

NOVEL PRIORITISED EGPRS MEDIUM ACCESS REGIME FOR REDUCED FILE TRANSFER DELAY DURING CONGESTED PERIODS

D. Todinca, P. Perry and J. Murphy
Dublin City University, Ireland

ABSTRACT

The goal of this paper is to investigate the efficiency of different algorithms used for resource allocation in data transfer over EGPRS networks. The focus is on the efficiency during congested periods in order to ensure reduced file transfer delay. A number of algorithms are presented for the resource allocation, and some generic mathematical results are presented for a two coding system. Simulation is relied on to produce results for the algorithms and it is found that Iterative Round Robin (IRR) and Oldest Queue (OQ) produce the best results. These are then proposed to be good candidates for implementing transmission control.

INTRODUCTION

The evolution from the voice based mobile telephone systems of today towards the third generation (3G) mobile telecommunications system will be based on the Global System for Mobile communications (GSM) Goodman (1) and will begin with the General Packet Radio Service (GPRS) Bettstetter et al (2), Gudding (3).

Enhanced Data-rates for Global Evolution (EDGE) offers an additional channel capacity, which will bring 3G packet switched data services to users of the EDGE systems. This final system is referred to as Enhanced GPRS (EGPRS) and is often regarded as part of the 3G family.

The allocation of radio resources in GPRS and EGPRS is an active research area and there are many papers that address this problem from different perspectives. In Foh et al (4) the authors build a complex mathematical model of the system that combines the packet data arrival traffic, the channel allocation between voice and data and the data packet transfer in GPRS. This approach is global, but the paper doesn't focus on the scheduling algorithms used for data transfer. Other works (Jiang et al (5), Pang et al (6), Ajib and Godlewski (7)) use simulations to investigate the algorithms used for resource allocation in GPRS or EGPRS networks. The work in (6) proposes a new scheduling method, the Earliest Deadline First (EDF), but that method appears to be rather complex to be efficiently implemented in GPRS (each packet receives a time stamp and the packets are scheduled according to their time stamp).

The authors of paper (7) use a detailed simulation model of the RLC and MAC layers involved in the GPRS transmission. They simulate two scheduling algorithms, the first one is executed at the connection establishment and it is used for the distribution of the traffic channels between mobile stations (MSs). The second algorithm is used to determine how to share the resources offered by such a traffic channel between the mobile stations that are connected to that channel. The paper (5) investigates the Weighted Fair Queueing (WFQ) algorithm for resource allocation, including the influence that TCP/IP has on the users' performance.

In the context of other works, we focus primarily on the resource allocation algorithms, rather than system details like the interaction with TCP/IP or network signaling. That is, we use relatively simple models for user behaviour and system performance.

THE PROBLEM

In this work we limit our investigations to a single cell with N users and a centralized controller, the Packet Control Unit (PCU). The PCU is the part of the Base Station Sub-system (BSS) which performs the arbitration mechanism to share the radio resources between users. The users can use nine different channel coding schemes, each one allowing a certain bit rate.

The users want either to send or to receive data and the PCU performs the arbitration mechanism in order to share the radio resources between the users. The network resources are represented by B parallel Packet Data Traffic Channels (PDTCH's) available every 20ms.

The users can have P levels of precedence and/or priority. The precedence level is assigned to a user according to his or her willingness to pay for a certain category of services, hence the precedence level will be constant in time, while the priority level of a user can change in time, being determined according to different criteria, for example according to the coding scheme of the user, which can change with the user's distance to the Base Station. The number of priority levels may vary, depending on the criteria used to determine the priority level.

From the N users the PCU will select a number of K users that will be allowed to send or receive data. We call them *active users*, and can have values in the range

$[I, N]$. When an active user has finished sending a file, its place will be taken by another user, from the $(N-K)$ users waiting to become active.

The existing problem can be split into two sub-problems: *transmission control*, means sharing the B PDTCH's between the K active users and *admission control*, means selecting the K active users from the N existing users. The admission control problem includes selecting an appropriate number of active users. The PCU will perform algorithms for admission control and for transmission control.

Transmission Control

Transmission control is implemented at the Medium Access Control (MAC) and Radio Link Control (RLC) level. At these levels the information available for the PCU are the number of waiting data blocks for each user and the priority and precedence levels of users. The algorithms used for transmission control must ensure low delay and high throughput and they have to be simple, fast and easy to implement, because they must provide results every 20ms. The transmission control algorithms should have the capability to implement priority and precedence levels in order to ensure different levels of benefit for different categories of users or services. This is important for implementing Quality of Service (QoS) over EGPRS networks Karagiannis and Heijenk (8), Rexhepi (9).

Admission Control

Although the problem of admission control is a more complex problem than transmission control, it is less constrained in time. Consequently, fast, simple algorithms are required for transmission control. Selecting the number of active users is part of the admission control problem. The value of K will influence the waiting times experienced by users. A large value of K (close to N) will lead to a situation when the users don't wait too long until they are connected, but then the sending or receiving process will be slow, because the network resources are shared between many users. If K is small, then the active users will have good times for the sending or receiving process, but there will be many users that have to wait a long time in order to become active. It is also important to find a good distribution for the K active users, according to the precedence and priority levels. Given n_i users with the precedence/priority level i , ($i=1, 2, \dots, P$ and $n_1 + n_2 + \dots + n_p = N$), the problem is choosing k_1, k_2, \dots, k_p such that $K = k_1 + k_2 + \dots + k_p$, where k_i corresponds to the i^{th} level of precedence or priority.

Combining the two sub-problems we aim to efficiently implement Quality of Service (QoS) over EGPRS networks and to obtain low delay for users and a high throughput for network.

QUEUING THEORY FOR GPRS

Based on the assumption that a single GPRS cell can be adequately modeled as an M/M/1 system we derive some simple relationships. As a first approximation we assume that the two queues, for each of the CS rates, can be treated independently.

$$\text{Delay for CS1 Users} = E[t_1] = 1/(\mu_1 - \lambda_1) \quad (1)$$

$$\text{Delay for CS2 Users} = E[t_2] = 1/(\mu_2 - \lambda_2) \quad (2)$$

We define the service rate of the system to be the rate of throughput of TCP packets, μ_i (packets/second). The service rate of the queue is proportional to the number of slots that are given to each user. For the algorithms that are used in this paper there is a weighting factor X , which allows the CS2 user to have priority over the CS1 users, by allowing them to transmit X slots before the CS1 users can transmit 1 slot.

$$\text{Service rate for CS1 Users} = \mu_1 = F_n \{X/(X+1)\} \quad (3)$$

$$\text{Service rate for CS2 Users} = \mu_2 = F_n \{1/X\} \quad (4)$$

Each user generates the same input load, therefore the average arrival rate for CS1 or CS2 users is proportional to the number of active users. N_2 is the number of users using CS2 and N is the total number of users.

$$\text{Arrival rate for CS1 Users} = \lambda_1 = F_n \{N - N_2\} \quad (5)$$

$$\text{Arrival rate for CS2 Users} = \lambda_2 = F_n \{N_2\} \quad (6)$$

Combining the equation (3) and (5) into equation (1) we get an equation for the average delay for CS1 users. Similarly for CS2 users we can use equations (4) and (6) into equation (2).

$$E[t_1] = \frac{1}{F_n (1/X) + F_n (N_2)} \quad (7)$$

$$E[t_2] = \frac{1}{F_n (X/X + 1) - F_n (N_2)} \quad (8)$$

It is possible to analytically predict the values for the users total delay in certain conditions, for example if the degree of correlation between users is high (all users start sending data in the same time).

OUR MODEL

The tool used for modeling our problem is SES/*workbench*, from Hyperformix. Our model consists of SES nodes corresponding to the users and to the controller, every user having a corresponding SES node. The controller node corresponds to the Packet Control Unit and has two parts, an Admission Controller and a Transmission Controller.

We also included a *network conditions node*, which models the modification of network resources used for

data transfer (the parameter B) due to the sharing of network resources between voice transfer and data transfer. A *user mobility node* describes the user characteristic that, in this case, is simply the coding scheme used for changing the bit rate of a user.

The number of users in the system, N , is a parameter in our SES model. The model for a user is shown in Figure 1 and consists of a file generator node, a file buffer, a node used to split a file into data blocks and a data packet buffer node.

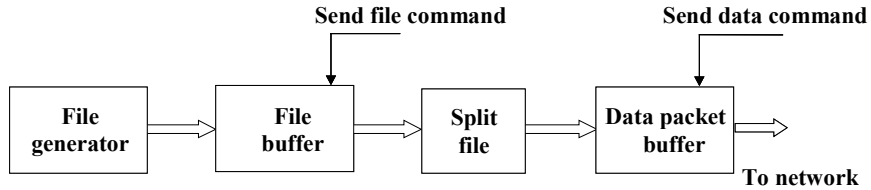


Figure 1. The SES model for a user

At certain time intervals a user can create a number of files, each file having a length denoted F_L . The period of the file creation process is randomly distributed around the parameter *mean_sending_interval*. A user attempts to send the files that it has created. The sending order is usually the order in which files have been created, but this order can be changed according to different criteria, for example the file type. The file that is currently being sent is put in the data packet buffer, while the other files are queued in the file buffer until the first file is completely sent.

To simplify the model, files are directly split into data blocks at the Radio Link Controller (RLC), so we do not model any of the higher layers in the system. The length of a data block is determined by the coding scheme used. When a user is allowed to send data, it sends a number of one or more data blocks.

The PCU uses to an arbitration algorithm to decide which users can send data and how many data blocks each of them can send during the current block period of 20ms. The decision is based on the state of the users and on the priority and precedence levels of the users. A maximum of eight data blocks can be sent in a block period on eight parallel Packet Data Traffic Channels (PDTCH's). As mentioned before, the *network conditions node* determines the number of PDTCH's available at a moment, which depends on the voice traffic and the policy used in the network to share resources between voice and data. The aim of this paper is to investigate different arbitration algorithms and to compare their performance for a given user activity scenario. By performance we mean the transfer delay, throughput and the probability of loss of data packets.

EXPERIMENTAL RESULTS

Conditions

In this paper we investigate the problem of transmission control and no admission control is performed and $N=K$. The network conditions do not change during simulation, the number of available PDTCH's always being eight ($B=8$), while the number of users is $N=K=10$ in our experiments.

To facilitate comparison of the different algorithms, we consider that the users can use only Coding Schemes 1

and 2 (CS1 and CS2). We also consider that the users are static, i.e. they do not change the coding scheme during simulation. We do not consider precedence levels for users in this paper. The priority level is based only on the coding scheme, which means that the users with coding scheme CS2 are high priority users and the users with CS1 have a low priority. Since priority is implemented using a weighting factor in each algorithm, this could equally well be used to provide the basis of QoS differentiation.

The network load is determined by the amount of data created by users per time unit. The user creates a *number of new files* at an interval given by a triangular probability distribution function centered on the value of the *mean_sending_interval*, which is 1000ms.

Algorithms

The algorithms investigated here are Iterative Round Robin (IRR), Oldest Queue (OQ), Longest Queue (LQ), Total Queue Length (TQL) and Total File Length (TFL).

In the first iteration of the IRR algorithm, each user receives a number of PDTCH's, according to its priority level: the low priority users will send only one data block, while a user with coding scheme CS2 will send a weighted number of data blocks. If there are more channels available after all users sent a number of data blocks, then the algorithm will perform another iteration, giving the remaining resources (channels) to the users that still have data blocks waiting in their queue. The algorithm stops when all the 8 available PDTCH's were allocated to the users or when all the users' queues are empty.

The other algorithms investigated in this paper (LQ, OQ, TQL and TFL) determine the highest value of a parameter (e.g. the queue length for LQ) and give all the available network resources to the user with the highest

value of the chosen parameter. If that user has less than 8 data blocks in its queue, then the next user receives the available resources and so on until there are no more resources available or all the users' queues are empty. For OQ the parameter for each user is the time elapsed since the user was served last time, for the LQ it is the queue length (the number of data blocks waiting in the data packet buffer). The TQL and TFL algorithms count the total amount of data that each user has in its two buffers (both the file buffer and the data packet buffer) expressed in number of data blocks for TQL, and respectively in bits for the TFL algorithm. In the case of high priority users, the weighting factor of the actual value of the chosen parameter is given by the parameter multiplied by the weight (e.g. effective Q length = $actual\ Q\ length \cdot weighting\ factor$), while for the low priority users the chosen parameter remains unchanged.

In our experiments the weighting factor, denoted by X , can have the values 8, 4, 2 and 1. If $X=1$, then no advantage is given to the high priority users, this case being considered only to compare the advantages given by larger weighting factors.

Results

The performance of the algorithms is expressed by different parameters, the most important being the total delay expected by users when they attempt to send a number of files. In our experiments the percentage of CS2 users among the all users is varied. As expected, the performance is better for both classes of users when the percentage of CS2 users is low and the weighting factor is high.

After all the CS2 users have sent a file, two situations are possible:

- The CS2 users continue sending files. In this case the sending delay for a CS1 user will have the same expression as for CS2 users with $X=1$. This case corresponds to the situation when the network load is 100% and the values for $N-N_2$ and N_2 are constant.
- All CS2 users had to send only one file.

For a *mean_sending_interval* of 1000ms, the following table presents the values of the file average sending delay for the CS2 users, for different network loads, when all users send only one file.

Load %	F 1	Average sending delay [ms]	
		X=4	X=8
40	16	140	90
60	24	220	150
80	32	300	210
100	40	380	270

In this case, the sending delay is the same as the total delay. While work is in progress to improve and extend the analytical models, we have to rely on simulations for more complex situations.

We varied the number of high priority users (CS2 users) from 1 to 10 and measured the average total delay

experienced by both high and low priority users. In Figure 2 (a), (b) and (c) the results for all the investigated algorithms for CS2 users are shown for a network load of 90%.

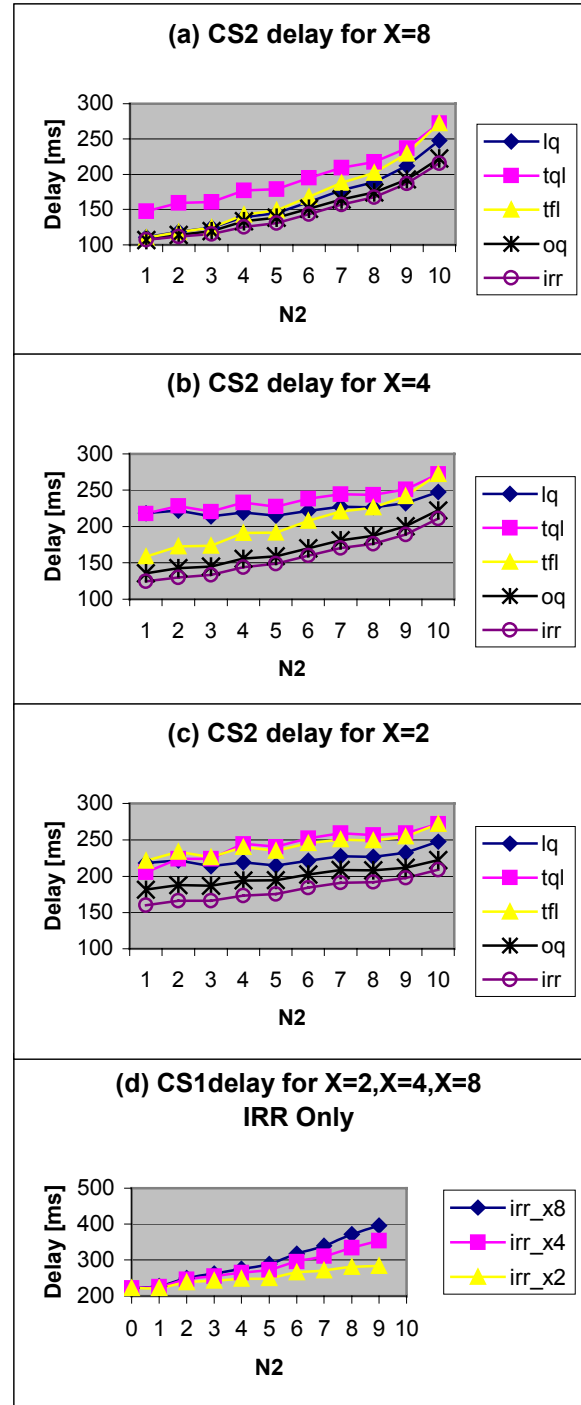


Figure 2: Delays for CS2 and CS1 against N_2

As can also be seen from Figure 2 (a), (b) and (c) the best results are obtained for IRR and OQ, followed by LQ, while TQL and TFL give poor performance. The IRR and OQ have the advantage that they are fair, that is the values for different users, within the same class of priority, are quite similar. This is not the case for LQ. The order of the algorithms is the same for network loads higher than 90%.

Figure 2 (d) shows the values for the CS1 delay for the IRR algorithm. This is plotted against the number of CS2 users for the relevant weighting factor X . The influence of the weighting factor is shown in Figure 3 for the IRR algorithm. The results for OQ are similar and we did not plot them. As expected, the performance increases with the value of the weighting factor and decreases when the percentage of high priority users increases.

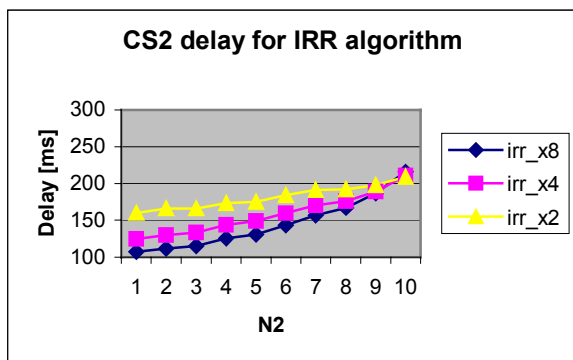


Figure 3: IRR Delay for CS2 Users

Based on equation (8) and the results shown in Figures 2 and 3, it is possible to see that they are of the correct shape. For CS2 users as X decreases the delay for CS2 should increase, and it does in the simulations. Also as N_2 increases then the delay for CS2 should increase. From equation (7) the delay for CS1 users as X decreases should also decrease and from Figure 2 (d) this can be seen to be the case. This is the opposite of what happens with CS2 users.

CONCLUSIONS AND FUTURE WORK

The best results are obtained for IRR and OQ, which are good candidates for implementing the problem of transmission control. Fair algorithms have better performance for the problem of transmission control, while the other algorithms can be useful for the problem of admission control. All algorithms investigated in this work can implement priority and precedence levels for users.

The ending points of the graphs in Figure 3 (when all users have high priority) are independent from the weighting factor X , which confirms the correctness of our simulation results. The slopes of the curves in Figure 3 depend on the value of X , which means that, changing the value of the weighting factor, we can obtain different levels of benefit for the high priority users. The different levels of benefit may be aligned with costs or service types. The fact that we can offer different service levels is important, because it means that we have a mechanism for implementing QoS.

In the current experiments the users are static, but in the future we will investigate a more detailed model, which

includes the user mobility, so that the coding scheme and hence the priority can change in time according to the distance to the BSS and other environmental conditions.

The work presented in this paper is part funded by a research grant from Enterprise Ireland (HE/2000/389) and Ericsson Systems Expertise, Ireland and part funded under the PRLTI Research Institute in Networks and Communications Engineering.

REFERENCES

1. Goodman D. J., 1997, "Wireless Personal Communications Systems", Addison-Wesley
2. Bettstetter C., Vogel H.-J. and Ebersacher J., 1999, "GSM Phase 2+ General Packet Radio Service GPRS: Architecture, Protocols, and Air Interface", IEEE Communications Surveys, Vol. 2, No. 3. URL: www.comsoc.org/pubs/surveys
3. Gudding H., 2000, "Capacity Analysis of GPRS. White Paper. Revised Edition of Master Thesis", Faculty of Electrical Engineering and Telecommunications, Norwegian Institute of Science and Technology. URL: www.mobileapplicationsinitiative.com/document/filer/capacity_analysis_gprs.pdf
4. Foh C.H., Meini B., Wydrowski B. & Zuckerman M., 2001, "Modeling & Performance Evaluation of GPRS", Proc. IEEE VTC 2001, Rhodes, Greece, 2108-2112
5. Jiang Z., J. Chang L.F. and Shankaranarayanan N.K., 2000, "Providing Multiple Service Classes For Bursty Data Traffic in Cellular Networks", Electronic Edition, in Proc. of IEEE INFOCOM 2000, Vol. 3, 1087-1096, URL: www.ieee-infocom.org/2000/papers/356.ps
6. Pang Q., Bigloo A., Leung V.C.M. and Scholefield, C., 1999, "Service scheduling for General Packet Radio Service classes", Proc. of IEEE Wireless Communications and Networking Conference WCNC'99, New Orleans, LA. URL: www.ece.ubc.ca/~vleung/conference_papers/wcnc99_pang.pdf
7. Ajib W. and Godlewski P., 2000, "Service Disciplines Performance for WWW Traffic in GPRS Systems", Proc. of IEE 3G 2000 Mobile Communications Technologies. URL: www.infres.enst.fr/~wajib/Papers/IEEMTS.pdf
8. Karagiannis G. and Heijnen G., 2000, "Technical Report University of Twente, TR-CTIT-00-29". URL: wwwhome.ctit.utwente.nl/~heijenk/publications.html
9. Rexhepi V., 2000, "Wireless Internet QoS, Master Thesis", University of Twente, Department of Computer Science and Electrical Engineering, URL: <http://www.ub.utwente.nl/webdocs/ctit/1/00000033.pdf>