

A Non-homogeneous QBD Approach for the Admission and GoS Control in a Multiservice WCDMA System*

Ioannis Koukoutsidis¹, Eitan Altman¹, and Jean Marc Kelif²

¹ INRIA, 2004 Route des Lucioles, BP 93,
06902 Sophia Antipolis Cedex, France
{gkoukout, altman}@sophia.inria.fr
² France Telecom R&D, Rue du Général Leclerc,
92794 Issy-les-Moulineaux Cedex 9, France
JeanMarc.Kelif@francetelecom.com

Abstract. We consider a WCDMA system with real time (RT) calls that have dedicated resources, and data non-real-time (NRT) calls that share system capacity. We apply reservation of some resources for the NRT traffic and assume that this traffic is further assigned the resources unused by RT calls. 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 uplink and downlink, and derive performance evaluation results regarding user-perceived QoS parameters, namely the blocking rates of RT calls and sojourn time of NRT calls. On what concerns the bandwidth-sharing policy of NRT traffic, we compare WCDMA behavior in the presence of a high data rate scheme. Finally, we extend our results to cover NRT admission control schemes and examine blocking behavior and transfer times of NRT traffic.

1 Introduction

In this paper, we are interested in analyzing resource sharing between RT (real time) and NRT (non-real time) traffic 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, based on a measure of capacity. In CDMA, capacity is rather a complex combination of cell parameters and channel conditions, being mostly interference-limited [4],[14]. However, existing models ([5],[8]) allow us to obtain the resources required by transmissions of a given class with a given GoS, both in the uplink as well as the downlink of a CDMA system.

RT traffic (conversational, streaming) has stringent QoS requirements with regards to transmission rate and/or duration. On the other hand, NRT traffic (transfer of files, web browsing, etc.) has no guaranteed bit rate and is apt for a processor-sharing setting. Based on these principles, we design the admission and rate control scheme. User-centric QoS parameters of interest are primarily the blocking probabilities for RT calls

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

and expected sojourn times for NRT calls. We use the inflected term ‘call’ to refer also to the sending of NRT data, including connectionless services. Further, the term ‘GoS’ refers specifically to the rate of transmission. We allow downgrading of this rate for RT calls during congestion epochs (i.e. having in mind adaptive real-time compression algorithms, or tolerable loss of quality). Overall, the proposed control policy combines admission control together with GoS control for both RT and NRT traffic. Fairness in transmission rate between users of the same service class is considered as a basic building block of the access mechanism.

This work is a follow-up of [5],[8] in which NRT traffic was scheduled using a time-sharing approach, as is the case in the High Speed Downlink Packet Access (HSDPA) system [13],[7]. This allowed to derive a tractable mathematical model based on a homogeneous QBD (Quasi Birth-Death) process [12],[11]. In this paper, we consider the standard case with simultaneous transmissions in a cell. The system analyzed then cannot be evaluated anymore with a homogeneous QBD and we present a more involved analysis based on a non-homogeneous QBD. A numerical investigation is conducted and key performance measures are computed.

2 Background: Computing the Transmission Rates

The analysis is based on radio models for the downlink (DL) and uplink (UL) introduced in [8],[5]. 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

$$(C/I)_k = \frac{E_s}{N_0} \frac{R_s}{W} \Gamma, \quad (1)$$

where E_s is the energy per transmitted bit of type s , N_0 is the background noise density, W is the spread bandwidth, R_s is the transmission rate of the type s call, and Γ is a constant related to shadow fading and imperfect power control (cf. [5]).

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 transmits at a total power $P_{tot,l}$ given by $P_{tot,l} = \sum_{j=1}^M P_{j,l} + P_{CCH}$, where P_{CCH} corresponds to the power transmitted for the orthogonal common control channels (CCH). Note that this last term is not power controlled and is assumed not to depend on l . Due to multipath propagation, a fraction α_k of the received own cell power is experienced as intracell interference (non-orthogonality factor). 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), $I_{k,intra} = \alpha_k \cdot (P_{CCH} + \sum_{j \neq k} P_{j,l})/g_{k,l}$ 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

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

where $\beta_k = \frac{(C/I)_k}{1 + \alpha_k(C/I)_k}$. We then consider two service classes, that will correspond to RT and NRT traffic. Let $(C/I)_s$ be the target SIR ratio for mobiles of service class s with a corresponding value of β_s . Let there be in a given cell M_s mobiles of class s . Using an average approximation¹, we substitute $F_{k,l}$, $g_{k,l}$, α_k by their sample averages over all $k = 1, \dots, M$. We denote these as F , G , α . We consider these parameters to be the same for all service groups. Then (2) gives the following value for $P_{tot,l}$ (we omit the index l):

$$P_{tot} = \frac{P_{CCH} + NG \sum_s \beta_s M_s}{1 - (\alpha + F) \sum_s \beta_s M_s}. \tag{3}$$

Further assuming that the power for the common channels is a fraction of the total power, $P_{CCH} = \psi P_{tot}$ and defining the downlink loading as $Y_{DL} = \sum_s \beta_s M_s$, this gives

$$P_{tot} = \frac{NG \sum_s \beta_s M_s}{Z_2}, \quad \text{where } Z_2 = (1 - \psi) - (\alpha + F)Y_{DL}. \tag{4}$$

Thus the maximum base station output power determines the maximum loading supported by the system. 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 capacity as $\Theta_\epsilon = 1 - \psi - \epsilon$, and the capacity required by a connection to be $\Delta(s) := (\alpha + F)\beta_s$. Combining this with (1) and substituting the expression for β_s we get the throughput of a connection s , that "uses" a capacity $\Delta(s)$.

$$R_s = \frac{\Delta(s)}{\alpha + F - \alpha\Delta(s)} \times \frac{N_0 W}{E_s \Gamma}. \tag{5}$$

A similar analysis can be followed to derive an expression in the case where macrodiversity is implemented in the downlink (cf. [10]).

¹ This is a standard approximation for downlink models, see [6, 7]; further, as was performed in [6], the accuracy of the single parameters can be improved by curve fitting, based on actual measurements for the total base station output power.

2.2 Uplink

We briefly recall the capacity notions from the case of the uplink from [5]. Define for $s = 1, 2$,

$$\tilde{\Delta}_s = \frac{E_s R_s}{N_0 W} \Gamma, \text{ and } \Delta'(s) = \frac{\tilde{\Delta}(s)}{1 + \tilde{\Delta}(s)}. \quad (6)$$

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 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 intracell interference, and M_s is the number of mobiles of type s in the cell). Here in order to maintain an equal rate, the smallest maximum received power amongst all mobiles in the cell determines the maximum uplink loading. Again, to avoid that Z_2 becomes too close to zero one imposes the constraint $Z_2 \geq \epsilon$ 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 (6) we get

$$R_s = \frac{\Delta(s)}{1 + f - \Delta(s)} \times \frac{N_0 W}{E_s \Gamma}. \quad (7)$$

3 Admission and Rate Control

We consider 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 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 $[R^{min}, R^{max}]$. Note that $\Delta(RT)$ is increasing with the transmission rate. Hence the achievable capacity set per RT mobile has the form $[\Delta^{min}, \Delta^{max}]$. The maximum number of RT calls that can be accepted is $M_{RT}^{max} = \lfloor L_{RT}/\Delta^{min} \rfloor$. We assign full rate R^{max} (and thus the maximum capacity Δ^{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 L_{RT}/M_{RT} and the rate is reduced accordingly.

We next describe the rate control scheme for NRT calls. 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 [8],[5], the capacity $C(M_{RT})$ unused by the RT traffic (which changes dynamically as a function of the number of RT connections present) was fully assigned to a single NRT mobile, this being time-multiplexed rapidly so that the throughput is

shared equally between the present NRT mobiles. This modeling is consistent with a fair implementation of a high data rate scheme. Specifically, schemes such as HDR [1], corresponding to the CDMA 1xEV-DO standard, and its 3GPP counterpart HSDPA [13] have been proposed for the downlink in order to achieve higher transmission 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. Here we consider the standard case where transmissions are simultaneous and available capacity is split equally between the NRT calls, in a fair rate sharing approach. Then according to the previous analysis and assuming that channel conditions do not change substantially, the total transmission rate R_{NRT}^{tot} 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

$$DL : R_{NRT}^{tot}(M_{NRT}, M_{RT}) = \frac{M_{NRT}C(M_{RT})}{M_{NRT}(\alpha + F) - \alpha C(M_{RT})} \times \frac{N_0 W}{E_s \Gamma},$$

$$UL : R_{NRT}^{tot}(M_{NRT}, M_{RT}) = \frac{M_{NRT}C(M_{RT})}{M_{NRT}(1 + f) - C(M_{RT})} \times \frac{N_0 W}{E_s \Gamma}.$$

The expression for the downlink with macrodiversity is similarly derived, albeit being 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} . RT call durations and NRT file sizes are all mutually independent. Note that since their mean duration is fixed, the evolution of RT calls is not affected by the process of NRT calls and can be studied independently as an Erlang loss system. However, the departure rate of NRT calls depends on the current number of RT and NRT calls:

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

The number of active sessions in the downlink and uplink models 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 $\pi Q = 0$, 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 π as $[\pi(0), \pi(1), \dots]$ 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, $\pi(i) = [\pi(i, 0), \pi(i, 1), \dots, \pi(i, M_{RT}^{\max})]$, for $i \geq 0$. In (i, j) , j is referred to as the *phase* of the state. The generator Q is given by

$$Q = \begin{bmatrix} B & A_0 & 0 & 0 & \dots \\ A_2^1 & A_1^1 & A_0 & 0 & \dots \\ 0 & A_2^2 & A_1^2 & A_0 & \dots \\ 0 & 0 & \ddots & \ddots & \ddots \end{bmatrix} \quad (8)$$

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

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

Of course, A_1^i is properly modified on the boundaries $j = 0$, $j = M_{RT}^{\max}$. We also have $B = A_1^i + A_2^i$. Due to the special structure of the matrix, this is independent of i .

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 e.g [11]). However, the number of states is often so large that the solution becomes untractable. For this reason, algorithmic approaches are usually sought. Here we use an extension of a method introduced in [3] 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 that reflect user-perceived QoS 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 model of the high data rate scheme in WCDMA. Continuing, we explore the extent to which intercell interference can deteriorate system behavior. Finally, numerical results are extended to the case of an NRT call admission control scheme.

5.1 Setting

Here we address the values of parameters used in the numerical evaluation. Common CDMA performance evaluation parameters (such as chip rate, energy-to-noise require-

ments, interference factors, etc.) are derived from equipment capabilities and field tests. 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 : Uplink 4.2 dB, Downlink 7.0 dB (12.2 kbps voice)
- E_{NRT}/N_0 : Uplink 2.2 dB (64 kbps data), Downlink 5.0 dB (144 kbps data)
- Average RT call duration: $1/\mu_{RT} = 125$ sec
- Mean NRT session size: $1/\mu_{NRT} = 160$ kbits
- Arrival rate of calls: $\lambda_{RT} = \lambda_{NRT} = 0.4$
- Interference factor: Uplink $f = 0.73$, Downlink $F = 0.55$
- Non-orthogonality factor: $\alpha = 0.64$
- Fraction of power for CCH channels: $\psi = 0.2$

The traffic characteristics for RT and NRT calls are chosen to correspond to heavy traffic conditions, whereupon performance evaluation must focus. We assume a chip pulse is rectangular, so that the chip rate equals the spread bandwidth. The parameter F , which accounts for shadow fading in the calculation of the system capacity, has been incorporated in the E_b/N_0 targets. These are set here according to §12.5 of [7] (3GPP performance requirements for a slow moving user, Tables 12.26, 12.27). 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 [9].

5.2 Uplink and Downlink Performance

Here we study the behavior in the uplink and the downlink of the WCDMA system. 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(a), for different values of the L_{NRT} threshold. As anticipated, the probability of rejection increases as more capacity is reserved for NRT calls. In the case of NRT traffic, performance evaluation results are portrayed in Fig. 1(b). 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. The behavior of NRT traffic reflects the general admission and rate control policy 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. In that sense, Fig. 1(b) shows the improvement in NRT traffic transfer time as the capacity reserved for it increases.

These results also permit to see the trade-off relationship between the performance of RT and NRT transmissions. However, we remark that although NRT improvement through capacity reservation comes at the expense of RT traffic, a region of L_{NRT} values can be selected where performance is satisfactory for both service classes. For

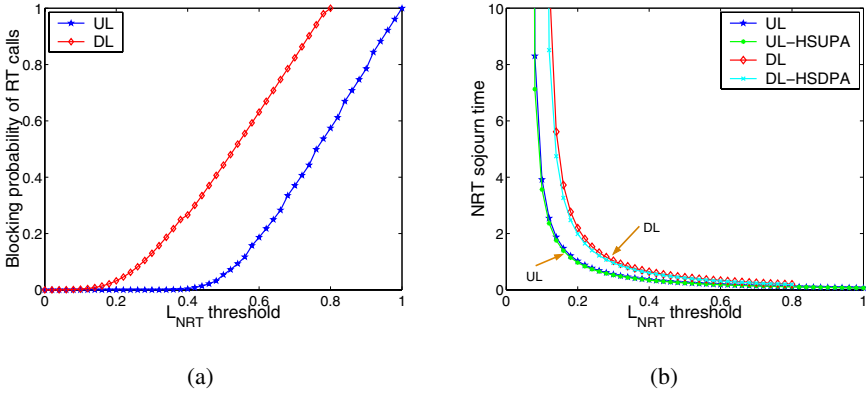


Fig. 1. RT call blocking probability (a) and mean NRT sojourn time (b) vs. L_{NRT} reservation, in the UL and DL cases. Comparison with high data rate services in (b)

example, in the results here a good operating region for both service classes can be chosen near $L_{NRT} = 0.2$ in the DL and between $0.2 \leq L_{NRT} \leq 0.4$ in the UL.

We also make the following observations regarding the determination of the system bottleneck. Although the downlink enjoys less interference, this can be largely eclipsed by the increased E_b/N_0 ratios that require more capacity for a given transmission rate, and the expended power for CCH channels. This is substantiated in the results here and is evidence that, with user mobility and intracell interference due to non-orthogonal channels, the downlink may be the bottleneck even with symmetric traffic transferred on both sides. On the other hand, further numerical evaluations can show that the UL is usually bottlenecked for static users and a smaller non-orthogonality factor. Of course one should keep in mind that in reality, with time-varying channel and traffic conditions, both sides may be the bottleneck at one time or another.

An ergodicity condition is essential for stability in the theoretical case of an unbounded number of NRT calls. As shown in Fig. 1(b), below a certain value of the L_{NRT} threshold (approximately² $L_{NRT} \approx 0.1$ 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 [11]:

$$\mu_{NRT} \cdot \mathbb{E}R_{NRT}^{tot} > \lambda_{NRT}. \tag{9}$$

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 $M_{NRT} \rightarrow \infty$, the total transmission rate reaches a limit in both the UL and DL cases. Therefore, the non-homogeneous LDQBD process 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. It can be formally shown that sta-

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

bility conditions are the same for both processes, i.e. it suffices to check the ergodicity of the limiting homogeneous process. The general theorem is deferred to the Appendix; due to limited space, only a sketch of the proof is presented. Interested readers may refer to the complete version in [10].

For NRT calls, Fig. 1(b) also presents a comparison of the standard WCDMA behavior with that of the scheme similar to HSDPA, mentioned in § 3. We also consider the corresponding scheme in the uplink —analogously named HSUPA (which has recently been added in 3GPP Release 6 [7]). An 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. 1(b) shows that the outperformance of the time-scheduling approach is non-negligible for small NRT reserved capacity. In the numerical results obtained, the difference reached up to 80 sec in the uplink, for $L_{NRT} \approx 0.06$.

5.3 Impact of the Interference Expansion Factor

As CDMA 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 received intercell-to-intracell power, F in the downlink. This is the analog of the ratio of intercell-to-intracell interference in the uplink. A more perceptive term for such ratios is the *interference expansion factor*. Increasing values of F can then be seen as increased intercell interference.

Numerical results are portrayed in Fig. 2. The value of the interference expansion factor depends on the traffic distribution of interfering cells and may well assume values greater than unity [15]; however we take selected values until $F = 1$ for our test cases here. We may deduce that intercell interference has a significant impact on performance. Concerning the blocking probability of RT calls in Fig. 2(a), for smaller values of F an initially good performance is observed; for the smallest value $F = 0.1$, the loss

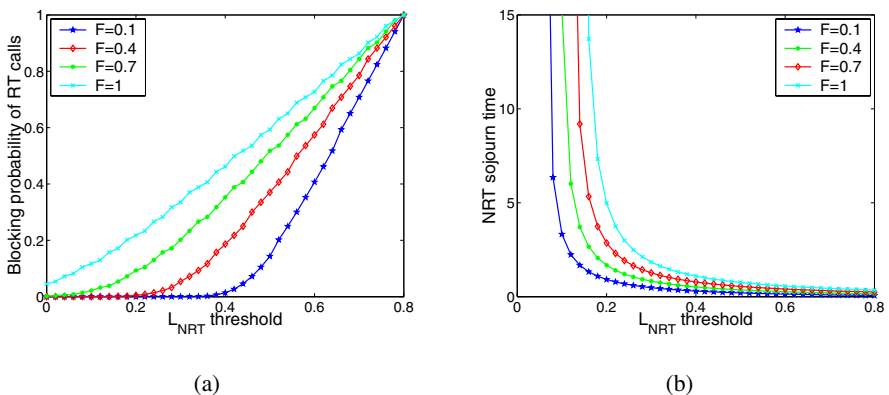


Fig. 2. RT call blocking probability (a) and mean NRT sojourn time (b) vs. L_{NRT} reservation, for different values of the interference ratio, F , in the downlink

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. 2(b) that the mean transfer time is greater as interference increases, as well as that the instability region is larger.

The deterioration of system behavior in all cases is due to the fact that more power, and hence more capacity is required by users to overcome interference. This means less resources available –even for the lowest quality RT calls– and smaller transfer rates for NRT sessions. Naturally, the same observations carry over to the uplink. Further, an analogous situation –due to power control– occurs in the uplink and downlink in case of increased intracell interference, and we expect similar observations to carry over to this case.

5.4 NRT Call Admission Control

Even though best-effort applications are considered to be elastic, we have seen that under a small reserved capacity and high loads, NRT rate calls can suffer severe performance degradation, in terms of very large transfer times. This could lead to unwanted renegeing, as a result of user impatience. Hence setting an upper bound on the number of admitted NRT sessions is required to ensure some minimal QoS in these cases.

The setting of an upper bound introduces call blocking for NRT traffic. Since we have assumed Poisson arrivals, the blocking probability of an incoming NRT call is

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

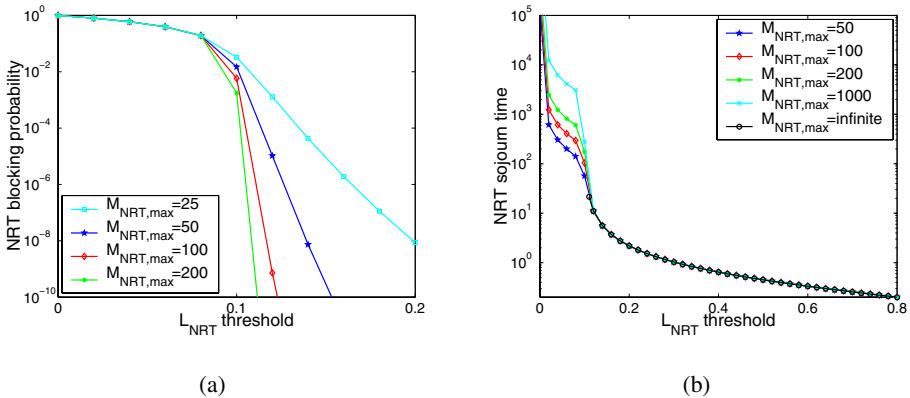


Fig. 3. NRT admission control scheme. NRT call blocking probability (a) and mean sojourn time (b) vs. L_{NRT} reservation, for different allowed maximum number of NRT calls (downlink)

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:

$$T_{NRT}^{soj} = \frac{E[M_{NRT}]}{\lambda_{NRT}(1 - P_B)}. \quad (10)$$

The impact of the number of allowed NRT calls is considered in the numerical evaluations of Fig. 3. Algorithm *Finite LDQBD* (Appendix A) is used to calculate the stationary distribution.

We observe in Fig. 3(b) that restricting access for data transmissions on the CDMA link can improve performance in critical congestion regions where resources for this traffic are limited. For example, for a resource reservation $L_{NRT} \approx 0.1$ and $M_{NRT}^{max} = 100$ we have $T_{NRT}^{soj} = 104$ s, which decreases by more than 45% if we restrict to $M_{NRT}^{max} = 50$. This improvement is traded-off with an increase in blocking probability for new calls; as anticipated, lowering the maximum number of admitted NRT calls increases blocking (Fig. 3(a)). However, we reason that this effect must be largely mitigated due to the fact that NRT calls then spend less time in the system. In any case, from a QoS perspective, ensuring acceptable quality to users already in the system is more important.

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 classes of traffic is determined by the actual traffic load, E_b/N_0 requirements for each class, as well as interference and physical power limitations. Besides that, the actual system behavior and QoS parameters are mirrored through the admission and rate control scheme applied. Here, we have studied a system with adaptive-rate RT calls and elastic NRT traffic. The general admission control scheme allows NRT calls to benefit from periods of low or intermittent RT traffic to attain an improved performance.

QoS management is introduced by varying the amount of capacity reservation for elastic traffic. Both for the uplink and downlink, it has been shown that capacity reservation can offer significant performance improvement to NRT sessions, at the expense of increased blocking of RT calls. However, the amount of reservation need not be very high; for the test cases considered, a reservation around 20% of the total capacity vastly improves the NRT performance, while not significantly harming RT behavior.

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 resources, can then reduce congestion and improve performance. Additionally, the impact of interference should be carefully considered in the choice of a capacity reservation.

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

References

1. Bender, P., Black, P., Grob, M., Padovani, R., Sindhushayana, N., Viterbi, A.: CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users. *IEEE Communications Magazine* **38** (2000) 70–77
2. Brandt, A., Last, G.: On the pathwise comparison of jump processes driven by stochastic intensities. *Mathematische Nachrichten* **167** (1994) 21–42
3. Gaver, D.P., Jacobs, P.A., Latouche, G.: Finite birth-and-death models in randomly changing environments. *Advances in Applied Probability* **16** (1984) 715–731
4. Gilhousen, K.S., Jacobs, I.M., Padovani, R., Viterbi, A.J., Weaver, A., Jr., Wheatley C.E.: On the capacity of a cellular CDMA system. *IEEE Transactions on Vehicular Technology* **40** (1991) 303–312
5. Hegde, N., Altman, E.: Capacity of multiservice WCDMA networks with variable GoS. In *Proc. of IEEE WCNC* (2003)
6. Hiltunen, K., De Brarnardi, R.: WCDMA downlink capacity estimation. In *Proc. IEEE VTC-Spring* (2000) 992–996
7. Holma H., Toskala, A. (eds.): *WCDMA for UMTS: Radio access for third generation mobile communications*. John Wiley & Sons, 3rd Ed. (2004)
8. Kelif, J.M., Altman, E.: Admission and Gos control in multiservice WCDMA system. In *Proc. ECUMN, LNCS 3262* (2004) 70–80
9. Kim, S.-L., Rosberg, Z., Zander, J.: Combined power control and transmission rate selection in cellular networks. In *Proc. IEEE VTC-Fall* (1999) 1653–1657
10. Koukoutsidis, I., Altman, E., Kelif, J.M.: A non-homogeneous QBD approach for the admission and GoS control in a multiservice WCDMA system. *INRIA Research Report No. RR-5358* (2004)
11. Latouche, G., Ramaswami, V.: *Introduction to matrix analytic methods in stochastic modeling*. ASA-SIAM (1999)
12. Neuts, M.F.: *Matrix-geometric solutions in stochastic models: an algorithmic approach*. The John Hopkins University Press (1981)
13. Parkvall, S., Dahlman, E., Frenger, P., Beming, P., Persson, M.: The high speed packet data evolution of WCDMA. In *Proc. 12th IEEE PIMRC* (2001)
14. Viterbi, A.M., Viterbi, A.J.: Erlang capacity of a power-controlled CDMA system. *IEEE J. Selected Areas in Communications* **11** (1993) 892–900
15. Viterbi, A.J., Viterbi, A.M., Zehavi, E.: Other-cell interference in cellular power-controlled CDMA. *IEEE Transactions on Communications*, **42** (1994) 1501–1504

Appendix

A LDQBD Algorithms

Consider the transition probability matrix for a LDQBD process as in § 4 but with a finite number of levels, K . Clearly we have only matrices A_2^K, A_1^K in the last level, with $A_1^K[j, j] = -\lambda_{RT} - j \cdot \mu_{RT} - \nu(K, j)$. We use the following algorithm from [3] to calculate the steady state distribution. The algorithm consists of the following steps:

Algorithm *Finite LDQBD* :

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

$$S_0 = B,$$

$$S_n = A_1^n + A_2^n (-S_{n-1}^{-1}) A_0, \quad 1 \leq n \leq K.$$

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

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

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 $k \rightarrow \infty$. It holds that $Q_0^{(1)} < Q_0^{(2)} < \dots < Q'_0$ and $Q_2^{(1)} > Q_2^{(2)} > \dots > Q'_2, \forall k \in \mathbb{Z}^+$. Transitions rates in $Q_1^{(k)}, Q'_1$ are identical within the same level. Then, if the homogeneous QBD process $X'(t)$ is ergodic, so is the non-homogeneous LDQBD process $X(t)$. 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.

Proof (sketch). We first proceed to show that $X(t) \leq_{st} X'(t)$, i.e. that $X'(t)$ stochastically dominates $X(t)$. Let (E, \leq) be a countable partially ordered set, and a set $F \subseteq E$ which is \leq -increasing. Denote by $q(i, j), q'(i, j)$ the transition intensities of $X(t), X'(t)$, respectively, where $\sum_{j \neq i} q_{ij} < \infty$ and $\sum_{j \neq i} q'_{ij} < \infty \forall i, j \in E$. Then from [2] $X(t) \leq_{st} X'(t)$ if and only if the following conditions hold, for all $x \leq y$ in E and all increasing sets, F :

- (i) if $x, y \in F$, $\sum_{z \notin F} q(x, z) \geq \sum_{z \notin F} q'(y, z)$
- (ii) if $x, y \notin F$, $\sum_{z \in F} q(x, z) \leq \sum_{z \in F} q'(y, z)$.

We define the partial order relation ($<$) by $(i, j) < (k, l)$ if $((i < k) \wedge (j \leq l)) \vee ((i \leq k) \wedge (j < l))$. It is then easy to show that our system satisfies conditions (i) and (ii), considering that transitions have the same structure and are *skip-free* in each direction. Therefore $X(t) \leq_{st} X'(t)$. Then considering the recurrence times $\sigma_\ell, \sigma'_\ell$ to the *smallest*³ state $\ell = (0, 0)$, we may prove that $E[\sigma_\ell] \leq E[\sigma'_\ell]$, from which we conclude that if $X'(t)$ is ergodic, both mean recurrence times are finite and $X(t)$ is also ergodic.

In the reverse part, we show that there exists a modified QBD process $X''(t)$ which is not ergodic and for which holds $X''_t \leq_{st} X^L_t$, where X^L_t is the *truncated* LDQBD for levels $k \geq L$, obtained by rerouting transitions from level L to $L - 1$ back to L . Then, using again mean recurrence times, we show that X^L_t is not ergodic from which we can also establish that the original LDQBD process is not ergodic. □

³ Note that due to the partial order here, the ‘smallest’ state is defined as $\ell = \{x \in E : \nexists x' \neq x \text{ with } x' > x\}$.