An Emulator Framework for a New Radio Resource Management for QoS Guaranteed Services in W-CDMA Systems

Oriol Sallent, *Member, IEEE*, Jordi Pérez-Romero, *Student Member, IEEE*, Fernando J. Casadevall, *Member, IEEE*, and Ramon Agustí, *Member, IEEE*

***Abstract*—In the context of third-generation (3G) systems a mix of services with different requirements are expected. Consequently, packet scheduling mechanisms for quality of service (QoS) guarantees 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 a strategy is assessed by system level simulations and, in order to gain more insight into the difficulties of this optimization problem, it is compared to other alternatives. This work is part of the Wineglass project, within the Fifth Framework Program of the European Commission (IST), where a real time demonstrator including the radio resource management tasks is being developed. Thus, an implementation approach of the proposed scheduling is also described. The implementation is based on lookup tables and this approach is validated by simulation.**

***Index Terms*—Mobile radio, quality of service, radio resource management, W-CDMA.**

I. Introduction

Second-generation (2G) wireless systems focused their effort on providing mobile voice applications to the end user with an acceptable quality. The evolution of the end users' needs toward multimedia applications has pushed the wireless community to conceive the so-called third-generation (3G) systems (such as UMTS or IMT-2000), where a very large amount of both circuit-switched services and packet-switched services for voice and data will be provided.

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 quality of service (QoS) for each one. The only way to harmonize these two contradictory points (scarce bandwidth and a stringent QoS) is with a proper management of the available radio resources.

In particular, Release 99 of UMTS from 3GPP [1] is mainly focused on an efficient provision of circuit-switched services. Packet-switched services are still considered as complementary, and, moreover, the optimization of capacity in a multiservice environment is not envisaged as foreground. However, Release 2000 of UMTS in 3GPP and future evolutions in 3G systems are focusing in an all-IP end-to-end architecture, and consequently the packet-switched services in the air interface will gain momentum and the optimization of capacity in a multiservice environment is going to be faced definitively. Furthermore, this optimization will be a consequence of efficient algorithms for radio resource and QoS management, which will not be standardized, but their specific realization is implementation dependent.

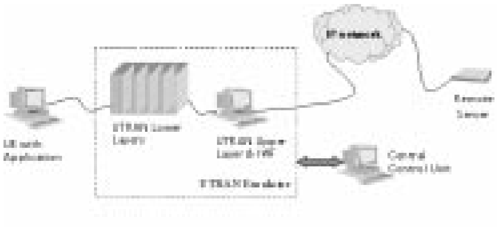


Fig. 1. Wineglass project architecture.

In order to test the algorithms, their simulation is not enough and therefore the services must be validated in an environment as close as possible to that as will be in the real world. That is why an actual implementation in a testbed within the framework of the IST Wineglass project [2] is envisaged. The IST Wineglass project aims to exploit IP-based techniques to support mobility and soft-guaranteed QoS, in a wireless Internet architecture incorporating a 3GPP-UMTS access network (UTRAN) and a WLAN access network. In the context of the Wineglass project, the UTRAN access network is an emulated W-CDMA FDD-based system. The emulation model considers the system to emulate as a black box, whose input–output behavior intends to reproduce the real system without requiring a knowledge of the internal structure and processes. That is, the internal structure of the model is normally not related to the internal structure of the real system. If the parameters of the emulation model have been adequately selected, this emulator allows the accurate operation in real time with a moderate implementation complexity.

The real time emulator (RTE) to be developed in Wineglass will allow us to test the performance of a mix of services under different scenarios and system conditions. As Fig. 1 shows, applications running in the user equipment (UE) will provide traffic flows to the RTE. The RTE will emulate in real time the behavior of the radio interface and will deliver the resulting information to the receiver application part.

The radio interface of the UMTS terrestrial radio access (UTRA) is layered into three protocol layers, as presented

Manuscript received December 1, 2000; revised June 1, 2001. This work is part of the Wineglass project supported in part by the European Commission under the Information Society Technology (IST) framework.

The authors are with Universitat Politècnica de Catalunya, 08034 Barcelona, Spain (e-mail: sallent@tsc.upc.es; jorperez@tsc.upc.es; ferranc@tsc.upc.es; ramon@tsc.upc.es).

Publisher Item Identifier S 0733-8716(01)08472-4.

0733–8716/01$10.00 © 2001 IEEE

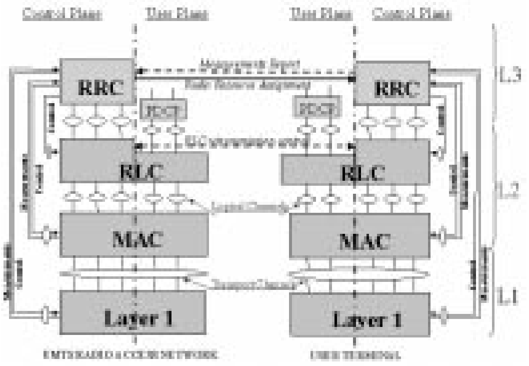


Fig. 2. Radio interface protocol architecture.

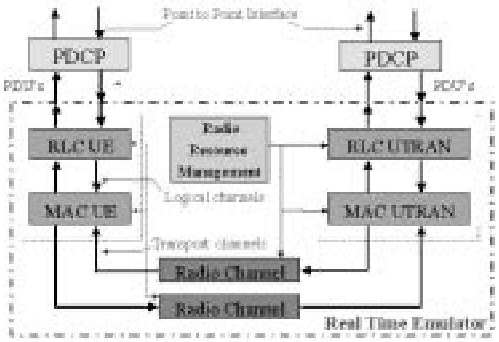


Fig. 3. Block diagram of the emulated UTRA-FDD radio interface.

in Fig. 2: the physical layer (L1), the data link layer (L2), and the network layer (L3). Additionally, layer 2 is split into two sublayers, the radio link control (RLC) and the medium access control (MAC). On the other hand, the RLC and layer 3 protocols are partitioned in two planes, namely the user plane and the control plane.

According to the structure of the radio interface protocol architecture above described, four main blocks must be taken into account to emulate the radio interface of the UTRAN (see Fig. 3).

Radio Channel: The radio-channel functionality deals with the emulation of the transmission chain at bit level. The RTE must emulate in real time the input–output behavior of the transport chain. Markov chains based on the hidden Markov model (HMM) are used to model this transport functionality. That is, for each testing scenario, the Markov chains are properly trained through adequate offline simulations, which reproduce (in terms of error distribution) the statistical behavior of the radio channel. Once this statistical behavior is known, the parameters of the HMM of the RTE must be properly tuned to reproduce this behavior with enough accuracy from a statistical viewpoint [3]. This approach allows the emulation of a great number of scenarios with different propagation conditions, environments (macrocell, microcell), mobile terminal speed, etc., provided that they are trained properly. Therefore, the main advantage of using an HMM model is the reduction in time, resources, and effort with regard to implementing a real system.

MAC and RLC Functionality: The functionality associated to the data link layer (layer 2) is split between two inner sublayers: MAC and RLC. The MAC sublayer maps the so-called logical channels, which characterize what type of data is transmitted, into transport channels. Moreover, the MAC sublayer is also responsible for selecting the appropriate transport format (TF) for the transport channels. The RLC sublayer provides segmentation and retransmission procedures for both user and control data. The RLC sublayer can work in a transparent mode as well as unacknowledged or acknowledged modes. In this late case an automatic repeat request (ARQ) mechanism is used for error correction purposes. Additionally when packet data services are considered, a new protocol, called packet data convergence protocol (PDCP) is used. The main objective of this protocol is the compression of the redundant control information allocated at the header of the IP packets.

In summary, the data link layer (layer2) is devoted to set up, reconfigure, and release reliable radio bearers in an appropriate way to send the user data through the air interface.

Radio Resource Management: The radio resource management (RRM) entity is 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 maximize the system throughput for mixed services with different bit rates and quality requirements. So, an important objective of the RTE will be the ability to demonstrate the system performance under different RRM strategies. Note that such evaluation requires considering multiple users in several cells and it would originate an unaffordable computation complexity for a real time operation. Consequently, a suitable RRM modeling must be introduced, allowing the operation in real time while maintaining the same statistical behavior as in a real situation.

In the last few 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 [4], [5], but thus far only able to offer best-effort services. In particular, future packet-switching networks will have to support information with QoS guarantees. An important issue in providing guaranteed performance service is the choice of packet service or scheduling disciplines at the switching node. In [6] this topic is extensively covered.

In wireless environments, the problem of QoS provisioning for multimedia traffic has also gained interest in the literature in recent years, as the problem arises in the context of 2.5G (i.e., GPRS) and 3G (i.e., UMTS, cdma2000) systems and is not present in 2G systems (i.e., GSM, IS-95). Thus, Naghshineh and Acampora [7], [8] introduced resource sharing schemes for QoS guarantees to different service classes in microcellular networks. Akyildiz *et al.* [9] proposed the so-called WISPER protocol, scheduling the transmissions according to their BER requirements. Das *et al.* [10] developed a general framework for QoS provisioning by combining call admission control, channel reservation, bandwidth reservation, and bandwidth compaction.

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. Thus, three main issues are covered: 1) 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; 2) the implementation of the scheduling algorithm in the context of the Wineglass demonstrator so that a real time operation is feasible; and 3) the validation of the proposed implementation approach from a statistical point of view. In particular, Section II presents the scheduling algorithm and Section III includes the performance evaluation of the algorithm by means of simulation. The implementation matters are presented in Section IV and the validation with the implemented RRM is discussed in Section V. Finally, Section VI summarizes the conclusions reached.

II. Scheduling Strategy

For multimedia traffic (voice, data, video) to be supported successfully, it is necessary to provide a 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 a 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 behavior 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 email could be served on a best effort basis.

The proposed RRM strategy, which applies on a frame-by-frame basis, focuses on a soft-QoS guarantee for interactive-like services (i.e., WWW browsing) and can be split into three different steps.

A. Prioritization

All users intended to transmit information must be classified according to the type of service (first prioritization level) and, for the same type of service, according to their QoS (second prioritization 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 (in this case WWW browsing), the smaller the deadline the higher the position of a particular user in the serving queue will be. Taking these parameters into account, the RRM is in charge of deciding which traffic flows should be transmitted first. As a result, a table containing all active users from higher to lower priority would be available.

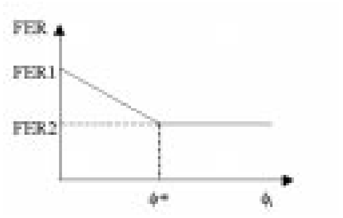


Fig. 4. Variation of the target FER as a function of the priority.

For WWW users, which is the focus of this paper, priority to the ith user could be assigned according to the following function:

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

with L_i being the packet length remaining to be transmitted, n an integer, and TO_i the time (in frames) left until the packet deadline for the considered service. Note that TO_i refers to packets arriving to the radio access network, as we are concerned with QoS for the radio interface segment. Thus, the specific deadline for the packets should be set as a result of an end-to-end QoS evaluation.

The above expression assigns a higher priority to those users with either a great deal information to transmit or much time already waiting for permission. When the delay threshold is overcome, the priority level is drastically increased. As the prioritization problem does not accept an optimal solution, by means of intensive priority functions search, (1) was found to be a suitable option, although other functions could be used.

B. 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 to whom resources are attempted to be allocated, some system capacity should be devoted to this user and a 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 behavior were as optimal 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 frame error rate (FER) 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 Fig. 4. This function indicates that the higher the priority level the lower the FER at the receiver side should be. A limit on the minimum FER requirement should be included, otherwise a single user could demand most of the system capacity. Again, there would be other FER variation functions to be used (linear variation, parabolic variation, etc.) and no optimal solution can be proved. Simulations were carried out in order to devise a suitable function and parameters of such a function before Fig. 4 was retained.

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 an FER.

The BER at the receiver output under perfect power control and Gaussian interference hypothesis is given by

$$\text{BER} = \frac{1}{2}\,\text{erfc}\left(\sqrt{\frac{E_b}{N_o}}\right). \tag{2}$$

The relationship between the FER and the (E_b/N_o) when no coding scheme is considered is given by

$$\text{FER} = 1 - \left[1 - \frac{1}{2}\,\text{erfc}\left(\sqrt{\frac{E_b}{N_o}}\right)\right]^{L/SF} \tag{3}$$

where $L = L_b \times SF_{\text{máx}}$, L_b is the number of bits per frame when the lowest transmission rate (highest spreading factor, $SF_{\text{máx}}$) is used. For UTRA-FDD, $L_b = 150$ bits/frame and $SF_{\text{máx}} = 256$ 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. We denote that for $SF = i$ $(i = 4, 8, 16, 32, 64, 128, 256)$ and a given FER the requirement is $(E_b/N_o)_i$.

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. In order to gain more insight into how this issue impacts the overall scheduling behavior, different approaches will be considered in the performance evaluation section.

C. Availability Check (Uplink Case)

Once the capacity requirement for the current user has been decided, the scheduler must check that a feasible solution does exist 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 us 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 Kth user. The scheduler has to evaluate the following inequalities:

$$\frac{P_k \times SF_k}{N'_o \dfrac{1}{T_c} + \chi + \rho \times [P_R - P_k]} \geq \left(\frac{E_b}{N_o}\right)_{k,SF_k} \qquad k = 1 \cdots K \tag{4}$$

$$P_R = \sum_{i=1}^{K} P_i \tag{5}$$

where

P_k kth user received power at the base station;
SF_k kth user spreading factor;
N'_o thermal noise spectral density;
χ intercell interference;
ρ orthogonality factor;
T_c chip duration.

$(E_b/N_o)_{k,SFk}$ stands for the kth user requirement assuming a spreading factor equal to SF_k. P_R is the total received power at the base station.

Implicitly in the above inequalities a certain received power level is assumed in each case

$$P_k \geq \frac{N'_O \dfrac{1}{T_c} + \chi + \rho \times P_R}{\dfrac{SF_k}{\left(\dfrac{E_b}{N_o}\right)_{k,SF_k}} + \rho} \qquad k = 1 \cdots K. \tag{6}$$

Adding all K inequalities it holds that

$$\sum_{i=1}^{K} P_i = P_R \geq \sum_{i=1}^{K} \frac{N'_O \dfrac{1}{T_c} + \chi + \rho \times P_R}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho} \tag{7}$$

$$P_R \geq \frac{\left(N'_O \dfrac{1}{T_c} + \chi\right) \displaystyle\sum_{i=1}^{K} \frac{1}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho}}{1 - \displaystyle\sum_{i=1}^{K} \frac{\rho}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho}}. \tag{8}$$

Note that, as P_R is inherently a positive magnitude, a condition appears that has only to do with the capacity of a CDMA multiple access system, since only the (E_b/N_o) requirements and the multiuser interference play a role

$$\sum_{i=1}^{K} \frac{\rho}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho} \leq 1. \tag{9}$$

Also note that when translating the conditions into power levels, reflecting that this is a key resource to be managed by the scheduler, physical layer parameters (for example the thermal noise) play a role in the process.

Additionally, physical limitations into the receiver and/or transmitted power levels must be taken into account. Focusing on the uplink, let us consider the restriction of a maximum transmitted power by the mobile $P_{T,\text{máx}}$ (typically 33 dBm).

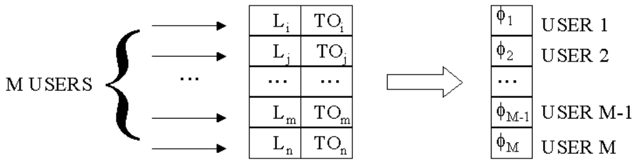


Fig. 5. Generation of the priority table.

Depending on the propagation conditions, the received power may vary and the upper bound on the received level may be found. So, for the kth user located at distance d_k from the base station, its maximum received power level, $P_{k,\max}$, is

$$P_{k,\max} = \frac{P_{T,\max} \times \Gamma}{L_p(d_k)} \tag{10}$$

where $L_p(d_k)$ is the kth user path loss including the shadowing component. Γ is a constant that includes the transmitter and receiver antenna gains. Consequently, an additional restriction arises in (6) which is

$$\frac{\Gamma \times P_{T,\text{máx}}}{L_p(d_k)} \geq P_k \geq \frac{N'_O \dfrac{1}{T_c} + \chi + \rho \times P_R}{\dfrac{SF_k}{\left(\dfrac{E_b}{N_o}\right)_{k,SF_k}} + \rho} \qquad k = 1 \cdots K. \tag{11}$$

Of course, the better the propagation conditions the easier to satisfy the condition. Consequently, the capacity will tend to be allocated to the users closer to the base station. This case reveals what is known as cell breathing.

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.

Figs. 5 and 6 summarize the proposed scheduling strategy. Assume that M users intend to transmit in a given frame. The first step is to generate the priority table according to the priority function given by (1). Without losing general applicability, one could consider that users are dynamically numbered for each service according to their priority level, so that user 1 has higher priority than user 2 and so on, as reflected in Fig. 5.

Users are considered one by one, from higher to lower priority. For the generic user i, a target FER is devised according to the priority level (Fig. 4). Then, the scheduler selects a preferred SF according to a suitable criterion. As pointed out in Section II-B, a possibility would be to allocate an SF as low as possible. So, it initially selects $SF = 4$, which together with the FER requirement is translated into a (E_b/N_o) requirement (3). Then, the capacity condition expressed by (9) is checked. Next, the scheduler verifies that there exists a feasible solution for the current user and the previously allocated users to satisfy all requirements by solving (8) first and then the set of equations expressed in (11). 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 ith 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 ith user at a lower rate by increasing the spreading factor. If all possible SF are tested and none leads to a feasible solution the ith user is not allowed to transmit in the next frame and the process skips to the next user in the priority table. For the next frame, the positions in the priority table may change due to either the arrival of new requests or the change in the priority level (TO_i is reduced in a unity for all users already in the table and L_i will be reduced, only for those users having transmitted in the frame, depending on the SF allocated).

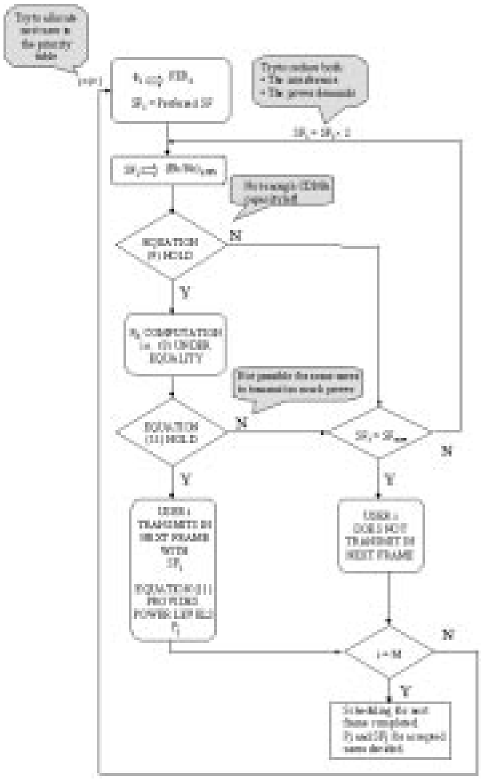

Try to allocate next user in the priority table
i=i+1
SF_i = Preferred SF
Try to reduce both:
• The interference
• The power demands
SF_i = SF_i × 2
EQUATION (9) HOLD
N
Y
P_i COMPUTATION i.e. (11) UNDER EQUALITY
EQUATION (11) HOLD
N
Y
N
Y
USER i TRANSMITS IN NEXT FRAME WITH SF_i
EQUATION (11) PROVIDES POWER LEVELS P_i
USER i DOES NOT TRANSMIT IN NEXT FRAME
i = M
N
Y
Scheduling for next frame completed. P_i and SF_i for accepted users decided

Fig. 6. Flow chart for the scheduling operation in uplink direction.

D. Downlink Formulation

The scheduling principles for the downlink operation are the same as explained in the previous section. However, in the case of downlink direction some differences in the formulation arise compared to the uplink case. In particular, the intercell interference will be user-specific since it depends on the user location. Thus the constraints for a given user will be expressed as

$$\frac{P_k \times SF_k}{N'_o \dfrac{1}{T_c} + \chi_k + \rho \times \left[\dfrac{P_T - P_{Tk}}{L_p(d_k)}\right] \times \Gamma} \geq \left(\frac{E_b}{N_o}\right)_{k,SF_k} \qquad k = 1 \cdots K \tag{12}$$

$$P_T = \sum_{i=1}^{K} P_{Ti} \tag{13}$$

$$P_k = \frac{P_{Tk} \times \Gamma}{L_p(d_k)} \tag{14}$$

where P_T is the base station transmitted power, P_{Tk} the power devoted to the kth user, and χ_k represents the intercell interference observed by the kth user. Similarly to (6), an equivalent expression is obtained

$$P_{Tk} \geq \frac{L_p(d_k)}{\Gamma} \, \frac{N'_O \dfrac{1}{T_c} + \chi_k + \rho \times \dfrac{P_T \times \Gamma}{L_p(d_k)}}{\dfrac{SF_k}{\left(\dfrac{E_b}{N_o}\right)_{k,SF_k}} + \rho} \qquad k = 1 \cdots K. \tag{15}$$

Again, adding all K inequalities it follows:

$$P_T \geq \frac{\displaystyle\sum_{i=1}^{K} \frac{\left(N'_O \dfrac{1}{T_c} + \chi_k\right)}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho} \, \frac{L_p(d_i)}{\Gamma}}{1 - \displaystyle\sum_{i=1}^{K} \frac{\rho}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho}}. \tag{16}$$

Additionally, physical limitations into the power levels are given by the maximum base station transmitted power $P_{T\max}$

$$P_T = \sum_{i=1}^{K} P_{Ti} \leq P_{T\max}. \tag{17}$$

Consequently, while for the uplink case (11) referred user specific conditions, in the downlink direction a similar condition appears as a unique expression because in this case the base station concentrates all users traffic

$$P_{T\max} \geq P_T \geq \frac{\displaystyle\sum_{i=1}^{K} \frac{\left(N'_O \dfrac{1}{T_c} + \chi_k\right)}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho} \, \frac{L_p(d_i)}{\Gamma}}{1 - \displaystyle\sum_{i=1}^{K} \frac{\rho}{\dfrac{SF_i}{\left(\dfrac{E_b}{N_o}\right)_{i,SF_i}} + \rho}}. \tag{18}$$

Noting that P_T is a positive magnitude it turns out again in (9), related to the CDMA capacity, by setting the denominator of (18) greater than zero. In case (18) holds, the power devoted to each user is readily provided by (15) with P_T being the minimum power satisfying (18).

E. Extension to Mixed Services Scenario

In case that other services in addition to WWW are supported, the above framework can be easily extended. Prioritization is initially carried out among services, i.e., conversational and streaming services first, then interactive services and finally background services. For the capacity requirement step, regardless of the strategy used, which could be different for different type of services, a target (E_b/N_o) and a certain SF would be eventually chosen. For example, the strategy followed for a conversation service could simply be a fixed SF and a fixed target for (E_b/N_o), instead of the more elaborated strategy to manage an interactive service. The availability check step would be the same as explained in previous sections, taking into account all users from all services and their respective SF and target (E_b/N_o).

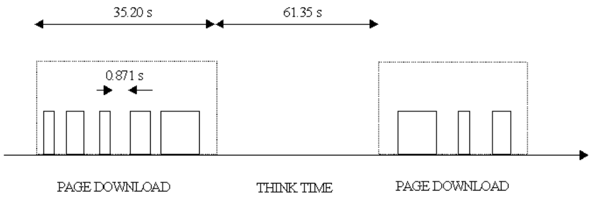


Fig. 7. Traffic model for the downlink WEB service.

III. 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 email (assumed a best effort service). The impact of including conversational and/or streaming services would mainly be a shift on the performance figures, as the prioritization step would give higher priority to conversational and/or streaming services and, consequently, less system capacity would be available for WWW and email.

For WWW service the traffic model presented in [11] is considered. This model considers for each user the download of WWW pages with a certain thinking time between them.

As Fig. 7 summarizes, the model parameters for the downlink direction are the following.

- *Page duration:* Lognormal distribution, average: 35.20 s, standard deviation: 134.11 s.
- *Thinking time between pages:* Lognormal distribution, average: 61.35 s, standard deviation: 144.72 s.
- *Number of packet arrivals per page:* It follows a lognormal distribution, with a delta of width 0.5 in $x = 1$, thus yielding a mean of 6.66 and a standard deviation of 24.31.
- *Number of bytes per packet in the downlink direction:* Lognormal distribution, average: 4827 bytes, standard deviation 41 008 bytes.
- *Time between packet arrivals:* Exponential distribution, average 0.871 s.

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. In this case, the number of bytes per packet follows a lognormal distribution with average 400 bytes and standard deviation 733 bytes.

For email service the model considered is a Poisson generation process with $\lambda = 0.001\,99$ mails/s. The packet length is lognormal with a mean of 11 877 bytes and a deviation 27 772 bytes [11].

The soft QoS-related values assumed for a typical WWW packet will be a “typical” delay below 0.5 s ($\tau_{\text{typ}} \leq$ 0,5 s) together with a maximum delay of 1.5 s in 95% of the cases ($\tau \leq$ 1.5 s). This delay threshold established for the radio interface should be suitably mapped from the desired WWW page delay, as WWW pages are related to upper layers. Thus, for a frame structure of 10 ms, the QoS parameters selected are $\tau_{\text{typ}} \leq 50$ frames and $\tau_{\text{máx}} \leq 150$ 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. In order to focus on the scheduling process, the WWW service is assumed to be supported through a Dedicated CHannel (DCH), although common packet channel (CPCH) could also be suitable. For the email service Forward Access CHannel (FACH) is assumed in the downlink direction and Random Access CHannel (RACH) in uplink direction, although in case emails include attachments CPCH could be more suitable [12] or even a transport channel type switching from RACH/FACH to DCH could be performed.

The cellular model includes the reference cell and six 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 neighboring cells, although performance statistics are only collected into the reference cell. The mobility model defined in [13] is considered with a mobile speed of 50 km/h. The propagation model used is also defined in [13].

The Gaussian hypothesis and perfect power control are assumed for CDMA interference characterization. For DCH closed-loop power control operates, while for RACH and FACH open-loop power control applies. 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 [14]. 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 $SF = 4$ (19 200 bits/frame) to $SF = 512$ (150 bits/frame) [12].

In order to show how the scheduling algorithm prioritizes at service level, Fig. 8 shows the average delay performance for the email and WWW users versus the throughput of email users in the downlink for different WWW load levels. Equation (1) with $n = 1$ is used as the priority function and Fig. 4 is considered with $\text{FER}_1 = 0.1$, $\text{FER}_2 = 0.01$, and $\phi*$ being the priority level when $TO = 1$ frame. As expected, when the number of WWW users is relatively low, email users are allocated the spare system capacity and, consequently, their performance is good. For heavy WWW loads, the best effort policy associated to the email 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.

On the other hand, Fig. 9 presents the percentage of DL WWW packets that experience a delay higher than the soft-QoS bound of 1.5 s. As 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.

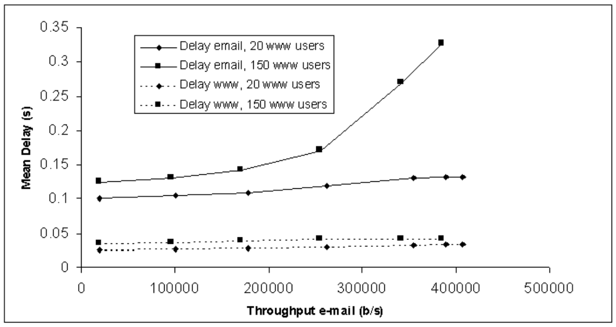


Fig. 8. Throughput delay for the email and WWW service with different numbers of WWW users.

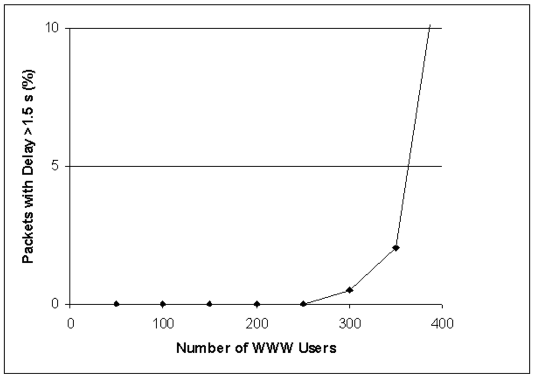


Fig. 9. Percentage of WWW packets that experience a delay higher than 1.5 s.

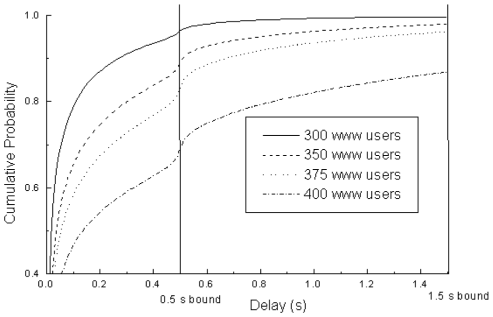


Fig. 10. Cumulative probability of the delay for different numbers of users.

In order to see how the scheduling algorithm operates, Fig. 10 presents the cumulative probability of the delay experienced by WWW packets for different numbers 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 0.5-s bound (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.

Impact of the Preferred SF Criteria: An important task of the scheduler is to decide the SF allocated to each user, which is a key parameter in the context of W-CDMA systems. In order

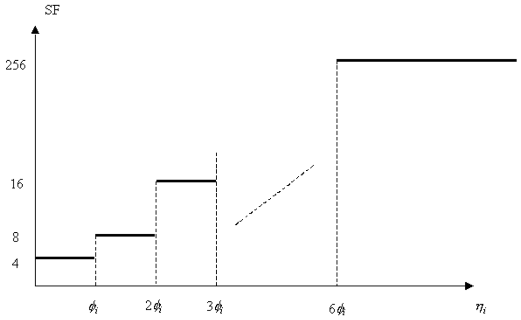


Fig. 11. Variation of the preferred SF as a function of the parameter η_I.

to gain some insight into the impact of such a decision many criteria could be proposed. For example, in Section II-B, a criterion trying to allocate the lowest possible SF has been introduced and will be referred as Criterion 1 in the following. This criterion tends to a "time scheduling" policy (high transmission rate and, consequently, ability to allocate few users).

For comparison purposes let define another criteria, referred to in the following as Criterion 2: let us assume that users are numbered according to the priority order, the scheduler has already allocated $(i-1)$ users for the next frame, and it is analyzing the ith user. Define

$$\eta_i = \sum_{j>i}^{K} \phi_j \tag{19}$$

where ϕ_j is the priority assigned to user j and K is the total number of users managed by the RRM in a given frame. Thus, η_i measures the added priority of the remaining users in the priority table. High η_i values indicate that after the current user there still remains high priority traffic to be served. Consequently, η_i tries to sense the status of the overall system load and choose a preferred SF as follows: for low values of η_i, the preferred SF will be low (high transmission rate) because the capacity required after this user is allocated is also low; for high values of η_i the reverse applies. So, from this parameter η_i the preferred SF could be derived from Fig. 11. Notice that the function must be discrete because the number of possible SF is limited. Also, as the SF selection function accepts many different forms, several simulations were carried out and Fig. 11 was finally retained as a suitable function.

In principle, the higher the transmission rate the better for the current user, and so the criterion will lead also to a table of ordered preferred spreading factors, from low SF (high transmission rate) to high SF (low transmission rate). According to the flow chart presented in Fig. 6, this criterion will select an initial spreading factor, $SF(\eta_i)$, instead of the initial selection of criterion 1 ($SF = 4$), and in case no feasible solution exists the scheduler will try to increase the spreading factor. This criterion tends to "time scheduling" for low loads situations (if there are few requests the transmission rate will be high) and to "code scheduling" for high loads situations (when there are many requests in the system the rates allocated are low in order to allocate as many users as possible).

Fig. 12 compares the performance achieved for WWW users in the uplink direction. Criterion 2 provides the best performance as it leads to an spreading factor adaptation to the overall system load conditions. In turn, criterion 1 is equivalent to criterion 2 for low loads (both tend to assign low SF) but for heavy loads it continues to assign low SF to the high priority users, which reveals not to be the most suitable policy.

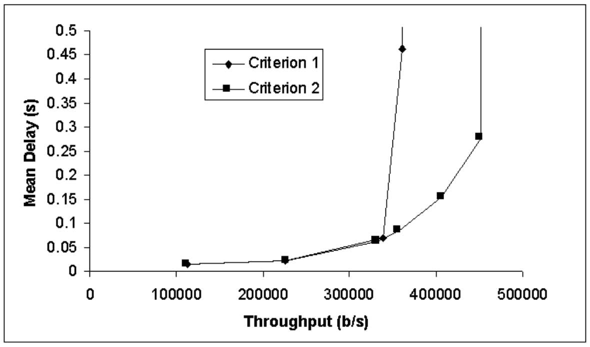


Fig. 12. Throughput delay of WWW users for different UL scheduling policies.

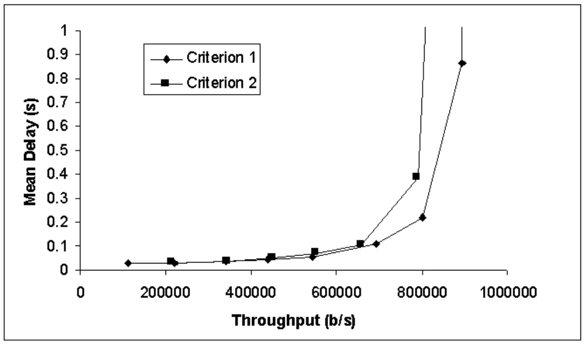


Fig. 13. Throughput delay of WWW users for different DL scheduling policies.

Fig. 13 shows the results for the downlink. In this case, criterion 1 offers better performance than criterion 2. The reason is that traffic characteristics in the downlink are quite different than in the uplink, as packets are much longer. As criterion 1 tends to allocate many bits per frame for high priority users regardless of the system load, for heavy loads this policy is better in order to meet the delay bound requirement. We note that for heavy loads criterion 2 tends to allocate few bits per frame to a large number of users and this leads to longer delays. So, differences in the conclusions reached for uplink and downlink cases reveal that a packet scheduler needs to consider a number of system aspects, for example the traffic characteristics. In this sense, it is worth noting that the propagation environment may have an impact on different scheduling strategy performances. For example, the orthogonality factor ρ affects the "code scheduling" capabilities in the downlink direction, so that the lower the multipath propagation in the environment the better the "code scheduling" will perform.

Impact of the Power Management Criteria: Another important issue in the scheduling process is the power management criterion, as the physical power restrictions have a straight impact on whether a feasible power allocation solution exists or not. In Section III and for the uplink case, the imposed conditions are given by the maximum transmitted power by the mobile terminal. Thus, users closer to the base station will be fa-

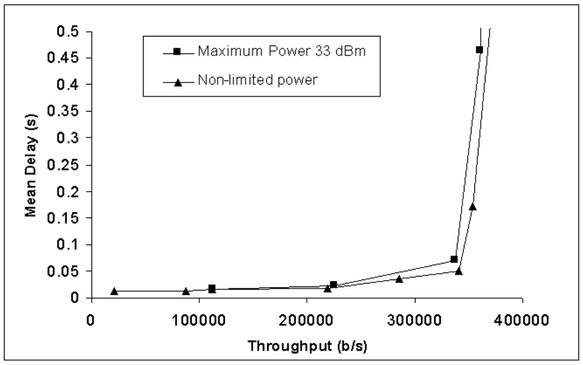


Fig. 14. Throughput delay for the UL with and without power restrictions.

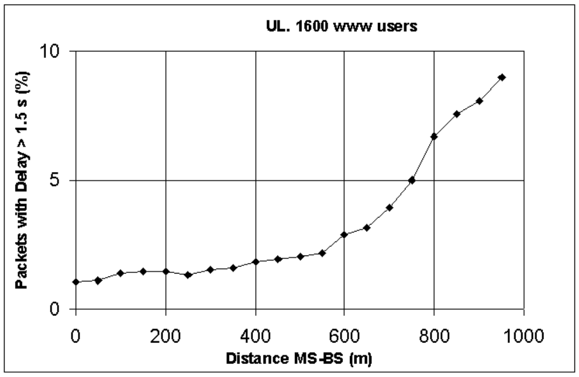


Fig. 15. Effect of cell breathing.

vored because of the lower path loss. According to Fig. 14, it should be noted that the performance achieved is not very far from the performance obtained without power limitations (assuming infinite transmitted power). This matter points out that the system tends to be limited in many cases by the inherent limit of CDMA (9), and consequently the considered maximum transmitted power does not limit seriously the system capacity.

The effect of cell breathing can be observed in Fig. 15, where the percentage of packets exceeding the delay threshold as a function of the distance to the base station is plotted. It can be observed how those users located close to the BS meet the soft-QoS criterion (5% of packets with a delay higher than 1.5 s), while those located at the cell edge do not meet the criterion.

IV. RRM Implementation

The emulation of the radio interface protocol stack has been presented in the introduction section. From the RTE point of view the input will be packet data units (PDU) generated by higher layers at the transmitter side of the reference user and the RTE will emulate its transmission through the radio channel (including the effect of all the other users sharing the access) under the scheduler control and will deliver the received message to the upper layers of the receiver side.

Offline link level simulations are carried out in order to obtain the statistical behavior of the radio channel, represented as a Markov chain that in the RTE will reproduce that behavior in terms of error distribution. Thus, long offline simulations are switched to fast transition probabilities coming out from the Markov transition matrix, able to run in real time within a DSP board.

TABLE I
Possible Transport Formats

Format 1	No transmission
Format 2	Transmission with SF=256 (150 bits/frame)
Format 3	Transmission with SF=128 (300 bits/frame)
Format 4	Transmission with SF=64 (600 bits/frame)
Format 5	Transmission with SF=32 (1200 bits/frame)
Format 6	Transmission with SF=16 (2400 bits/frame)
Format 7	Transmission with SF=8 (4800 bits/frame)
Format 8	Transmission with SF=4 (9600 bits/frame)

In the case of the RRM sublayer a similar approach is followed. We note that the scheduling algorithm is in charge of deciding whether a specific user is allowed to transmit or not in the next frame and, in the former case, which is the spreading factor to be used for this packet transmission. The spreading factor together with the transmitted power in order to achieve a certain SIR target (set by the outer loop power control) are the relevant aspects to be taken into account. Thus, offline system level simulations including the scheduling algorithm proposed in Section II are carried out in order to capture the statistical behavior that characterizes dynamic spreading factors allocation to users, the interference statistics derived from these spreading factors assignments, and the outer loop power control operation. The resulting statistical behavior, which clearly embeds the criteria followed in the scheduling algorithm, will be first retained in terms of lookup tables and later on applied to drive the permission to transmit in real time for a reference user.

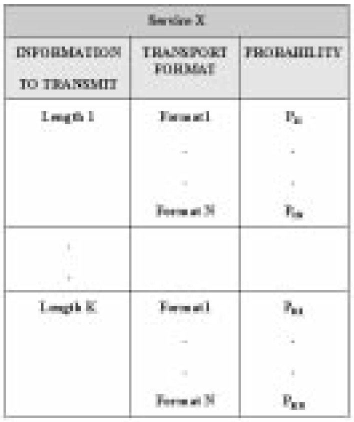
Thus, from the emulation point of view, the output of the RRM sublayer should be to decide the transport format to apply to the reference user. Possible transport formats for the case of a DCH in the UL are shown in Table I.

More specifically, two lookup tables are generated to capture the statistical RRM behavior.

- The first lookup table contains the transport format probability as a function of the number of bits per packet to be transmitted, quantified in K levels, each level covering a certain range of information length maintained in the transmitter buffer. Then, given a specific service and the amount of information, the probability that any of the N transport formats is assigned by the scheduling mechanism in the next frame is kept in a lookup table as P_{ij} (see Table II), corresponding to the probability of applying Format j (see Table I) whenever the amount of information to be transmitted is in the range of Length i.
- The second lookup table (Table III) contains the histograms of the E_b/N_o distribution as a function of the transport format decided by the RRM module.

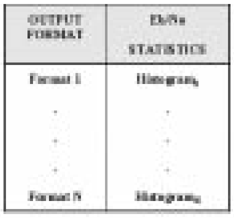
Note that different scheduling algorithms would lead to different values in the lookup tables. Thus, the RTE will be able to test different services under different RRM strategies by simply switching the memory area in the DSP board.

TABLE II
LOOKUP TABLE FOR DECIDING THE TRANSPORT FORMAT



Service X		
INFORMATION TO TRANSMIT	TRANSPORT FORMAT	PROBABILITY
Length 1	Format1 . . Format N	P_{11} . . P_{1N}
. .		
Length K	Format1 . . Format N	P_{K1} . . P_{KN}

TABLE III
LOOKUP TABLE FOR DECIDING THE E_b/N_o LEVEL



OUTPUT FORMAT	Eb/No STATISTICS
Format 1 . . . Format N	Histogram_1 . . . Histogram_N

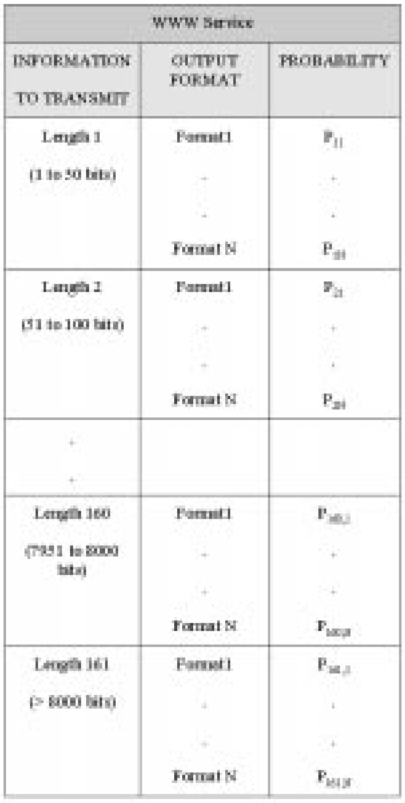
In summary, the RRM implementation is based on system level simulations to derive the transport formats statistics in conjunction with the interference statistics. During the emulation process, packets from higher layers will arrive at the emulator. According to the RRM lookup tables a transport format as well as an interference level will be selected for the current frame. These parameters select a specific Markov chain (physical layer performance obtained after link level simulations) and the packet is delivered with a certain number of errors (if any) on it. The packet is transferred to higher layers that may ask for a retransmission (RLC sublayer) or pass the packet up to the application layer to view the performance of the emulated service.

V. IMPLEMENTATION VALIDATION

In order to validate that the statistical behavior of the scheduling algorithm is suitably reproduced in the emulation by means of lookup tables, according to the methodology proposed in Section IV, a validation procedure is carried out. It consists of obtaining several performance measurements in the offline system level simulations and then comparing them with those provided by the RTE.

When constructing the probability tables, a parameter to be defined in the implementation approach is the granularity value on the packet length definition (range on packet length). For this purpose, different length granularities have been checked: 50, 100, 200, 500, 1000, 2000, 4000, and 8000 bits (i.e., a range of 100 bits, for example, means that each row of the lookup table covers 100 bits) in the case of WWW service. A maximum packet length of 8000 bits has been considered (note that from Section III the mean packet length in the UL is 400 bytes = 3200 bits, with a deviation of 733 bytes = 5864 bits). Thus, for the case of 50 bits granularity the table would look like Table IV.

TABLE IV
LOOKUP TABLE FOR DECIDING THE TRANSPORT FORMAT



WWW Service		
INFORMATION TO TRANSMIT	OUTPUT FORMAT	PROBABILITY
Length 1 (1 to 50 bits)	Format1 . . Format N	P_{11} . . P_{1N}
Length 2 (51 to 100 bits)	Format1 . . Format N	P_{21} . . P_{2N}
. .		
Length 160 (7951 to 8000 bits)	Format1 . . Format N	$P_{160,1}$ . . $P_{160,N}$
Length 161 (> 8000 bits)	Format1 . . Format N	$P_{161,1}$ . . $P_{161,N}$

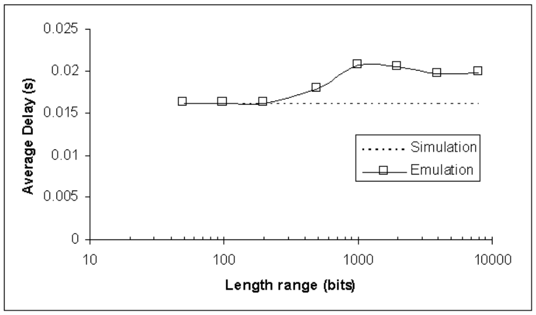


Fig. 16. Average delay comparison for different length ranges in the emulator.

For a system load of 600 users, the average packet delay in the UL depending on the granularity in the lookup table is shown in Fig. 16. It can be observed that for granularity ranges below 200 bits (40 to 160 entries per table) the first moment is suitably reproduced in the emulator. For the second moment, again granularities up to 200 bits show the same value obtained from the emulation and the simulation (see Fig. 17).

The cumulative probability derived for 200 bits granularity agrees with the simulated one (see Fig. 18), whereas when the granularity is 2000 bits (see Fig. 19) significant differences appear.

Thus, the maximum length range for which the behavior of the emulated RRM remains good and agrees with the simulated

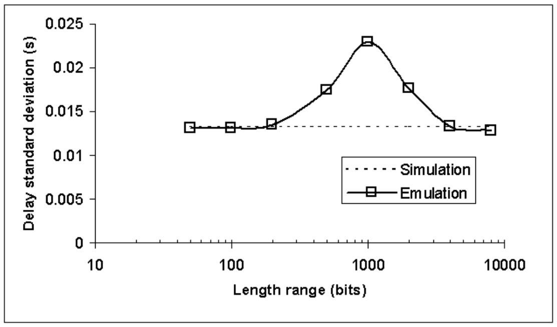


Fig. 17. Standard deviation comparison for different length ranges in the emulator.

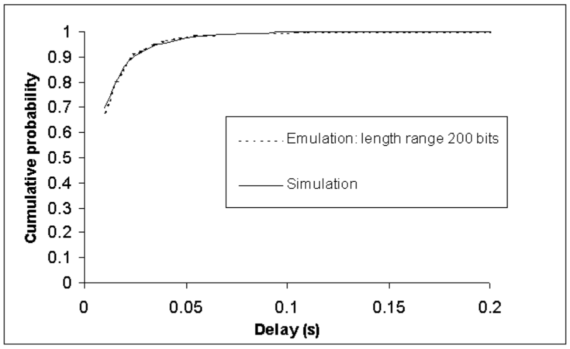


Fig. 18. Comparison between simulation and emulation with range 200 bits.

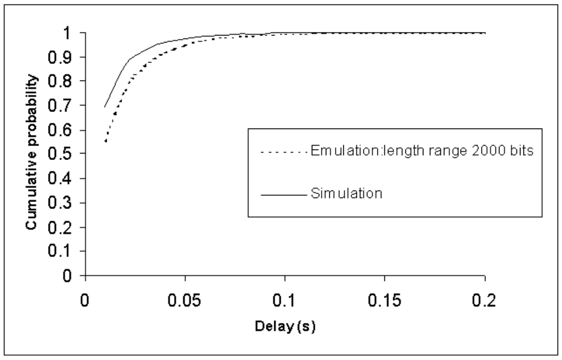


Fig. 19. Comparison between simulation and emulation with range 2000 bits.

RRM is around 200 bits. This leads to a very affordable memory requirement in the RRM implementation within the RTE framework.

VI. Conclusion

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 maximize 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 environments, in particular in the W-CDMA scenario selected in UMTS, 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. Results reveal that so many different issues impact on the scheduling process that the optimization process is a difficult task. In particular, it has been shown that different traffic characteristics for uplink and downlink, in the case of WWW, lead to different suitable criteria into the scheduling algorithm. While for long packets (downlink) a "time scheduling" policy offers a better performance, for shorter packets (uplink) a "code scheduling" policy is more suitable.

This work is part of the IST Wineglass project, where the implementation of a testbed is envisaged in order to test and validate services in an environment as close as possible to the real world. An important objective of the RTE will be the ability to demonstrate the system performance under different RRM strategies. So, an RRM implementation procedure by means of lookup tables allowing the real time emulation of the scheduling algorithm has also been proposed in the paper. The implementation approach has been validated comparing statistical results from the RTE and simulations, concluding that the lookup tables properly capture the statistical behavior at RRM level.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments.

References

[1] [Online]. Available: http://www.3gpp.org.
[2] [Online]. Available: http://www.cordis.lu/ist.
[3] A. Umbert and P. Díaz, "A radio channel emulator for WCDMA, based on the hidden Markov model (HMM)," in *Proc. IEEE Vehicular Technol. Conf.* , Boston, MA, Sept. 2000, pp. 2173–2179.
[4] S. Blake *et al.*, "An architectrure for differenciated services,", IETF RFC 2475, Dec. 1998.
[5] Y. Bernet, "The complementary roles of RSVP and differential services in the full-service network," *IEEE Commun. Mag.*, Feb. 2000.
[6] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proc. IEEE*, vol. 83, pp. 1374–1396, Oct. 1995.
[7] M. Naghshineh and A. S. Acampora, "QoS provisioning in micro-cellular networks supporting multiple classes of traffic," *Wireless Networks*, vol. 2, pp. 195–203, 1996.
[8] ——, "Design and control of microcellular networks with QoS provisioning for data traffic," *Wireless Networks*, vol. 3, pp. 249–256, 1997.
[9] I. F. Akyldiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. Networking*, vol. 7, no. 2, pp. 146–158, Apr. 1999.
[10] S. K. Das *et al.*, "A call admission and control scheme for QoS provisioning in next generation wireless networks," *Wireless Networks*, vol. 6, pp. 17–30, 2000.
[11] M. Bartoli, G. Foddis, D. Minervini, and M. Molina, "Modelli di traffico E-mail e WWW per il servizio GPRS," Internal Rep. CSELT, Feb. 2000.
[12] 3GPP TS 25.211 v3.2.0, "Physical channels and mapping of transport channels onto physical channels (FDD),".
[13] ETSI TR 101.112 v3.2.0, "Selection procedures for the choice of radio transmission technologies of the UMTS,".
[14] K. Parsa, S. S. Ghassemzadeh, and S. Kazeminejad, "Systems engineering of data services in UMTS W-CDMA systems," in *IST Proc.*, Galway, Ireland, Oct. 2000, pp. 435–444.

Oriol Sallent (M'98) received the Engineer and Doctor Engineer degrees in Telecommunication from the Universitat Politècnica de Catalunya (UPC), Spain, in 1994 and 1997, respectively.

He joined the Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona, where he became Assistant Professor in 1994 and Associate Professor in 1998. His research interests include the field of mobile communication systems, especially packet radio techniques and spread-spectrum systems.

Dr. Sallent received the Doctorate Award from the Telecommunication Engineer Association of Spain in 1997.

Jordi Pérez-Romero (S'98) was born in Barcelona, Spain, in 1974. In 1997, he received the Engineer degree in Telecommunications in the Escola Tècnica Superior d'Enginyeria de Telecomunicació, from the Universitat Politècnica de Catalunya, UPC, and the Ph.D. degree from the same university in 2001.

He is currently an Assistant Professor in the field of radio communications. He has been involved in different European projects and his research interests include the packet transmission mechanisms and the radio resource management strategies for CDMA mobile communications networks.

Fernando J. Casadevall (M'87) received the Engineer of Telecommunication and Dr. Engineering degrees from the Universitat Politécnica de Catalunya (UPC), Spain, in 1977 and 1983, respectively.

In 1978 he joined UPC, where he was an Associate Professor from 1983 to 1991. He is currently a Full Professor in the Signal Theory and Communications Department. His current research interests include the performance analysis and development of digital mobile radio systems, in particular cellular and personal communication systems, multipath transceiver designs (including Software Radio techniques), mobility and radio resources management. He has published more the 50 technical papers in both international conferences and magazines. From October 1992 to January 1996, he was responsible for the Information Technology Area in the National Agency for Evaluation and Forecasting (Spanish National Research Council).

Ramon Agustí (M'78) was born in Riba-roja d'Ebre (Tarragona), Spain, on August 15, 1951. He received the Engineer of Telecommunications degree from the Universidad Politécnica de Madrid, Spain, in 1973, and the Ph.D. degree from the Universitat Politècnica de Catalunya, Spain, 1978.

In 1973 he joined the Escola Tècnica Superior d'Enginyers de Telecomunicació de Barcelona, Spain, where he became a Full Professor in 1987. For the last 15 years his research interests have included performance analysis and development of planning tools and equipment for mobile communication systems, and he has published nearly 100 papers in these areas. He has also been an Advisor to Spanish and Catalonian Governmental Agencies (DGTel, CICYT, ANEP, and CIRIT) on issues concerning mobile communications.

Dr. Agustí received the Catalonia Engineer of the Year Prize in 1998. He is part of the editorial board of several scientific international journals.