A Scheduling Algorithm for Soft-QoS Guarantee in 3G Systems

O. Sallent, J. Pérez-Romero, R. Agustí, F. Casadevall
Universitat Politècnica de Catalunya
c/ Jordi Girona 1-3, 08034 Barcelona, Spain
email: [oriol, jorperez, ferranc, ramon @ tsc.upc.es]

Abstract

In the context of 3G systems a mix of services with different requirements are expected. Consequently, packet scheduling mechanisms for QoS guarantee will play a key role. This paper proposes a new scheduling strategy that makes consistent the target quality in the radio link with the priority level assigned to each user. The performance of such strategy is assessed by system level simulations.

1.- INTRODUCTION

One of the main goals of the future 3rd generation mobile communication systems will be the provision of different kinds of multimedia services with a certain degree of Quality of Service (QoS) that is to be guaranteed depending on the considered application and its interactivity requirements. In this framework, CDMA packet based networks, such as the considered in the UTRA proposal, provide an inherent flexibility to handle the provision of these services.

In any wireless system there is clearly a very scarce resource, which is the available bandwidth. 3G systems intend to provide several classes of services at different bit rates while assuring a QoS for each one. The only way to harmonise these two contradictory points (scarce bandwidth and a stringent QoS) is with a proper management of the available radio resources.

The Radio Resource Management (RRM) entity is the responsible for the correct use of the air interface resources in order to guarantee the quality to the offered services. For the UMTS system new and specific radio resource management algorithms related to load control, packet scheduling and admission control are required to guarantee QoS and to maximise the system throughput for mixed services with different bit rates and quality requirements.

In the last years a vast amount of literature has been produced concerning the proposals of architectures suitable for an adequate resource management with some kind of guaranteed QoS in fixed line environments, mostly related with Internet scenario [1, 2] so far only able to offer best-effort services. In wireless environments, the problem of QoS provisioning for multimedia traffic has also gained interest in the literature in recent years [3, 4]. However, little effort has been devoted to date in addressing the RRM topic to guarantee a given QoS in a packet driven environment such as the above mentioned in the framework of the 3G systems and in particular in the W-CDMA scenario selected in UMTS.

In the above framework, this paper focuses on the RRM sublayer, which in the context of 3G systems plays a key role. In order to gain more insight in how the packet scheduler impacts on the system performance, a proposal of a packet-oriented scheduling strategy for soft-QoS guarantee including new concepts in how the traffic flows are managed is presented and comparisons with other alternatives are carried out. In particular, section 2 presents the scheduling algorithm and section 3 includes the performance evaluation of the algorithm by means of simulation. Section 4 summarizes the conclusions reached.

2.- SCHEDULING STRATEGY

For multimedia traffic (voice, data, video) to be supported successfully, it is necessary to provide QoS guarantee. QoS requirements can be specified by many different parameters: transmission rate, delay or reliability being some of the most common. The QoS provisioning at the radio link layer means that the multimedia traffic should get predictable service from the wireless system. However, in packet radio communications several issues of random nature make this task especially difficult to achieve: packet generation from many different sources that must be multiplexed within a limited set of shared resources, variable propagation characteristics and others. Consequently, the policy followed by the scheduling algorithm should lead to a system behaviour as close as possible to the desired one.

Since QoS in mobile environments is hard to guarantee in strict sense, the concept of soft-QoS arises. Thus, the soft-QoS requirements for delay sensitive services can be established in terms of a certain desired delay bound and a certain percentage of packets arriving later than a given threshold. WWW service could be included in this category, whereas a delay tolerant service as e-mail could be served on a best effort basis.

The proposed RRM strategy, which applies on a frame by frame basis, focuses on soft-QoS guarantee and can be split in three different steps:

1) Prioritisation

All users intended to transmit information must be classified according to the type of service (first prioritisation level) and, for the same type of service, according to their QoS (second prioritisation level). This QoS is established in terms of the amount of information to be transmitted and the deadline or time remaining to keep the delay below the desired one. That is, for the same type of service, the smaller the deadline the higher the position in the serving queue will be. Taking these parameters into account, the RRM is in charge to decide which traffic flows should be transmitted first. As a result, a table containing all active users from higher to lower priority would be available.

For Non Real Time with Quality of Service requirement users, which is the focus of this paper, priority to the i-th user could be assigned according to the following function:

$$\phi_i = \begin{cases} \dfrac{L_i}{TO_i} & TO_i > 0 \\ L_i(2-TO_i)^n & TO_i \le 0 \end{cases} \qquad (1)$$

L_i being the packet length remaining to be transmitted and TO_i the time (in frames) left until the packet deadline for the considered service.

The above expression assigns higher priority to those users with either much information to transmit or much time already waiting for permission. When the delay threshold is overcome, the priority level is drastically increased.

2) Capacity requirement

This step is carried out user by user, from highest to lowest priority. In order to satisfy the current user QoS requirement, i.e. the user whom resources are tried to be allocated, some system capacity should be devoted to this user as well as certain transmission rate (or, equivalently, spreading factor SF) should be allocated to this user. Since there is some degree of freedom in the choice, it should be carried out in a smart way by the scheduler so that the overall system behaviour were as optimum as possible.

For WWW-like services there are many factors to consider before a decision about the amount of capacity to be assigned to that user can be taken. For example, one should make consistent the priority level to the radio link quality required for the information transmission, in the sense that a high priority indicates urgency to transmit that information. In this case, a low FER (Frame Error Rate) should be guaranteed, to avoid as much as possible retransmissions that would increase delay.

A possible variation of the FER requirement as a function of the user priority is given by Figure 1. This function indicates that the higher the priority level the lower the FER at the receiver side should be. A limit on the FER requirement should be included, otherwise a single user could demand most of the system capacity.

If in the first step we set up a condition on the FER, which could be reasonable, we must also know the relationship between the spreading factor used in the packet transmission and the required E_b/N_o (energy per bit to noise spectral density, where N_o accounts for the thermal noise plus the multiuser interference) to achieve such FER.

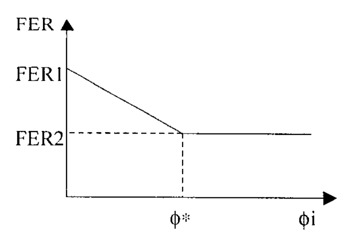

Figure 1. Variation of the target FER as a function of the priority.

The relationship between the FER (Frame Error Rate) and the (E_b/N_o) when no coding scheme is considered, under perfect power control and gaussian interference hypothesis is given by:

$$FER = 1 - \left[1 - \frac{1}{2} erfc\left(\sqrt{\frac{E_b}{N_o}}\right)\right]^{L/SF} \qquad (2)$$

where $L=L_b \times SF_{máx}$, L_b being the number of bits per frame when the lowest transmission rate (highest spreading factor, $SF_{máx}$) is used. For UTRA-FDD, L_b =150 bits/frame and $SF_{máx}$ =256 in the uplink. Then, for a given FER requirement according to the user priority, different (E_b/N_o) values would be required depending on the spreading factor used.

At this point, a target FER has been selected for the user but the scheduler still has the freedom to decide which SF will be used. For example, a criterion for the preferred SF could be trying to allocate the lowest possible SF (highest transmission rate) to the current user.

3) Availability check

Once the capacity requirement for the current user has been decided, the scheduler must check that a feasible solution do exists to satisfy at the same time both the current user requirements and those of the users with higher priority that have already been accepted for transmission in the next frame.

At this point the scheduler must include an interference model in order to devise the expected (E_b/N_o) for all users. Let consider that (K-1) users have already been allocated for transmission in the next frame and the scheduler is trying to allocate resources to the K-th user. For the uplink case, the scheduler has to evaluate the following inequalities [5]:

$$\frac{\Gamma \times P_{T,máx}}{L_p(d_k)} \geq P_k \geq \frac{N'_o \dfrac{1}{T_c} + \chi + \rho \times P_R}{\dfrac{SF_k}{\left(\dfrac{E_b}{N_o}\right)_{k,SF_k}} + \rho}$$

k=1..K (3)

where P_k is the k-th user received power at the base station, SF_k is the k-th user spreading factor, N'_o is the thermal noise spectral density, χ is the intercell interference, ρ is the orthogonality factor and T_c is the chip duration. $(E_b/N_o)_{k,SFk}$ stands for the k-th user requirement assuming a spreading factor equal to SF_k. P_R is the total received power at the base station. $P_{T,máx}$ is the maximum transmitted power by the mobile (typically 33 dBm). d_k is the distance from the k-th user to the base station and $L_p(d_k)$ being the k-th user path loss including also the shadowing component. Γ is a constant that includes the transmitter and receiver antenna gains. Of course, the better the propagation conditions the easier to satisfy the condition will be. Consequently, the capacity will tend to be allocated to the users closer to the base station, revealing what is known as cell breathing.

For the downlink direction a similar formulation arises [5]:

$$P_{T\max} \geq P_T \geq \frac{\displaystyle\sum_{i=1}^{K} \frac{\left(N'_O \dfrac{1}{T_c} + \chi_k\right)}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho} \frac{L_p(d_i)}{\Gamma}}{1 - \displaystyle\sum_{i=1}^{K} \frac{\rho}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho}} \tag{4}$$

P_T being the base station transmitted power Additionally, physical limitations into the power levels are given by the maximum base station transmitted power, P_{Tmax}.

We note that, once the spreading factor is fixed, the above conditions set the received power level depending on the required quality, which in the UTRA context is known as outer loop power control. During the frame transmission this target power is tracked by the fast power control (one command per slot or, equivalently, 1500 Hz update rate), which in the UTRA context is known as inner loop power control.

The scheduler verifies that there exists a feasible solution for the current user and the previously allocated users to satisfy all requirements. A feasible solution means that there are physically achievable power levels for all users and that all (E_b/N_o) requirements are met. In this case, the K-th user is accepted for transmission in the next frame with the current parameters and the process will be applied to the next user in the priority table. Otherwise, the scheduler tries to allow the transmission of the K-th user at a lower rate by increasing the spreading factor. If all possible SF are tested and none leads to a feasible solution the K-th user is not allowed to transmit in the next frame and the process skips to the next user in the priority table.

3.- PERFORMANCE EVALUATION

In order to assess the performance of the proposed scheduling algorithm some simulations have been carried out. The simulation model assumes that the two reference services are WWW (assumed a service with soft-QoS requirements) and e-mail (assumed a best effort service). For WWW service the traffic model presented in [6] is considered. This model considers for each user the download of WWW pages with a certain thinking time between them. For the uplink direction, the traffic model is essentially the same as in the downlink case except for the number of bytes per packet, which is lower as it only accounts for TCP/IP control commands and page addresses. For e-mail service the model considered is a Poisson generation process [6].

The soft QoS-related values assumed for a typical WWW packet will be a "typical" delay below 0,5 s ($\tau_{typ} \leq 0{,}5$ sec.) together with a maximum delay of 1,5 s in 95% of the cases ($\tau_{máx} \leq 1{,}5$ sec.). Thus, for a frame structure of 10 ms, the QoS parameters are $\tau_{typ} \leq 50$ frames and $\tau_{máx} \leq 150$ frames. The "typical" delay is defined as the desired delay bound to be experienced by an arbitrary packet. Then, the deadline in the scheduling algorithm will be defined according to this parameter.

The cellular model includes the reference cell and 6 interfering cells with cell radius 1 km, corresponding to a macrocell scenario. Since the scheduling algorithm has a strong impact on the interference statistics, it is applied also to the neighbouring cells, although performance statistics are only collected into the reference cell. The mobility model defined in [7] is considered with a mobile speed of 50 km/h. The propagation model used is also defined in [7].

The gaussian hypothesis and perfect power control are assumed for CDMA interference characterisation. For (E_b/N_o) calculations, the orthogonality factor ρ is obtained through link level simulations. Values for ρ are 0.67 in downlink and 1 in uplink [8]. The intercell interference considered for the scheduling process calculations in a given frame is the intercell interference observed in the previous frame. For the uplink case, the possible spreading factors range from SF=4 (9600 bits/frame) to SF=256 (150 bits/frame). For the downlink case, the spreading factors range from SF=4 (19200 bits/frame) to SF=512 (150 bits/frame) [9].

In order to show how the scheduling algorithm prioritises at service level, Figure 2 shows the average delay performance for the e-mail and WWW users versus the throughput of e-mail users in the downlink for different WWW load levels. As expected, when the number of WWW users is relatively low, e-mail users are allocated the spare system capacity and, consequently their performance is good. For heavy WWW loads, the best effort policy associated to the e-mail service is reflected as a high increase in the delay performance, since priority is given to WWW users. In any case, the increase in the number of email users does not have a significant impact over the performance of WWW users.

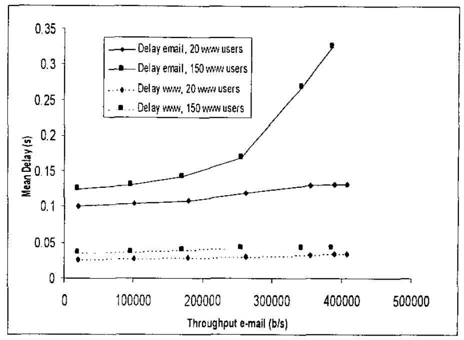

Figure 2. Throughput-delay for the e-mail and WWW service with different number of WWW users.

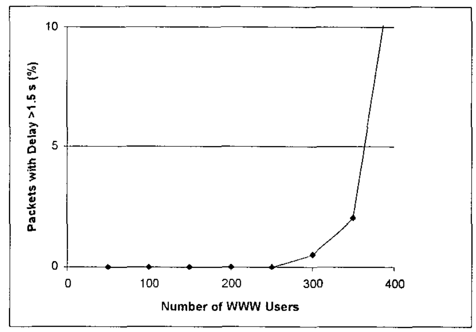

Figure 3. Percentage of WWW packets that experience a delay higher than 1.5 s.

On the other hand, Figure 3 presents the percentage of DL WWW packets that experience a delay higher than the soft-QoS bound of 1.5 s. As it can be observed, the maximum number of users that guarantees a maximum of 5% arriving later than this bound is around 370 users. This limit introduces a criterion to be considered by the admission control.

In order to see how the scheduling algorithm operates, Figure 4 presents the cumulative probability of the delay experienced by WWW packets for different number of users. As the system load increases the scheduling algorithm begins to play a role that can be observed by a change in the delay statistics: the scheduling algorithm is able to increase the cumulative probability around the 50 frames threshold, indicating that as the packets approach the timeout the scheduling algorithm tends to assign a high priority level and a suitable transport format in order to meet the delay requirements.

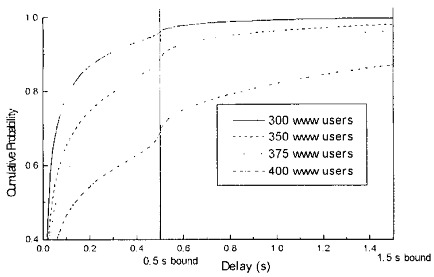

Figure 4. Cumulative probability of the delay for different numbers of users.

4.- CONCLUSIONS

For 3G systems new and specific radio resource management algorithms related to load control, packet scheduling and admission control are required to guarantee QoS and to maximise the system throughput for mixed services with different bit rates and quality requirements. This paper has focused on the problem of QoS provisioning for packet driven environment, in particular in the W-CDMA scenario selected in UMTS, as little effort has been devoted to date in addressing this topic.

A new scheduling algorithm that makes consistent the target quality in the radio link with the priority level assigned to each user has been proposed and evaluated through system level simulations.

5.- ACKNOWLEDGEMENTS

This work is part of the Wineglass project partially funded by the European Commission under the Information Society Technology (IST) framework.

6.- REFERENCES

[1] S. Blake et al. "An Architectrure for Differenciated Services" IETF RFC 2475, December 1998.

[2] Yoram Bernet, "The complementary Roles of RSVP and Differential services in the full-Service Network", IEEE Comm. Magazine, February 2000.

[3] M. Naghshineh, A. S. Acampora, "Design and control of microcellular networks with QoS provisioning for data traffic", Wireless Networks 3 (1997) , pp. 249-256.

[4] I. F. Akyldiz, D. A. Levine, I. Joe, "A Slotted CDMA Protocol with BER Scheduling for Wireless Multimedia Networks", IEEE/ACM Transactions On Networking, Vol. 7, No2, April 1999, pp. 146-158.

[5] O. Sallent, J. Pérez-Romero, F. Casadevall, R. Agustí, "An Emulator Framework for a New Radio Resource Management for QoS Guaranteed Services in W-CDMA Systems", submitted to IEEE Journal on Selected Areas in Communications.

[6] M. Bartoli, G. Foddis, D. Minervini, M. Molina, "Modelli di traffico E-mail e WWW per il servizio GPRS", Internal Report from CSELT, February 2000.

[7] ETSI TR 101.112 v3.2.0, "Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS"

[8] K. Parsa, S. S. Ghassemzadeh, S. Kazeminejad, "Systems Engineering of Data Services in UMTS W-CDMA Systems", IST Proceedings, Galway, Ireland, October 2000, pp. 435 - 444.

[9] 3GPP TS 25.211, "Physical channels and mapping of transport channels onto physical channels (FDD)"