Packet Scheduling Algorithms for Interactive and Streaming Services under QoS Guarantee in a CDMA System

Luis Almajano, Jordi Pérez-Romero
Dept. Signal Theory and Communications. Universitat Politècnica de Catalunya (UPC)
c/ Jordi Girona 1-3. Campus Nord. Building D4. 08034-Barcelona. Spain
e-mail: luis.almajano@vodafone.com, jorperez@tsc.upc.es

Abstract – **One of the major issues for the deployment of future third-generation mobile communication systems with high-bandwidth and real-time multimedia applications is the specification of acceptable Quality of Service (QoS) requirements for mobile users. This paper presents scheduling algorithms focused on giving QoS guarantee in terms of bit rate in a packet switched CDMA scenario by taking into account the service classes that are defined in the UMTS system, with a special emphasis on the streaming and interactive services. The algorithm carries out three steps, namely prioritisation, resource allocation and availability check. Different alternatives are presented for these steps. Particularly, by extending the leaky bucket algorithm to the CDMA scenario for the prioritisation and the allocation process, we show that a non work conserving allocation strategy appears to be more efficient when aiming at guaranteeing a specific average bit rate with lower deviations around the mean value thus fitting better the allocated resources to the users' requirements. On the other hand, work conserving strategies achieve higher peak bit rates at the expense of exhibiting a much higher deviation around the mean value.**

I. INTRODUCTION

While second generation telecommunication systems enabled voice traffic to go wireless, third generation systems will make multimedia communication possible by supporting higher data rates and new flexible communication capabilities. So as to successfully allow a wide range of applications with very different QoS requirements, efficient and reliable Radio Resource Management (RRM) techniques are needed to satisfy at the same time all the QoS requirements while maximizing the system capacity [1]. On the other hand, most of the new services to be supported are of a very bursty nature where high activity periods are alternated with low activity periods. In that sense, packet oriented strategies that share the resources among users according to scheduling algorithms claim for a more efficient use of the scarce radio resources, since the information to be transmitted is taken into account. Similarly, the inherent flexibility of the CDMA multiple access strategy to support different data rates and its ability to multiplex different traffic sources makes it an excellent option to be combined with appropriate packet scheduling mechanisms.

Although packet switched services are still considered as complementary, 3GPP Release 2000 for UMTS and future evolutions in 3G systems are focusing on an all-IP end-to-end architecture. This implies that packet switched services in the air interface will gain momentum and the optimization of capacity will strongly depend on how algorithms for Radio Resource and QoS Management are used.

Within this framework, in this paper we present proper scheduling algorithms focused on giving QoS guarantee in terms of bit rate in a packet switched CDMA scenario by taking into account the service classes that are defined in the UMTS system (i.e., conversational, streaming, interactive and background). Special emphasis will be given to streaming and interactive services since they are those which differ most from the circuit switched services (like the conversational ones) while meeting certain QoS guarantees (unlike the background services).

Based on the *leaky bucket* algorithm, this paper extends the concept of tokens to a CDMA packet switched mobile communications system. This goal is achieved by introducing a new parameter called "Service Credit" (SCr) [2], which measures the difference between the bit rate requested by a user and the bit rate that the system has offered him.

This paper is organised as follows. We start in section II with a general description of the algorithm and a deeper analysis on the Service Credit. In section III we outline the proposed prioritisation criteria whereas in section IV some new resource allocation techniques are described. The last step in the RRM algorithm dealt in section V leads to present computer simulations results in section VI, showing the scheduler performance under different work conditions. Finally, section VII is devoted to the conclusions.

II. ALGORITHM DESCRIPTION

In future 3G systems, services with different requirements will be demanded and an algorithm able to manage the available resources will be needed. Contrary to TDMA or FDMA, where time slots or frequency channels are the available resources to manage, in a CDMA environment, power level, spreading factor (SF) and channelization codes have become the new resources to allocate.

The framework of this paper is based on a cell with frames of 10 ms duration, where the power level and spreading factor can be changed on every frame. A centralized strategy has been considered, where the Radio Network Controller (RNC) is the main responsible for the resource management. The main objective of the proposed RRM strategy will be then to determine the necessary spreading factor (bit rate) and power level of each connection in order to give QoS guarantees.

We consider the coexistence of the following representative services for each service class: voice service for conversational users, WWW browsing for interactive users and a two-layered video service consisting of a basic layer that offers the minimum required quality and an enhancement layer that offers a better quality for streaming users (the enhancement can only be transmitted if the basic layer is transmitted, too).

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

- *Prioritisation.* All users intended to transmit information must be somehow classified. Several prioritisation criteria are introduced in this paper.

- *Resource allocation.* In order to fulfil users QoS requirements, system capacity (i.e. power level, spreading factor) must be devoted to each user in such a way that the overall performance is as optimum as possible.

- *Availability check.* Once the capacity requirement for each user has been decided, the scheduler must check that a feasible solution exists to satisfy at the same time both the current user requirements and those of the users that have already been allocated for transmission.

The criteria followed in the current paper in terms of QoS evaluation has been focused on bit rate measurements, since the aim of this work is to give bit rate guarantees. In order to monitor the QoS received by a certain connection, we make use of the Service Credit (SCr) concept, that extends the idea of tokens from the leaky bucket algorithm to a CDMA packet switched mobile communications system. The SCr is defined as a real number, which is associated with an active link or user and it computes the difference between the bit rate requested by the user and the bit rate that the system has provided to him. So it accounts for the amount of service the system owes to the user.

The SCr value of each active connection must be updated every 10 ms, that is, every frame, following the expression:

$$SCr_i = SCr_{i-1} + (b_{min} / b_{basic}) - num\_of\_tx\_ok \qquad (1)$$

where SCr_i [transport blocks/frame] represents the SCr value for the current frame i, SCr_{i-1} is the SCr value in the previous frame, b_{min} [bits/frame] is the minimum bit rate requested by the user, b_{basic} [bits/transport block] is the basic bit rate (which depends on the class of service) and num_of_tx_ok shows the number of transport blocks that have been successfully sent during the current frame. Thus, positive SCr values indicate that the system is "in debt" with a certain user, meaning that the offered bit rate stands below the requested bit rate, whereas negative values reflect those links whose offered bit rate stand above the requested bit rate.

The information that the Service Credit (SCr) provides will be used by the algorithm to schedule users system access requests and data information transmissions. In the following, the three steps of the algorithm are explained in more detail and different possibilities for each of them are considered.

III. PRIORITISATION CRITERIA

Prior to resource allocation, as explained in II, the scheduler must order the different requests according to a certain criterion. This paper considers the prioritisation at two levels:

- *Prioritisation according to the class of service.* The scheduler proceeds to order the different requests depending on the class of service they belong to. From highest to lowest priority level: conversational, streaming, interactive, streaming enhancement and background class. Logically, for this classification QoS requirements in terms of delay have been taken into account.

- *Prioritisation according to the SCr.* In case two or more users belong to the same class of service, a second prioritisation level based on the SCr is considered. Under this situation, the SCr value of each request will lead to determine the priority of each connection, so that higher values get higher priority levels. It must be reminded that the SCr measures the difference between the demanded service and the offered service. Thus, the SCr stands as a suitable parameter to determine priority among different requests. Again, it may occur that two or more links have the same SCr value. Given that case, the ratio b_{min} / b_{basic} , which indicates how much does the SCr increase (in case there are no resources available), will be used as determinant to decide which user must be allocated first.

IV. RESOURCE ALLOCATION

Once all users intended to transmit in a frame are prioritised, the scheduler is responsible for the resource allocation. Particularly, an appropriate spreading factor and a suitable power level should be devoted for the user in such a way that the overall performance is as optimum as possible. This step is of paramount importance, since the capacity allocated to each request is strongly related with the QoS that the end-user will perceive.

IV.A. Spreading Factor Determination

According to expression (1), it is possible that the SCr may reach a negative value. This means that the user has been offered more service than the minimum required. In this paper we evaluate the following possibilities based on [3]:

a) Non-work-conserving strategy:

Since the SCr shows the service "debt" or amount of information that a user has to transmit, and due to the importance of this parameter in prioritisation process, the idea to assign a spreading factor according to the SCr arises. To this end, even if radio resources are available, only those requests with positive SCr values are served. If the user has no service credit no transmission is allowed. Therefore, in this case, the spreading factor is selected in such a way that the minimum required resources are used in order to reach exactly the desired bit rate.

By means of this strategy, a relationship between the SCr value for each kind of service and the spreading factor needed

to be allocated in order to transmit the stored data is established.

Following the expression below:

$$(SCr \cdot b_{basic}) = number\ \ of\ \ bits\ \ to\ transmit \qquad (2)$$

and taking the WCDMA physical layer into account, as Table I shows for the uplink [4], the spreading factor assignment is straightforward.

TABLE I WCDMA uplink case

SF	4	8	16	32	64	128	256
kbps	960	480	240	120	60	30	15
Bits/ frame	9600	4800	2400	1200	600	300	150

b) Work-conserving strategy:

By means of this strategy, as long as radio resources are available, all requests are served, even if their SCr is negative. This could allow a higher bit rate than expected at the expense of a higher interference level and a higher power consumption.

Two alternatives have also been considered in this case to decide the spreading factor:

b.1) SF=SFmin

Under this strategy the system always tries to concede the highest bit rate, that is, the lowest spreading factor (SF). Thus, the scheduler will first try to assign a spreading factor of 4 to the user. In case power requirements are not reached, a higher spreading factor will be retried. Once tried the highest spreading factor, if there are still no available resources, the user will have to wait for the next frame for transmission. This strategy tends to a "time scheduling" policy, where a reduced number of users access the system with very high bit rates [5].

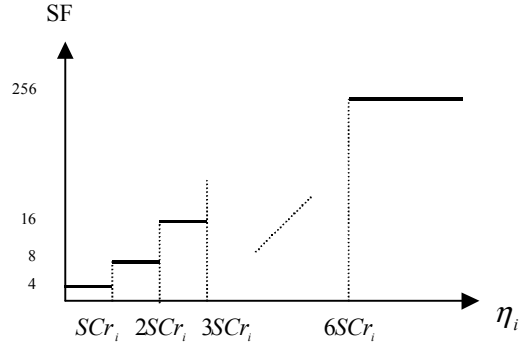

Figure 1: SF assignment for SF=f(η) strategy

b.2) SF=f(η)

Let η_i be a measure of the priority of the users remaining to be allocated in the system after user i:

$$\eta_i = \sum_{j>i}^{K} SCr_j \qquad (3)$$

so that high/low η_i values mean that after assigning capacity to user i a high/low load still has to be served, a strategy as depicted in Figure 1 could be beared in mind, where the spreading factor is decided according to the value of η_i. Thus, the bit rate offered to a user will be reduced as the load to the system grows.

Contrary to **SF=SFmin**, the current strategy tends to a "code scheduling" policy, where the system assigns low bit rates in order to allocate as much users as possible [5].

IV.B. Required Eb/No

For each transmission, a suitable power level should be adjusted to meet the required Eb/No target. While this target can be somewhat relaxed for interactive services, it becomes specially critical for streaming services, that have little margin for retransmissions when a packet is received in error. In this case, there are two parameters that play a key role, namely the FER (Frame Error Rate) target and the number of allowable retransmissions. As the delay requirements for streaming services are stringent, whenever a packet is received in error it should be retransmitted as fast as possible, together with the rest of packets that have arrived during its transmission. Consequently, the instantaneous bit rate after a retransmission should be set higher than in the previous transmission, which would lead to a low number of users simultaneously transmitting. To deal with this inconvenient, in this paper we propose to adaptively vary the FER target as a function of the number of retransmissions for streaming packets, in the sense that the Eb/No target is decreased in each retransmission.

The proposed function to adaptively vary the FER target for streaming packets according to the number of retransmissions is depicted in Figure 2, where the values for FER_1 and FER_2 are set by simulation. Notice that the higher the number of retransmissions the lower the spreading factor that should be used to transmit the packets that have been delayed. Therefore, in order to not pose a limitation in the number of scheduled transmissions, the FER requirement is somewhat relaxed.

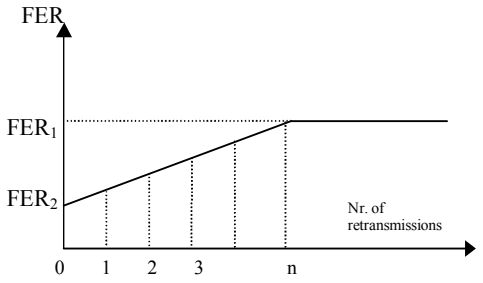

Figure 2: FER_{target} variation as a function of the number of retransmissions

It is worth mentioning that FER and Eb/No target requirements are tightly coupled depending on link layer characterisation. For instance, if no channel coding is assumed:

$$FER = 1 - \left[1 - \frac{1}{2} erfc\left(\sqrt{\frac{E_b}{N_0}}\right)\right]^L \qquad (4)$$

L being the number of bits per frame depending on spreading factor SF and Eb/No the relation energy per bit to noise spectral density, it becomes clear that for different Eb/No and SF values, a given FER requirement can be reached.

Whenever certain channel coding characteristics are assumed the relationship between FER, Eb/No and SF should be obtained by means of link layer simulations.

V. AVAILABILITY CHECK

Once spreading factor and Eb/No targets have been decided for the different users, the required power level is calculated and the decision about the acceptance or rejection of each transmission is carried out. Particularly, the same procedure as given in [5] is considered. Essentially, for each user it checks the following inequation regarding CDMA capacity:

$$\frac{P_k \times SF_k}{P_N + \chi + \rho \times [P_R - P_k]} \geq \left(\frac{E_b}{N_o}\right)_{k,SF_k} \qquad (5)$$

where P_k is the k-th user received power at the base station, SF_k is the k-th user spreading factor, P_N is the thermal noise power, 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 and finally P_R is the total received power at the base station.
Before a new transmission is accepted, the condition (5) is checked for all the already accepted transmissions and for the new one and P_R is computed. If no solution exists that satisfies at the same time all the requests with their corresponding power limitations, the new transmission is not accepted with the specific SF and SF is increased by a factor of 2 until a feasible solution exists. In the case that SF=SFmax and no solution exists, transmission is postponed and the algorithm goes to next request.

VI. SIMULATION RESULTS

As simulation scenario a single macro-size cell of 1 km of radius has been deployed. However, since a CDMA system strongly depends on inter- and intracell interference, the effect of an interfering ring, as explained in [6], has been taken into account. Propagation and mobility models are taken from [7] including shadowing fading with 10 dB standard deviation and 20 m decorrelation length. The simulation parameters for each service are defined in Table II.

In order to present how the algorithm schedules the different service classes, Figure 3 shows the percentage of lost packets for conversational and streaming packets (basic and enhancement layer) with 5 conversational users and 3 streaming users in the central cell when priority handling criteria are based on the two steps described in point 3. It is worth noting that neither conversational user nor the basic streaming layer degrades when increasing the number of www users. On the contrary, the bursty nature of WWW traffic allows the scheduling algorithm to assign resources for the streaming enhancement level during the inactivity periods, improving then the quality of the streaming service.

TABLE II Simulation parameters

Conversational	
Traffic model	ON-OFF
Minimum bit rate (b_{min})	15 kbps
Basic bit rate (b_{basic})	15 kbps
FER	0.01
Streaming	
Traffic model	CBR model
Minimum bit rate (b_{min})	30 Kbps
Basic bit rate (b_{basic})	30 kbps
Maximum packet delay	60 ms
FER_{min}	0.01
FER_{max}	0.1
Streaming enhancement	
Minimum bit rate (b_{min})	60 Kbps
Interactive	
Traffic model	[8]
Minimum bit rate (b_{min})	10 Kbps
Basic bit rate (b_{basic})	30 kbps
FER_{min}	0.01
FER_{max}	0.1

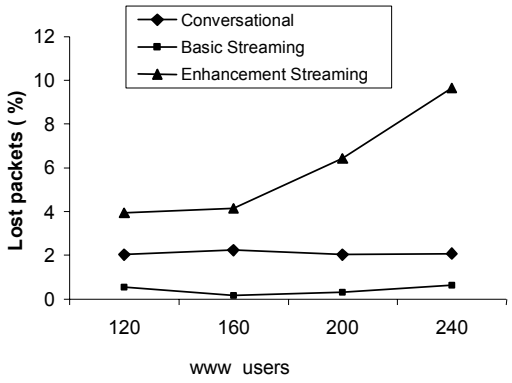

Figure 3 Packet loss for conversational and streaming services

Regarding the spreading factor proposals for the resource allocation step, the performance of the work-conserving strategies SF=SFmin and SF=f(η) is depicted in Figure 4 and Figure 5 for streaming and interactive users, respectively.

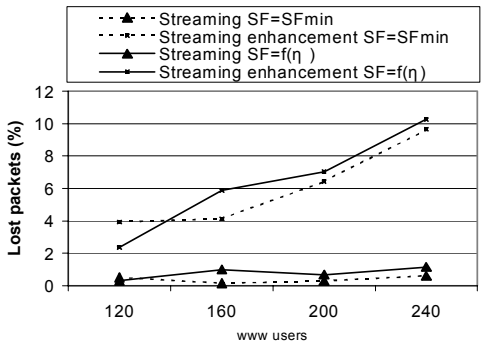

Figure 4 Packet loss comparison for streaming services under SF=SFmin and SF=f(η) strategies

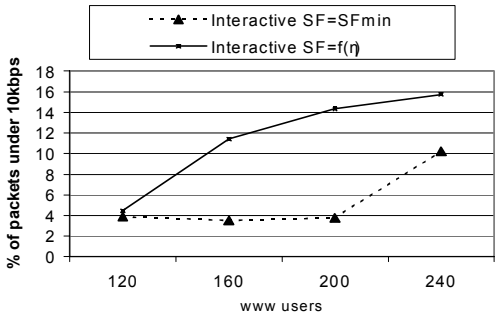

Figure 5 Packets under 10kb/s for interactive services for SF=SFmin and SF=f(η) strategies

It can be observed that, while for streaming users the selection of one strategy has not great impact in terms of packet loss (i.e., a packet is lost if the maximum number of 6 retransmissions is reached), the use of SF=SFmin policy allows a greater number of interactive users to access the system whilst guaranteeing the 10 kbps minimum bit rate in the 95% of the cases.

Bit rate cumulative distribution functions for interactive users in the uplink are shown in Figure 6 to evaluate the different allocation policies. When a work-conserving strategy with SF=SFmin is deployed, high peak data rates are achieved with grand deviation around the mean value. On the contrary, the non-work-conserving strategy is able to adjust more properly the offered bit rate to the user's requirements by avoiding unnecessary high peak rates.

Regarding streaming services, the possibility of varying the FER target as a function of the number of retransmissions is studied. An example of this situation is depicted in Figure 7, where the packet loss for the basic streaming layer is shown as a function of the number of users, when the maximum number of retransmissions is set to 6. A certain gain is observed.

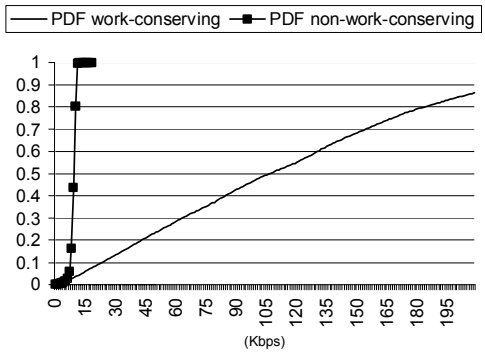

Figure 6 Cumulative distribution functions of the bit rate for work-conserving and non-work-conserving strategies

VII. CONCLUSIONS

In the present paper new radio resource management techniques for packet switched CDMA based mobile systems have been proposed. The Service Credit (SCr) arises as a new parameter to be considered for the scheduler in order to assure QoS guarantee in terms of bit rate for streaming and interactive services. The main conclusions that can be extracted are summarised in the following points:

- The proposed algorithm is able to handle appropriately the different transmissions to allow the coexistence of the considered services. The priority handling mechanism enables the use of the spare capacity left by the WWW users in their OFF periods to increase the performance achieved by the streaming services through the transmission of their enhancement layer.
- From the point of view of WWW services, the proposed spreading factor allocation policies are compared. The non-work-conserving strategy appears to be more efficient when aiming at guaranteeing a specific average bit rate, with lower deviations in the performance achieved around the mean value, thus fitting better the allocated resources to the users' requirements. On the other hand, the work conserving strategies achieve higher peak bit rates at the expense of exhibiting a much higher deviation around the mean value.
- For streaming services, the possibility of varying the Eb/No target as a function of the number of retransmissions allows a certain increase in the system capacity.

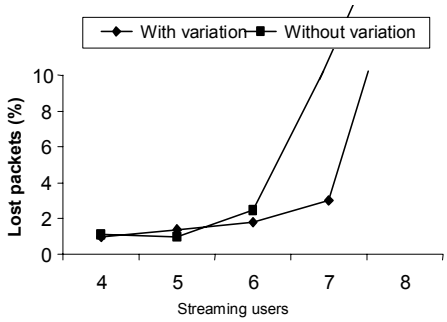

Figure 7 Impact of an adaptive variation function for the FER target

ACKNOWLEDGEMENTS

This work has been funded by the Spanish Research Council under grant TIC2001-2222.

REFERENCES

[1] Harri Holma, Antti Toskala, *WCDMA for UMTS. Radio Access For Third Generation Mobile Communications*, John Wiley & Sons, 2000.

[2] Luis G. Alonso, Ramón Agustí, "MAC Protocols and Scheduling Algorithms for QoS Guarantee in CDMA Based Mobile Communication Systems", PhD dissertation, February 2001

[3] H. Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks", Proceedings of the IEEE, Vol.83, No.10, October, 1995, pp.1374-1396

[4] 3GPP TS 25.211 v3.5.0 "Physical channels and mapping of transport channels onto physical channels (FDD)", Release 1999, December, 2000.

[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", IEEE Journal on Selected Areas in Communications, October, Vol.19, No. 10, October 2001, pp. 1893-1904. 2001.

[6] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver, C.E. Wheatley III, "On the Capacity of a Cellular CDMA System", IEEE Trans. Veh. Tech., Vol. 40, Nr 2, May 1991.

[7] 3G TR 25.942 v 2.1.3. "RF System Scenarios", Release 1999, March, 2000.

[8] 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.