Wireless Access to the World Wide Web in an Integrated CDMA System

Cristina Comaniciu   Narayan B. Mandayam   David Famolari   Prathima Agrawal
EMail: ccris,narayan@winlab.rutgers.edu; fam,pagrawal@research.telcordia.com

Index Terms: *CDMA, power control; admission control; packet scheduling; web browsing sessions.*

Abstract

In this paper we propose a wireless access mechanism for web traffic packets in an integrated wireless CDMA system, that has both voice and web traffic. The proposed scheme is a MAC/LL (Medium Access Control layer/Link layer) scheduling algorithm that consists of a two level control: admission control and packet scheduling. The admission control restricts the number of users in the system such that QoS requirements (target signal to interference ratio, and delay) for both voice and web traffic can be met. The packet scheduling balances the system interference on a slot by slot basis such that the target signal to interference ratios (SIRs) can be achieved for all users (voice and web browsing sessions) with a higher scheduling priority for voice. Designing admission control for web users based on the average offered rate per session is difficult due to the high variations in the offered load generated by heavy tailed distributions for web traffic. To overcome this problem, we propose an admission control algorithm that adaptively estimates the aggregate average load based on load measurements using a sliding observation window.

Parts of this paper were presented at ICME'00 and VTC'00 fall. This work is supported in part by the NSF under Grant No. NCR 97-06036 and through the NSF-KDI program under grant number IIS-98-72995, and by Telcordia Technologies.

1 Introduction

In recent years, there has been a great amount of activity for developing the next generation wireless networks, which are expected to provide a wide range of services, such as voice, data, video, and web traffic. Effective MAC (Medium Access Control) protocols have to be implemented to handle conflicting QoS requests from different types of traffic, and to achieve efficient system resource utilization. There has been a lot of interest in designing access protocols for wireless multimedia services (e.g. [2, 3, 6, 9, 11, 12] and the references therein). However, very few papers consider non Poisson traffic in wireless systems. For wireline networks, WAN traffic is modeled as a self-similar process [4]. Self-similarity is obtained as the effect of multiplexing many on/off sources with heavy-tailed On/Off period lengths [4]. For the case of a single cell CDMA system, the number of web users may not be large enough for the assumption of self-similarity to hold. Hence, it is more appropriate to design an access mechanism by considering models for the web browsing sessions [19], rather than using a self-similarity model for the aggregate traffic. To provide QoS for both voice and web users in an integrated CDMA system, we propose a two level access control: at a call arrival time scale (admission control) and at a time slot scale (packet scheduling).

Designing admission control for web users based on the average offered rate per session is difficult due to the high variations in the offered load generated by heavy tailed distributions for web traffic. To overcome this problem, the proposed admission control adaptively estimates the average web load aggregated from all www sessions currently admitted in the system. The aggregate average load estimation is based on load measurements using a sliding observation window. Our work is motivated by [7] which discusses admission control for ATM networks. [7] proves that significant statistical multiplexing gains can be achieved if the admission criterion is based on instantaneous load measurements and the admission threshold is derived using a decision theoretic approach.

At a slot time scale, resources are shared among web users according to the residual capacity left over after subtracting the voice contribution. Two different schemes for predicting the future voice load are proposed and compared. The “one step prediction” method is similar with the prediction technique proposed in [6]. In [6] it is shown that a simple predictive scheme as the one used in this paper outperforms other low complexity non-predictive methods such as the one discussed in [2]. However, the analysis presented in [2] and [6] is not suitable for www admission control. The "mean value prediction"scheme represents the case with no explicit access control and it is analyzed for comparison purposes.

We design admission and packet scheduling based on the "noise rise" [1] condition, which was proved in [10] to satisfy the power control feasibility condition as well. In this paper we focus on the single cell case with perfect power control. Under more realistic assumptions (imperfect power control and multiple cells), the admission control imposes supplementary constraints [3, 17, 18]. We also show how the analysis can be extended for the imperfect power control and the multiple cell scenarios as well. We show that, although the analysis and the system capacity changes, the proposed admission and packet scheduling mechanisms remain the same. As a final observation, we note that the analysis is general enough to include a variety of elastic data traffic models. In a related paper [8], we have analyzed a more general case, where voice, data and video traffic are present in the CDMA system.

The paper is organized as follows: in section 2 we present the system and data models (voice and web traffic) used in this work; section 3 describes the proposed access control with emphasis on the admission control in section 3.1, and packet scheduling in 3.2; section 3.3 extends the analysis for the multiple cell case; section 3.4 presents simulation results; and section 4 presents the conclusions.

2 System and Data Models

We consider the scenario in which a set of power controlled mobile terminals are transmitting packets to a base station, in a single cell, direct sequence CDMA system. Admission and packet scheduling algorithms are proposed for the uplink, which is considered to be slotted, with a slot duration equal to a frame duration. Two main categories of traffic are discussed: web browsing sessions and voice. Each web session can vary its transmission rate by increasing or decreasing the spreading gain. A web browsing session (see Figure 1) [19] is comprised of a number of packet calls, N_{pc}, (Geometrically distributed) separated by reading times (Geometrically distributed inter-arrival time between packet calls). Each packet call consists of a number of IP packets (N_p), which is Pareto distributed, and has a Geometrically distributed inter-arrival time between packet arrivals (T_i). The length of a web traffic packet is fixed. Web browsing sessions arrive according to a Poisson distribution with arrival rate λ_S. The above description is proposed in [19] for modeling large file uploads, and also the reverse traffic corresponding to www downloads. In our simulations we assume that the new session requests also include sessions handed off from other cells (arriving with a Poisson distribution as in [13]). No admission priority is assigned to handed off web session connection requests. The dwell time of a www session in the cell is modeled by an exponential distribution with mean μ_d [13].

The QoS requirements for both voice and web users are bit error rate (BER) and the access delay for packets. The BER target is mapped into a Signal to Interference Ratio target. The performance measure of interest is the SIR outage probability: the probability that the target SIR cannot be achieved. In terms of delay, voice users are delay intolerant and are given priority over web calls which are delay insensitive. Only an average delay performance is required for the web traffic. Using an analogy with a queueing system, the average delay experienced by web packets is directly determined by the aggregate web load in the system. As a consequence, we consider as QoS requirements for web browsing sessions the target SIR and the aggregate www traffic load.

As in [2] and [6], the voice activity is modeled as a continuous ON/OFF Markov process, with a transition rate μ from ON to OFF and λ from OFF to ON states. The voice and data transmission occurs in slots of duration T_s. T_s is equal to a frame duration. Therefore, the cumulative voice activity process can be approximated as a discrete Markov chain. The transition probabilities for N_v voice users in the system and $v(n)$ active voice users in the n^{th} time slot are given by:

$$\begin{cases} P\{v(n+1) = k-1 | v(n) = k\} = \frac{\mu_k}{\lambda_k + \mu_k}(1 - \exp(-(\lambda_k + \mu_k)T_s)) \\ P\{v(n+1) = k | v(n) = k\} = \exp(-(\lambda_k + \mu_k)T_s) \\ P\{v(n+1) = k | v(n) = k\} = \exp(-(\lambda_k + \mu_k)T_s). \end{cases} \tag{1}$$

where: $\lambda_k = \lambda(N_v - k)$, $\mu_k = \mu k$.

The stationary probability of state k, in which k users are active is given by:

$$P_V(k) = \begin{pmatrix} N_v \\ k \end{pmatrix} \left(\frac{\lambda}{\mu}\right)^k \Big/ \sum_{j=0}^{N_v} \begin{pmatrix} N_v \\ j \end{pmatrix} \left(\frac{\lambda}{\mu}\right)^j . \tag{2}$$

3 Access Control for QoS Guarantees

The role of the access control is to balance the system utilization such that QoS requirements are met for all users. We assume that SIR requirements for all calls in the system can be met if the total system load is maintained below some threshold [1], [10]. Denoting the system bandwidth as W, the "noise rise" condition can be expressed as:

$$Z \triangleq \sum_{i=1}^{N_v} SIR_i^v R_i^v + \sum_{i=1}^{N_w} SIR_i^w R_i^w \leq W(1-\eta), \tag{3}$$

where N_v and N_w, represent the number of voice calls and www calls, R_i^v and R_i^w are the transmission rates for the i^{th} voice and www call, SIR_i^v and SIR_i^w are the target SIRs for voice and www, and Z is the system load. The noise rise coefficient η represents the ratio of the background noise power spectral density to the total received interference power density. Note that we do not distinguish between handed off web sessions and new web session initiations in the cell, while accounting for the number of www calls, N_w. Condition (3) should hold at all time scales. A two level access control, implemented via admission control and packet scheduling is proposed. The packet scheduling is based on (3) expressed at the time scale of frames. It assigns resources for the web traffic according to the residual capacity available after the voice traffic contribution has been subtracted. An imperfect scheduling resulting in higher system load constitutes an SIR outage, since SIR targets cannot be met for the slots in which the "noise rise" condition does not hold. Thus, the SIR outage probability characterizes the performance of the packet scheduling scheme.

The admission control guarantees the feasibility of the QoS specifications for all users (SIRs and delay). It is based on the "noise rise" condition in (3) specified at the call arrival time scale. The admission control for web users is characterized by a new performance measure, the load outage probability, which represents the probability that, after admission, the obtained aggregate www traffic load in the cell is higher than a prespecified target value.

3.1 Admission Control

The admission control reserves resources for the aggregate web traffic, based on voice traffic descriptors: average and peak transmission rates, and on QoS requirements for the web sessions: average throughput and minimum throughput. In order to reserve a minimum throughput for the www users (R^w_{min}), while guaranteeing zero delay for the voice traffic for the worst case in which all voice users transmit at peak rates (R^v_{max}), the maximum load condition should hold:

$$Z_{max} \triangleq \sum_{i=1}^{N_v} SIR^v_i (R^v_{max})_i + \sum_{i=1}^{N_w} SIR^w_i (R^w_{min})_i \leq W(1-\eta). \tag{4}$$

An average rate reservation for the i^{th} www session ($(R^w_{av})_i$) can be guaranteed if the average load condition holds:

$$Z_{av} \triangleq \sum_{i=1}^{N_v} SIR^v_i (R^v_{av})_i + \sum_{i=1}^{N_w} SIR^w_i (R^w_{av})_i \leq W(1-\eta), \tag{5}$$

where $(R^v_{av})_i$ represents the average offered voice rate for user i.
As in [8], a new call can be admitted in the system if (4) and (5) simultaneously hold.
By analyzing (5), it can be seen that it requires the apriori knowledge of R^w_{av} for each www session. In order to determine R^w_{av}, its relationship with the offered throughput per web session has to be first established. Assuming that all web sessions have the same SIR target, we can define the aggregate rate reservation available for www users (R^*_w) as:

$$R^*_w = \sum_{i=1}^{N_w} (R^w_{av})_i = \frac{W(1-\eta) - \sum_{i=1}^{N_v} SIR^v_i (R^v_{av})_i}{SIR^w}. \tag{6}$$

Using an analogy with a queueing system, the web traffic will be serviced by a queue having a rate of service R^*_w, and input offered traffic $\sum_{i=1}^{N_w} \Gamma^w_i$, with Γ^w_i being the offered rate for the i^{th} www session. The average delay experienced by the www packets is directly influenced by the aggregate www load in the system (ρ_w). To avoid very large delays, a target load (ρ_t) for the queue has to be imposed:

$$\rho_w = \frac{\sum_{i=1}^{N_w} \Gamma^w_i}{R^*_w} \approx \rho_t < 1. \tag{7}$$

Note that condition (7) is based on (5) and represents the admission condition for the web users. If a minimum throughput requirement is specified by the web sessions, (7) will be used in conjunction with (4) for web traffic admission. Without loss of generality, $R^w_{min} = 0$ throughout the remainder of the paper.
We notice that condition (7) requires the apriori knowledge of offered throughput per session. However, in a www session, the number of offered packets per packet call is Pareto distributed, therefore Γ^w is a random

variable with a very large variance. As a consequence, the sample mean estimate obtained for the life of a session will be far away from the mean of the random variable Γ^w. Thus, we cannot rely on the statistical mean to describe the offered rate per www session: the expected value for the average offered rate per www session will greatly over or under reserve resources for that particular session. Instead of statistically describing the aggregate www offered load, we estimate ρ_w by measuring at each time slot the average load for the last W_{size} slots. If the difference between the target load and the current measured average load ($\rho^{meas}(n)$) for the n^{th} time slot is greater than some threshold T_ρ, new sessions can be admitted. This threshold is chosen such that, with high probability, the load for the next slot is less that the target load plus an acceptable performance tolerance parameter, ϵ.

We summarize the admission control algorithm as follows:

At the end of each time slot n:

- Measure the new total offered load averaged over the last W_{size} slots. W_{size} is the length of a moving average window over which samples are collected for average load measurement for slot n. The measured average load for slot n ($\rho^{meas}(n)$), is given by:

$$\rho^{meas}(n) \triangleq \frac{1}{min\{n, W_{size}\}} \sum_{j=\delta}^{n} (N_{arr})_j, \tag{8}$$

  where $\delta = max\{0, n - W_{size}\}$, $(N_{arr})_j =$ the number of IP packets arriving in slot j. $(N_{arr})_j$ is computed at the base station by adding up all arrival rate measurements received from all active www terminals.

- Admit all new connection requests in the next time slot, if

$$\Delta\rho(n) \triangleq (\rho_t - \rho^{meas}(n)) \geq T_\rho, \quad T_\rho \geq 0; \tag{9}$$

  Otherwise, reject all new connections for the next time slot.

The admission control performance is characterized by the load outage probability, i.e., the probability that, after admission, the obtained aggregate www load in the cell is higher than a prespecified value $\rho_t + \epsilon$. The load for the next time slot $n+1$ can be computed as the load previously measured in slot n plus the new offered load (denoted as $\rho_{new}(n+1)$) from all sources starting to transmit in slot $n+1$:

$$\rho(n+1) = \rho^{meas}(n) + \rho_{new}(n+1). \tag{10}$$

To simplify the computation in (10), we make the conservative assumption that no active packet call in slot n will end in slot $n+1$. Defining the admission of new sessions in the n^{th} time slot as the event $A = \{(\Delta\rho(n) \geq T_\rho) \text{ and } (T_\rho \geq 0)\}$, the load outage probability can be expressed as:

$$P_{out\_load|A} \triangleq P\{\rho_{new}(n+1) \geq \Delta\rho(n) + \epsilon|A\}. \tag{11}$$

The admission threshold T_ρ is determined such that the load outage probability for any slot n, is maintained below a prescribed value ψ.

$$T_\rho \triangleq min\{\Delta\rho(n) : (P_{out\_load|A} \leq \psi)\}. \tag{12}$$

We now derive an expression for the load outage probability in (11) as follows. $\rho_{new}(n+1)$ can be expressed as the product of the number of new packet calls offered in slot $n+1$ (denoted as $N_{new}(n+1)$), and the average rate offered per packet call (R_{pc}), divided by the reserved rate R_w^*:

$$\rho_{new}(n+1) = \frac{N_{new}(n+1)R_{pc}}{R_w^*}. \tag{13}$$

The rate offered per packet call is a random variable specified as:

$$R_{pc} = \frac{N_p m}{(N_p - 1)T_i}, \tag{14}$$

where N_p is the number of IP packets in a packet call, m is the length of an IP packet (measured in number of basic packet lengths), and T_i represents the interarrival time between IP packets.

The number of new packet calls offered in slot $n+1$ comprises of new session arrivals, and of new packet call initiations from previously admitted sessions that are currently inactive. The number of new sessions in slot $n+1$ (denoted as $N_s(n+1)$) is Poisson distributed with rate $\lambda_s T_s$ sessions per unit slot T_s, and the number of new packet call initiations in the next time slot is a binomial distributed random variable ($N_b(n+1)$). For a given number N_i of inactive web calls in the system, $N_b(n+1)$ is characterized by the probability mass function:

$$P_{N_b}(l) = \begin{pmatrix} N_i \\ l \end{pmatrix} p_c^l (1-p_c)^{N_i - l}, \tag{15}$$

where p_c represents the probability of a new packet call initiation in the next time slot.

Therefore, $N_{new}(n+1)$ is a random variable expressed as:

$$N_{new}(n+1) = N_b(n+1) + N_s(n+1). \tag{16}$$

Equation (13) can now be rewritten as:

$$\rho_{new}(n+1) = \frac{(N_b(n+1) + N_s(n+1))}{R_w^*} \frac{N_p m}{(N_p - 1)T_i}. \tag{17}$$

N_p is Pareto distributed [19] with the characteristic function given by

$$F_{N_p}(x) = 1 - \left(\frac{b}{x}\right)^{\alpha}, \; x \geq b \; ; \; \alpha = 1.1 \; ; \; b = 2.27. \tag{18}$$

The load outage probability, conditioned on N_i (the number of inactive web sessions in the CDMA system), N_b, T_i and N_s can be expressed as:

$$P_{out\_load|(A,N_i,N_b,T_i,N_s)} = \begin{cases} 1 \, , & \text{if } \frac{m(N_b+N_s)}{D^*(\Delta\rho)T_i} \geq 1 \\ 0 \, , & \text{if } \left(1 - \frac{m(N_b+N_s)}{D^*(\Delta\rho)T_i}\right) \geq \frac{1}{b} \\ 1 - b^{\alpha}\left(1 - \frac{m(N_b+N_s)}{D^*(\Delta\rho)T_i}\right)^{\alpha}, & \text{otherwise} \end{cases} \tag{19}$$

where $D^*(\Delta\rho)$ is defined as:

$$D^*(\Delta\rho) = (\Delta\rho + \epsilon)R_w^*. \tag{20}$$

The derivation of (19) is presented in the Appendix.

By averaging over the probability mass functions for the number of new session arrivals (P_{N_s}, with parameter $\lambda_s T_s$), the interarrival time between packets (P_{T_i}, with parameter p_i) and the number of new packet call initiations from inactive sessions (P_{N_b} with parameters N_i and p_c), we can express the load outage probability conditioned on the number of inactive web sessions in the system:

$$P_{out\_load|(A,N_i)} = \sum_{l=0}^{N_i}\sum_{j=1}^{\infty}\sum_{k=0}^{\infty} P_{out\_load|(A,N_i,N_b,T_i,N_s)} P_{N_s}(k) P_{T_i}(jt_i) P_{N_b}(l), \tag{21}$$

where t_i is the discrete unit of time for the packet interarrival. $P_{out\_load|(A,N_i)}$ depends on the particular value that $\Delta\rho(n)$ may take, and on the number of inactive sessions currently in the system (N_i). For notation simplicity, we denote $P_{out\_load} \triangleq P_{out\_load|(A,N_i)}$.

Analyzing (19), we notice that, due to a very small interarrival time between packets, variations in the factor $D^*(\Delta\rho)$ have a negligible impact on the load outage probability value. Thus, the dependence of P_{out_load} on $\Delta\rho(n)$, for any time slot n, is negligible. In Figure 2, P_{out_load} is plotted as a function of both the number of inactive sessions currently admitted in the cell (N_i) and the arrival rate of new sessions (λ_s). For a fixed constraint for P_{out_load} ($P_{out\_load} \leq 0.025$), the admission threshold T_ρ is plotted in Figure 3. It can be seen that T_ρ is either 0 (for the feasible region) or 1 (infeasible region). For the first case, in an arbitrary slot n, all calls can be admitted regardless of how close the current measurement is to the target load, provided that $\Delta\rho(n) \geq 0$, whereas for the second case no call can be admitted for the given pair (λ_s, N_i). The load outage probability is predominantly affected by the arrival rate of new sessions. To limit the arrival rate of new sessions to a desired target $\lambda_t \triangleq max\{\lambda : P_{out\_load}(\lambda) \leq \psi\}$, we need an auxiliary admission mechanism. The base station continuously estimates $\widehat{\lambda_s}$, and broadcasts an admission probability:

$$p_{adm} \triangleq min\left\{1, \frac{\lambda_t}{\widehat{\lambda_s}}\right\}. \tag{22}$$

Each arriving new session generates a Bernoulli random variable with this probability. If a "1" is generated, the session sends a new call request to the base station; otherwise it refrains to request a new connection and it is blocked. In our discussion we assume that blocked arrivals are not allowed to retry to reconnect, although the case of infinite number of retries as well as finite number of retries can easily be incorporated in the analysis. After the preliminary arrival rate control, the number of new connection requests is a Poisson random variable with arrival rate $min\{\lambda_s, \lambda_t\}$.

The admission scheme proposed in this section, controls only connection requests from new session arrivals. The inactive calls already admitted in the system are automatically granted access at the time they become active. Future work may consider a tighter www load control for which new packet call initiation requests (from previously inactive sessions) are rejected whenever $\rho^{meas}(n) \geq \rho_t + \epsilon$.

We also note that admission priorities for the handed off calls may be implemented by setting different admission thresholds for the new calls initiated in the current cell and the handed off calls. This also represents future work and it is not addressed in this paper.

If imperfections in the power control loops are taken into account, the previous analysis has to be modified to derive a new expression for R_w^*. In this case, the received SIR for both voice and www calls is a random variable. Applying the central limit theorem, the average load and the maximum load can be approximated to be Gaussian random variables, and therefore the conditions (4) and (5) should hold in a probabilistic sense. Thus, we derive a new expression for R_w^* by imposing the condition:

$$Q\left(\frac{W(1-\eta) - E[Z_{av}^*]}{\sqrt{var[Z_{av}^*]}}\right) \leq \phi, \tag{23}$$

where ϕ represents the outage probability target.

Let us denote $E[Z_{av}^*] = \xi$ and $var[Z_{av}^*] = \zeta$. We assume that the reserved rate for www sessions is equally shared among the active web users, therefore $R_w^* \approx N_w^a R_{av}^w$ (R_{av}^w is the rate reservation for a particular web session). Denoting R as the voice transmission rate, the mean and variance for the average load can be expressed as:

$$\begin{cases} RN_vE[SIR^v] + R_{av}^w N_w^a E[SIR^w] = \xi; \\ R^2 N_v var[SIR^v] + (R_{av}^w)^2 N_w^a var[SIR^w] = \zeta \end{cases} \tag{24}$$

Denoting $\beta = R_{av}^w E[SIR^v] var[SIR^w] - RE[SIR^w] var[SIR^v]$, $A = W(1-\eta)$ and $\Delta = (AE[SIR^w] + 1/2((Q^{-1}(\phi))^2 R_{av}^w var[SIR^w])^2 - A^2 E[SIR^w]^2 - E[SIR^w](Q^{-1}(\phi))^2 RN_v\beta$, conditions (24) and (23) yield the quadratic inequality:

$$\zeta = \frac{1}{E[SIR^w]} \left(\xi R_{av}^w var[SIR^w] - RN_v\beta\right) \leq \frac{(A-\xi)^2}{(Q(\phi)^{-1})^2}. \tag{25}$$

The maximum value for ξ, such that $(A - \xi) > 0$ is found to be:

$$\xi = \frac{AE[SIR^w] + 1/2((Q^{-1}(\phi))^2 R_{av}^w var[SIR^w] - \sqrt{\Delta}}{E[SIR^w]}. \tag{26}$$

Since $R_{av}^w N_w^a \approx R_w^*$ and given ξ defined as in (26), from the first equation in (24) we derive the new reservation rate for the aggregate web traffic in the cell:

$$R_w^* = \frac{\xi - RN_vE[SIR^v]}{E[SIR^w]}. \tag{27}$$

3.2 Packet Scheduling

The packet scheduling multiplexes the voice and web traffic such that the system resources are used as efficiently as possible and all delay requirements are met. Voice sessions are always given the highest priority and they are allowed to transmit without delay. The web packets are scheduled to transmit according to the available capacity (residual capacity) obtained by subtracting the voice traffic contribution from the total system capacity. The www transmission rate is adaptively modified from one slot to another, based on resource availability. The resource scheduling is done such that the "noise rise" [1] condition holds at the slot (frame duration) time scale, for each slot. Resource availability for www is determined by estimating the voice load contribution for the next time slot. For the voice calls, the fluctuations in the load contribution are the result of changes in voice activity. We propose and compare two schemes for predicting the voice load for the next slot:

- (i) *One step prediction*. The total estimated load $\widehat{Z_v}(n)$ for the voice users at slot n, is determined based on the load measured in the previous time slot ($Z_v^m(n-1)$). A simple one step estimator is used: $\widehat{Z_v}(n) = Z_v^m(n-1)$.
- (ii) *Mean value prediction*. This method assumes that the number of active voice users in each time slot is exactly equal to the expected value for the number of active users, and computes the predicted voice load accordingly.

Using the above described prediction methods, the residual capacity available for the www users is determined. Based on its value, the base station assigns transmission rates for the next time slot to each www user, by traversing a circular list which contains all active www users' ids, together with their last rate request. The list is updated whenever a new rate request from a www user is received. The rate requested by a www user (denoted as R^{req}) is computed every time slot by the requesting terminal according to the current number of basic www packets waiting for transmission. A basic www packet is equivalent to a voice packet transmitted with rate $R = R_{max}^v$. Denoting the number of basic www packets as N_q, the new requested rate can be expressed as:

$$R^{req} = min\{N_q R, R_{MAX}\}, \tag{28}$$

where R_{MAX} is the limit imposed on the maximum rate that can be requested.
A NULL rate request means that the www call became inactive, and as a consequence, the base station will remove the user from the circular list used for scheduling.

If perfect power control is assumed, the residual load available for www transmission in slot n is computed as a function of the estimated load for the voice calls ($\widehat{Z_v}(n)$) as:

$$\widehat{Z}_{residual}(n) = W(1-\eta) - \widehat{Z_v}(n); \tag{29}$$

The voice load estimate is obtained for the "one step prediction" method as:

$$Z_v(n) = N_v^a(n-1) R_{max}^v SIR^v, \tag{30}$$

and for the "mean value prediction" method as:

$$Z_v(n) = N_v a R_{max}^v SIR^v. \tag{31}$$

$N_v^a(n-1)$ is the number of active voice users in the previous slot, and a is the voice activity coefficient.

The total aggregate rate that the www users can transmit in the next time slot ($R^w_{tot}(n)$), given the fact that all www users have the same BER requirement, is given by:

$$R^w_{tot}(n) = \frac{1}{SIR^w} \widehat{Z}_{residual}(n). \tag{32}$$

$R^w_{tot}(n)$ can be also determined for the imperfect power control case by using a similar approach to that described in the previous section. For this case, the instantaneous load condition should hold in a probabilistic sense, and the residual capacity available for data transmission at each time slot can be derived in a similar manner as the derivation for R^*_w in section 3.1.

The choice of the maximum burst rate assignment (R_{MAX}) parameter affects the system performance. For practical reasons, R_{MAX} cannot be driven to infinity. Further, for a cellular system, high rate burst transmission may cause excessive interference to the neighboring cells, and the scheduling procedure has to account for this effect [3]. If $R_{MAX} \to \infty$, the system is equivalent to a queueing system for which the service rate (R^*_w) is obtained as a long term average of the total aggregate rate available for www traffic in a particular slot n, i.e, $R^*_w = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} R^w_{tot}(n)$. Under the assumption that the data queue service process is ergodic, R^*_w can also be expressed as: $R^*_w = \mathbf{E}_{N^a_v}[R^w_{tot}(N^a_v)]$, where $\mathbf{E}_{N^a_v}[.]$ represents the mean with respect to N^a_v, the number of active voice users.

If R_{MAX} is too small, the system resources will be used inefficiently. This happens because the total rate request for the www traffic in slot n is limited to R_{MAX} multiplied by the number of active www users ($N^a_w(n)$). Even though more www packets are waiting in the queues and more resources might be available, the web transmission rate in slot n is at most equal to $N^a_w(n)R_{MAX}$.

Let N^i_q be the number of www packets in user i's queue. If $N^a_w(n)R_{MAX} < R^w_{tot}(n)$ and $\sum_{i=1}^{N^a_w(n)} N^i_q > R_{MAX}/R$, the system resources are wasted for the time slot n. By analogy with a queueing system, this has the effect of reducing the service rate of the queue (R^*_w) to an effective service rate $R^*_{eff} = \mathbf{E}_{N^a_v}[min\{N^a_w(n)R_{MAX}, R^w_{tot}(n)\}]$.

We wish to study the efficiency in allocating the system resources for both "one step prediction" and "mean value prediction" schemes. Specifically, we measure the efficiency as:

$$\eta_R = \frac{R^*_{eff}}{R^*_w}. \tag{33}$$

We parameterize η_R as a function of $\theta = R_{MAX}/R$, where R is the transmission rate of an active voice user (basic rate), $R = R^v_{max}$. For the perfect power control case, we determine R^*_{eff} for a given number of voice terminals (N_v) and a given number of active www terminals (N^a_w) in the system, considering that all voice terminals have SIR target requirement SIR^v, and all www sessions require an SIR target $SIR^w = KSIR^v$:

$$\text{(i) “one step prediction':} \quad R^{*(i)}_{eff} = \mathbf{E}_{N^a_v}\left[\min\left\{\theta R N^a_w, \frac{W(1-\eta)}{SIR} - \frac{N^a_v R}{K}\right\}\right], \tag{34}$$

$$\text{(ii) “mean value prediction''} \quad : R^{*(ii)}_{eff} = \min\left\{\theta R N^a_w, \frac{W(1-\eta)}{SIR} - \frac{N_v a R}{K}\right\}. \tag{35}$$

In figure 4 the efficiency in allocating the system resources (η_R) is shown for the two proposed packet scheduling mechanisms, for a range of values for θ. The system's parameters were chosen as: $W = 1.25MHz$, $\eta = 0.1$, $N_v = 20$ voice calls, $a = 0.4$, $SIR^v = 7dB$, $SIR^w = 10dB$, and $N^a_w = 5$ and $N^a_w = 2$ active www calls. It can be seen that a small θ reduces the efficiency η_R. Since only an average throughput requirement is specified for the www users, the system should operate with the smallest θ that gives $\eta_R = 1$. As we can see by comparing figure 4 (a) and (b), this optimal θ value depends on the number of active www users in the system. Therefore, if less www users are active, we will use the system resources more efficiently by allowing higher www transmission rates.

The packet scheduling performance is characterized by two performance measures: SIR outage probability and www traffic throughput. If perfect power control is assumed, the worst case SIR outage probability can be determined for $\rho = 1$, when www queues are full and there are always www packets to be transmitted. For a system having both voice and www users, the SIR outage is given by the errors in predicting the voice load for the next time slot. Thus, considering the voice model described by equations (1)-(2), for the “one step prediction” we have:

$$P^{max}_{out\_SIR} = Prob\left\{N^a_v(n+1) > N^a_v(n)\right\} \approx \sum_{k=0}^{N_v-1} P\{v(n+1) = k+1 | v(n) = k\} P_V(k). \tag{36}$$

For the “mean value prediction” method an outage occurs whenever the number of active voice users is greater than the mean number:

$$P^{max}_{out\_SIR} = \sum_{k=\lceil E\{N^a_v\}\rceil}^{N_v} P_V(k). \tag{37}$$

Figure 5 illustrates worst case SIR outage plots for the two prediction methods, as a function of the number of voice users in the system.

3.3 Extension to Multiple Cells

If the multiple cell case is considered, the impact of users in neighboring cells have to be accounted for. As in [15], the average interference caused by an out-of-cell interferer, aggregated from all cells is equivalent to the interference caused by ν in-cell users. For a path loss exponent $\delta = 4$, $\nu \approx 0.55$. However, the interference caused by the data users varies according to their location and their particular burstiness pattern. These variations are averaged out for the admission control condition, since this condition characterizes the system performance on the long term run. Hence the interference created by N_w^* out-of-cell sessions is equivalent, on average with the interference created by νN_w^* in cell users. However, for the packet scheduling, the interference variability has to be accounted for. A high www rate user in a neighboring cell is equivalent to $\nu_w(r)$ www users in the host cell, where r is the distance from the mobile user to its host cell base station [3]. As in [3], $\nu_d(r)$ is expressed as the ratio of the path loss of the mobile to the host cell to the path loss to the neighboring cell. Note that the packet scheduling also relies on instantaneous rate measurements for data users in the neighboring cells.

We revisit the admission control conditions to account for the out-of-cell interference. As in the previous section, we assume that no minimum throughput requirement is specified by the www users, i.e., $R_{min}^w = 0$. Therefore, for N_v^* equal to the number of out-of-cell voice calls, the maximum load condition limits the number of voice connections that can be supported by the CDMA system:

$$Z_{max}^* \triangleq \sum_{i=1}^{N_v} SIR_i^v (R_{max}^v)_i + \nu \sum_{i=1}^{N_v^*} SIR_i^v (R_{max}^v)_i \leq W(1-\eta). \qquad (38)$$

The average load condition has to account for the average out-of-cell interference created by the voice users as well as for the average aggregate out-of-cell interference generated by the web users. A reasonable assumption is that all voice connections require the same target SIR, SIR^v, and all web connections same target SIR, SIR^w. Assuming that all the cells guarantee the same service rate R_w^* for the web traffic, and that the admission control scheme maintains the traffic load close to the target load ρ_t, the average aggregate interference created by the out-of cell www users can be expressed as: $\nu SIR^w \sum_{i=1}^{N_w^*} (R_{av}^w)_i \approx \nu SIR^w N_{cell} \rho_t R_w^*$. N_w^* is the number of out-of-cell www sessions, and N_{cell} represents the number of neighboring cells considered in the analysis. Hence, the average load condition can be derived to be:

$$Z_{av}^* \triangleq SIR^v R_{av}^v \left[N_v + \nu N_v^*\right] + SIR^w R_w^* \left[1 + \nu \rho_t N_{cell}\right] \leq W(1-\eta). \tag{39}$$

From (39), we can solve for the maximum reservation rate per cell available for www users:

$$R_w^* = \frac{W(1-\eta) - SIR^v R_{av}^v \left[N_v + \nu N_v^*\right]}{SIR^w \left[1 + \nu \rho_t N_{cell}\right]} \tag{40}$$

The admission control algorithm is identical with the single cell case, with R_w^* defined as in (40).

Since the interference caused by data users depends on the users' positions in the cell and varies according to the burstiness pattern, the packet scheduling is based on neighbor coordinated access control as in [3]. The residual load available for www transmission in slot n is estimated as:

$$\widehat{Z}_{residual}(n) = W(1-\eta) - \widehat{Z_v}(n) - \nu N_v^* aRSIR^v - SIR^w \sum_{i=1}^{N_w^{*a}} R_i^w(n) \frac{z_i}{z_0}, \tag{41}$$

where $\widehat{Z_v}(n)$ was defined in the previous section, a represents the voice activity factor, R is the voice transmission rate, $R_i^w(n)$ is the transmission rate of the i^{th} out-of-cell www interferer, N_w^{*a} represents the number of active out-of-cell web sessions and $\frac{z_i}{z_0}$ is the ratio of the path loss of the i^{th} out-of-cell www user to its host cell to the path loss to the analyzed cell. N_w^{*a}, $R_i^w(n)$ and $\frac{z_i}{z_0}$ are determined based on the information received from the neighboring base stations.

3.4 Simulation Results and Observations

The system performance is experimentally determined by using a system simulator implemented in OPNET [20]. The parameters of interest are: average www load, access delay for www traffic packets and percentage of SIR outage, measured for sample paths of duration $T = 40$ minutes. For our simulations, we abstract the physical channel model as follows. We simulate the system under the assumption of perfect power control, for which each user's received SIR is deterministic and equal to its target SIR. Imperfections in the power control loops can be handled by modeling the received SIR as a random variable (see equations (23) - (27) in section 3.1). A reasonable assumption is to consider the received SIR to be a lognormal random variable [1]. This assumption is motivated by the fact that, in practice, SIR estimates are obtained using averages over a couple of frames [16]. Thus, the effect of Rayleigh fading is averaged out and only shadow fading has to be considered. While this impacts the analysis, it does not alter the admission control and packet scheduling mechanisms proposed here. Therefore, to illustrate the performance of the proposed web admission control/packet scheduling scheme, the simplified model based on the perfect power control assumption suffices. The numerical values used for the system parameters in the simulations are mostly based on the cdma2000 specifications [19] and are given in Table 1.

The admission control restricts the number of new connection requests allowed in the system. New session requests are comprised of two types of arrivals: sessions handed off from other cells (arriving with a Poisson distribution as in [13]) and new sessions initiations in the current cell (Poisson distributed). The total number of new session requests is therefore Poisson distributed with a cumulated arrival rate $\lambda_s = 1$ session/sec. The www sessions that become inactive are allowed to maintain the ppp (point-to-point protocol) connection and are automatically granted access at the time they become active. The time necessary to establish a new connection for both inactive calls and new admitted sessions is neglected in simulation. The residence time (dwell time) of a www session in the cell is assumed to be modeled by an exponential distribution with mean μ_d. As in [13], for our simulations we consider $\mu_d = 3$ minutes. In order to show that the mobility model has no impact on the proposed access control performance, another value $\mu_d = 10$ minutes is considered as well. Both proposed packet scheduling methods are implemented. Figures 6, 7, 8, 9, 10 illustrate the simulation results as sample paths for: (from top to bottom) the number of www sessions currently admitted in the cell, the number of currently active www sessions, the measured aggregate data offered load and the number of www packets arriving in a slot.

Figure 6 represents the case in which the admission control is used in conjunction with the ”one step prediction” method, and figure 7 illustrates the case in which the “mean value prediction” method is used as a packet scheduling mechanism. For both cases, $\mu_d = 3$ minutes and the length of the moving average window is $W_{size} = 30000$ slots = 10 minutes. Comparing the results obtained for these two cases in terms of percentage of outage and average aggregate offered load (Table 2), it can be seen that for both cases, the measured average load for a simulation time of 40 minutes is close to the specified target $\rho_t = 0.1$. Also, although both cases give a smaller outage percentage than the worst case outage probability derived in section 3.2 (because of the small average www load), the “one step prediction” performs better, resulting in a better control of the SIR outage probability, compared with the “mean value prediction” method.

Figures 8 and 9 illustrate the performance of the admission control for different lengths of the moving average window, $W_{size} = 3000$ slots = 1 minute, and $W_{size} = 60000$ slots = 20 minute, respectively. Both cases implement the ”one step prediction” method and use $\mu_d = 3$ minutes. We notice that a smaller length for the averaging window leads to more variability in the measured load, and to a less tight admission control, resulting in a higher average offered load measured for the duration of the simulation (see Table 2). Also, comparing the average offered load obtained at the end of the simulation for the three cases illustrated in figures 6, 8 and 9, we notice that the best results are given for the largest window size. The tradeoff is that the largest window size requires more memory to store and process the samples.
Finally, comparing figures 6 and 10, for the same simulation conditions but for different values for μ_d, 3 and 10 minutes respectively, we see that the system performance is similar, but for the case illustrated in figure 10, more inactive calls are maintaining ppp connections in the simulated cell.

Analyzing the web packets’ access delays (scheduling delays) (see Table 2) obtained from simulations, it can be seen that the delay is influenced by both the average load value obtained for the duration of the simulation and by the particular average load trace obtained. Although in general, it may seem that a lower average load value gives a lower access delay, in some cases, higher instantaneous measured loads for longer periods of time lead to higher access delays. This happens because a larger percentage of www packets experience higher delays due to increased congestion (see figure 11). By comparing figures 6 and 9 we observe that although higher load variations in figure 9 lead to a higher average load value measured at the end of the simulation, congestion occurs for shorter periods of time yielding lower access delays for the web packets. In conclusion, the www access delay performance will be highly dependent on the particular realization (trace) of the simulation, but a tight control on the average aggregate load will keep the delay close to some specified target, which can be experimentally determined.

We mention that the access delay represents only delay introduced by the packet scheduling scheme. We have not considered the interaction of the proposed admission/packet scheduling scheme with TCP. Any access scheme will introduce a variability in the RTT (round trip time) which may cause session time-outs and therefore decrease the overall throughput due to unnecessary retransmissions [14]. Thus, the end-to-end delay experienced by www packets is greatly influenced by the interaction of the access scheme with TCP, which is beyond the scope of this work.

4 Conclusion

In this paper we proposed a wireless access mechanism for web traffic packets in an integrated wireless CDMA system, that has both voice and web traffic. The access scheme is based on a two level control: admission control and packet scheduling. At the call arrival time scale, the admission control guarantees the QoS requirements in terms of bit error rates and delays for all users. At a slot time scale, a packet scheduling mechanism delivers QoS by scheduling the available resources among voice and www users. The difficulty in designing admission control for www traffic comes from the fact that an accurate estimate for the average offered rate per web session is hard to determine. The web traffic model comprises heavy tailed distributions which result in high variations in the offered load per session. Thus, the expected value for the average offered rate per www session will greatly over or under reserve resources for that particular web session. As a solution to this problem, we propose an admission control algorithm that adaptively estimates the aggregate average offered load for the web traffic, based on load measurements using a sliding observation window. The performance of the admission control is measured by the load outage probability, which is the probability that, after admission, the average load is maintained below a prescribed value. Two different packet scheduling mechanisms are proposed and compared. The performance of the packet scheduling scheme is characterized by the SIR outage probability, which gives the percentage of time slots in which the SIR targets could not be met for all users (voice and www) in the system.

The simulation results showed that the proposed admission control maintains the average aggregate load close to a desired target, therefore guaranteeing bounded average delay for the web packets. At the expense of increased implementation complexity, a tighter control of the average aggregate www load can be obtained if a larger moving average window is used. In terms of SIR outage probability, both analysis and simulations suggest that the SIR outage probability is better controlled by the "one step prediction" based packet scheduling mechanism.

APPENDIX

Derivation of the conditioned load outage probability formula, when the conditioning is with respect to N_i (the number of inactive web sessions in the CDMA system), N_b, T_i and N_s:

We start our derivation using the expressions for the load outage probability (11) and for the new offered load (17). With respect to the random variables N_b, T_i, N_s and N_p, the outage condition becomes:

$$\frac{(N_b + N_s)}{R_w^*} \frac{N_p m}{(N_p - 1)T_i} \geq \Delta\rho + \epsilon. \tag{42}$$

Rearranging expression (42) and denoting $D^*(\Delta\rho) = (\Delta\rho + \epsilon)R_w^*$, the outage condition is given as:

$$N_p[D^*(\Delta\rho)T_i - m(N_b + N_s)] \leq D^*(\Delta\rho)T_i. \tag{43}$$

We distinguish two cases:

1. $D^*(\Delta\rho)T_i - m(N_b + N_s) \leq 0$ Since $D^*(\Delta\rho)T_i > 0$, condition (43) is valid for all $N_p \geq b$. Thus:

$$P_{out\_load|(A,N_i,N_b,T_i,N_s)} = 1, \text{ if } \frac{m(N_b + N_s)}{D^*(\Delta\rho)T_i} \geq 1 \tag{44}$$

2. $D^*(\Delta\rho)T_i - m(N_b + N_s) > 0$ (43) can be further expressed as:

$$N_p \leq \frac{D^*(\Delta\rho)T_i}{D^*(\Delta\rho)T_i - m(N_b + N_s)} \tag{45}$$

Since N_p is Pareto distributed, if $\frac{D^*(\Delta\rho)T_i}{D^*(\Delta\rho)T_i - m(N_b+N_s)} < b$, $P_{out\_load|(A,N_i,N_b,T_i,N_s)} = 0$. Otherwise, using (18), we compute the conditioned load outage probability as:

$$P_{out\_load|(A,N_i,N_b,T_i,N_s)} = F_{N_p}\left(x = \frac{D^*(\Delta\rho)T_i}{D^*(\Delta\rho)T_i - m(N_b + N_s)}\right). \square \tag{46}$$

References

[1] A.M. Viterbi, A.J. Viterbi. *Erlang Capacity of a Power Controlled Cellular CDMA System,* IEEE Journal on Selected Areas in Communications, vol 11, No 6, August, 1993, pp. 892-900.

[2] A. Sampath, J.M. Holtzman. *Access Control of Data in Integrated Voice/Data CDMA Systems: Benefits and Tradeoffs,* IEEE Journal on Selected Areas in Communications, vol 15, No 8, October 1997, pp. 1511-1526.

[3] Sarath Kumar, Sanjiv Nanda, *High Data-Rate Packet Communications for Cellular Networks Using CDMA: Algorithms and Performance*, IEEE Journal on Selected Areas in Communications, vol 17, No 3, March 1999, pp. 472-492.

[4] M.E. Crovela and A. Bestavros, *Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes*, IEEE/ACM Transaction on Networking, vol 5, No 6, December 1997, pp. 835-846.

[5] A. Erramilli, O. Narayan, W. Wilinger. *Experimental Queueing Analysis with Long-Range Dependent Packet Traffic*, IEEE/ACM Transaction on Networking, vol 4, No 2, April 1996, pp. 209-223.

[6] C. Comaniciu, N.B. Mandayam. *Delta Modulation based Prediction for Access Control in Integrated Voice/Data CDMA Systems*, IEEE Journal on Selected Areas in Communication (JSAC), vol 18, No 1, January 2000, pp. 112-122.

[7] R.J. Gibbens, F.B. Kelly, P.B. Key. *A Decision-Theoretic Approach to Call Admission Control in ATM Networks*, IEEE Journal on Selected Areas on Communications, vol 13, No 6, August 1995, pp 1101-1113.

[8] C. Comaniciu, N.B. Mandayam, D. Famolari, P. Agrawal. *QoS Guarantees for Third Generation (3G) CDMA Systems via Admission and Flow Control*, IEEE Vehicular Technology Conference (VTC), September 2000, Boston, vol. 1, pp. 249 -256.

[9] N. Dimitriou, R. Tafazolli, G. Sfikas. *Quality of Service for Multimedia CDMA*, IEEE Communications Magazine, vol. 38, no. 7, July 2000, pp. 88-94.

[10] A. Sampath, N.B. Mandayam, J.M. Holtzman. *Erlang Capacity of a Power Controlled Integrated Voice and Data CDMA System,* IEEE Vehicular Technology Conference (VTC), 1997, vol 3, pp. 1557-1561.

[11] Y. Bao, A. Sethi. *Performance-driven Adaptive Admission Control for Multimedia Applications*, IEEE International Conference on Communications (ICC),Vancouver, June '99, pp. 199-203.

[12] S. Choi, K.G. Shin. *An Uplink CDMA System Architecture with Diverse QoS Guarantees for Heterogeneous Traffic*, IEEE/ACM Transaction on Networking, vol 7, No 5, October 1999, pp. 616-628.

[13] L. Ortigoza-Guerrero, A.H. Aghvami. *A prioritized Handoff Dynamic Channel Allocation Strategy for PCS*, IEEE Transaction on Vehicular Technology, vol 48, no. 4, July 1999, pp. 1203-1215.

[14] Sudhir Ramakrishna and Jack M. Holtzman. *Interaction of TCP and Data Access Control in an Integrated Voice/Data CDMA System* , Mobile Networks and Applications (MONET), vol 3, issue 4, 1998, pp. 409-417.

[15] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver, and C.E. Wheatley III, *On the capacity of a cellular CDMA system*, IEEE Trans. Veh. Technol., vol 40, No 3, May 1991, pp. 303-312.

[16] Deepa Ramakrishna, Narayan B. Mandayam, Roy D. Yates. *Subspace Based Estimation of the Signal-to-Interference Ratio for CDMA Cellular Systems*, IEEE Transactions on Vehicular Technology, vol. 49, No. 3,September 2000, pp. 1732 -1742.

[17] M. Andersin, Z. Rosberg, J. Zander. *Soft and Safe Admission Control in Power-Controlled Mobile Systems*, IEEE/ACM Transactions on Networking, vol. 5, no. 2, April 1997, pp. 255-265.

[18] D. Kim. *Efficient Interactive Admission Control in Power-Controlled Mobile Systems*, IEEE Transactions on Vehicular Technology, vol. 49, no. 3, May 2000, pp. 1017-1028.

[19] The cdma2000 ITU-R Candidate Submission, April 1998.

[20] OPNET, Modeler Manuals - Modeling, vol 1 & 2, version 6.

Table 1: *Numerical values for simulations*

Parameter	Notation	Value
System bandwidth	W	1.25 MHz
noise rise coefficient	η	0.1
target SIR for web users	SIR^w	10 dB
target SIR for voice users	SIR^v	7 dB
number of voice users in the system	N_v	20
mean packet interarrival time	t	0.01 seconds
unit discrete time for packet interarrival	t_i	0.001 seconds
probability of a new packet arrival	p_i	0.2
probability of a new packet call initiation in the next time slot	p_c	0.0001667
mean number of packet calls per session	$\mu_{N_{pc}}$	5
mean interarrival time between packet call initiations	$\mu_{T_{pc}}$	120 seconds
new session arrival rate per second	λ_s	1 session/second
Pareto distribution parameter	α	1.1
Pareto distribution parameter	b	2.27
length of a packet (measured in number of basic packets)	m	20
time slot = frame duration	T_s	0.02 seconds
performance tolerance	ϵ	0.1
load outage probability constraint	ψ	0.025
target www load	ρ_t	0.1
i-th voice call maximum rate	$(R^v_{max})_i$	9.6 Kb/s
voice activity factor	a	0.4
SIR outage probability constraint	ϕ	0.01
max rate burst assignment	R_{MAX}	32*9.6 Kb/s
mean dwell time	μ_d	3, 10 minutes
length of moving average window	W_{size}	1, 10, 20 minutes

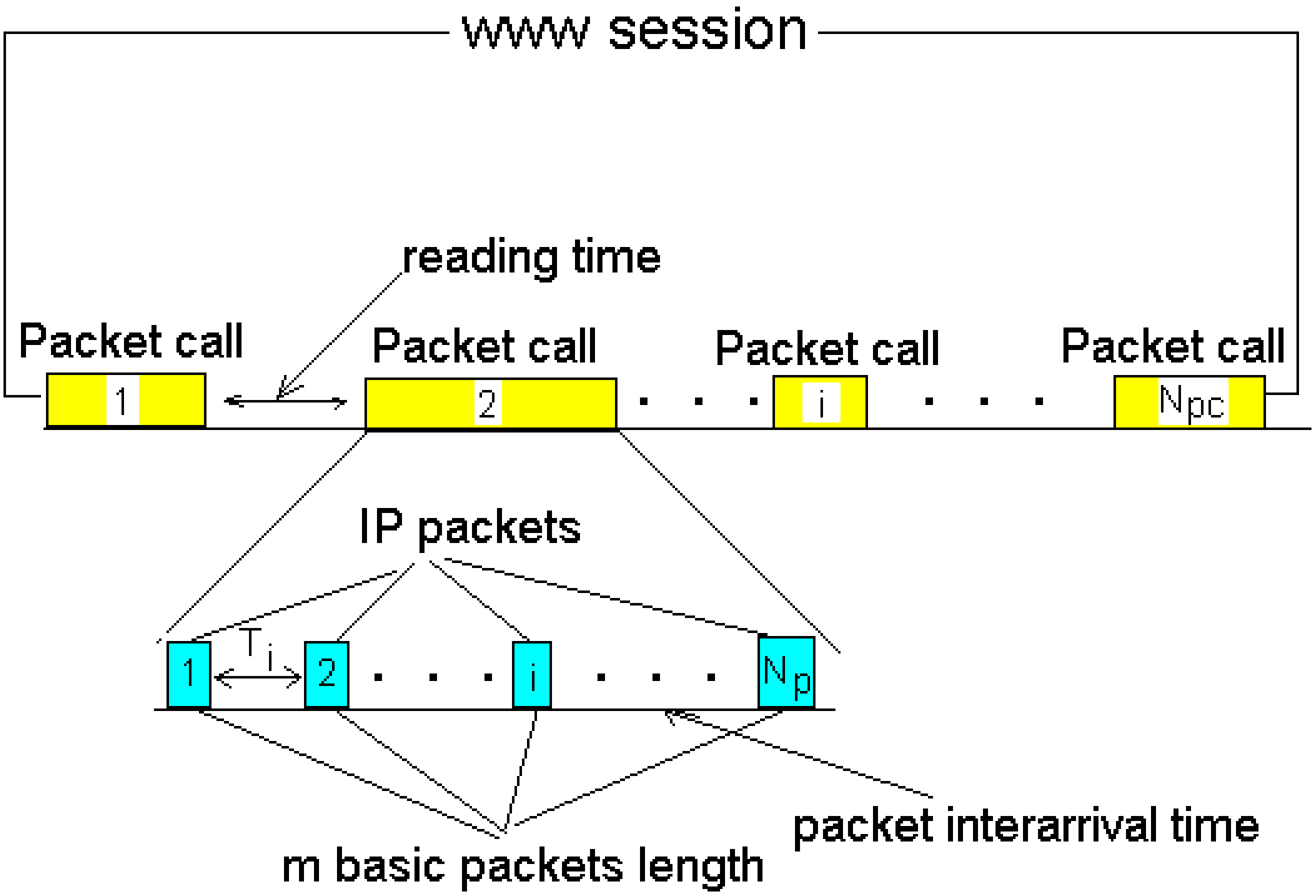


Figure 1: Web browsing session

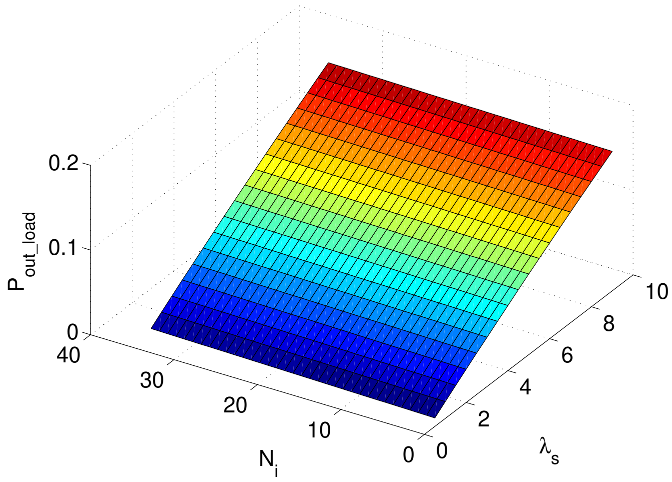


Figure 2: Dependence of load outage probability on the rate of new session arrivals (λ_s) (sessions/second), and on the number of inactive sessions in the cell (N_i).

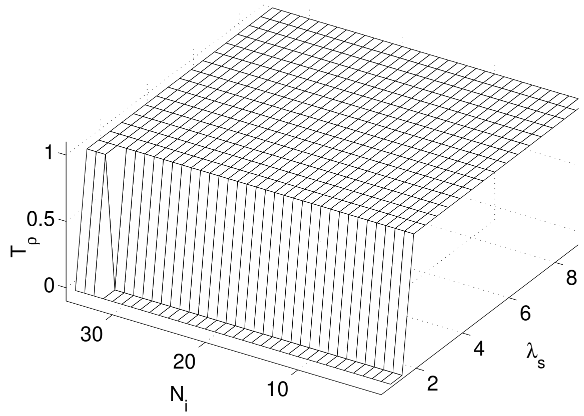
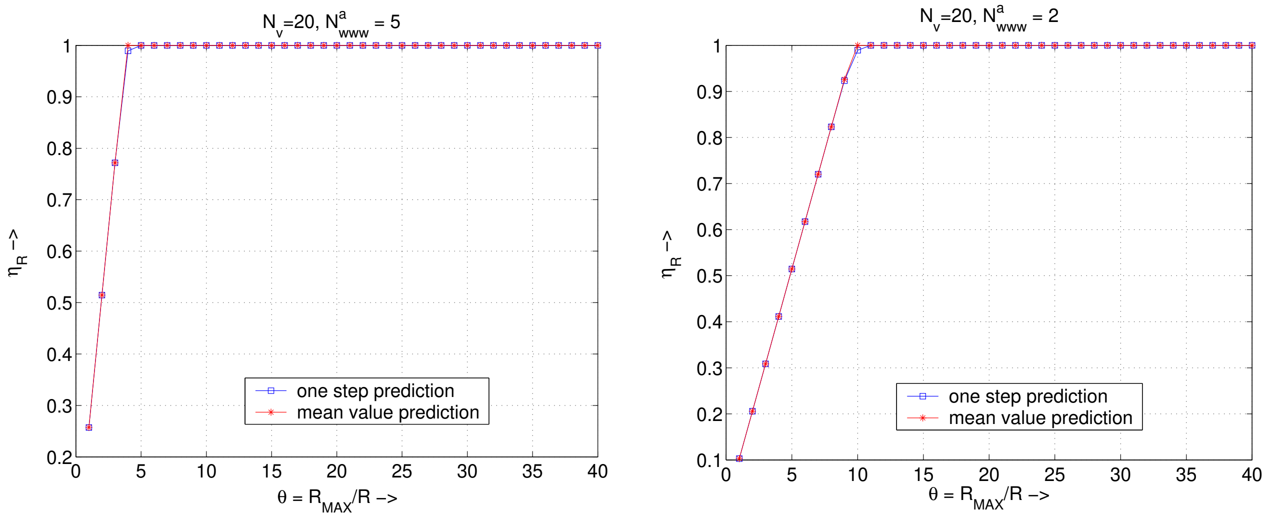


Figure 3: Dependence of the admission threshold on the rate of new session arrivals (λ_s) (sessions/second), and on the number of inactive sessions in the cell (N_i).



(a) (b)

Figure 4: Efficiency in allocating the system resources as a function of the R_{MAX} constraint

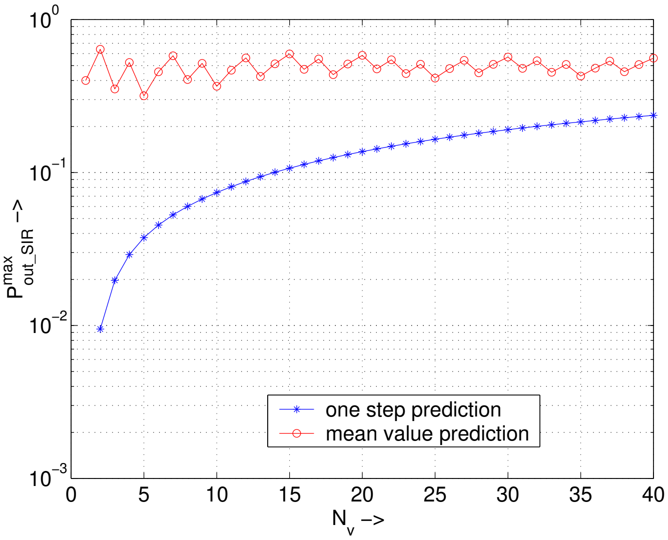


Figure 5: Worst case SIR outage probability for perfect power control case

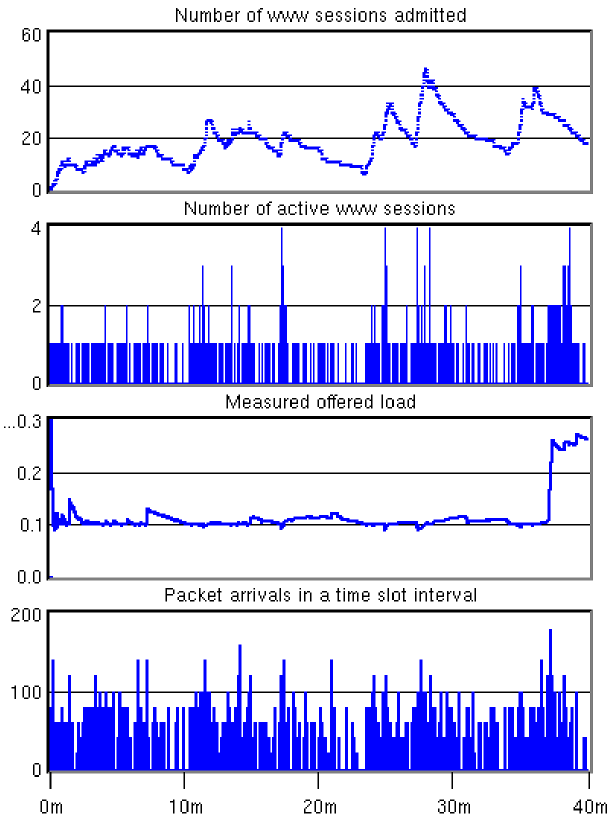


Figure 6: Sample path simulation: "one step prediction", $\mu_d = 3$ minutes, $W_{size} = 30000$ slots

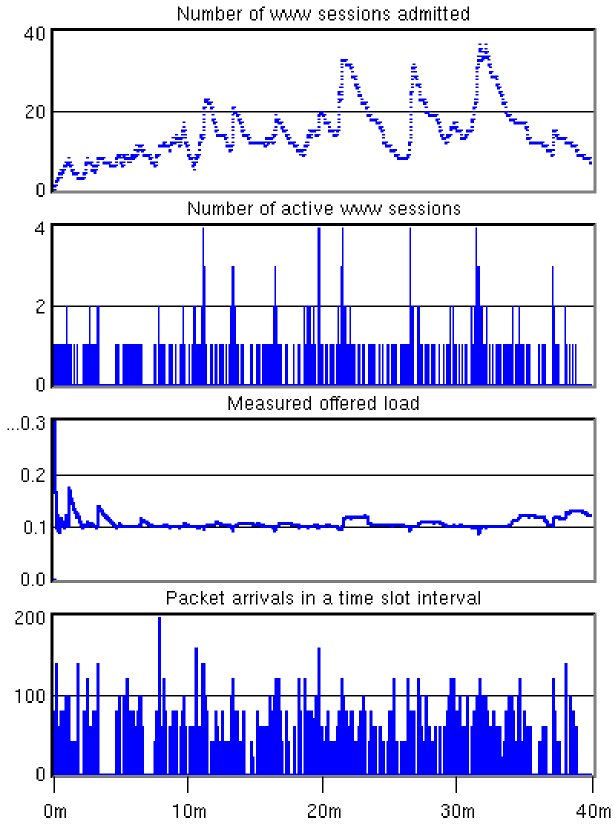


Figure 7: Sample path simulation: "mean value prediction", $\mu_d = 3$ minutes, $W_{size} = 30000$ slots

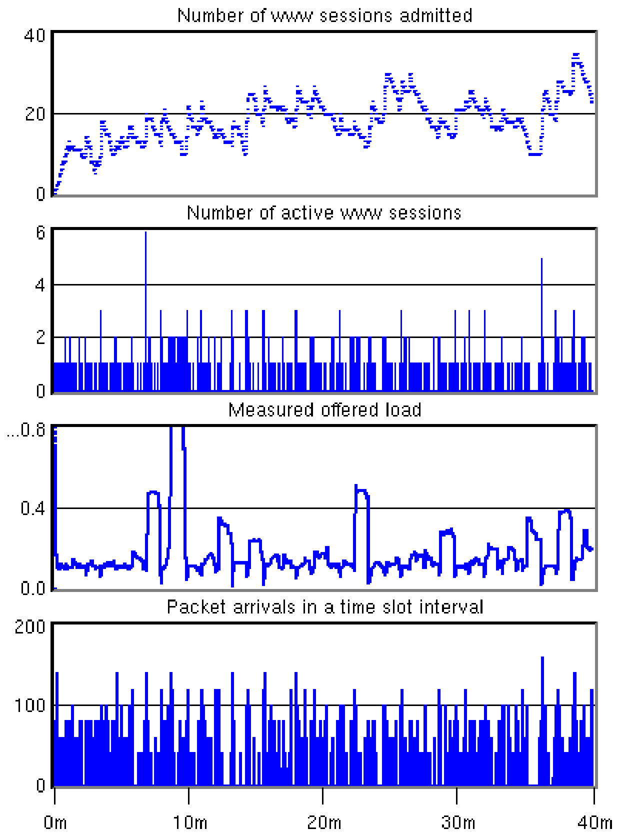


Figure 8: Sample path simulation: “one step prediction”, $\mu_d = 3$ minutes, $W_{size} = 3000$ slots

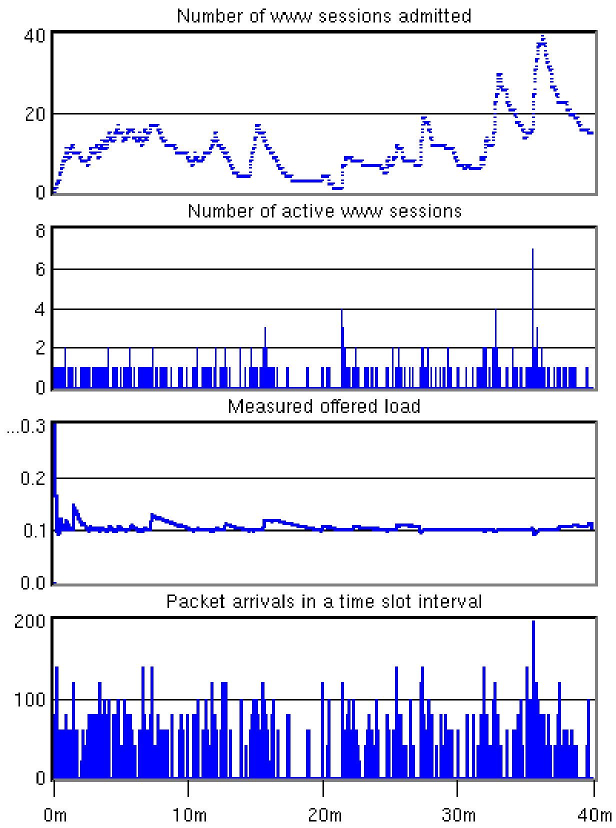


Figure 9: Sample path simulation: “one step prediction”, $\mu_d = 3$ minutes, $W_{size} = 60000$ slots

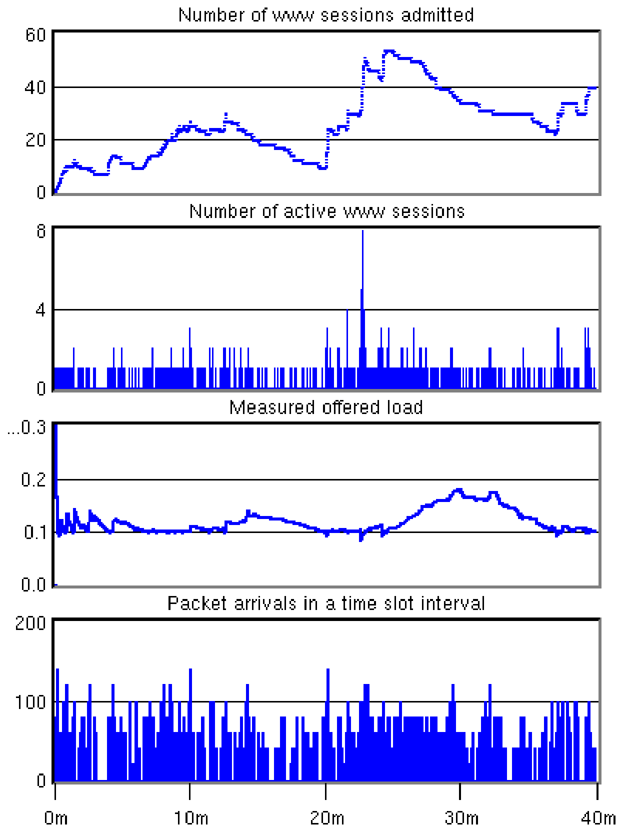


Figure 10: Sample path simulation: "one step prediction", $\mu_d = 10$ minutes, $W_{size} = 30000$ slots

Table 2: *System performance*

Simulation case illustrated in	Average aggregate www load	Percentage of SIR outage	Average delay for web traffic packets
Figure 6	0.1495	0.0103	15.11 sec.
Figure 7	0.1067	0.0290	1.35 sec.
Figure 8	0.1907	0.0135	7.03 sec.
Figure 9	0.1057	0.0075	2.2 sec
Figure 10	0.1203	0.0085	1.35 sec.

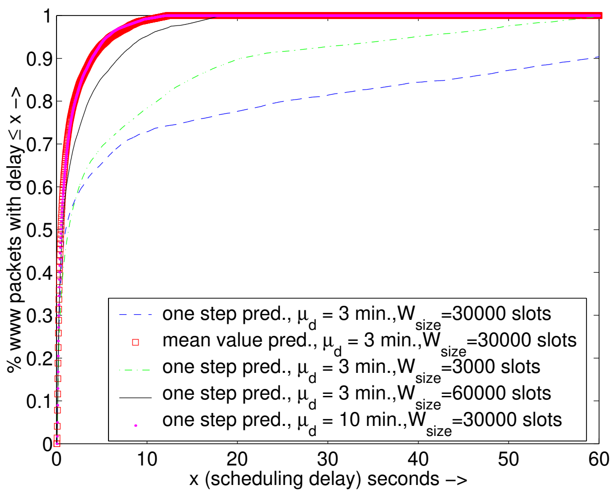


Figure 11: Percentage of web traffic packets experiencing a scheduling delay less than the value on the abscisa