A non-homogeneous QBD approach for the admission and GoS control in a multiservice WCDMA system

Ioannis Koukoutsidis, Eitan Altman, Jean Marc Kelif

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A non-homogeneous QBD approach for the admission and GoS control in a multiservice WCDMA system

Ioannis Koukoutsidis — Eitan Altman — Jean Marc Kelif

N° 5358

Novembre 2004

Thème COM

ISSN 0249-6399 ISRN INRIA/RR--5358--FR+ENG

INRIA
SOPHIA ANTIPOLIS

A non-homogeneous QBD approach for the admission and GoS control in a multiservice WCDMA system

Ioannis Koukoutsidis*, Eitan Altman*, Jean Marc Kelif †

Thème COM —Systèmes communicants
Projet MAESTRO

Rapport de recherche n° 5358 — Novembre 2004 —25 pages

Abstract: We consider a WCDMA system with two types of calls: real time (RT) calls that have dedicated resources, and data non-real-time (NRT) calls that share system capacity. 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 derive performance evaluation results regarding transmission rates, blocking of RT calls and sojourn time of NRT calls, under different traffic characteristics. On what concerns the bandwidth-sharing policy of NRT traffic, we compare WCDMA behavior in the presence of high data rate schemes. Finally, we extend our results to cover NRT admission control schemes and examine blocking behavior and sojourn time of NRT traffic.

Key-words: CDMA, multiservice traffic, admission and GoS control, quasi-birth-death process

This work was supported by a CRE research contract with France Telecom R&D and by the EURO NGI network of excellence.

* {Giannis.Koukoutsidis}{Eitan.Altman}@sophia.inria.fr
† {JeanMarc.Kelif}@francetelecom.com, France Telecom R&D, Rue du Général Leclerc, 92794 Issy les Moulineaux Cedex 9

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 —Télécopie : +33 4 92 38 77 65

Une approche de QBD non-homogène pour le contrôle d'admission et du GoS dans un système multiservice WCDMA

Résumé : Nous considérons un système WCDMA (*Wideband Code Division Multiple Access*) avec deux types de trafic: les appels en temps réel (TR), qui ont des ressources dédiées, et l'envoi des données non temps-réel (NTR), qui partagent la capacité du système. Nous considérons la réservation de quelques ressources pour le trafic NTR; de plus, les ressources laissées par le trafic TR lui sont également allouées. Le degré de service (GoS) est contrôlé afin de permettre le traitement d'un plus grand nombre d'appels TR pendant les périodes de congestion, au détriment des débits de transmission qui se retrouvent dégradés. Nous analysons le sens descendant (avec ou sans macrodiversité) aussi bien que le sens montant du système WCDMA et évaluons les performances des points de vue des débits de transmission, du blocage des appels TR et du temps de séjour des appels NTR, en fonction des caractéristiques des différents trafics. En ce qui concerne la politique de partage de la capacité allouée au trafic NTR, nous comparons le comportement du WCDMA en présence des protocoles de transmission des données à haut débit. Finalement, nous étendons notre analyse pour inclure le contrôle d'admission du trafic NTR et examinons le blocage et le temps de séjour des appels NTR.

Mots-clés : CDMA, trafic multiservice, contrôle d'admission et du degré de service, processus de quasi naissance et de mort

Contents

1 Introduction

In this paper, we are interested in analyzing resource sharing between RT (real time) and NRT (non-real time) calls in a cellular CDMA network, as well as the attained QoS (quality of service) and GoS (grade of service). 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. [6],[15]. 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 trade-offs between QoS parameters of the different classes of mobiles. However, unlike TDMA or FDMA systems in which the notions of channels and capacity are clear, the capacity of a CDMA system is rather a complex combination of cell parameters and channel conditions, being mostly interference-limited [8],[21]. This largely differentiates the analysis and complicates dimensioning and planning of network resources.

The QoS parameters of interest are primarily the blocking probabilities for RT calls and expected sojourn times for NRT calls. We allow downgrading of transmission rates (which is viewed as the GoS) of RT calls[1] during congestion epochs. The main factors influencing bandwidth sharing are then the call admission policy for RT calls as well as their downgrading policy.

This paper is a follow-up of [9],[12] in which NRT traffic was scheduled using a time-sharing approach, as is the case in the High Speed Downlink Packet Access (HSDPA) [20],[11] system. This allowed to derive a tractable mathematical model based on a homogeneous QBD (Quasi Birth-Death) process [19],[14]. The network analyzed in the current paper cannot be evaluated anymore with a homogeneous QBD and we present a more involved analysis based on a non-homogeneous QBD.

In order to study the system's performance we first model the downlink case which allows to obtain the capacity required by a call of a given class with a given GoS. In particular we also consider the case of macrodiversity. We then introduce a corresponding uplink model. We propose a control policy that combines admission control together with the control of the transmission rate for RT traffic. Key performance measures are then computed.

2 Background: Computing the transmission rates

The analysis is based on radio models for the downlink and uplink introduced in [12],[9]. For completeness we recall in this section the derivation of capacities and transmission rates.

2.1 Downlink

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

$$(C/I)_k = \frac{E_s}{N_o}\frac{R_s}{W}\Gamma, \tag{1}$$

where E_s is the energy per transmitted bit of type s, N_o is the thermal noise density, W is the WCDMA chip rate from which the modulation bandwidth is derived, R_s is the transmission rate of the type s call, and Γ is a constant that is related to shadow fading and imperfect power control, and is derived in the same way as in the uplink case [9].

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

[1] UMTS uses the Adaptive Multi-Rate (AMR) codec that offers eight different transmission rates of voice that vary between 4.75 Kbps to 12.2 Kbps, and that can be dynamically changed every 20 msec.

$\sum_{j=1}^{M} P_{j,l}+P_{SCH}+P_{CCH}$ where P_{SCH} , P_{CCH} correspond to the power transmitted for the non-orthogonal synchronization channel (SCH) and the orthogonal common channels (CCH), respectively. Note that these two terms are not power controlled and are assumed not to depend on l. Due to the multipath propagation, a fraction a of the received own cell power is experienced as intracell interference. 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), $I_{k,intra} = \alpha \cdot (P_{SCH} + P_{CCH} + \sum_{j\neq k} P_{j,l})/g_{k,l}$ and $I_{k,inter} = \sum_{j=1,j\neq l}^{S} P_{tot,j}/g_{k,j}$. 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

$$\beta_k = \frac{P_{k,l}/g_{k,l}}{(F_{k,l}+\alpha)P_{tot,l}/g_{k,l} + N}, \tag{2}$$

$$\text{where } \beta_k = \frac{(C/I)_k}{1+\alpha(C/I)_k}. \tag{3}$$

We then consider two service classes, that will correspond to real time (RT) and non-real time (NRT) traffic. Let $(C/I)_s$ be the target SIR ratio for mobiles of service class s and β_s be the corresponding value in (3). Let there be in a given cell M_s mobiles of class s. We shall use the following approximations. First we replace $F_{k,l}$ by a constant (e.g. its average value, as in [10]; this is a standard approximation, see [11]). Secondly, we approximate $g_{k,l}$ by their averages. More precisely we define G_s to be the average over all mobiles k belonging to class s, $s = 1, 2$. With these approximations (2) gives the following value for $P_{tot,l}$ (we omit the index l):

$$P_{tot} = \frac{P_{SCH} + P_{CCH} + N\sum_s \beta_s M_s G_s}{1-(\alpha+F)\sum_s \beta_s M_s}. \tag{4}$$

Further assuming that the system is designed so as to have $P_{SCH} + P_{CCH} = \psi P_{tot}$ and defining the downlink load as $Y_{DL} = \sum_s \beta_s M_s$, this gives

$$P_{tot} = \frac{N\sum_s \beta_s M_s G_s}{Z_2}, \quad \text{where} \quad Z_2 = (1-\psi) - (\alpha+F)Y_{DL}. \tag{5}$$

In practice, to avoid instabilities and due to power limitation of the base stations, one wishes to avoid that Z_2 becomes too close to zero, thus one poses the constraint $Z_2 \geq \epsilon$ for some $\epsilon > 0$. We can thus define the system's capacity as $\Theta_\epsilon = 1 - \psi - \epsilon$, and the capacity required by a connection to be

$$\Delta(s) := (\alpha+F)\beta_s. \tag{6}$$

Combining this with (1) and with (3) we get

$$R_s = \frac{\Delta(s)}{\alpha+F-\alpha\Delta(s)} \times \frac{N_o W}{E_s \Gamma}. \tag{7}$$

2.2 Downlink with macrodiversity

Our approach is inspired by [10] who considered the single service case. A mobile i in macrodiversity (MD) is connected to two base stations, b and l. Following [10] we assume that the Maximum Ratio Combining (MRC) is used and hence the power control tries to maintain

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

where γ_k is given by the constant in (1). We additionally define

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

We set b to be the station with larger SIR so that we always have $\Omega_i \leq 1$. We get for the combined C/I [10]:

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

The transmission power becomes

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

where

$$\kappa_i = \frac{(C/I)_i}{1+\Omega_i+\alpha(C/I)_i}. \tag{8}$$

Let there be M mobiles in a cell b (we shall omit this index) of which a fraction μ is in macrodiversity. Then the total base station output power can be written as

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

Note that P_i, the power for a single link user is calculated the same way as in the previous case. We now consider two classes of service $s = 1, 2$ corresponding to RT and NRT mobiles. For a given service class $s = 1, 2$, Ω_i is replaced by a constant Ω_s (its average over all mobiles of the same service as i); we also replace $F_{i,b}$ by one of four constants F_s^{NMD} and F_s^{MD}, $s = 1, 2$, where F_s^{NMD} (resp. F_s^{MD}) corresponds to an average value of $F_{i,b}$ over mobiles in service s which are not in macrodiversity (and which are in macrodiversity, resp.). Finally, we replace $g_{i,b}$ by one of the four constants G_s^{NMD} and G_s^{MD}, $s = 1, 2$, where G_s^{NMD} (resp. G_s^{MD}) corresponds to an average value of $g_{i,b}$ over mobiles in service s which are not in MD (and which are in MD, resp.). 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

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

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

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

$$\begin{aligned}\Delta(s) &= (1-\mu)\cdot\beta_s(a + F_s^{NMD}) + 2\mu\cdot k_s(a + F_s^{MD}) = \\ &= (1-\mu)\cdot\frac{R_s\cdot\delta_s}{1+aR_s\delta_s}(a + F_s^{NMD}) + 2\mu\cdot\frac{R_s\cdot\delta_s}{1+\Omega_s+aR_s\delta_s}(a + F_s^{MD}).\end{aligned} \tag{9}$$

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

2.3 Uplink

We briefly recall the capacity notions from the case of uplink from [9]. Define for $s = 1, 2$,

$$\tilde{\Delta}_s = \frac{E_s}{N_o}\frac{R_s}{W}\Gamma, \text{ and } \Delta'(s) = \frac{\tilde{\Delta}(s)}{1+\tilde{\Delta}(s)}. \tag{10}$$

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 [9] 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). To avoid instability one requires that $Z_2 \geq \epsilon$ for some $\epsilon > 0$. We can thus define the system's 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 (10) we get

$$R_s = \frac{\Delta(s)}{1+f-\Delta(s)} \times \frac{N_o W}{E_s \Gamma}. \tag{11}$$

3 Admission and rate control

We assume that there exists a capacity L_{NRT} reserved for NRT traffic. The RT traffic can use up to a capacity of $L_{RT} := \Theta_\epsilon - L_{NRT}$. We also introduce GoS by providing RT calls with a variable transmission rate. In such a case, we may allow more RT calls at the expense of a reduced transmission rate.

Assume more generally that the set of available transmission rates for 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 $[\Delta^{\min}, \Delta^{\max}]$. Note that the maximum number of RT calls that can be accepted is $M_{RT}^{\max} = \lfloor \Theta_\epsilon/\Delta^{\min} \rfloor$. We assign full rate $R^{\max}$ (and thus the maximum capacity $\Delta^{\max}$) for each RT mobile as long as $M_{RT} \leq N_{RT}$, where $N_{RT} = \lfloor L_{RT}/\Delta^{\max} \rfloor$. 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.

We consider that NRT calls make use of the reserved system capacity, as well as any capacity left over from RT calls. Thus the available capacity for NRT calls is a function of M_{RT} as follows:

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

In [12],[9] the capacity $C(M_{RT})$ unused by the RT traffic (which dynamically changes as a function of the number of RT connections present) was fully assigned to a 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. This modeling is more appropriate for a high data rate scheme. Specifically, schemes such

as HDR [1], corresponding to the CDMA 1xEV-DO standard, and its 3GPP counterpart HSDPA [20] have been proposed for the downlink in order to achieve higher asymmetric rates. These schemes implement a complex scheduler which evaluates channel conditions and pending transmissions for each connection, using additionally fast retransmission and multicoding to improve throughput. The scheduling decisions permit the system to benefit from short-term variations and allow most of the cell capacity to be allocated to one user for a very short time, when conditions are favorable.

The modeling in this optimum scenario follows a homogeneous QBD approach, as the transmission rate is independent of the number of on-going NRT sessions[2]. Here we consider the case where available capacity is split equally between the NRT calls, thus employing a *fair rate* sharing approach. According to the previous analysis, if there are k NRT calls present the transmission rate of a single NRT call is given by:

$$d.l.: \quad R_{NRT}(M_{RT}) = \frac{C(M_{RT})/k}{\alpha + F - \alpha C(M_{RT})/k} \times \frac{N_o W}{E_s \Gamma},$$

$$u.l.: \quad R_{NRT}(M_{RT}) = \frac{C(M_{RT})/k}{1 + f - C(M_{RT})/k} \times \frac{N_o W}{E_s \Gamma}.$$

Then, in contrast to [12],[9] the total transmission rate R^{tot}_{NRT} of NRT traffic for the downlink and uplink depends on the number M_{RT} of RT calls as well as the number M_{NRT} of NRT calls and is given respectively by

$$d.l.: \quad R^{tot}_{NRT}(M_{RT}, M_{NRT}) = \frac{M_{NRT} C(M_{RT})}{M_{NRT}(\alpha + F) - \alpha C(M_{RT})} \times \frac{N_o W}{E_s \Gamma},$$

$$u.l.: \quad R^{tot}_{NRT}(M_{RT}, M_{NRT}) = \frac{M_{NRT} C(M_{RT})}{M_{NRT}(1 + f) - C(M_{RT})} \times \frac{N_o W}{E_s \Gamma}.$$

The expression for the downlink with macrodiversity is similarly derived, albeit being much more cumbersome.

4 Traffic model and the LDQBD approach

We assume that RT and NRT calls arrive according to independent Poisson processes with rates λ_{RT} and λ_{NRT}, respectively. The duration of an RT call is exponentially distributed with parameter μ_{RT}. The size of an NRT file is exponentially distributed with parameter μ_{NRT}. Interarrival times, RT call durations and NRT file sizes are all independent. Note that the departure rate of NRT calls depends on the current number of RT and NRT calls:

$$\nu(M_{RT}, M_{NRT}) = \mu_{NRT} R^{tot}_{NRT}(M_{RT}, M_{NRT}).$$

Under these assumptions, the number of active sessions in all three models (downlink with and without macrodiversity and uplink) can be described as a *non-homogeneous* or *level-dependent* (LD) QBD process, and we denote by Q its generator. Upon a stable system, the stationary distribution π is calculated by solving

$$\pi Q = 0, \tag{12}$$

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 lexicographically. We may thus partition

[2]Note however that the mathematical modeling does not take into consideration the delay caused by the scheduling operation and the corresponding throughput decrease that can be induced. Further, the modeling is idealistic because it does not consider the random fluctuations in signal conditions and assumes that all users are, at any moment, capable of transmitting at the peak rate. For more details, interested readers are referred to the works in [3],[2].

π as $[\pi(0), \pi(1), \ldots]$ with $\pi(i)$ 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 $i \geq 0$. In (i,j), j is referred to as the *phase* of the state. The generator Q is given by

$$Q = \begin{bmatrix} B & A_0 & 0 & 0 & \cdots \\ A_2^1 & A_1^1 & A_0 & 0 & \cdots \\ 0 & A_2^2 & A_1^2 & A_0 & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots \end{bmatrix} \tag{13}$$

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

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

Of course, $A_1^j[i,i]$ are properly modified on the boundary $i = M_{RT}^{max} + 1$. We also have $B = A_1^j + A_2^j$. Due to the special structure of the matrix, this is independent of j.

As in the QBD case, there exist matrix-geometric methods to calculate the equilibrium distribution of a LDQBD process. These involve the solution of a system of matrix recurrence equations, see [14]. However, the number of states is often so large that the solution becomes untractable. For this reason, algorithmic approaches are usually seeked. Here we extend a method introduced in [7] for a finite non-homogeneous QBD process. The implementation is simple and converges to the equilibrium distribution in a relatively small number of steps. Details of the algorithm are deferred to the Appendix.

5 Numerical Evaluation

In this section, the major performance evaluation results are presented for a system with integrated RT and NRT calls. First the uplink and downlink performance is analyzed and the system bottleneck is determined. Comparisons are then carried out against our –idealistic– model of the high data rate HSDPA scheme in WCDMA. Continuining, we explore the extent to which intercell interference can deteriorate system behavior. Next, the macrodiversity behavior under maximum ratio combining is exhibited. We also present results by varying traffic characteristics on the bottleneck side. Finally, a discussion of NRT call admission control is given followed by evaluation results.

5.1 Setting

First we address the values of parameters used in the numerical evaluation. Common CDMA performance evaluation parameters (such as chip rate, energy-to-noise requirements, interference factors, etc.) are derived from equipment capabilities and field tests. The actual traffic characteristics (rate of arrivals, service times) can be modified more flexibly to reflect differrent traffic conditions. The parameters initially used for the numerical evaluations in our setting are as follows:

- Chip rate: $W = 3.84$ Mcps
- Transmission rate of RT mobiles: max 12.2 Kbps, min 4.75 Kbps

- E_{RT}/N_0 (12.2 Kbps voice service): Uplink 4.2 dB, Downlink 7.0 dB
- E_{NRT}/N_0 (144 Kbps data service): Uplink 2.2 dB, Downlink 5.0 dB
- Average RT call duration: 125 s
- Arrival rate of RT calls: $\lambda_{RT} = 0.4$
- Mean NRT session size: 160 Kbits
- Arrival rate of NRT calls: $\lambda_{NRT} = 0.4$
- Interference factor: Uplink $f = 0.73$, Downlink $f = 0.55$, Downlink with macrodiversity $f = 0.65$
- Fraction of received own cell power experienced as intra-cell interference: $a = 0.64$
- Fraction of total power transmitted in the downlink for SCH and CCH channels: $\psi = 0.2$
- Safety margin for capacity: $\epsilon = 10^{-5}$

The traffic characteristics for RT and NRT calls are initially chosen to correspond to heavy traffic conditions, where by default performance evaluation is more challenging. In addition, we note that the evaluation of results is particularly sensitive to the E_b/N_0 requirements; increased values lead to enhanced capacity requirements and thus to an extreme load in the system. Here, the E_b/N_0 targets are set according to §12.5 of [11] (3GPP performance requirements for a slow moving user, Table 12.26). Values are greater in the downlink, the reason being smaller receiver sensitivity and antenna gain in the mobile units. In addition, antenna diversity is not usually assumed in the downlink. We have also made the simplifying assumption that these values remain approximately constant for different transmission rates. This generally holds when the same type of modulation is used for all rates [13].

The parameter Γ, which accounts for shadow fading in the calculation of the system capacity has been incorporated in the values of E_b/N_0. Also, note the value chosen for interference in the case of the downlink with macrodiversity, $f = 0.65$; this is increased compared to the case without MD, as a mobile being in the edge between two base stations would experience more interference (although not as much as in the uplink).

5.2 Uplink and downlink performance

Here we study the behavior in the uplink (UL) and the downlink (DL) of the WCDMA system. This responds to our first major concern, i.e. which side represents the bottleneck of the system. We have, on the one hand, that the downlink enjoys less interference; however, this is largely eclipsed by the increased E_b/N_0 ratios that require more capacity for a given transmission rate, and the expendited power for SCH and CCH channels. Hence, the downlink is expected to be the bottleneck of the system.

This is confirmed in both RT and NRT call behavior. For RT traffic, the major performance metric is the blocking probability of a new call, since QoS bounds are otherwise guaranteed. This is calculated and shown graphically in Fig. 1, for different values of the L_{NRT} threshold, ranging from 0 to $(0.8 - 10^{-5})$ in the downlink (due to SCH and CCH channels) and 0 to $(1.0 - 10^{-5})$ in the uplink. As anticipated, the probability of rejection increases as more capacity is reserved for NRT calls. However, as a distinctive difference we append that a blocking probability $P_B > 10^{-2}$ can be induced by $L_{NRT} = 0.44$ in the uplink, while in the downlink the NRT reserved capacity may be as low as $L_{NRT} = 0.16$.

In the case of NRT traffic, performance evaluation results are portrayed in Fig. 2. Here, quality of service is manifested essentially by the time it takes to complete the document transfer, i.e. the mean sojourn time in the system. In addition, the mean transmission rate of a single NRT call is of equal interest, since no

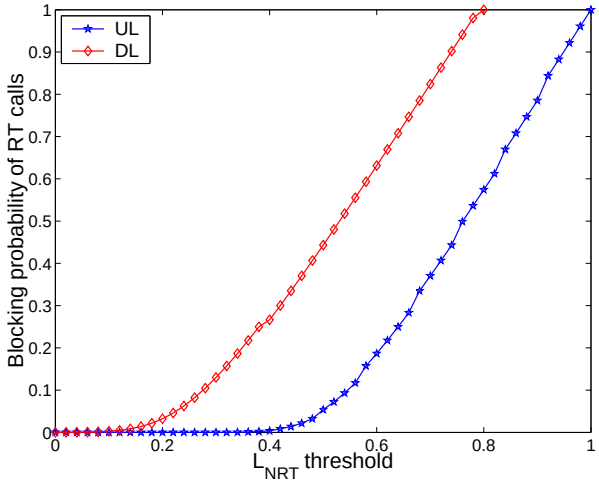
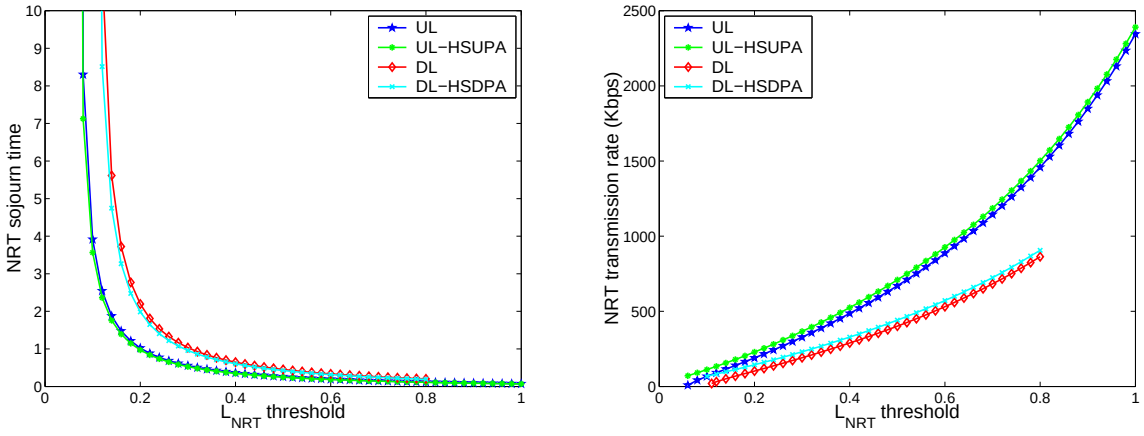


Figure 1: Blocking probability of RT calls vs. L_{NRT} reserved capacity, in the UL and DL cases.



(a) (b)

Figure 2: Mean NRT sojourn time (a) and transmission rate (b) vs. L_{NRT} reserved capacity, in the UL and DL cases. Comparison with high data rate services.

constraints are imposed. Again we ascertain the performance deterioration in the downlink. It should also be noted at this point that differences between the uplink and downlink are much higher if we take examples with large asymmetries in transmitted traffic in each direction, which is commonly the case.

The behavior of NRT traffic deserves to be explained, since it reflects the general admission and rate control modeled previously: given the same NRT file size distribution and in availability of a lot of resources, the NRT calls that "come into" the system transmit at a higher rate and then leave. Therefore, the corresponding sojourn time can be smaller. On the other hand, if there are only few resources, the NRT calls that join in transmit at a very low rate and stay in the system longer. It follows that the mean number of NRT calls decreases in the first case, while it increases in the latter.

An ergodicity condition is essential for stability in the theoretical case of an unbounded number of NRT calls. As shown in Fig. 2(a), below a certain value of the L_{NRT} threshold (approximately[3] $L_{NRT} \approx 0.1$ in the DL case), the sojourn time tends to infinity and the system becomes unstable. That is, below a certain capacity the NRT transmission rate becomes too small, which leads to a very high number of such calls in the system. In the system under consideration, the stability condition is:

$$\mu_{NRT} \cdot \mathbb{E}R_{NRT}^{tot} > \lambda_{NRT}. \tag{14}$$

Here, the calculation of $\mathbb{E}R_{NRT}^{tot}$ is problematic, since it also depends on the number of NRT calls which is unbounded. However, we observe that as $M_{NRT} \to \infty$, the total transmission rate reaches a limit in all UL and DL cases. For example, in the UL case we have

$$\lim_{M_{NRT}\to\infty} R_{NRT}^{tot} = \frac{C(M_{RT})}{1+f} \cdot \frac{N_0 W}{E_s \Gamma}.$$

Therefore, the non-homogeneous LDQBD process asymptotically converges to a homogeneous QBD process. Moreover, the departure rates of NRT calls in the LDQBD process are greater for smaller levels, and always greater than those of the limiting process[4]. It can be formally shown that stability conditions are the same for both processes, i.e. it suffices to check the ergodicity of the limiting homogeneous process. The proof is deferred to the Appendix, as its applicability is more general and its scope can be extended beyond the main theme of the paper.

The divergent performance of the uplink and downlink is revisited in Fig. 2(b), in terms of the mean NRT transmission rate. For small NRT allocated capacity, the transmission rate is in any case small. However, the difference becomes more pronounced as the L_{NRT} threshold increases. For $L_{NRT} \approx 0.8$, the difference amounts to 595.2 Kbps. The uplink transmission rate can attain even larger values as more capacity is being allocated, reaching 2.34 Mbps for $L_{NRT} \approx 1$ (UL).

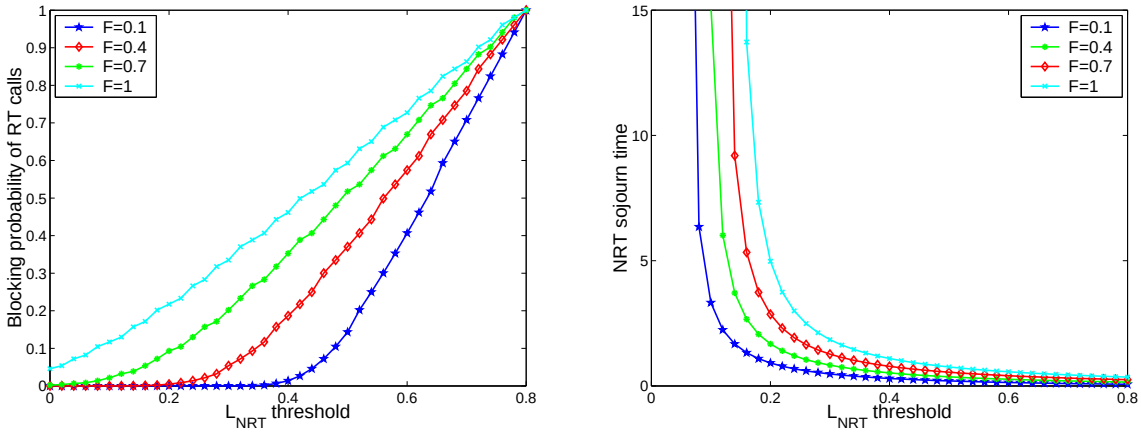
In addition, Fig. 2 presents a comparison of the standard WCDMA behavior with that of the HSDPA scheme. We also consider the corresponding scheme in the uplink –analogously named HSUPA– which has recently been added in 3GPP Release 6 [11]. The numerical results underlying Fig. 2(b) reveal that the high data rate scheme can increase the cell throughput in case of small NRT reserved capacity, the observed increase becoming proportionately smaller for higher values. Comparing the normal WCDMA and HSDPA cases, we note an increase of 162% for $L_{NRT} = 0.12$, and 4.97% when $L_{NRT} = 0.8$. The attainable performance improvement is then apparent under system congestion conditions, namely very high load or very small allocated capacity. Indeed, in terms of the mean sojourn time, Fig. 2(a) shows that the outperformance of the time-scheduling approach is non-negligible for small NRT reserved capacity (approx. regions $L_{NRT} < 0.14$ in the uplink, $L_{NRT} < 0.2$ in the downlink). In the numerical results obtained, the difference reached up to 80 sec in the uplink, for $L_{NRT} = 0.06$.

[3] A granularity of 10^{-2} is taken in the numerical results.

[4] NRT arrival rates are the same; refer to $Q_2^{(k)}$ inequality relations in the Appendix.

5.3 Impact of the interference expansion factor

As CDMA system capacity is primarily limited by interference, we would like to know to what extent this affects system behavior. Here numerical results are taken by varying the ratio of intercell-to-intracell interference, F in the downlink[5]. A more perceptive term for this is the *interference expansion factor*. Increasing values of F can then be seen as increased intercell interference.



(a) (b)

Figure 3: Blocking probability of RT calls (a) and mean NRT sojourn time (b) vs. L_{NRT} reserved capacity, for different values of the interference ratio, F, in the downlink.

Numerical results are portrayed in Fig. 3. The value of the interference expansion factor depends on the traffic distribution of interfering cells and may well assume values greater than unity [22]; however we take selected values until $F = 1$ for our test results here. We may deduce that intercell interference has a significant impact on performance. Concerning the blocking probability of RT calls in Fig. 3(a), for smaller values of F an initially good performance is observed; for the smallest value $F = 0.1$, the loss rate remains insignificant until $L_{NRT} < 0.4$. However, blocking severely increases for higher interference ratio; for $F = 1$, a blocking probability of $P_B = 5 \cdot 10^{-2}$ occurs even for no allocated NRT capacity and is almost linearly increased to the value of 1 as the L_{NRT} threshold increases.

The NRT behavior is similarly affected. We observe in Fig. 3(b) that the mean transfer time is greater as interference increases, as well as that the instability region is larger. We are able to make an illustrative comparison for the value of $L_{NRT} = 0.14$ where all systems are stable. We have, in that case, that $T_{soj}(F = 0.1) = 1.67\ s$, $T_{soj}(F = 0.4) = 3.71\ s$, $T_{soj}(F = 0.7) = 9.19\ s$, and $T_{soj}(F = 1) = 96.86\ s$. In a realistic setting, the first three values may be tolerable, however the last value certainly isn't, especially in view of the mean size of the document in transfer (160 Kbits).

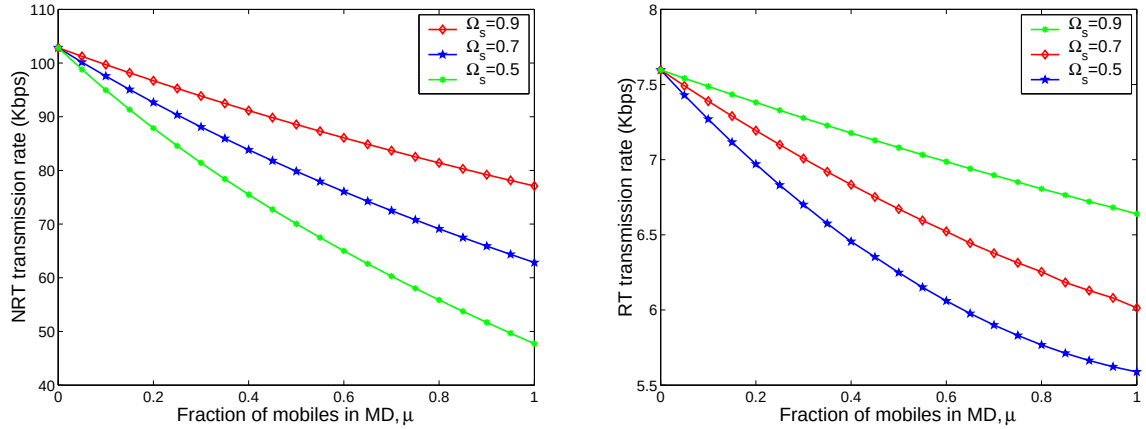
Similar tests conducted in the uplink as well as the high data rate scheme lead to the same conclusions. The deteriotation of system behavior in all cases is due to the fact that more power, and hence more capacity is requested by users to overcome interference. This means less bandwidth available –even for the lowest quality RT calls– and smaller transfer rates for NRT sessions. Finally, the same situation –due to power control– occurs in the uplink, in case of increased intracell interference, and we expect the same observations to carry over to this case.

[5]Note that in the downlink this is mitigated by the effect of multipath loss, as $F_{k,l} = a \cdot \frac{I_{k,inter}}{I_{k,intra}}$. However this does not affect the generality of results.

5.4 Macrodiversity behavior

Macrodiversity in the downlink refers to the maintainance of an on-going connection between the mobile terminal and the network by more than one base stations, through maximum ratio combining. It is employed in *soft* and *softer* handover techniques, in order to combat fading and improve signal quality; specifically, as the propagation conditions are different at the same instance of time, a combination of the received signals is always better or equal than the received signal. The study of macrodiversity has an added significance in our performance evaluation here since it has not yet been contemplated for packet data services.

We draw attention to the trade-off that arises in the downlink analysis presented in § 2.2: for those mobiles in MD the transmitted power can be smaller, since MRC is used (this being the macrodiversity gain). However, the base station generally expends more power to maintain additional links to those mobiles.



(a) (b)

Figure 4: Downlink with macrodiversity. Mean NRT (a) and RT (b) transmission rates vs. the fraction of mobiles in MD, for different received C/I ratios, Ω_s. Results are taken for $L_{NRT} = 0.2$ (a) and $L_{NRT} = 0$ (b).

Numerical results are displayed in Fig. 4, where both NRT and RT throughput are shown to deteriorate from the non-MD case ($\mu = 0, \Omega_s = 0$). Note that the throughput decreases in a sublinear fashion as μ increases. However, we may achieve better performance as the reception ratio Ω_s between the two base stations is improved. In Fig. 4(a), NRT transmission rate is reduced by more than half (53.6%, $\Omega_s = 0.5$) for fairly poor relevant reception conditions, while for comparable reception ($\Omega_s = 0.9$) the relevant decrease is 25.05%.

Analogous remarks can be made for the RT throughput (Fig. 4(b)). Observe that due to the imposed constraints, the transmission rate is always kept within $4.75 \leq R_{RT} \leq 12.2$. The mean rate is closer to the lower bound because of heavy RT call arrival rate. It is added that, were there no transmission constraints on RT traffic, the maximum transmission rate can reach up to $R_{RT}^{max} = 997.4$ Kbps with full bandwidth allocation.

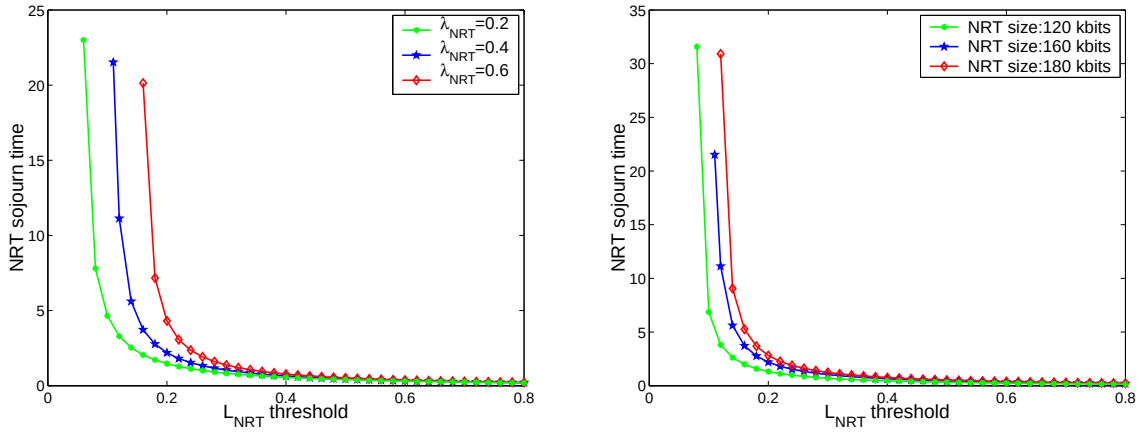
5.5 Varying traffic characteristics

Traffic parameters are primarily related to the arrival rate and file size for NRT calls, and the arrival rate and mean session time for RT calls. We vary each of these and show that they may influence to a greater or

lesser extent the overall performance of the system. The RT call behavior can be studied independently as an $M/M/c/c$ system, and the impact of traffic parameters is reduced to studying the load λ_{RT}/μ_{RT}. Hence we will be concerned with the NRT behavior, which may be affected by all parameters in the complex system. Note that because of the interaction of NRT and RT calls and their competition for system resources, the notion of system load is not straightforward. Hence we study parameters separately and refer to 'very high load' as those sets of values that drive the system in a region towards instability. Hereafter results refer also to the downlink (without macrodiversity), since it has been identified as the bottleneck of the system.

5.5.1 NRT traffic parameters

Results by modifying NRT traffic parameters are reported in Fig. 5. Both increasing the rate of arrivals and the mean NRT file size have a downgrading effect on the mean sojourn time. The instability region is clearly of utmost importance, since for larger reserved capacity sojourn times become extremely small and almost negligible. From the numerical results underlying Fig. 5(a), instability regions occur approximately at $L_{NRT} < 0.16$ for $\lambda_{NRT} = 0.6$, $L_{NRT} < 0.1$ for $\lambda_{NRT} = 0.4$ and $L_{NRT} < 0.06$ for $\lambda_{NRT} = 0.2$. Hence, as it anticipated, for smaller loads the system can be stable for smaller values of reserved capacity.



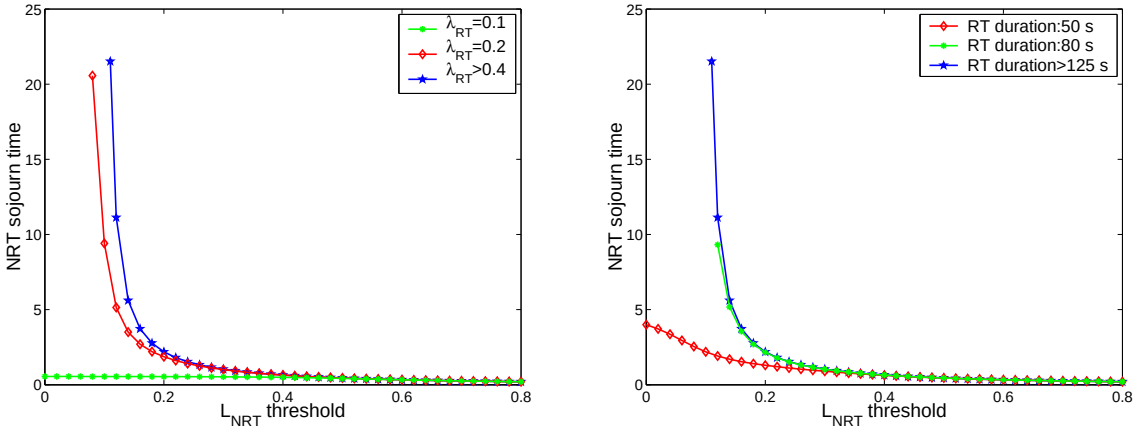
(a) (b)

Figure 5: Varying NRT traffic parameters (DL, non-MD). Mean NRT sojourn time vs. L_{NRT} reserved capacity, for differrent (a) arrival rates of NRT calls and (b) NRT file sizes.

5.5.2 RT traffic parameters

RT traffic indirectly influences NRT behavior. A primary observation from the numerical results here is that increasing parameters beyond a certain value does not have any effect on performance whatsoever. Such 'edge' behavior is seen in Fig. 6(a), for $\lambda_{RT} > 0.4$ and Fig. 6(b) for $\frac{1}{\mu_{RT}} > 125$ sec. We can straightforwardly conclude that for these load values, the system is in saturation, so that not only the minimum capacity (i.e. L_{NRT}) is attributed to NRT calls, but additionaly the number of RT sessions (as viewed by NRT arrivals) is almost constant so that NRT calls receive the exact same service.

Additionally, albeit indirect, the impact of RT traffic on NRT calls can be substantial for low load values. Reduced RT traffic frees more capacity and thus improves performance and stability of NRT. For example, in Fig. 6(a) the sojourn time reaches a very low value for $\lambda_{RT} = 0.1$; what's more, the system is then stable



(a) (b)

Figure 6: Varying RT traffic parameters (DL, non-MD). Mean NRT sojourn time vs. L_{NRT} reserved capacity, for differrent (a) arrival rates of RT calls and (b) RT session durations.

even for no L_{NRT} reserved capacity. The same behavior is shown in Fig. 6(b), where for very small RT duration the system exhibits good NRT performance and stability even for $L_{NRT} = 0$. It is trivially added that the same overall stability can be observed for sufficiently low NRT loads.

5.6 NRT call admission control

So far we have only considered bounds on the number of RT applications, leaving NRT traffic unconstrained to make use of the available bandwidth. Nevertheless, even though best-effort applications are considered to be elastic, we have seen that under a small reserved capacity and high loads, available rate calls can suffer severe performance degradation. Hence some form of call admission control (CAC) might be required to ensure some minimal quality of service in these cases.

It is more difficult to define a policy that immediately corresponds to a QoS criterion, since the notion of capacity is more implicit in CDMA and large variances in transmission rates can occur because of the dynamic resource allocation scheme[6]. Roughly, we can base an NRT admission control policy on a minimum allowed, or desirable rate for the transmission of these calls. For instance, the minimal capacity that corresponds to a given rate R_{NRT}^{min} for a single NRT call is (downlink, no macrodiversity):

$$\Delta_{NRT}^{min} = \frac{(a+f)R_{NRT}^{min}}{aR_{NRT}^{min} + \frac{wN_0}{E_s\Gamma}}. \tag{15}$$

Then, we have that the maximum number of allowed NRT calls when only the reserved L_{NRT} capacity is left over (worst case) is:

$$M_{NRT}^{max} = \lfloor \frac{L_{NRT}}{\Delta_{NRT}^{min}} \rfloor.$$

Table 1 represents indicative values of the minimum transmission rate and the corresponding maximum number of allowed NRT calls, given various values of the L_{NRT} threshold. The number of allowed NRT

[6]It is worth noting that, in contrast with FDMA or TDMA systems and provided that $M_{NRT}(t) > 0$, resource utilization is not affected by CAC in this scheme. It is the GoS that is affected.

calls for a given R_{NRT}^{min} rate grows with reserved capacity; equivalently, for the same M_{NRT}^{max} the minimum guaranteed NRT rate increases. Clearly, R_{NRT}^{min} values represent worst case bounds here, as more available capacity may be allocated to NRT traffic. Moreover, the mean transmission rate is generally much higher.

	R_{min}^{NRT}		
M_{max}^{NRT}	$L_{NRT} = 0.02$	$L_{NRT} = 0.2$	$L_{NRT} = 0.4$
2000	$6.45 \cdot 10^{-3}$	$6.45 \cdot 10^{-2}$	0.129
1000	$1.29 \cdot 10^{-2}$	0.129	0.258
200	$6.45 \cdot 10^{-2}$	0.646	1.29
100	0.129	1.29	2.58
50	0.258	2.59	5.19
25	0.517	5.19	10.42

Table 1: Minimum NRT transmission rate (Kbps) for a given maximum number of NRT calls, under different L_{NRT} reserved capacity (DL, non-MD).

The setting of an upper bound introduces call blocking for NRT traffic. The blocking probability will be the main parameter under examination here. Since we have assumed Poisson arrivals, the blocking probability of an incoming NRT call is

$$P_B = Pr\{M_{NRT} = (M_{NRT}^{max})\} = \sum_{i=0}^{M_{RT}^{max}} \pi(M_{NRT}^{max}, i).$$

Then the average sojourn time of an NRT session can be calculated using Little's law, considering the portion of NRT calls that are admitted into the system:

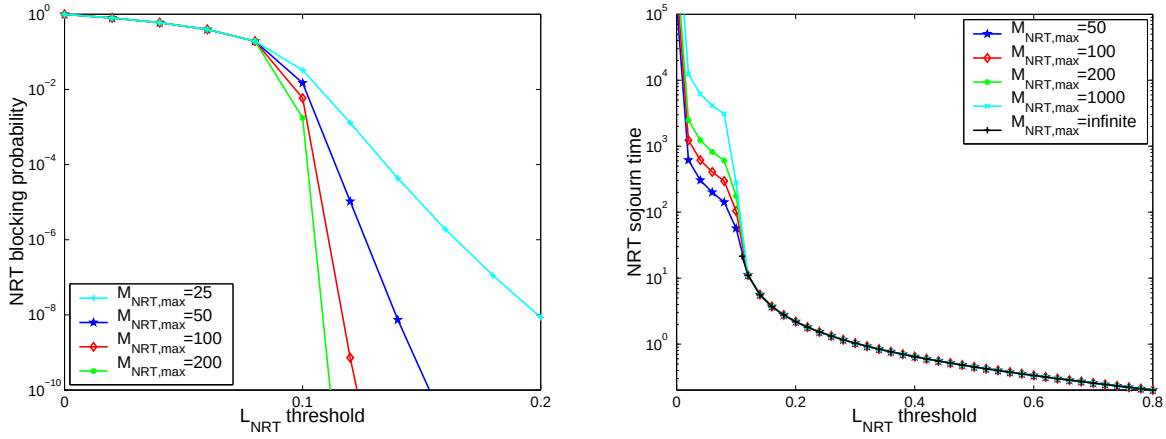
$$T_{NRT}^{soj} = \frac{E[M_{NRT}]}{\lambda_{NRT}(1 - P_B)}. \tag{16}$$

The direct impact of the number of allowed NRT calls is considered in the numerical evaluation of Fig. 7. Algorithm *Finite LDQBD* (Appendix A) is used to calculate the stationary distribution. As anticipated, raising the number of NRT calls decreases blocking (Fig. 7(a)). However, this effect must be largely mitigated due to the fact that NRT calls then spend more time in the system. For small values of the reserved capacity ($L_{NRT} \approx 0.1$), the blocking probability reaches values where the loss in performance becomes apparent. Further, for smaller reservations blocking is dominant even for fairly large M_{max}^{NRT} values. On the contrary, we may observe the drastic drop in blocking for small increases after $L_{NRT} > 0.1$, which once again points out the significance of the role of capacity reservation on CAC.

Fig. 7(b) also depicts the impact of the number of allowed NRT calls on the individual mean sojourn times. Increasing M_{NRT}^{max} logically increases the time spent in the system. Remark here that the sojourn time will assume extremely high values under congestion conditions, or equivalently in regions where the unconstrained system would be unstable (approx. $L_{NRT} < 0.1$ in the graph). Therefore, the gain obtained from less blocking reflects the loss of performance of those served, and this constitutes the trade-off we should consider in the design of the admission control scheme.

6 Summary and conclusions

We end by recapitulating the major conclusions drawn from this research. The performance of an integrated CDMA system with RT and NRT traffic is determined by the actual traffic load, E_b/N_0 requirements for



(a) (b)

Figure 7: NRT admission control scheme (DL, non-MD). NRT call blocking probability (a) and mean sojourn time (b) vs. L_{NRT} reserved capacity, for different allowed maximum number of NRT calls.

each class, as well as interference and the amount of available capacity. Besides that, the actual system behavior is mirrored through the call admission and GoS control scheme applied. Here, we have studied a system with adaptive-rate RT calls and elastic NRT traffic. The general CAC scheme allows NRT calls to benefit from low or intermittent RT traffic to attain an improved performance. Both for the uplink and downlink, it has been shown that bandwidth reservation can offer significant performance improvement to NRT calls, at the expense of increased blocking of RT sessions. However, the amount of reservation need not be very high; for the test cases considered, a reservation smaller than 20% of the total capacity vastly improves the NRT performance, while leaving RT behavior intact.

In case of overload conditions, the behavior of the system can severely degrade. High data rate methods such as HSDPA, which employ a complex scheduling of the different user transmissions each making use of the whole available bandwidth, can then reduce congestion symptoms and improve performance.

In addition, the use of macrodiversity techniques deteriorates transmission capacity in the downlink because of the requirement for a base station to maintain additional links to mobile units. This performance degradation is mitigated in the case of better reception conditions from the two base stations.

Finally, stricter admission control policies might be imperative to reduce the service time of NRT calls, especially under high load conditions. In this scope, we have demonstrated how the setting of an admission control policy on NRT traffic is a trade-off between the number of calls allowed and the GoS offered to those served.

Appendix

A LDQBD algorithms

Consider the transition probability matrix for a LDQBD process with a finite number of levels, K.

$$Q = \begin{bmatrix} B & A_0 & 0 & \cdots & \cdots \\ A_2^1 & A_1^1 & A_0 & 0 & \cdots \\ 0 & A_2^2 & A_1^2 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & A_2^K & A_1^K \end{bmatrix}$$

where again $B = A_1^j + A_2^j$ and $A_1^K[i,i] = -\lambda_{RT} - i \cdot \mu_{RT} - \nu(i,K)$. We use the following algorithm from [7] to calculate the steady state distribution. The algorithm is similar to later introduced stochastic complementation methods [18] and consists of the following steps:

Algorithm *Finite LDQBD* :

1) Compute the stochastic S_i matrices using the following recursion:

$$\begin{aligned} S_0 =& B, \\ S_n =& A_1^n + A_2^n(-S_{n-1}^{-1})A_0, \quad 1 \le n \le K. \end{aligned}$$

2) Find the stationary distribution of the S_K stochastic matrix by solving

$$\begin{aligned} \pi_K \cdot S_K &= 0, \\ \pi_K \cdot e &= 1. \end{aligned}$$

3) Recursively compute the remaining stationary distributions

$$\pi_n = \pi_{n+1} \cdot A_2^{n+1} \cdot (-S_n^{-1}), \quad \text{for } 0 \le n \le K.$$

4) Renormalize to obtain the steady-state distribution

$$\pi = \frac{\pi}{\pi \cdot e}.$$

In order to solve the infinite system, the objective is to find a value for the number of level K^* such that $\pi(k) \approx 0 \; \forall \; k > K^*$. Thus we may extend the previous algorithm as follows:

set $K^* = K_{init}$
while $\pi(K^*) \cdot e > \epsilon$
    $K^* = K^* + h$,
    run algorithm *Finite LDQBD*
end

The values of ϵ, h define the tolerance and step size, respectively and determine the accuracy and rate of convergence of the algorithm. An appropriate value of K_{init} can be readily available from runs in the finite case, which give an indice on how big the number of levels should be. Provided the system is stable, the algorithm will converge to the steady-state distribution.

B Ergodicity theorem

Theorem 1 *Consider a stochastic irreducible LDQBD process $X(t)$ whose submatrices $Q_0^{(k)}$, $Q_1^{(k)}$, $Q_2^{(k)}$ converge to level independent submatrices, Q'_0, Q'_1, Q'_2 of a homogeneous QBD process $X'(t)$ as the level number $k \to \infty$, i.e. $\lim_{k\to\infty} Q_s^{(k)} = Q'_s$, $\{s = 0, 1, 2\}$. The number of phases at each level is finite, denoted by m. The LDQBD is, in matrix-block form:*

$$Q = \begin{pmatrix} Q_1^{(0)} & Q_0^{(0)} & 0 & 0 & \cdots \\ Q_2^{(1)} & Q_1^{(1)} & Q_0^{(1)} & 0 & \cdots \\ 0 & Q_2^{(2)} & Q_1^{(2)} & Q_0^{(2)} & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots \end{pmatrix}$$

The matrices $Q_0^{(k)}$, $Q_2^{(k)}$ determine transitions up and down one level, respectively, and it holds that[7] *$Q_0^{(0)} < Q_0^{(1)} < \cdots < Q'_0$, and $Q_2^{(1)} > Q_2^{(2)} > \cdots > Q'_2$, for every defined $k \in \mathbb{Z}^+$. Further, we assume that transitions are skip-free in each direction*[8]*, and that transition rates in matrices $Q_1^{(k)}$, Q'_1 are identical within the same level. Then, if the homogeneous QBD process $X'(t)$ is ergodic, the non-homogeneous LDQBD process $X(t)$ also is. Conversely, if process $X'(t)$ is not ergodic with a positive expected drift, i.e. $d = \pi Q'_0 e - \pi Q'_2 e > 0$, process $X(t)$ is also not ergodic. For these cases, both processes satisfy the same ergodicity condition.*

Proof Denote by $X(t)$, $X'(t)$ the stochastic processes determined by Q, and its counterpart Q', respectively. It is reminded that the ergodicity condition for the homogeneous QBD is [14]

$$\pi Q'_2 e > \pi Q'_0 e. \tag{17}$$

In the first part of the proof, we proceed to show that $X(t) \leq_{st} X'(t)$, i.e. that $X'(t)$ stochastically dominates $X(t)$. For this we need the following Lemma, initially reported in [17] and later explored in [4],[16].

Lemma 1 *Define $(E, \leq)$ to be a countable partially ordered set, and a set $F \subseteq E$ which is $\leq$-increasing. Let $X(t)$, $X'(t)$ be Markovian skip-free processes on E with transition intensities $q(i, j)$, $q'(i, j)$, respectively, s.t. $\sum_{j\neq i} q_{ij} < \infty$ and $\sum_{j\neq i} q'_{ij} < \infty$ for every $i \in E$. Then $X'(t)$ stochastically dominates $X(t)$ if and only if the following conditions hold, for all $x \leq y$ in E and all increasing sets, F:*

(i) if $x, y \in F$,

$$\sum_{z\notin F} q(x, z) \geq \sum_{z\notin F} q'(y, z)$$

and

(ii) if $x, y \notin F$,

$$\sum_{z\in F} q(x, z) \leq \sum_{z\in F} q'(y, z).$$

□

[7] Notice that, in our paradigm, we have level-independent matrices Q_0. However, it is trivial to modify the proof in that case.

[8] A *skip-free* process is one that cannot skip adjacent states. This refers to departures and arrivals of NRT and RT calls in our system.

It is obvious that the first condition refers to the case where the sum of transition rates towards 'smaller' states is always less or equal for the dominating process, while the second one states that the sum of transition rates towards 'larger' states is always greater or equal for the dominating process.

In order to prove the theorem we need to show that conditions (i), (ii) hold for the stochastic processes given by the matrices $Q_0^{(k)}, Q_1^{(k)}, Q_2^{(k)}$ and Q'_0, Q'_1, Q'_2. The proof follows similar steps with those of Bright and Taylor [5].

First we define the partial order relation $(<)$ by $(i,j) < (k,l)$ if:

$$((i < k) \wedge (j \leq l)) \vee ((i \leq k) \wedge (j < l))$$

Since transitions are *skip-free* in each direction and matrices Q, Q' have the exact same structure, it suffices to prove stochastic dominance for this order.

We start by examining condition (i) of the Lemma and consider increasing sets F on the state space E, according to the partial order. We only consider non-trivial states[9], i.e. $x, y \in F \; s.t. \sum_{z \notin F} q(x,z) \neq 0 \wedge \sum_{z \notin F} q'(y,z) \neq 0$. With great generality[10], there exists a set of *boundary* states $B = \{x : \sum_{z \notin F} q(x,z) \neq 0\}$, i.e. states for which there exists at least one transition to the complementary set, F^c. It follows that non-trivial states are boundary states.

We examine different cases for condition (i) to hold. Consider any boundary element $x = (i,k) \in F$; in the general case, there may exist transitions to:

(a) $(i-1,k)$ and $(i,k-1) \notin F$, or
(b) $(i-1,k) \notin F$, or
(c) $(i,k-1) \notin F$.

We treat the cases $x = y \in F$ and $x < y \in F$ separately.

$x = y \in F$. Assume first that (a) holds. We have for $X(t)$

$$\sum_{z \notin F} q(x,z) = (Q_1^{(i)})_{k,k-1} + (Q_2^{(i)})_{k,k}$$

and for $X'(t)$

$$\sum_{z \notin F} q'(x,z) = (Q'_1)_{k,k-1} + (Q'_2)_{k,k}.$$

Transition rates within the same level are identical, so that $(Q'_1)_{k,k-1} = (Q_1^{(i)})_{k,k-1}$. Also by definition $(Q'_2)_{k,k} < (Q_2^{(i)})_{k,k}$. Therefore, condition (i) is satisfied.

If case (b) holds, i.e. there is only one transition to $(i-1,k) \notin F$, we have

$$\sum_{z \notin F} q(x,z) = (Q_2^{(i)})_{k,k}$$
$$\text{and} \quad \sum_{z \notin F} q'(x,z) = (Q'_2)_{k,k}$$

where $(Q'_2)_{k,k} < (Q_2^{(i)})_{k,k}$ and (i) is satisfied.

[9] Summations equal zero in any other case. Also, since $x \leq y \in F$ and transitions are skip-free in each direction we cannot have the case $\sum_{z \notin F} q(x,z) = 0 \wedge \sum_{z \notin F} q'(y,z) \neq 0$.

[10] if $F = E$, then $B = \varnothing$ and $\sum_{z \notin F} q(x,z) = \sum_{z \notin F} q'(y,z) = 0$, i.e. again condition (i) is satisfi ed.

Finally, in (c) if there exists only a transition to $(i, k-1) \notin F$, we get

$$\sum_{z \notin F} q(x,z) = (Q_1^{(i)})_{k,k-1} = (Q_1')_{k,k-1} = \sum_{z \notin F} q'(x,z)$$

so that (i) again holds.

$x < y \in F$. Consider again a boundary state $x = (i,k)$ where cases (a), (b), or (c) may hold. Then the only non-trivial $y > x$ can be either $y = (i+1,k)$ or $(y = i, k+1)$. If $y = (i+1,k)$ either case (a) or (c) will exist for x and we have

$$\begin{aligned} & \sum_{z \notin F} q'(y,z) = (Q_1')_{k,k-1} \\ \text{and} \quad & \sum_{z \notin F} q(x,z) = (Q_1^{(i)})_{k,k-1} + (Q_2^{(i)})_{k,k}, \quad \text{in case (a)} \\ \text{or} \quad & \sum_{z \notin F} q(x,z) = (Q_1^{(i)})_{k,k-1}, \quad \text{in case (c).} \end{aligned}$$

In any case, it holds that $\sum_{z \notin F} q(x,z) \geq \sum_{z \notin F} q'(y,z)$ so that condition (i) is fulfilled.

Similarly, for $y = (i, k+1)$ we have

$$\sum_{z \notin F} q'(y,z) = (Q_2')_{k,k+1}$$

and either (a) or (b) will hold for x. We then get

$$\begin{aligned} & \sum_{z \notin F} q(x,z) = (Q_1^{(i)})_{k,k-1} + (Q_2^{(i)})_{k,k}, \quad \text{in case (a)} \\ \text{and} \quad & \sum_{z \notin F} q(x,z) = (Q_2^{(i)})_{k,k}, \quad \text{in case (b).} \end{aligned}$$

where again we always have $\sum_{z \notin F} q(x,z) \geq \sum_{z \notin F} q'(y,z)$.

The proof that condition (ii) of the Lemma is fulfilled is derived in a similar manner, considering transitions to larger states based on the matrices $Q_0^{(k)}$, $Q_1^{(k)}$ and the set of boundary elements $B = \{y : \sum_{z \in F} q'(y,z) \neq 0\}$ for $x \leq y \notin F$.

Thus we arrive at $X(t) \leq_{st} X'(t)$. We shall use this to establish the stability of the non-homogeneous process $X(t)$. To this end, we consider the mean recurrence time to the *smallest*[11] state $\ell = (0,0)$, defined by

$$\sigma_\ell = inf\{t > 0 : X(t) = \ell | t > \rho_\ell\},$$

where ρ_ℓ is the first exit time from ℓ. Then for the stochastic processes X_t,X_t' it must hold that $\sigma_\ell \leq_{st} \sigma_\ell'$.

We prove this by contradiction; assume that $\sigma_\ell >_{st} \sigma_\ell'$. Then it must hold that $E[\sigma_\ell] > E[\sigma_\ell']$, from which $\Pr[X = \ell] < \Pr[X' = \ell]$. Since ℓ is the smallest state, we conclude that $\Pr[X > \ell] > \Pr[X' > \ell]$. But this contravenes the stochastic order relation.

Therefore, we deduce that

$$\sigma_\ell \leq_{st} \sigma_\ell'$$

[11] Note that due to the partial order here, the 'smallest' state is defi ned as $\ell = \{x \in E : \nexists\, x' \neq x \text{ with } x' > x\}$.

from which

$$E[\sigma_\ell] \leq E[\sigma'_\ell].$$

Hence, if the homogeneous process $X'(t)$ is ergodic, the mean recurrence time of process $X(t)$ to state ℓ is finite and thus ℓ is positive recurrent. Since $X(t)$ is irreducible, it follows that all other states are positive recurrent and the process is ergodic.

We also proceed to show that if $X'(t)$ is not ergodic with a positive[12] expected *drift*, i.e. $\pi Q'_2 e < \pi Q'_0 e$, then $X(t)$ is also not ergodic. We may then say that condition (17) is close to being necessary for the ergodicity of the LDQBD process.

Since elements of matrices Q'_2,Q'_0 are real, there exist appropriate values such that the resulting modified QBD process, $X''(t)$ has $Q''_2 > Q'_2$ and $Q''_0 < Q'_0$, and it still holds that[13] $\pi Q''_2 e \leq \pi Q''_0 e$, i.e. the process is not ergodic.

Next we define the *L-embedded* chain of the LDQBD process, consisting of all levels $i \geq L$. This is the truncated LDQBD process obtained by rerouting transitions from level L to $L-1$ back to the same level, i.e. $Q_2^{(L)} = 0$ and $Q_1^{(L)} + Q_0^{(L)} = 0$. It is straightforward to show that if the L-embedded process $X^L(t)$ is not ergodic, then the original LDQBD process $X(t)$ is also not ergodic. Consider a state x in the state space of the L-embedded process, S. Since transitions of the two processes are identical beyond level L, we have

$$E[_{S^c}T_{x\to x}] = \infty$$

for the mean recurrence time of process $X(t)$ to state x, avoiding states in the complementary set S^c. Since the L-embedded process is also irreducible[14], there exists a probability $0 <_{S^c}P_{x\to x} < 1$ that $X(t)$ does not pass through S^c during its first recurrence[15] to x. Hence the following inequality holds:

$$E[T_{x\to x}] \geq E[_{S^c}T_{x\to x}] \cdot {}_{S^c}P_{x\to x},$$

from which we conclude that $E[T_{x\to x}] = \infty$ and thus LDQBD is not ergodic.

Consider now the sequence of L-embedded submatrices, $\{L = 0, 1, \dots\}$. As L increases, the matrices $Q_2^{(L)}$ ($Q_0^{(L)}$) become smaller (larger). Therefore, there exists a level L after which $Q_2^{(L)} < Q''_2$, $Q_0^{(L)} > Q''_0$. Then we can follow a similar procedure as in the first part of the proof to show that for processes $X''(t)$, $X^L(t)$ defined on the same state space S, it holds

$$X''_t \leq_{st} X^L_t.$$

Then, we conclude for the mean recurrence time to the smallest state $\ell \in E_L$, that

$$\begin{aligned} &\sigma^L_\ell \geq \sigma''_\ell \\ \Rightarrow\ &E[\sigma^L_\ell] \geq E[\sigma''_\ell]. \end{aligned}$$

Since $X''(t)$ is not ergodic, $E[\sigma''_\ell] = \infty$ and thus $X^L(t)$ is also not ergodic. Then from the preceding argument we can conclude that the LDQBD process is not ergodic, which completes the proof.

■

[12] We do not treat the case $\pi Q'_2 e = \pi Q'_0 e$ here; for this, the QBD process is also not ergodic (since (17) is a necessary and suffi cient condition, cf. [14]), but we cannot examine the behavior of LDQBD by the analogous argument used in the reverse part of the proof.

[13] The stationary probability vector π of transitions within the same level is invariant to changes in Q_2,Q_0.

[14] This follows immediately from the structure of transition probabilities.

[15] Since the whole process is irreducible, there exists a positive probability to return to x in fi nite time. Moreover, the probability ${}_{S^c}P_{x\to x}$ cannot be equal to 1.

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 Communications Magazine*, 70–77, July 2000.

[2] T. Bonald and A. Proutière, "Wireless downlink data channels: User performance and cell dimensioning", *Proc. ACM Mobicom*, San Diego, USA, September 2003.

[3] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks", *Proc. IEEE Infocom*, San Fransisco, USA, March/April 2003.

[4] A. Brandt, G. Last, "On the pathwise comparison of jump processes driven by stochastic intensities", *Mathematische Nachrichten*, 167, 21–42.

[5] L. Bright, P. Taylor, "Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes", *Commun. Statist.-Stochastic Models*, 11(3), 497–525, 1995.

[6] Y. Fang, Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks", *IEEE Transactions on Vehicular Technology*, 51(2), 371–382, March 2002.

[7] D.P. Gaver, P.A. Jacobs, G. Latouche, "Finite birth-and-death models in randomly changing environments", *Advances in Applied Probability*, 16, 715–731, 1984.

[8] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, A. Weaver, Jr., C.E. Wheatley , "On the capacity of a cellular CDMA system", *IEEE Transactions on Vehicular Technology*, 40(2), 303–312, May 1991.

[9] N. Hegde, E. Altman, "Capacity of multiservice WCDMA Networks with variable GoS", *Proc. of IEEE WCNC*, New Orleans, Louisiana, USA, March, 2003.

[10] K. Hiltunen, R. De Brarnardi, "WCDMA downlink capacity estimation", *Proc. IEEE VTC-Spring*, 992–996, Tokyo, Japan, 2000.

[11] H. Holma and A. Toskala, Eds.,*WCDMA for UMTS: Radio access for third generation mobile communications*, John Wiley & Sons, 3rd Edition, 2004.

[12] J.M. Kelif, E. Altman, "Admission and Gos control in multiservice WCDMA system", *Proc. ECUMN '04*, Porto, Portugal, October 2004.

[13] S.-L. Kim, Z. Rosberg, J. Zander, "Combined power control and transmission rate selection in cellular networks", *Proc. IEEE VTC-Fall*, 1653–1657, Amsterdam, The Netherlands, 1999.

[14] G. Latouche, V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, ASA-SIAM, 1999.

[15] C.W. Leong, W. Zhuang, "Call admission control for voice and data traffic in wireless communications", *Computer Communications*, 25(10), 972–979, 2002.

[16] J.F. López, S. Martínez, G. Sanz, "Stochastic domination and Markovian couplings", *Advances in Applied Probability*, 23, 1064–1076, 2000.

[17] W.A. Massey, "Stochastic orderings for Markov processes on partially ordered spaces", *Mathematics of Operations Research*, 12(2), 350–367, 1987.

[18] C.D. Meyer, "Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems", *SIAM Review*, 31(2), 240–272, 1989.

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

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

[21] A.M. Viterbi, A.J. Viterbi, "Erlang capacity of a power-controlled CDMA system", *IEEE J. Selected Areas in Communications*, 11(6), 892–900, August 1993.

[22] A.J. Viterbi, A.M. Viterbi and E. Zehavi, "Other-cell interference in cellular power-controlled CDMA", *IEEE Transactions on Communications*, 42(2/3/4), 1501-1504, Feb./March./April 1994.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifi que
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399