Admission and GoS control in a multiservice WCDMA system [⋆]

Jean Marc Kelif

France Telecom R&D, 38-40 Rue du General Leclerc, 92794 Issy-les-Moulineaux Cedex 9, France

Eitan Altman [*] and Ioannis Koukoutsidis

INRIA BP93, 2004 Route des Lucioles, 06902 Sophia-Antipolis Cedex, France.
Tel: +33 4 92 38 77 86. Fax: +33 4 92 38 77 65

Abstract

We consider in this paper a WCDMA system with two types of calls: real time (RT) calls that have dedicated resources, and data non-real time (NRT) calls that are treated using a time-shared channel (such as the HDR or the HSDPA). We consider reservation of some resources for the NRT traffic and assume that this traffic is further assigned the resources left over from the RT traffic. The grade of service (GoS) of RT traffic is also controlled in order to allow for handling more RT calls during congestion periods, at the cost of degraded transmission rates. We consider both the downlink (with and without macrodiversity) as well as the uplink and study the blocking probabilities of RT traffic as well as the expected sojourn time of NRT traffic. We further study the conditional expected sojourn time of a data connection given its size and the state of the system. Finally, we extend our framework to handle handover calls.

Key words: WCDMA, call admission control, GoS, HSDPA, HDR, handover.

[⋆] This work was supported by a CRE research contract with France Telecom R&D

[*] Corresponding Author
Email address: altman@sophia.inria.fr (Eitan Altman).
URL: www.inria.fr/mistral/personnel/Eitan.Altman (Eitan Altman).

1 Introduction

Important performance measures of call admission control policies in systems with heterogeneous service classes are the probability of rejection of calls of different classes as well as the sojourn time of non-real time transfers. In order to be able to compute these and to design the call admission control policies, a dynamic stochastic approach should be used based on statistical assumptions regarding the call arrival processes and durations, as well as data transfer sizes.

In this context, a classical approach widely used in wireless networks is based on adaptively deciding how many channels (or resources) to allocate to calls of a given service class, see e.g. [3,9,10]. Then one can evaluate the performance as a function of some parameters (thresholds) that characterize the admission policy, using Markov chain analysis. This allows to optimize and to evaluate tradeoffs between QoS parameters of the different classes of mobiles. This approach, natural to adopt in TDMA or FDMA systems, can also be followed in the case of a CDMA system, even though the notion of capacity is much more complex to define. For the uplink case in CDMA, the capacity required by a call has been studied in the context of call admission, see e.g. [15,7,4].

We focus here on two types of calls, real-time (RT) and non-real time (NRT) data transfers. Whereas all calls use CDMA, we assume that NRT calls are further time-multiplexed (which diminishes the amount of interference, thus increasing the available average throughputs). This combination of time multiplexing over CDMA is typically for high speed downlink data channels, such as the High Speed Downlink Packet Access (HSDPA) [13] and the High Data Rate (HDR) in CDMA-2000 systems [1].

Similarly to the uplink analysis [4], we propose a simple model that allows us to define in the downlink case the capacity required by a call of a given class when it uses a given grade of service (transmission rate). In particular, we also consider the case of macrodiversity. We then propose a control policy that combines admission control together with a control of the grade of service (GoS) of real-time traffic. Key performance measures are then computed by modeling the CDMA system as a quasi-birth-and-death (QBD) process. We obtain the call blocking probabilities and expected transfer times, already available for the uplink case in [4]. We further obtain (both for the uplink and downlink) another important performance measure: the expected transfer time of a file conditioned on its size. We study the influence of the control parameters on these performance measures. We finally extend the model to handle handover calls.

The structure of the paper is as follows. We begin by introducing in Sections

2, 3 and 4 the frameworks corresponding to the downlink, with and without macrodiversity, as well as the uplink of a CDMA system. Using power control arguments, we obtain for all three cases the transmission rates for various classes of calls which are compatible with given signal to noise and interference ratios. We then introduce in Section 5 the basic control actions: call admission and control of GoS. The statistical modeling of the system is presented in Section 6. It is then used in Section 7 for an extensive numerical investigation. The extension of the model and the analysis to handover traffic is given in Section 8, and we conclude the paper in Section 9.

2 Downlink

We use a model similar to the one presented in [5]. Let there be S base stations. The minimum power received at a mobile k from its base station l is determined by a condition concerning the signal to interference ratio, which should be larger than some constant

$$\gamma := \frac{E_s}{N_0}\frac{R_s}{W}\Gamma, \tag{1}$$

where E_s/N_0 is the energy per transmitted bit (of type s) to interference density, W is the WCDMA modulation bandwidth, and R_s is the transmission rate of the type s call. The constant Γ accounts for the random behavior of a signal due to shadow fading and imperfect power control; more specifically, to account for this randomness we study a probabilistic condition

$$\Pr\{(C/I)_k > \gamma\} > 1 - \chi,$$

where χ is a small outage probability. Considering a log-normal distribution of the SIR, $(C/I)_k = 10^{\frac{\xi_k}{10}}$, where $\xi_k \sim N(\mu_\xi, \sigma_\xi)$, it can be derived that Γ fulfills the probabilistic condition by satisfying [**?**,4]:

$$\log\Gamma = \frac{\sigma_\xi^2}{20h} - \frac{Q^{-1}(1-\chi)\sigma_\xi}{10}, \tag{2}$$

where $h = 10/ln10$ and $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

Any other causes of randomness, most notably fast fading, can be taken into account the same way by considering a different distribution, e.g. a Rayleigh fading distribution [14].

Now let $P_{k,l}$ be the power transmitted to mobile k from base station l. Assume that there are M mobiles in cell l; the base station of that cell transmits at a total power $P_{tot,l}$ given by

$$P_{tot,l} = \sum_{j=1}^{M} P_{j,l} + P_{SCH} + P_{CCH}, \tag{3}$$

where P_{SCH}, P_{CCH} correspond to the power transmitted for the non-orthogonal synchronization (SCH) and orthogonal common control channels (CCH), respectively [6]. Note that these last two terms are not power controlled, and so they can be modeled by adding a constant power. Also they are assumed not to depend on l. Due to the non-orthogonality of the synchronization channel but also to the multipath propagation, a fraction α_k of the received own cell power is experienced as intracell interference by mobile k. We generally call this parameter as the *non-orthogonality factor* in the downlink. Let $g_{k,l}$ be the attenuation between base station l and mobile k. Denoting by $I_{k,inter}$ and $I_{k,intra}$ the intercell and intracell interferences, respectively, we have

$$\left.\frac{C}{I}\right|_k = \frac{P_{k,l}/g_{k,l}}{I_{k,inter} + I_{k,intra} + N},$$

where N is the receiver noise floor (assumed not to depend on k),

$$I_{k,intra} = \alpha_k \cdot (P_{SCH} + P_{CCH} + \sum_{j \neq k} P_{j,l})/g_{k,l}$$

and

$$I_{k,inter} = \sum_{j=1, j \neq l}^{S} P_{tot,j}/g_{k,j}.$$

We define

$$F_{k,l} = \frac{\sum_{j=1,j\neq l}^{S} P_{tot,j}/g_{k,j}}{P_{tot,l}/g_{k,l}},$$

i.e. the ratio between the received intercell and intracell power. It then follows that

$$\beta_k = \frac{P_{k,l}/g_{k,l}}{(F_{k,l} + \alpha_k)P_{tot,l}/g_{k,l} + N}, \tag{4}$$

$$\text{where } \beta_k = \frac{(C/I)_k}{1 + \alpha_k (C/I)_k}. \tag{5}$$

Basic algebra yields the following:

$$P_{tot,l} = \frac{P_{SCH} + P_{CCH} + N \sum_k g_{k,l}\beta_k}{1 - \sum_k (F_{k,l} + \alpha_k)\beta_k} \tag{6}$$

In downlink CDMA dimensioning, it is important to estimate the total amount of base station power required, which limits the capacity. In order to calculate the total base station power in our problem, we should normally solve for the transmitted powers to each mobile separately in (4), and substitute them in (3). However, even if we know the parameters for each individual connection this does not result in any useful dimensioning on the link. In order to obtain a simplified formula, we replace $g_{k,l}$, $F_{k,l}$, α_k by single parameters G, F, α. These values can initially be chosen by taking the sample average for $g_{k,l}$, $F_{k,l}$, α_k over all $k = 1, \ldots, M$. Their accuracy can then be improved based on actual measurements for the mean total base station output power, which should be close-by in the simplified and realistic models. We will refer to this, albeit somewhat imprecisely, as an "average approximation". Such an approximation has been used in many downlink dimensioning models for CDMA, see e.g. [6,5,16], as it provides an easy way to estimate the pole capacity.

We next consider two service classes, denoted by $s = \{1, 2\}$ (that will correspond to RT and NRT traffic, respectively). Let $(C/I)_s$ be the target SIR ratio for mobiles of service class s and let β_s be the corresponding value in (5). Let there be in a given cell M_s mobiles of class s.

Using the average approximation, we finally get for the total output power of base station l (we shall omit the index l):

$$P_{tot} = \frac{P_{SCH} + P_{CCH} + NG \sum_s \beta_s M_s}{1 - (\alpha + F) \sum_s \beta_s M_s}. \tag{7}$$

We now further assume that the power in the synchronization and common control channels is a fraction ψ of the total output power, i.e., $P_{SCH} + P_{CCH} = \psi P_{tot}$ and defining the downlink loading as $Y_{DL} = \sum_s (\alpha + F)\beta_s M_s$, this gives

$$P_{tot} = \frac{NG \sum_s \beta_s M_s}{Z_2}, \text{ where } Z_2 = (1 - \psi) - Y_{DL}. \tag{8}$$

In most cases, the maximum base station output power determines the maximum loading supported by the system. Then, according to the power limitation of the base station, one poses the constraint $Z_2 \geq \epsilon$ for some $\epsilon > 0$. Consequently, we can define the system's *nominal* capacity as $\Theta_\epsilon = 1 - \psi - \epsilon$, and the *nominal* capacity required by a connection to be

$$\Delta(s) := (\alpha + F)\beta_s. \tag{9}$$

We note that β_s will allow to depend on M_s, $s = 1, 2$. Combining this with (1) and with (5) we get the throughput of a connection s, that "uses a capacity $\Delta(s)$":

$$R_s = \frac{\Delta(s)}{\alpha + F - \alpha \Delta(s)} \times \frac{N_0 W}{E_s \Gamma}. \tag{10}$$

3 Downlink with macrodiversity

In this section, we demonstrate an extension of the previous model by which the case of macrodiversity can be considered in the downlink. Our approach is inspired by [5] who considered the single service case[1] . A mobile i in macrodiversity is connected to two base stations, b and l. b is defined to be the station with larger SIR. Following [5] we assume that the Maximum Ratio Combining is used and hence the power control tries to maintain

$$\gamma_i = \frac{C}{I}\bigg|_i = \frac{C}{I}\bigg|_{i,b} + \frac{C}{I}\bigg|_{i,l},$$

where γ_i is given by the constant in (1). We have $\Omega_i \leq 1$ where

$$\Omega_i := \frac{C/I|_{i,l}}{C/I|_{i,b}}.$$

This gives for the combined C/I [5]:

$$\frac{C}{I}\bigg|_i = \frac{(1+\Omega_i)P_{i,b}/g_{i,b}}{\alpha_b(P_{tot,b} - P_{i,b})/g_{i,b} + F_{i,b}P_{tot,b}/g_{i,b} + N}.$$

The transmission power becomes

$$P_{i,b} = \kappa_i(\alpha P_{tot,b} + F_{i,b}P_{tot,b} + g_{i,b}N),$$

where

$$\kappa_i = \frac{(C/I)_i}{1 + \Omega_i + \alpha(C/I)_i}. \tag{11}$$

Let there be M mobiles in a cell b (we shall omit this index) of which a fraction μ is in macrodiversity. We assume that by symmetry, the base station of that cell transmits also to a number μM of mobiles that are geographically situated in neighboring cells. Then the total base station output power can be calculated as

$$P_{tot} = \sum_{i=1}^{(1-\mu)M} P_i + \sum_{j=1}^{2\mu M} P_j + P_{SCH} + P_{CCH},$$

where the notations i,j in the sums should be understood to refer to single link and macrodiversity mobiles, respectively. The power for a single link user should be calculated the same way as in § 2.

We now consider two classes of services $s = \{1, 2\}$ corresponding to RT and NRT mobiles. We make the following "average approximations", similarly to the previous section: For a given service class $s = \{1, 2\}$ Ω_i is replaced by a

[1] We extend the context here to refer more generally to macrodiversity, and not only the soft handover procedure.

constant Ω_s (its average over all mobiles of the same service as i); we also replace $F_{i,b}$ by one of two constants F^{NMD} and F^{MD}, where F^{NMD} (resp. F^{MD}) corresponds to an average value of $F_{i,b}$ over mobiles which are not in macrodiversity (and which are in macrodiversity, resp.). Likewise, we replace $g_{i,b}$ by one of the two constants G^{NMD} and G^{MD}. This gives the total power of a base station b:

$$P_{tot} = \frac{Z_1}{Z_2}$$

as long as Z_2 is strictly positive, where

$$Z_1 := (1-\mu) \sum_{s=1,2} M_s \beta_s G^{NMD} N + 2\mu \sum_{s=1,2} M_s \kappa_s G^{MD} N$$

and

$$Z_2 := (1-\psi) - (1-\mu) \sum_{s=1,2} M_s \beta_s (\alpha + F^{NMD}) - 2\mu \sum_{s=1,2} M_s \kappa_s (\alpha + F^{MD}).$$

Again we avoid that Z_2 becomes too close to zero by posing the constraint $Z_2 \geq \epsilon$ for some $\epsilon > 0$. We can thus define the system's *nominal* capacity as $\Theta_\epsilon = 1 - \psi - \epsilon$, and the capacity required by a connection of type $s = 1, 2$ to be

$$\Delta(s) = (1-\mu)\beta_s(\alpha + F^{NMD}) + 2\mu\kappa_s(\alpha + F^{MD}).$$

Combining this with (1) and (11), we get

$$\Delta(s) = (1-\mu) \cdot \frac{R_s \cdot \delta_s}{1 + \alpha R_s \delta_s}(\alpha + F^{NMD}) + 2\mu \cdot \frac{R_s \cdot \delta_s}{1 + \Omega_s + \alpha R_s \delta_s}(\alpha + F^{MD}). \tag{12}$$

Here, $\delta_s = \frac{E_s \Gamma}{N_0 W}$ and we have considered the rate R_s of a connection equal, irrespective if a mobile is in macrodiversity or not. Solving for R_s, this leads to a quadratic equation giving two values, of which we retain the positive.

4 Uplink

We briefly recall the capacity notions from the case of uplink from [4]. Define for $s = 1, 2$,

$$\tilde{\Delta}_s = \frac{E_s}{N_o} \frac{R_s}{W} \Gamma, \text{ and } \Delta'(s) = \frac{\tilde{\Delta}(s)}{1 + \tilde{\Delta}(s)}. \tag{13}$$

The power that should be received at a base station originating from a type s service mobile in order to meet the QoS constraints is given by Z_1/Z_2 [4] where

$$Z_1 = N\Delta'(s)$$

and

$$Z_2 = 1 - (1+f) \sum_{s=1,2} M_s \Delta'(s)$$

(N is the background noise power at the base station, f is some constant describing the average ratio between inter and intra cell interference, and M_s is the number of mobiles of type s in the cell). Also in this case $Z_2 \geq \epsilon$ for some $\epsilon > 0$. We can thus define the system's *nominal* capacity as $\Theta_\epsilon = 1 - \epsilon$, and the capacity required by a connection of type $s = 1, 2$ to be $\Delta(s) = (1+f)\Delta'(s)$. Combining this with (13) we get

$$R_s = \frac{\Delta(s)}{1 + f - \Delta(s)} \times \frac{N_o W}{E_s \Gamma}. \tag{14}$$

5 Admission and rate control

In the design of an admission and rate control scheme for heterogeneous services we will consider that RT calls, which have more stringent QoS requirements, have priority over system resources. NRT traffic, on the other hand, has no guaranteed bit rate and can be served in a processor-sharing fashion. However, to prevent RT calls from overwhelming the link we will also assume that a portion of the system resources is reserved for NRT traffic. Further, to also achieve a multiplexing gain for RT calls, we will allow a limited rate degradation for such traffic.

We consider here a fair transmission rate scheme, such that mobiles which belong to the same service class (NRT or RT) transmit or receive at the same rate. For NRT traffic, for which fast-time multiplexing will be considered, this can be viewed as a fair implementation of an HSDPA or HDR scheme where transmission to each mobile takes place at the same *average* rate. For this, an underlying scheduler is also assumed that assigns time slots in proportion to the peak feasible rates of mobiles, in order to achieve the same average rate.

These basic principles of admission and rate control are made more explicit in the following. One must also have in mind that either the uplink or the downlink can be the bottleneck of a CDMA system at one time or another; so from an engineering perspective one should focus only on the more restrictive direction when accepting calls. In our paper, all the notations will be understood to relate to that direction.

Capacity reservation. We assume that there exists a capacity L_{NRT} reserved for NRT traffic. The RT traffic can use up to a capacity of $L_{RT} := \Theta_\epsilon - L_{NRT}$.

GoS control of RT traffic. UMTS will use the Adaptive Multi-Rate (AMR) codec that offers eight different transmission rates of voice that vary between

4.75 kb/s to 12.2 kb/s, and that can be dynamically changed every 20 ms. The lower the rate is, the larger the amount of compression is, and we say that the grade of service (GoS) is lower. For simplicity we shall assume that the set of available transmission rates of RT traffic has the form $[R^{\min}, R^{\max}]$. We note that $\Delta(RT)$ is increasing with the transmission rate. Hence the achievable capacity set per RT mobile has the form $[\Delta_{RT}^{\min}, \Delta_{RT}^{\max}]$. Note that the maximum number of RT calls that can be accepted is $M_{RT}^{\max} = \lfloor L_{RT}/\Delta_{RT}^{\min} \rfloor$. We assign full rate $R_{RT}^{\max}$ (and thus the maximum capacity $\Delta_{RT}^{\max}$) for each RT mobile as long as $M_{RT} \leq N_{RT}$ where $N_{RT} = \lfloor L_{RT}/\Delta_{RT}^{\max} \rfloor$. For $N_{RT} < M_{RT} \leq M_{RT}^{\max}$ the capacity of each present RT connection is reduced to $\Delta_{MR} = L_{RT}/M_{RT}$ and the rate is reduced accordingly (e.g. by combining (1), (5) and (9) for the case of downlink).

Fast time multiplexing for NRT traffic. The capacity $C(M_{RT})$ unused by the RT traffic (which dynamically changes as a function of the number of RT connections present) is fully assigned to one single NRT mobile, and the mobile to which it is assigned is time multiplexed rapidly so that the throughput is shared equally between the present NRT mobiles. The available capacity for NRT mobiles is thus

$$C(M_{RT}) = \begin{cases} \Theta_\epsilon - M_{RT}\Delta_{RT}^{\max}, & \text{if } M_{RT} \leq N_{RT}, \\ L_{NRT}, & \text{otherwise.} \end{cases}$$

The total transmission rate R_{NRT}^{tot} of NRT traffic for the downlink and uplink is then given respectively by

$$\begin{aligned} DL: \quad & R_{NRT}^{tot}(M_{RT}) = \frac{C(M_{RT})}{\alpha + F - \alpha C(M_{RT})} \times \frac{N_o W}{E_s \Gamma}, \\ UL: \quad & R_{NRT}^{tot}(M_{RT}) = \frac{C(M_{RT})}{1 + f - C(M_{RT})} \times \frac{N_o W}{E_s \Gamma}. \end{aligned} \tag{15}$$

The expression for downlink with macrodiversity is derived similarly, albeit being more complex.

Remark 1. The expressions that we have obtained for the total throughput available for NRT traffic may be in practice non-accurate due to the many approximations we use, such as using the averaged values f and F in the above equations. Since these expressions are used later in a dynamic context, the price of changing the expressions to complex ones can render the later Markovian analysis unfeasible. To be able to have better precision, we need to sacrifice the generality of the model. As an example, a trivial sufficient condition for replacing $F_{k,l}$ by their average is that there is a single cell.

A simulation example. We ran a simulation to test some of the simplifications that we used in this paper concerning the downlink to check the value of R_{tot}.

Some mobiles using a voice service are randomly generated in a cell located in a UMTS network environment. For each random generation, the simulation calculates the local downlink interference factor parameter f_i for each mobile. The number of mobiles of a cell N_{MS}^{cell} is limited by the total load of the cell which has to be inferior to 100%. We calculate the total throughput of the cell as follows:

$$R_{tot} = 12.2 N_{MS}^{cell} kb/s$$

We obtain an average number of mobiles for the cell, as the ratio between the total number of mobiles generated N_{ms}^{tot} considering all the generations, to the number of generations N_{gen}. We obtained

$$N_{MS}^{cell} = \frac{N_{MS}^{tot}}{N_{gen}} = 27.6 mobiles.$$

We obtain a total throughput of

$$R_{tot} = 237 kb/s.$$

We obtain an average interference factor per mobile as follows:

$$F = \frac{\sum_{i=1}^{N_{MS}^{tot}} f_i}{N_{MS}^{tot}} = 1.13.$$

We then use the following analytical calculation:

$$R_{tot} = N_{MS}^{cell} \frac{1-\phi}{N_{MS}^{cell}(\alpha + F) - \alpha(1-\phi)}$$

ϕ is the fraction of the BS power dedicated to common channels. 1-ϕ is the capacity of the cell. In our simulations we have $\phi = 0.14$ and $\alpha = 0.79$.

With the analytical method we obtain $R_{tot} = 379 kb/s$. The difference between the two values is

$$\frac{379 - 339}{379} = 11\%$$

The total throughput is thus close the simulated one.

6 Stochastic model and the QBD approach

In this section we proceed to study a stochastic traffic model and examine steady-state performance measures of the system. We consider the total nominal capacity to remain fixed throughout the system lifetime. This is true in case capacity is limited by base station (DL) output power or interference (UL) and the channel environment conditions do not change very extremely, so that given the rate and power adaptation the same maximum loading is achieved at any time instant. We also assume that for the time-multiplexing of NRT calls an appropriate scheduling scheme is feasible such that each mobile transmits or receives *instantaneously* at a rate given by (15), and they all obtain the same average rate.

Model. We assume that RT and NRT calls arrive according to independent Poisson processes with rates λ_{RT} and λ_{NRT}, respectively. The duration of a RT call is exponentially distributed with parameter μ_{RT}. The size of a NRT file is exponentially distributed with parameter μ_{NRT}. Interarrival times, RT call durations and NRT file sizes are all independent.

The departure rate of NRT calls depends on the current number of RT calls:

$$\nu(M_{RT}) = \mu_{NRT} R_{NRT}^{tot}(M_{RT}).$$

QBD approach. Under these assumptions, the number of active sessions in all three models (downlink, with and without macrodiversity and uplink) can be described as a QBD (quasi-birth-and-death) process, and we denote by Q its generator. We shall assume that the system is stable. The stationary distribution of this system, π, is calculated by solving:

$$\pi Q = 0, \tag{16}$$

with the normalization condition $\pi e = 1$ where e is a vector of ones of proper dimension. The vector π represents the steady-state probability of the two-dimensional process. We partition π as $[\pi(0), \pi(1), \ldots]$ with the vector $\pi(i)$ for level i, where the levels correspond to the number of NRT calls in the system. We may further partition each level into the number of RT calls, $\pi(i) = [\pi(i,0), \pi(i,1), \ldots, \pi(i, M_{RT}^{\max})]$, for $i \geq 0$. The entries of π are ordered lexicographically, i.e. $\pi(k,i)$ precedes $\pi(l,j)$ if $k < l$, or if $k = l$ *and* $i < j$. The generator matrix Q is given by

$$Q = \begin{bmatrix} B & A_0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots \end{bmatrix} \tag{17}$$

where the matrices B, A_0, A_1, and A_2 are square matrices of size $(M_{RT}^{\max} + 1)$. The matrix A_0 corresponds to a NRT connection arrival, given by $A_0 = \text{diag}(\lambda_{NRT})$. The matrix A_2 corresponds to a departure of a NRT call and is given by $A_2 = \text{diag}(\nu(j); 0 \leq j \leq M_{RT}^{\max})$. The matrix A_1 corresponds to the arrival and departure processes of RT calls. A_1 is tri-diagonal as follows:

$$\begin{aligned} A_1[j, j+1] &= \lambda_{RT} \\ A_1[j, j-1] &= j\mu_{RT} \\ A_1[j, j] &= -\lambda_{RT} - j\mu_{RT} - \lambda_{NRT} - \nu(j) \end{aligned}$$

We also have $B = A_1 + A_2$. We follow a matrix-geometric approach for the solution of the QBD process. Assuming a steady-state solution exists, π is given by [11]:

$$\pi(i) = \pi(0)\mathbf{R}^i \tag{18}$$

where $\mathbf{R}$ is the minimal non-negative solution to the equation:

$$A_0 + \mathbf{R}A_1 + \mathbf{R}^2 A_2 = 0. \tag{19}$$

The vector $\pi(0)$ is obtained from the normalization condition, which in matrix notation writes as: $\pi(0)(I - \mathbf{R})^{-1}e = 1$.

Note that the evolution of number of RT calls is not affected by the process of NRT calls and the Erlang formula can be used to compute their steady state probability, and in particular, the probability of blocking of a RT call:

$$P_B^{RT} = \frac{(\rho_{RT})^{M_{RT}^{\max}} / M_{RT}^{\max}!}{\sum_{j=1}^{M_{RT}^{\max}} (\rho_{RT})^j / j!},$$

where $\rho_{RT} = \lambda_{RT}/\mu_{RT}$. This is the main performance measure for the RT traffic. For NRT calls the important performance measure is expected sojourn time which is given by Little's law as

$$T_{NRT} = E[M_{NRT}]/\lambda_{NRT}.$$

Conditional expected sojourn times. The performance measures so far are similar to those already obtained in the uplink case in [4]. We wish however to present more refined performance measures concerning NRT calls: the expected sojourn times conditioned on the file size and the state upon the arrival of the call. We follow [12] and introduce a non-homogeneous QBD process with the following generator Q^* and the corresponding steady state probabilities π^*:

$$Q^* = \begin{bmatrix} B & A_0 & 0 & 0 & \cdots & \\ (1/2)A_2 & A_1^{(2)} & A_0 & 0 & 0 & \cdots \\ 0 & (2/3)A_2 & A_1^{(3)} & A_0 & 0 & \cdots \\ 0 & 0 & (3/4)A_2 & A_1^{(4)} & A_0 & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \tag{20}$$

where the matrices A_0, A_2, B are the same as introduced before, and $A_1^{(k)}, k \geq 2$ is the same as A_1 defined before except that the diagonal element is chosen to be minus the sum of the off-diagonal elements of Q^*, i.e.

$$A_1^{(k)}[i,i] = -\lambda_{RT} - i\mu_{RT} - \lambda_{NRT} - \frac{k-1}{k}\nu(i).$$

The conditional expected sojourn time of a NRT mobile given that its size is v, that there are i RT mobiles and $k-1$ NRT mobiles upon it's arrival, is obtained from [12, Corollary 3.3 and remarks in § 8.3]:

$$T_{k,i}(v) = \frac{v}{R^* - \rho^*} + \overline{1_{k,i}}\left[I - \exp\left(v\mathcal{R}^{-1}Q^*\right)\right]\overline{w}, \tag{21}$$

where

$$R^* := \sum_{k,i} \pi^*(k,i)R_{NRT}^{tot}(i), \qquad \mathcal{R} = diag\left[\frac{1}{k}R_{NRT}^{tot}(i)\right],$$

$$\rho^* := \frac{\lambda_{NRT}}{\mu_{NRT}},$$

$\overline{1_{k,i}}$ is a vector of proper dimension whose entries are all zero except for the (k,i)-th entry whose value is 1, and $\overline{w}$ is the solution of

$$Q^*\overline{w} = \frac{1}{R^* - \rho^*}\mathcal{R} \cdot \overline{1_{k,i}} - \overline{1_{k,i}}.$$

The entries of $\mathcal{R}$ along the diagonal are ordered lexicographically in (k,i).

Remark 2. Expression (21) simplifies considerably in case the capacity allocated to NRT calls is fixed. Suppose that the number of RT sessions stays fixed to i throughout the system lifetime (this can be used as an approximation when the average duration of RT sessions is very large). The service rate is constant, $R_{NRT}^{tot}(i)$. Then we can study the system as an M/M/1 queue with processor sharing, for which we can easily derive from [2]:

$$T_k(v) = \frac{v/R_{NRT}^{tot}(i)}{1-\rho'} + [k(1-\rho') - \rho']\frac{1 - e^{-(1-\rho')\mu_{NRT} \cdot v}}{\mu_{NRT}R_{NRT}^{tot}(i)(1-\rho')^2} \tag{22}$$

provided that $\rho' := \dfrac{\lambda_{NRT}}{\mu_{NRT} R^{tot}_{NRT}(i)} < 1$ (ergodicity condition).

The equation is obtained by translating to time units: A job of v size units requires a service time $v/R^{tot}_{NRT}(i)$, if it were served alone in the system. Furthermore, the distribution of service time requirement is also exponential with mean $1/(\mu_{NRT} R^{tot}_{NRT}(i))$.

To illustrate the role of the conditional sojourn time, we use (22) to compute the maximum number k of NRT calls present at the arrival instant of an NRT call (we include in this number the arriving call) such that the expected sojourn time of the connection, conditional on its size and on k, is below 1 sec. This is depicted in Fig. 1. For example, if the mean size of the file is 100 kb then its conditional expected sojourn time will be smaller than 1 sec as long as the number of mobiles upon arrival (including itself) does not exceed 12. This figure is obtained with $R^{tot}_{NRT}(0) = 1000$ kbps, $\lambda_{NRT} = 1$ (we took no RT calls, i.e. $i = 0$).

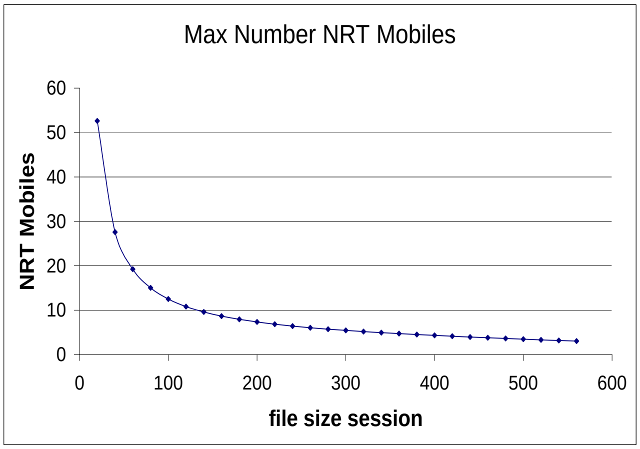

Fig. 1. Maximum number of NRT connections upon arrival such that the conditional expected sojourn time is below 1 sec, as a function of the mean size of the file

7 Numerical results

In this section, we examine basic performance parameters when RT and NRT traffic is integrated in the link, according to our transmission and rate control scheme. We consider the following setting illustrated in Table 1, based on standard WCDMA parameter values (cf. [6]). Unless stated otherwise, the data are for both the downlink (DL) and uplink (UL).

Table 7: Numerical Values.

Transmission rate of RT mobiles	*min* 4.75 kbps	*max* 12.2 kbps
E_{RT}/N_0	7.9 dB (12.2 kbps, *UL)*	
	11.0 dB (12.2 kbps, *DL*)	
E_{NRT}/N_0	4.5 dB (144 kbps, *UL*)	
	4.8 dB (384 kbps, *DL*)	
Mean NRT session size	$1/\mu_{NRT} = 160$ kbits	
Mean RT call duration	$1/\mu_{RT} = 125$ s	
Call arrival rates	$\lambda_{RT} = \lambda_{NRT} = 0.4$	
Intercell interference factors	$UL\colon f = 0.73$	$DL\colon F = 0.55$
Non-orthogonality factor (DL)	$a = 0.64$	
Chip rate	$W = 3.84$ Mcps	
Fraction of power for SCH, CCH channels	$\psi = 0.2$	

Moreover, in our numerical investigation we have chosen a very small value of ϵ ($\epsilon = 10^{-5}$), such that with negligible background noise an average mobile is located a few hundred meters from the base station [2] . Of course, in a more realistic application the value of ϵ must be selected more carefully and separately for the uplink and downlink.

The constant Γ is computed so as to guarantee that the probability of exceeding the target C/I ratio is 0.99. It corresponds to a standard deviation constant $\sigma = 0.5$ (see [4]).

Influence of NRT reservation on RT traffic In Fig. 2 we depict the average cell capacity in terms of the average number of RT mobiles for both uplink and downlink as a function of the reservation threshold for NRT traffic. We see that it remains almost constant (50 mobiles per cell) for up to 50% of the load.

[2] We have considered a background noise level of -100 dBm, and a path loss exponent 4 in an urban environment [6]. This roughly yields a power of about 20 Watt in the downlink (transmitted output power of a base station), and 1 Watt in the uplink (transmitted power from a mobile).

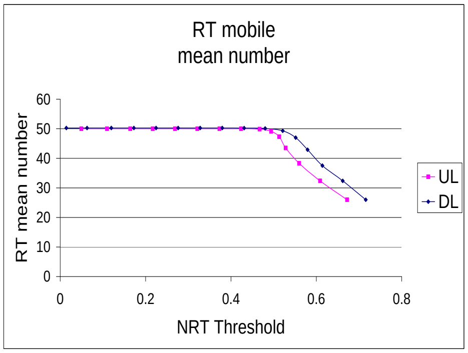

Fig. 2. Mean number of RT calls in a cell as a function of the reservation level for NRT traffic

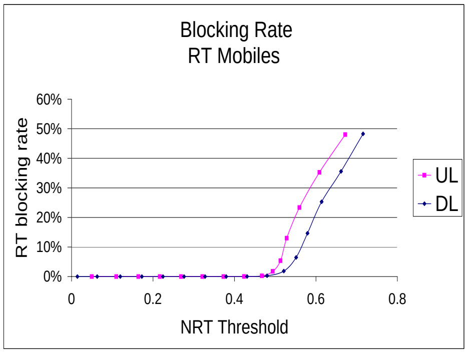

Fig. 3. Blocking rate for RT calls as a function of the reservation level for NRT traffic

In Fig. 3 we present the blocking rate of RT traffic. At a reservation L_{NRT} of 50% of the maximum load, the dropping rate is still lower than 1%.

Influence of NRT reservation on NRT traffic Fig. 4 shows the impact of the reservation threshold L_{NRT} on the expected sojourn time of NRT calls both on uplink and downlink. We see that the expected sojourn times become very large as we decrease L_{NRT} below 0.15% of the load. This demonstrates well the need for such a reservation. In the whole region of loads between 0.16 to 0.5 the NRT expected sojourn time is low and at the same time, as we saw before, the rejection rate of RT calls is very small. Thus, this is a good operating region for both RT and NRT traffic.

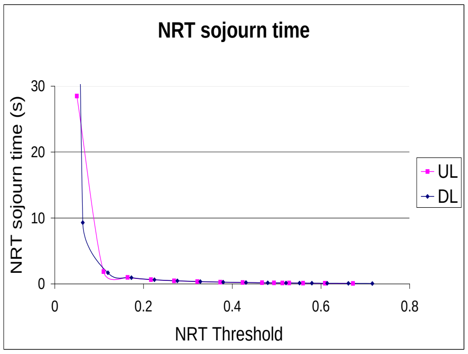

Fig. 4. Expected sojourn times of NRT traffic as a function of the NRT reservation

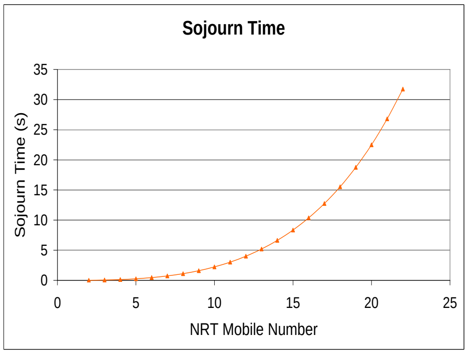

Fig. 5. Conditional expected sojourn time of an NRT mobile as a function of the number of mobiles in the cell

Conditional expected sojourn time The reservation limit L_{NRT} is taken to be 0.27 in Fig. 5, 6. In Fig. 5 we depict the expected sojourn time conditioned on the number of NRT and RT calls found upon the arrival of the call both being k and on the file being of the size of 100 kbits. The number k is varied in this figure.

Fig. 6 depicts for various file sizes, the maximum number k such that the conditional expected sojourn time of that file with the given size is below 1 sec. k is defined to be the total number of RT calls as well as the total number of NRT calls (including the call we consider) in the cell. We thus assume that the number of NRT and of RT calls is the same, and seek for the largest such number satisfying the limit on the expected sojourn time. Note, in comparison to Fig. 1, the decrease in the maximum number of mobiles, since RT calls now exist in the system.

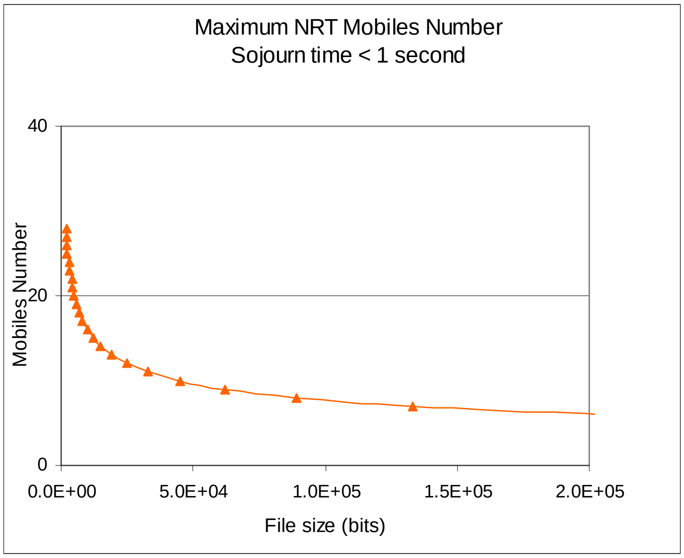

Fig. 6. Maximum number of NRT connections upon arrival such that the conditional expected sojourn time is below 1 sec, as a function of the size of the file

8 Extension to handover calls

So far in the paper, arrivals of new and handover calls in the CDMA cell had been succinctly incorporated in a single rate and thus not treated differently. We now wish to differentiate between these calls. We assume that RT new calls (resp. NRT new calls) arrive with a rate of λ_{RT}^{New} (resp. λ_{NRT}^{New}) where as the handover calls arrive at rate λ_{RT}^{HO} (resp. λ_{NRT}^{HO}). We assume that RT calls remain at a cell during an exponentially distributed duration with parameter μ_{RT}.

From a QoS perspective, avoiding blocking of handover calls is considered more important than avoiding blocking of new ones. So we define a new threshold $\overline{M}_{RT}^{New} < M_{RT}^{\max}$. New RT calls are accepted as long as $M_{RT} \leq \overline{M}_{RT}^{New}$, whereas handover RT calls are accepted as long as $M_{RT} \leq M_{RT}^{max}$. The behavior of NRT calls is as before. Define $\rho_{RT} = \lambda_{RT}/\mu_{RT}$ (the same as before, corresponding to the total arrival rate) and $\rho_{RT}^{HO} = \lambda_{RT}^{HO}/\mu_{RT}$. Let $p_{RT}(i)$ denote the number of RT mobiles in steady state. It is given by

$$p_{RT}(i) = \begin{cases} \frac{(\rho_{RT})^i}{i!} p_{RT}(0), & \text{if } 0 \le i \le \overline{M}_{RT}^{New} \\ \\ \frac{(\rho_{RT})^{\overline{M}_{RT}^{New}} (\rho_{RT}^{HO})^{i-\overline{M}_{RT}^{New}}}{i!} p_{RT}(0), & \text{if } \overline{M}_{RT}^{New} \le i \le M_{RT}^{\max} \end{cases}$$

where $p_{RT}(0)$ is a normalizing constant given by

$$\left(\sum_{i=0}^{\overline{M}_{RT}^{New}} \frac{(\rho_{RT})^i}{i!} + \sum_{i=\overline{M}_{RT}^{New}}^{M_{RT}^{\max}} \frac{(\rho_{RT})^{\overline{M}_{RT}^{New}} (\rho_{RT}^{HO})^{i-\overline{M}_{RT}^{New}}}{i!} \right)^{-1}.$$

The QBD approach introduced before can be directly applied again to compute the joint distribution of RT and NRT calls, and in particular, the expected sojourn time of NRT calls.

A numerical example We consider the uplink case of the CDMA system. The input data are the same as before, except that now a fraction of 30% of arriving RT calls are due to handovers. In Fig. 7 we present the impact of the choice of the NRT threshold on the blocking rate of RT mobiles. We also illustrate the impact of the differentiation between new and handover calls. The middle curve is obtained with no differentiation. The total dropping rate of the model with handover differentiation is larger, but the dropping rate of calls already in the system (that arrive through a handover) is drastically diminished (the curve called "Dropping").

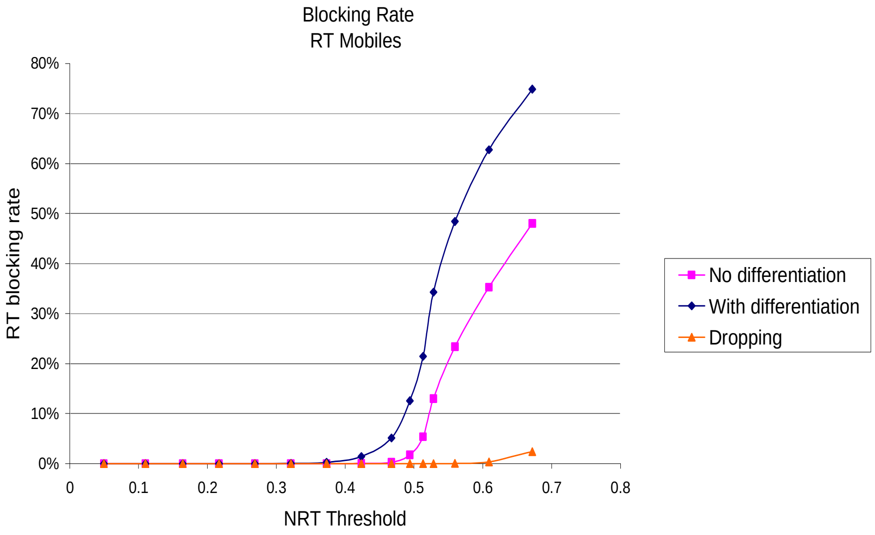

Fig. 7. RT dropping probabilities

9 Summary and conclusions

We have developed a simplified mathematical model that allowed us to analyze the performance of call admission control combined with GoS control in a WCDMA environment with integrated RT and NRT traffic. RT traffic has limited adaptive rate functionalities and priority over resources whereas NRT traffic obtains by time sharing the capacity left over by the RT traffic.

As performance measures we studied the blocking rate of RT traffic and the sojourn times of NRT traffic. We illustrated through numerical examples the importance of adding reserved capacity L_{NRT} for NRT traffic and demonstrated that this reservation can be done in a way not to significantly affect RT traffic. More specifically, we saw that the blocking rate of RT traffic was small and quite robust to the choice of L_{NRT}, over a large interval of values. For NRT traffic, we investigated not only the average sojourn time but also the conditional expected sojourn time given the file size and the number of RT and NRT mobiles present at the cell upon arrival.

Finally, we provided an extension of the multiservice system model to handle handover RT calls. It was shown that differentiating the admission control policy for such calls can greatly reduce their blocking probability, and therefore provide better QoS.

References

[1] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users", *IEEE Commun. Magazine*, 70–77, July 2000.

[2] E.G. Coffman Jr., R.R. Muntz, H. Trotter, "Waiting time distributions for processor-sharing systems", *Journal of the ACM*, vol. 17, no. 1, pp. 123–130, Jan. 1970.

[3] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks", *IEEE Trans. Vehicular Technology* vol. 51, no. 2, pp. 371–382, March 2002.

[4] N. Hegde and E. Altman, "Capacity of multiservice WCDMA Networks with variable GoS", in *Proc. of IEEE WCNC*, New Orleans, Louisiana, USA, March 2003. The full version is to appear in *Wireless Networks*, 2005.

[5] K. Hiltunen and R. De Bernardi, "WCDMA downlink capacity estimation", in *Proc. IEEE VTC*, pp. 992–996, Tokyo, Japan, May 2000.

[6] H. Holma and A. Toskala (Eds.), *WCDMA for UMTS: Radio access for third generation mobile communications*, 2nd Edition, J. Wiley & Sons, 2002.

[7] I. Koo, J. Ahn, H. A. Lee, K. Kim, "Analysis of Erlang capacity for the multimedia DS-CDMA systems", *IEICE Trans. Fundamentals*, vol E82-A, no. 5, pp. 849–855, May 1999.

[8] J. Laiho and A. Wacker, "Radio network planning process and methods for WCDMA", *Ann. Telecommun.*, vol 56, no. 5-6, pp. 317–331, May/June 2001.

[9] C. W. Leong and W. Zhuang, "Call admission control for voice and data traffic in wireless communications", *Computer Communications*, vol. 25, no. 10, pp. 972–979, 2002.

[10] B. Li, L. Li, B. Li, X.-R. Cao "On handoff performance for an integrated voice/data cellular system" *Wireless Networks*, vol. 9, no. 4, pp. 393 – 402, July 2003.

[11] M. F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach.* The John Hopkins University Press, 1981.

[12] R. Núnez Queija. "Sojourn times in non-homogeneous QBD processes with processor sharing". *Stochastic Models*, vol. 17, pp. 61–92, 2001.

[13] S. Parkvall, E. Dahlman, P. Frenger, P. Beming and M. Persson, "The high speed packet data evolution of WCDMA", in *Proc. of the 12th IEEE PIMRC*, San Diego, USA, 2001.

[14] B. Sklar, "Rayleigh fading channels in mobile digital communication systems — Part I: Characterization", *IEEE Commun. Magazine*, pp. 90–100, July 1997.

[15] X. Tang and A. Goldsmith, "Admission control and adaptive CDMA for integrated voice and data systems", in *Proc. IEEE VTC*, pp. 506–510, Rhodes, Greece, May 2001.

[16] K. Sipilä, Z.-C. Honkasalo, J. Laiho-Steffens, A. Wacker, "Estimation of capacity and required transmission power of WCDMA downlink based on a downlink pole equation", in *Proc. IEEE VTC*, pp. 1002–1005, Tokyo, Japan, May 2000.