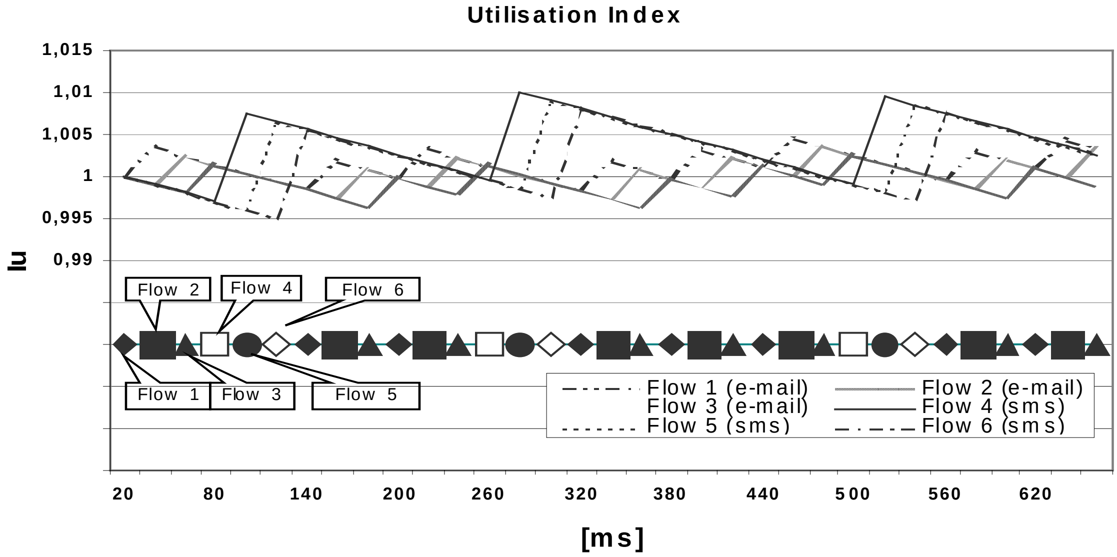

Figure 2. Evolution of the Utilization Index

II.B. Virtual Clock

The Virtual Clock (VC) scheduling strategy was proposed by L. Zhang [5] as a traffic control algorithm for high-speed packet switching networks. The VC strategy aims to emulate the behavior of an ideal time division multiplexing (TDM) system. For each packet a virtual timestamp is assigned. This virtual timestamp indicates the time that the packet would be transmitted assuming an ideal TDM strategy. The packets are transmitted in increasing order of their virtual timestamp.

The VC algorithm analyzed in this paper queues the incoming packets according to their virtual timestamp, called finish number, F(i). Every session (flow) has its own finish number. The finish number is determined using the negotiated end-to-end transmission rate, taking into account both the needs of applications and the available resources.

Each flow has its own clock that advances for every new incoming packet in the flow. At specified time intervals, the algorithm compares the virtual clock of the flows with the real system clock. If the virtual clock is faster, it means that the flow is sending too fast. In that case, the flow penalizes itself. The packets of this flow are queued at the end of the transmission queue and they would be eventually dropped.

When adapting the algorithm for GPRS, the finish number is calculated in a radio block basis. That is, the radio blocks stored in the queue are sorted according to their finish numbers.

Denote as f(i) the assigned bandwidth of session i and P(i,k,t) is the k-th radio block length of the flow i. The finish number of the k-th radio block allocated in the i flow at time t is defined by:

$$F(i,k,t) = \max\{F(i,k-1,t), real \quad time\} + \{P(i,k,t)/\phi_i\} \qquad (2)$$

where F(i,k-1,t) is the finish number of last transmitted radio block of the flow, which is compared with the real time for flow rate testing purposes.

III. SIMULATION RESULTS

In order to examine the performance of the proposed scheduling algorithms we have created a GPRS software simulator, using a MATLAB (Matrix Laboratory) simulation tool. Four different types of traffic sources, namely, e-mail, WWW-WAP, fleet management and short message applications are considered in our analysis.

The model used for e-mail application is described in reference [8], and it considers that incoming user messages are stored in a dedicated E-mail server. Choi and Limb presented a comprehensive model for Web traffic, called the “behavioral model”, [9]. In our study, we have adapted this model modifying the mean value of the IP packet length to emulate the comportment of WAP (Wireless Application Part) in the context of a GPRS system. The Fleet Management model is based on statistics obtained from fleet management applications using the Mobitex wireless packet data network in Sweden. The average message length on the uplink is 30 bytes and 115 bytes on the downlink [7]. Finally, the SMS, which is one of the most popular services in the actual GSM system, has been also included in our study. For SMS traffic, the length of the unit short message is limited to 140 bytes in order to take into account the restrictions in the Mobile Application Part (MAP) signaling layer. For all the models, the packet or session inter-arrival time distribution is assumed to be exponentially distributed.

In this paper only the downlink performance have been analyzed. It is assumed that there are 2 time slots (TS) exclusively dedicated for GPRS system in the cell. The maximum number of simultaneous active mobile stations is 32. The mobiles terminals allocate (2:1) TS per connection. No IP header compression is used within the SNDCP. Service scheduling is performed at radio blocks level in the RLC/MAC layer. It is assumed that this layer operates in acknowledge mode with a maximum number of retransmissions equal three. On the other hand, the LLC layer operates in non-acknowledge mode with maximum payload size of 1520 bytes. A TU50-noFH radio channel is assumed and the used channel coding is CS-1. All the MS are random distributed in the radio cell (R=3 Km). The time simulation is one hour. Table 2 depicts the values for QoS and traffic specified.

Services	Direction	Delay Class	Portion
WWW[2]	Downlink	QoS 2	20%
E-mail	Downlink	QoS 1	27%
Mobitex	Downlink	QoS 1	40%
SMS	Downlink	QoS 2	13%

Table 1. Simulation scenario

The assigned flow bandwidth depends on the type of the assumed QoS. The total link output bandwidth is distributed among all different classes. We assume that the negotiated rate for class (I) is 50% of available bandwidth, while for class (II) is 30% and the rest is for class III and Best Effort.

Figure 3 shows the evolution of the normalized delay, [7], versus the link utilization for both proposed scheduling algorithms. Notice that for both algorithms the QoS profile is guaranteed, because they could achieve up to the 90% of link utilization while meeting the delay requirements. When link utilization increases, the main difference between both scheduling algorithms is that VC takes in consideration the history of all flow, while in the MAT case the utilization index decrease when the flow is inactive. Therefore, after a long inactivity period, when a new packet arrives at the flow it is the firstly served. When the network traffic increases, QoS_2 is obviously penalized and for VC strategy this event appears earlier. In summary, for the simulated environment, it could be concluded that the main system limitation is the maximum number of simultaneous TFIs (Temporary Flow Identity) available rather than the packet delays, even for high load traffic conditions.

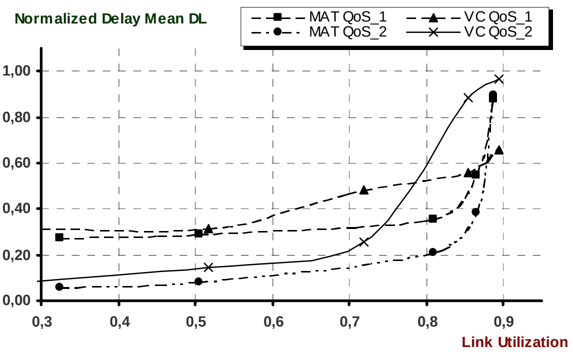

Figure 3. Normalized delay downlink

For both MAT and VC strategies, Figure 4 shows the evolution of the throughput versus the total number of sessions/hour (l). Here, the throughput is defined for each user packet as the number of bits divided by the transmission time (queuing time + air interface transmission). At the end of simulation the mean throughput is calculated for all application packets. Besides, the possible retransmissions at the RLC/MAC level are not taken into consideration. From the figure, it can be concluded that the VC algorithm provides a soft throughput degradation[3] when the traffic load increases. This is due because every connection has its independent finish number and, therefore, a packet from a given flow is served if it has the lowest finish number. This also explains the better throughput of the e-mail service because the e-mail packets are smaller that the Web ones. However, for the MAT case, the throughput significantly decreases when the traffic increases, falling to values much lower than 10 Kbps for the Web service.

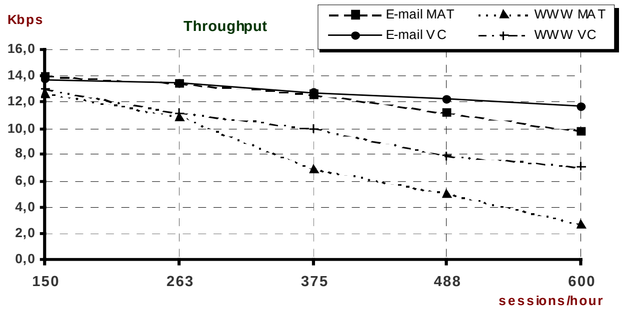

Figure 4. Throughput E-mail and WWW

For MAT strategy, Figure 5 shows as the throughput values for SMS and Fleet Management (Mobitex)[4] applications are maintained almost constant although load increases. For this type of applications, there is only one packet per session (maximum 8 RB), and as a result the transmission starts and finishes very quickly once the packet arrives to the transmission queue. A similar result could be observed when the VC strategy is considered.

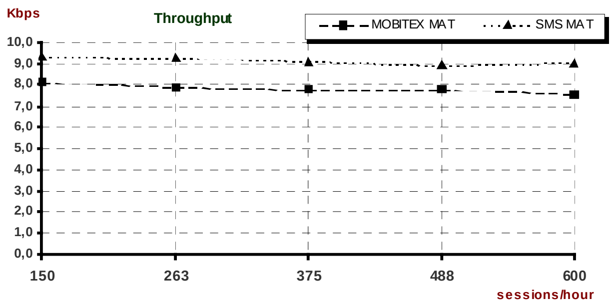

Figure 5. Throughput for SMS and Mobitex

On the other hand, it is interesting to compare the behaviour of these two strategies in terms of the mean waiting time to transfer the first radio block. Assuming QoS_1, Figure 6 shows the delay of the first radio-block for both MAT and VC strategies. Notice as that the delay for the VC strategy is higher than the obtained for MAT. This is due to the evolution of the utilization index parameter in the MAT strategy. Certainly, for every inactive flow, the MAT algorithm decreases its flow index (equivalent to the timestamp). Then when a new packet of a given flow arrives to the

[2] For the WWW the Get Request message has been considered in the uplink direction

[3] Decrease only 1,5 Kbps in the link utilization evolution QoS 1 and 5 Kbps for QoS 2

[4] For SMS and Mobitex the link utilization it is very small (under 1%).

PCU (Packet Control Unit), the first radio block of this flow will immediately be served, minimizing the delay. However, for a VC strategy when a new packet of a given flow arrives to the PCU, the timestamp, which finally will fix the delay for transmitting the first radio block, becomes affected by previous transmissions of this flow.

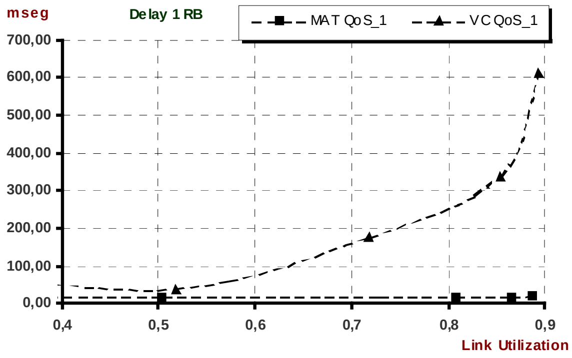

Figure 1. Delay of the first radio-block

To obtain a complete characterization of the performance of both strategies the 95% delay, as defined by ETSI, has also analysed in this paper. For both strategies, Figure 7 shows the evolution of this parameter versus the network load (link utilization). Notice that the delay values are less restrictive than the mean delay values and therefore both strategies perform much better in comparison to Figure 4. When link utilization is low (< 50%) both scheduling algorithms perform in a similar way, however when the link utilization increases the MAT strategy is slightly better than the VC strategy because MAT takes into consideration the previous evolution of the flow.

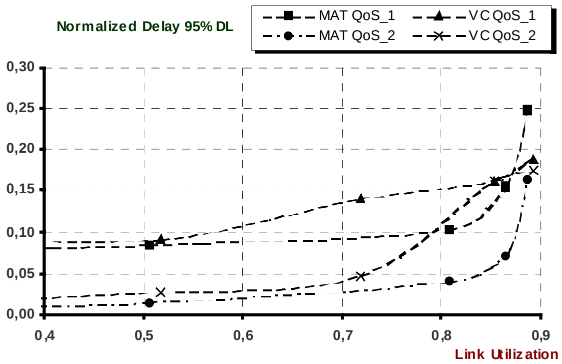

Figure 2. Normalized delay 95% downlink

Finally, for each pattern of traffic and both scheduling strategies, the maximum number of the sessions/hour admissible meeting QoS delay is reported in table 2.

IV. CONCLUSIONS

In this paper, and for the downlink of a GPRS system, the performances of two new scheduling methods, which allow guaranteeing the QoS, have been evaluated. The simulation results show that both VC and MAT scheduling strategies are very fair and they perform well enough in the context of a GPRS system with heterogeneous traffic, even for high network load conditions. Moreover, the implementation complexity is low in both cases.

Discipline	VC	MAT
Traffic	λ (CS1)	λ (CS1)
WWW	< 120	< 130
E-mail	< 160	< 175
SMS	< 80	< 90
Mobitex	< 240	< 250

Table 2. Comparative sessions/hour for MAT and VC strategies

ACKNOWLEDGEMENTS

This work has been carried out in the framework of the research project TIC2001-2222, founded by the Spanish Science and Technology Ministry

REFERENCES

[1] Digital cellular telecommunications system (Phase 2+) Universal Mobile Telecommunications System (UMTS), General Packet Radio Service (GPRS); Service description; stage 2; 3GPP TS 23.060 version 3.6.0 Release 1999.

[2] Digital cellular telecommunications system (Phase 2+);General Packet Radio Service (GPRS); Mobile Station (MS) - Base Station System (BSS) interface; Radio Link Control/Medium Access Control (RLC/MAC) protocol, GSM 04.60 version 8.4.1 Release 1999

[3] Digital cellular telecommunications system (Phase 2+) Universal Mobile Telecommunications System (UMTS), General Packet Radio Service (GPRS); Service description; stage 1, 3GPP TS 22.060 version 3.5.0 Release 1999.

[4] E.Monteiro, F. Boavida, V. de Freitas, "A Fairness Analysis of LAN/WAN Protocol Relays", Proceedings of the 4th Joint European Networking Conference, European Research Networking in a Global Context , May 10-13, 1993, Trondheim, Noruega,

[5] L. Zhang, “Virtual Clock: A New Traffic Control Algorithm for Packet Switching Networks”, Proceedings of SIGCOMM'90, pp.19-29, Sep. 1990.

[6] H. Zhang, “Service disciplines for guaranteed performance service in packet-switching network”, Proceedings of the IEEE, vol. 83, no 10, pp. 1374-96, Oct. 1995.

[7] Pang Q., Bigloo A., Leung V. Sccholefield C., “Service Scheduling for General Packet Radio Service”, IEEE Wireless Communications and Networking Conference 1999, September 1999, New Orleans, pp. 1229-1232

[8] H. Joseph, Z. Yixin and M. Seshu “Throughput and Buffer Analysis for GSM General Packet Radio Service (GPRS)” IEEE Wireless Communications and Networking Conference 1999, September 1999, New Orleans, pp. 1427-1431

[9] Choi, H., Limb, J. A behavioral model of web traffic. ICNP 99, Sep 1999