Capacity of multiservice WCDMA Networks with variable GoS

Nidhi Hegde and Eitan Altman
INRIA 2004 route des Lucioles,B.P.93 06902 Sophia-Antipolis, France
Email:{Nidhi.Hegde, Eitan.Altman}@sophia.inria.fr

Abstract— **Traditional definitions of capacity of CDMA networks are either related to the number of calls they can handle (pole capacity) or to the arrival rate that guarantees that the rejection rate (or outage) is below a given fraction (Erlang capacity). We extend the latter definition to other quality of service (QoS). We consider best-effort (BE) traffic sharing the network resources with real-time (RT) applications. BE applications can adapt their instantaneous transmission rate to the available one and thus need not be subject to admission control or outages. Their meaningful QoS is the average delay. The delay aware capacity is defined as the arrival rate of BE calls that the system can handle such that their expected delay is bounded by a given constant. We compute both the blocking probability of the RT traffic having an adaptive Grade of Service (GoS) as well as the expected delay of the BE traffic for an uplink multicell WCDMA system. This yields the Erlang capacity for former and the delay capacity for the latter.**

I. Introduction

Third generation mobile networks such as the Universal Mobiles Telecommunications System (UMTS), will provide a wide variety of services to users, including multimedia applications and interactive real-time applications as well as best-effort applications such as file transfer, Internet browsing, and electronic mail. These services have varied quality of service (QoS) requirements; real time applications (RT) needs some guaranteed minimum transmission rate as well as delay bounds which requires reservation of system capacity. We assume that RT traffic is subject to Call Admission Control (CAC) in order to guarantee the minimum rates for accepted RT calls. This implies that RT traffic may suffer rejections whose rate is then an important QoS for such applications. In contrast, Best-effort (BE) applications can adapt their transmission rate to the network's available resources and is therefore not subject to CAC. The relevant QoS measure for BE traffic is then the expected sojourn time (or delay) of a call in the system (e.g. the expected time to download a file).

We consider BE traffic sharing the network resources with RT applications. Our aim is to compute both the blocking (or rejection) probability of the RT traffic as well as the expected delay of the BE traffic for an uplink multicell WCDMA system. Although RT calls need a minimum guaranteed transmission rate, they are assumed to be able to adapt to network resources in a way similar to the BE traffic. For example, in the case of voice applications, UMTS will use 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. Although both RT and BE traffic have adaptive rates, we identify a key difference between the two: The *duration* of a RT call does not depend on the instantaneous assigned rate it gets (only the quality may change), whereas for BE calls, the *total volume transmitted* during the call does not depend on the assigned rate; the duration of BE calls therefore does depend on the dynamic rate assignment. We propose a probabilistic model that takes these features into account and enables to compute the performance measures of interest: we compute the blocking probabilities and the average throughput per RT calls, the expected average number of RT and BE calls in the system, and the expected delay of BE call.

We extend the notion of capacity in order to describe the amount of traffic for which the system can offer reasonable QoS. Traditional definitions of capacity of networks are either related to the number of calls they can handle (pole capacity) or to the arrival rate that guarantees that the rejection rate (or outage) is below a given fraction (Erlang capacity, see [11]). We extend the latter definition to other QoS. The delay aware capacity, suitable in particular for the BE traffic, is defined as the arrival rate of BE calls that the system can handle such that their expected delay is bounded by a given constant. We compute it as a function of other parameters of the system (rate of arrival and characteristics of RT traffic, the CAC and downgrading policy applied to RT traffic).

We briefly mention related work. In [10], an uplink CDMA with two classes is considered, the RT traffic is transmitted all the time, the non real time mobiles (NRT) are time-shared. A related idea has also been analyzed in [6]. The benefits of time sharing is studied and conditions for silencing some are obtained. The capacity of voice/data CDMA systems is also analyzed in [7] where both classes are modeled as VBR traffic. Adaptive features of transmission rates are not considered in the above references. In [1], the author considers the influence of the value of a fixed (not-adaptive) bandwidth per BE calls on the Erlang capacity of the system (that includes also RT calls), taking into account that a lower bandwidth implies longer call durations. A limiting capacity (as the fixed bandwidth vanishes) is identified and computed. Related work [2], [9] has also been done in wireline ATM networks (although without the power control aspects and without the downgrading features of wireless).

The structure of this paper is as follows. Next section introduces the model and preliminaries. Section III computes the performance of RT and BE traffic in the case of a single sector using a matrix geometric approach. This is then extended in Section IV to the multisector multicell case using a fix point argument. In Section V we provide numerical examples and we end with a concluding section.

II. Preliminaries

We consider the uplink of a multi-service WCDMA system with K service classes. Let X_j be the number of ongoing calls of type j in some given sector, and $\mathbf{X} = (X_1, \ldots, X_K)$. In CDMA systems, in order for a signal to be received, the ratio of it's received power to the sum of the background noise and interference must be greater than a given constant. For some given $\mathbf{X}$, this condition is as follows [5]:

$$\frac{P_j}{N + I_{\text{own}} + I_{\text{other}} - P_j} \triangleq \gamma_j \geq \tilde{\Delta}'_j, \; j = 1, \ldots, K, \quad (1)$$

where N is the background noise, and I_{own} and I_{other} are the total power received from the mobiles within the considered sector, and within the other sectors or cells, respectively. γ_j is the ratio of received power to total receive noise and interference at the base station, SIR, and $\tilde{\Delta}'_j$ is the required SIR for a call of class j, given by $\tilde{\Delta}'_j = E_j/N_{\text{o}} R_j W$ where E_j is the energy per transmitted bit of type j, N_{o} is the thermal noise density, W is the WCDMA modulation bandwidth, and R_j is the transmission rate of the type j call.

The interference received from mobiles in the same sector is simply $I_{\text{own}} = \sum_{j=1}^{K} X_j P_j$. When X_j for all $j = 1, \ldots, K$ is fixed, we also make the standard assumption [5] that the other-cell interference is proportional to interference for own cell, by some constant f, as such:

$$I_{\text{other}} = f I_{\text{own}}. \quad (2)$$

Note that the above assumes perfect power control. Due to inaccuracies in the closed-loop fast power control mechanism, mainly due to shadow fading of the radio signal, the γ_j may not be equal to $\tilde{\Delta}'_j$ at all times. We now define γ_j to be a random variable of the form $\gamma_j = 10^{\xi_j/10}$, where $\xi_j \sim N(\mu_\xi, \sigma_\xi)$ includes the shadow fading component and σ_ξ is the standard deviation of shadow fading with typical values between 0.3 and 2 dB [4], [11]. It follows then that γ_j has a lognormal distribution given by: $f_{\gamma_j}(x_j) = \frac{h}{x_j \sigma_\xi \sqrt{2\pi}} \exp\left(-\frac{(h \ln(x_j) - \mu_\xi)^2}{2\sigma_\xi^2}\right)$ where $h = 10/\ln 10$.

Since γ_j is now a random variable, we can write the condition (1) in terms of $\bar{\gamma}_j$, the average received SIR. We would now like to determine the required SIR, $\tilde{\Delta}_j$ such that $\bar{\gamma}_j = \tilde{\Delta}_j$ where $\tilde{\Delta}_j$ includes power control errors and replaces $\tilde{\Delta}'_j$ in (1). We determine $\tilde{\Delta}_j$ for the outage condition: $\Pr[\gamma_j \geq \tilde{\Delta}'_j] = \beta$ [12]. The reliability, β, is typically set to 99%. We have:

$$\Pr[\gamma_j \geq \tilde{\Delta}'_j] = \beta = \int_{\tilde{\Delta}'}^{\infty} f_{\gamma_j}(x) dx = Q\left(\frac{h \ln \tilde{\Delta}' - \mu_\xi}{\sigma_\xi}\right)$$

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

By inverting the above Q-function, we have:

$$\tilde{\Delta}'_j = 10^{\left(\frac{Q^{-1}(\beta)\sigma_\xi}{10} + \frac{\mu_\xi}{10}\right)} \quad (3)$$

Since γ_j is a lognormal random variable, its expectation is given by: $\bar{\gamma}_j = \exp\left(\frac{\sigma_\xi^2}{2h^2} + \frac{\mu_\xi}{h}\right)$. We solve for μ_ξ, to obtain:

$$\mu_\xi = h \ln \bar{\gamma}_j - \frac{\sigma_\xi^2}{2h} \quad (4)$$

We use (3) and (4) to get:

$$\tilde{\Delta}'_j = \bar{\gamma}_j 10^{\frac{Q^{-1}(\beta)\sigma_\xi}{10} - \frac{\sigma_\xi^2}{20h}}$$

We then have the SIR condition in (1) modified as follows:

$$\bar{\gamma}_j \geq \tilde{\Delta}'_j \Gamma = \frac{E_j}{N_{\text{o}}} \frac{R_j}{W} \Gamma \triangleq \tilde{\Delta}_j \quad (5)$$

where

$$\Gamma = 10^{\frac{\sigma_\xi^2}{20h} - \frac{Q^{-1}(\beta)\sigma_\xi}{10}}.$$

Note that Γ is independent of service class. The value of Γ is a function of the standard deviation of the shadow fading of users, σ_ξ, whose value varies with user mobility. Differences in the signal fading due only to user mobility are not considered in this paper. The above modified required SIR now includes a correction for imperfect power control.

Revisiting (1), we notice that in order serve a large of number of ongoing calls, that is to keep the X_js high, we must keep the P_js as low as possible. We then solve for the minimum required received power P_j satisfying (5) which is known to be the one that gives strict equality $\bar{\gamma}_j = \tilde{\Delta}_j$ in (5):

$$P_j = \frac{N \Delta_j}{1 - (1+f) \sum_{j=1}^{K} X_j \Delta_j} \quad (6)$$

where $\Delta_j = \frac{\tilde{\Delta}_j}{1+\tilde{\Delta}_j}$ turns out to be the signal-to-total-power ratio, STPR (see [1, eq 4]).

Define the loading as:

$$\theta = \sum_{j=1}^{K} X_j \Delta_j(\mathbf{X}). \quad (7)$$

This definition reflects the fact that Δ_j is a function of the number of each type of call in the system (since it depends on the transmission rate R_j and since R_j will be determined as a function of the system state). In this paper we consider both real time (RT) and best-effort (BE) services that receive a variable rate. As explained in Section III, the rate received by RT calls, and thus Δ_{RT}, depends on the number of RT calls. The rate received by BE calls depends on both X_{RT} and X_{BE}. We maintain this dependence throughout the paper, however for notational convenience we will sometimes drop the argument $(\mathbf{X})$.

Now we may define the integer capacity of the cell as the set X^* of vectors $\mathbf{X}$ such that the received powers of the mobiles stays finite, i.e. the denominator of (6) does not vanish [1]. In the equation for minimum received power shown in (6), this implies the condition $\theta(1+f) < 1$. The system prevents, through Call Admission Control (CAC), that the denominator vanishes; more generally, it is desirable to be even more conservative and to impose a bound on the capacity, $\Theta_\epsilon = 1-\epsilon$ where $\epsilon > 0$. Thus the CAC will ensure that $\theta \le \Theta_\epsilon/(1+f)$. Later on we shall consider special combined policies for RT traffic that combine CAC with some rate adaptation, along with a rate adaptation for NRT traffic, which will result in a further restriction on the number of RT calls that the system can handle (which will also be called, with some abuse of notation, the integer capacity of RT traffic).

III. Single Sector in Isolation

Let us first consider a single sector, so that we may exclude interference from other sectors and other cells in the calculations, thereby setting $f = 0$, in this section. We consider a base station with uplink capacity such that

$$\theta \le \Theta_\epsilon. \tag{8}$$

Here we define *capacity* in terms of the sum of Δ's, STPR, of all users. We denote by individual normalized bandwidth, the individual required STPR that corresponds to a particular rate. For example, a call that requires a rate of y bps requires a normalized bandwidth of $\Delta = \frac{E/N_o}{W/y+E/N_o}$ where E/N_o is the requirement specified for the given service type of the call.

A. Real Time Calls

We assume a single type of RT calls capable of accepting a variable rate, with a requested transmission rate $R^{\mathrm{r}}_{\mathrm{RT}}$. From (5) and the definition of Δ_j that follows (6), we derive the required bandwidth $\Delta^{\mathrm{r}}_{\mathrm{RT}}$ that corresponds to rate $R^{\mathrm{r}}_{\mathrm{RT}}$:

$$\Delta^{\mathrm{r}}_{\mathrm{RT}} = \frac{E_{\mathrm{RT}}/N_o}{W/R^{\mathrm{r}}_{\mathrm{RT}} + E_{\mathrm{RT}}/N_o}.$$

We now introduce the parameters of the call admission control for the RT traffic. All BE calls in the sector share equally the capacity remaining after RT calls have been allocated the required normalized bandwidth. In addition, we assume that some portion of the capacity is reserved for BE calls, thus the RT calls have a maximum capacity, denoted by L_{RT}. Let us denote L_{BE} to be the minimum portion of the total capacity available for BE calls. We then have $L_{\mathrm{BE}} = \Theta_\epsilon - L_{\mathrm{RT}}$. We have the following condition for the capacity bound on RT calls:

$$X_{\mathrm{RT}}\Delta_{\mathrm{RT}} \le L_{\mathrm{RT}} \tag{9}$$

where Δ_{RT} is the normalized bandwidth received by each RT call. Note that this value will depend on the number of RT calls, and thus may vary.

The integer capacity for RT calls, such that they all receive the requested rate $R^{\mathrm{r}}_{\mathrm{RT}}$ and bandwidth $\Delta^{\mathrm{r}}_{\mathrm{RT}}$, is then given by $N_{\mathrm{RT}} = \left\lfloor \frac{L_{\mathrm{RT}}}{\Delta^{\mathrm{r}}_{\mathrm{RT}}} \right\rfloor$.

1) CAC and GoS control: In a strict call admission control scheme for RT calls, new RT call arrivals would be blocked and cleared when there are N_{RT} RT calls in the sector. However, in UMTS, we can control the GoS, by providing RT calls with a variable transmission rate [3]. In such a case, we may allow more than N_{RT} RT calls, at the expense of reducing the transmission rate of all RT calls, thus keeping the total normalized bandwidth occupied by all RT calls within the limit. Let us then define a second threshold for admission of RT calls, $M_{\mathrm{RT}} > N_{\mathrm{RT}}$. Call admission control for RT calls then is as follows. As long as the number of RT calls is less than N_{RT}, all RT calls receive the requested normalized bandwidth $\Delta^{\mathrm{r}}_{\mathrm{RT}}$. When the number j of RT calls is more than N_{RT} but not more than M_{RT}, all RT calls receive with equality a modified (reduced) normalized bandwidth, denoted here as Δ^{j}_{RT}, such that (9) is still satisfied. If there are M_{RT} RT calls in the sector, new RT call arrivals are blocked and cleared. M_{RT} may be chosen so that RT calls receive a minimum transmission rate of $R^{\mathrm{m}}_{\mathrm{RT}}$, with normalized bandwidth $\Delta^{\mathrm{m}}_{\mathrm{RT}}$, even in the worst case. The integer capacity for RT calls then is $M_{\mathrm{RT}} = \left\lfloor \frac{L_{\mathrm{RT}}}{\Delta^{\mathrm{m}}_{\mathrm{RT}}} \right\rfloor$, where $\Delta^{\mathrm{m}}_{\mathrm{RT}} = \frac{E_{\mathrm{RT}}/N_o}{W/R^{\mathrm{m}}_{\mathrm{RT}}+E_{\mathrm{RT}}/N_o}$, as derived from (5). The bandwidth received by each RT call at some time t is thus a function of $X_{\mathrm{RT}}(t)$ as follows:

$$\Delta_{\mathrm{RT}}(X_{\mathrm{RT}}(t)) = \begin{cases} \Delta^{\mathrm{r}}_{\mathrm{RT}} & 1 \le X_{\mathrm{RT}}(t) \le N_{\mathrm{RT}}; \\ L_{\mathrm{RT}}/X_{\mathrm{RT}}(t) & N_{\mathrm{RT}} < X_{\mathrm{RT}}(t) < M_{\mathrm{RT}}. \end{cases} \tag{10}$$

2) RT Traffic Model: We assume that RT calls arrive according to a Poisson process with rate λ_{RT}. The duration of an RT call is assumed to have an exponential distribution with mean $1/\mu_{\mathrm{RT}}$, and is not affected by the allocated bandwidth. Let $X_1(t)$ and $X_2(t)$ represent the number of RT customers and BE customers respectively, at time t in the given sector. The number of RT calls in the system is not affected by the BE calls. Therefore, $X_1(t)$ follows a birth and death process, with birth rate λ_{RT} and death rate μ_{BE}. The steady-state probabilities $\pi_{\mathrm{RT}}(x)$ of the number of RT calls x in the system are given by:

$$\Pr[X_{\mathrm{RT}} = x] = \lim_{t\to\infty} \Pr[X_{\mathrm{RT}}(t) = x] = \frac{\rho^x_{\mathrm{RT}}/x!}{\sum_{i=0}^{M_{\mathrm{RT}}} \rho^i_{\mathrm{RT}}/i!} \tag{11}$$

where $\rho_{\mathrm{RT}} = \lambda_{\mathrm{RT}}/\mu_{\mathrm{RT}}$. For RT calls, we are interested in the call blocking probability and the average throughput. The call blocking probability is given by:

$$P^{\mathrm{RT}}_{\mathrm{B}} = \pi_{\mathrm{RT}}(M_{\mathrm{RT}}) = \frac{\rho^{M_{\mathrm{RT}}}_{\mathrm{RT}}/M_{\mathrm{RT}}!}{\sum_{i=0}^{M_{\mathrm{RT}}} \rho^i_{\mathrm{RT}}/i!} \tag{12}$$

We define $r(x)$ to be the transmission rate received by RT calls when there are x RT calls in the sector, as follows

$$r(X_{\mathrm{RT}}) = \frac{\Delta_{\mathrm{RT}}(X_{\mathrm{RT}})\, W}{(1-\Delta_{\mathrm{RT}}(X_{\mathrm{RT}}))\, E_{\mathrm{RT}}/N_o}$$

Since the transmission rate of RT calls is affected by the number of RT calls, we would like to include in our definition of expected throughput, a measure of the number of RT calls in the sector. We define the expected throughput per call as the ratio of the expected global throughput to the expected number of RT calls in the sector, as follows:

$$\mathbb{E}[r(X_{\mathrm{RT}})] = \frac{\sum_{x=1}^{M_{\mathrm{RT}}} Pr[X_{\mathrm{RT}} = x]\, x\, r(x)}{\sum_{x=1}^{M_{\mathrm{RT}}} Pr[X_{\mathrm{RT}} = x]\, x} \tag{13}$$

B. Best-Effort Calls

We define $C(x)$ to be the capacity available to BE calls when there are x RT calls, as follows:

$$C(x) = \begin{cases} \Theta_\epsilon - x\Delta_{\mathrm{RT}}^{\mathrm{r}} & , \quad x \leq N_{\mathrm{RT}}; \\ L_{\mathrm{BE}} & , \quad N_{\mathrm{RT}} < x \leq M_{\mathrm{RT}}. \end{cases}$$

All BE calls in the sector share equally the available bandwidth. We can then model BE service by a processor sharing(PS) discipline with a random service capacity. We study two performance metrics for BE calls: the average sojourn time of a BE call for given values of RT and BE load, and the maximum BE arrival rate such that the average delay is always bounded by a given constant.

Best-effort calls arrive according to a Poisson process with rate λ_{BE}. The required workload of BE classes, i.e. file sizes, are i.i.d exponentially distributed with mean $1/\mu_{\mathrm{BE}}$. The departure rate of BE calls is given by $\nu(X_{\mathrm{RT}}) = \mu_{\mathrm{BE}} R_{\mathrm{BE}}(X_{\mathrm{RT}})$, where $R_{\mathrm{BE}}(X_{\mathrm{RT}})$ is the total BE rate corresponding to the available BE capacity $C(X_{\mathrm{RT}})$, as follows:

$$R_{\mathrm{BE}}(X_{\mathrm{RT}}) = \frac{C(X_{\mathrm{RT}})\, W}{(1 - C(X_{\mathrm{RT}}))\, E_{\mathrm{BE}}/N_o}.$$

We assume no call admission control for BE calls. The process $(X_2(t), X_1(t))$ is an irreducible Markov chain. It is ergodic if and only if the average service capacity available to BE calls is greater than the BE load (as in [2]):

$$\mu_{\mathrm{BE}} \mathbb{E} R_{\mathrm{BE}}(X_{\mathrm{RT}}) > \lambda_{\mathrm{BE}}. \tag{14}$$

Specifically, the process $(X_2(t), X_1(t))$ is a homogeneous quasi birth and death process(QBD) with the generator Q. The stationary distribution of this system, π, is calculated by $\pi Q = 0$, with the normalization condition $\pi e = 1$ where e is a vector of ones of proper dimension. π represents the steady-state probability of the two-dimensional process lexicographically: we partition π as $[\pi(0), \pi(1), \ldots]$ with the vector $\pi(i)$ for level i, where the levels correspond to the number of BE calls in the system. We may further partition each level into the number of RT calls, $\pi(i) = [\pi(i,0), \pi(i,1), \ldots, \pi(i, M_{\mathrm{RT}})]$, for $i \geq 0$.

The generator Q has the form:

$$Q = \begin{bmatrix} B & A_0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots \end{bmatrix} \tag{15}$$

where the matrices B, A_0, A_1, and A_2 are square matrices of size $(M_{\mathrm{RT}} + 1)$. A_0 corresponds to a BE connection arrival, given by $A_0 = \mathrm{diag}(\lambda_{\mathrm{BE}})$. A_2 corresponds to a departure of a BE call. The departure rate for BE calls is $\nu(X_{\mathrm{RT}})$. Thus $A_2 = \mathrm{diag}(\nu(i); 0 \leq i \leq M_{\mathrm{RT}})$ A_1 corresponds to the arrival and departure processes of the RT calls. A_1 is tri-diagonal as follows:

$$A_1[i, i+1] = \lambda_{\mathrm{RT}}$$
$$A_1[i, i-1] = i\mu_{\mathrm{RT}}$$
$$A_1[i, i] = -\lambda_{\mathrm{RT}} - i\mu_{\mathrm{RT}} - \lambda_{\mathrm{BE}} - \nu(i)$$

We also have $B = A_1 + A_2$.

The steady-state equations can be written as:

$$0 = \pi(0)B + \pi(1)A_2 \tag{16}$$

$$0 = \pi(i-1)A_0 + \pi(i)A_1 + \pi(i+1)A_2 \quad i \geq 1 \tag{17}$$

We follow the matrix-geometric solution to this QBD [8]. Assuming stability as shown in (14), the steady-state solution π exists, and is given by:

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

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

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

In order to solve for $\mathbf{R}$, we find it efficient to write $A_1 = T - S$ where S is a diagonal matrix and T has a zero diagonal. The diagonal matrix S is positive and invertible, and we may write (19) as $\mathbf{R} = (A_0 + \mathbf{R}T + \mathbf{R}^2 A_2)S^{-1}$. This equation can then be solved by successive iterations starting with $\mathbf{R} = 0$, a zero matrix.

Once the matrix $\mathbf{R}$ is known, we may find $\pi(0)$ using the boundary condition (16) and the normalization $\pi e = 1$ which using (18) is equivalent to $\pi(0)(I - R)^{-1}e = 1$. The marginal distribution of the number of RT calls can easily be obtained by using (11). The marginal probability of the number BE calls is

$$\Pr[X_{\mathrm{BE}} = i] = \sum_{j=0}^{M_{\mathrm{RT}}} \pi(i,j) = \pi(i)e = \pi(0)\mathbf{R}^i e.$$

One way to compute the above is by finding the $M_{\mathrm{RT}} + 1$ eigenvalues and corresponding eigenvectors of the matrix $\mathbf{R}$. All $M_{\mathrm{RT}} + 1$ eigenvalues of the matrix $\mathbf{R}$ are distinct [9] and therefore $\mathbf{R}$ is diagonalizable. Define D to be a diagonal matrix containing the eigenvalues of $\mathbf{R}$, r_i, on the diagonal, and V to be a matrix containing the corresponding eigenvectors, v_i as columns. We then have:

$$\Pr[X_{\mathrm{BE}} = i] = \pi(0)\mathbf{R}^i e = \pi(0)VD^iV^{-1}e = \sum_{k=0}^{M_{\mathrm{RT}}} a_k r_k^i$$

where $a_k = \pi(0)v_k e_k' V^{-1} e$ and e_k' is a zero vector of proper dimension with the kth element equal to one. The expectation of X_{BE} is as follows:

$$\mathbb{E}[X_{\mathrm{BE}}] = \sum_{k=0}^{M_{\mathrm{RT}}} a_k \frac{r_k}{(1 - r_k)^2} \tag{20}$$

We can now use Little's Law to calculate the average sojourn time of a BE session, $T_{\mathrm{BE}} = \mathbb{E}[X_{\mathrm{BE}}]/\lambda_{\mathrm{BE}}$. Having obtained the expected delay of BE traffic in terms of the system parameters, one can now obtain the delay aware capacity of BE traffic, i.e. the arrival rate of BE calls that the system can handle such that their expected delay is bounded by a given constant.

IV. Extension to Multiple Sectors

In this section we provide an analysis for the multi-sector multi-cell case, by including an approximation for the other-sector interference, $I_{\rm other}$. Above in (2), we have made the assumption that $I_{\rm other}$ is proportional to $I_{\rm own}$ by a constant f. Such a definition of other sector interference and the subsequent derivation of minimum required received power in (6) holds for a static network with a fixed number of mobiles. However, in our dynamic model of stochastic arrivals and holding times, such a definition may not hold at all times. We then approximate the instantaneous interference $I_{\rm other}$ by its average $\mathbb{E}[I_{\rm other}]$. We modify (2) to $I_{\rm other} = f\mathbb{E}[I_{\rm own}] = \sum_{j=1}^{K} \mathbb{E}[X_j \Delta_j(\mathbf{X})]$. The minimum required received power in (6) is now as follows:

$$P_j = \frac{N\Delta_j}{1 - \sum_{j=1}^{K} X_j\Delta_j - f\mathbb{E}[X_j \Delta_j(\mathbf{X})]}$$

Let G denote the expected other-sector (and cell) interference, $G = f\sum_{j=1}^{K} \mathbb{E}[X_j \Delta_j(\mathbf{X})]$. The equation for P_j, above then implies the condition $\theta \leq 1 - G$. This condition is equivalent to (8) with $\Theta_G = 1 - G$ replacing Θ_ϵ.

The expected interference due to RT calls is calculated as follows:

$$f\mathbb{E}[X_{\rm RT}\Delta_{\rm RT}(X_{\rm RT})] = f\sum_{i=0}^{M_{\rm BE}} \pi_{\rm RT}(i) i \Delta_{\rm RT}(i)$$

where we use (11) for $\pi_{\rm RT}(i)$. For BE calls, we need not calculate the steady state distribution π. Since BE calls use all of the remaining capacity, the sum of the STPRs of the BE calls, where there is at least one BE call, is simply the available BE capacity, $C(X_{\rm RT})$. The expected interference due to BE calls is given by:

$$f\mathbb{E}[X_{\rm BE}\Delta_{\rm BE}(\mathbf{X})] = f(1 - \pi(0)e)\sum_{i=0}^{M_{\rm RT}} \pi_{\rm RT}(i)C(i)$$

where $\pi(0)e$ is the probability that there are no BE calls in the sector, and can be calculated using only (16) and the normalization condition $\pi e = 1$. For each fixed value of G, say g, we can obtain the probabilities $\pi_{\rm RT}$ and $\pi(0)$ using Θ_g instead of Θ_ϵ. We denote these values by $\pi^g_{\rm RT}$ and $\pi^g(0)$ respectively, and the expectation operator corresponding to these probabilities as $\mathbb{E}^g$. Define $F(g) = f\sum_{j\in K} \mathbb{E}^g[X_j\Delta_j(\mathbf{X})]$. G then is the solution of the fixed point equation:

$$g = F(g) \tag{21}$$

We can now set the BE threshold as $L^g_{\rm BE} = \Theta_g - L_{\rm RT}$. Under such a definition, for a given $L_{\rm RT}$, $F(g)$ can be shown to be continuous in g. $F(g)$ also maps onto itself, and thus by the Brower Fixed Point Theorem, there exists a solution. $F(g)$ can be shown to be nonincreasing in g, implying uniqueness of the solution to (21).

V. Numerical Results

In this section we perform numerical experiments to evaluate the performance of RT and BE calls. The rate requested by the RT calls is 12.2kbps(the maximum rate for AMR speech service in UMTS [3]). For the results shown here we have assumed a minimum acceptable rate of 7.95kbps, which is one of the eight possible rates for the AMR speech class. We assume that the set of rates acceptable to RT calls is continuous. We assume no minimum rate for BE calls. The average file size of a BE call is assumed to be 20kBytes. We assume $E_{\rm RT}/N_o = 4.1$dB, $E_{\rm BE}/N_o = 3.1$dB [3], a chip rate $W = 3.84Mcps$ and $\Theta_\epsilon = 1 - 10^{-5}$. We define the load in terms of the total RT rate available, $R_{\rm T}$. The total RT rate is in turn defined as the product of the minimum RT rate and the integer capacity for RT calls if there were no BE threshold, $R_T = \left\lfloor \frac{\Theta_\epsilon}{\Delta^{\rm m}_{\rm RT}} \right\rfloor R^{\rm m}_{\rm RT}$. The normalized load for RT calls is defined by $\tilde{\rho}_{\rm RT} = \frac{\lambda_{\rm RT}}{\mu_{\rm RT}}\frac{R^{\rm r}_{\rm RT}}{R_{\rm T}}$, and the BE normalized load is $\tilde{\rho}_{\rm BE} = \frac{\lambda_{\rm BE}}{\mu_{\rm BE}R_{\rm T}}$.

We consider the heavy traffic regime, where $\tilde{\rho}_{\rm RT} = 0.5$ and $\tilde{\rho}_{\rm BE} = 0.55$. We keep the normalized loads constant and vary the holding time of the RT calls. We evaluate the performance metrics of interest as a function of the BE reserved capacity, $L_{\rm BE}$.

Figure 1 shows the change in RT call blocking probability, computed using (12), as the BE Threshold, $L_{\rm BE}$ is varied from 0 to Θ_ϵ. As expected, as $L_{\rm BE}$ is increased, there is less capacity available for RT calls, and their call blocking probability increases. We may observe the tradeoff between the service qualities of BE and RT calls in Figures 2 and 3. These figures show the expected RT throughput and expected BE sojourn time, respectively. In Figure 2 we see that the expected RT throughput, computed using (13), is close to the requested rate of 12.2kbps up to a BE threshold of approximately $L_{\rm BE} = 0.35$. As $L_{\rm BE}$ is increased further, the expected RT throughput gradually drops, always remaining above the minimum rate of 7.95kbps.

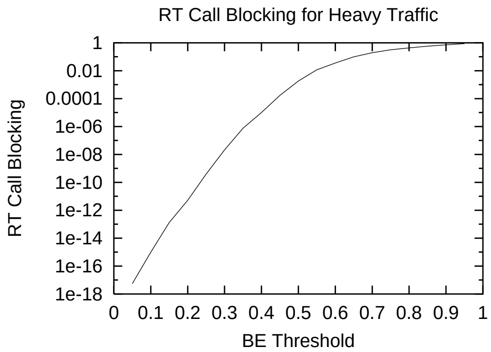

Fig. 1. RT Call Blocking for heavy traffic

The sensitivity of BE service quality is seen in Figures 3 and 4 with respect to not only the BE threshold, but also the RT call duration. In Figure 3 the expected BE sojourn time, computed using (20) and Little's Law, decreases as $L_{\rm BE}$ is increased.

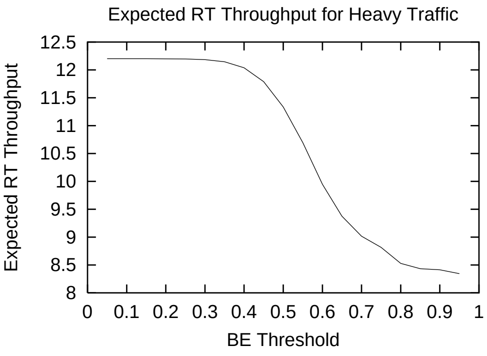

Fig. 2. Expected RT Throughput

For small values of $L_{\rm BE}$ we see that the expected BE sojourn time varies greatly with increasing $L_{\rm BE}$, when the duration of RT calls is large(smaller values of $\mu_{\rm RT}$). The duration of the RT calls determines the time scale of the evolution of the number of RT calls in the system, and thus the available capacity for the BE calls. When the mean duration of RT calls is small, the number of RT calls evolves much faster relative to the BE calls, and thus we would expect the BE calls to obtain a capacity that is fairly constant. When the mean duration of RT calls is large, the changes in capacity received by BE calls might cause the BE queue to build up for long periods during which there are many ongoing RT calls, thus resulting in higher average sojourn times. For related results for non-variable RT GoS, see [2] and [9]. We observe from the figure that this effect can be diminished by increasing the BE threshold. An increase in $L_{\rm BE}$ means that for BE calls the reserved capacity is substantial compared to the capacity remaining after RT calls are served, an effect similar to having a constant capacity.

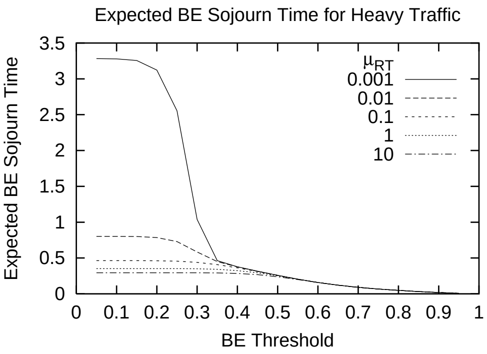

Fig. 3. Expected BE Sojourn Time

The delay aware capacity of BE calls for a fixed RT load is shown in Figure 4. Here, we find the maximum BE arrival rate such that $T_{\rm BE} \leq c$, where c is a constant, set to 0.25 in this figure. As expected, the maximum BE arrival rate increases as $L_{\rm BE}$ increases allowing a larger portion of the total capacity for BE calls. We note again the sensitivity to mean RT call duration at smaller values of $L_{\rm BE}$, where the delay capacity approximately doubles when $\mu_{\rm RT}$ is changed from 10 to 0.001.

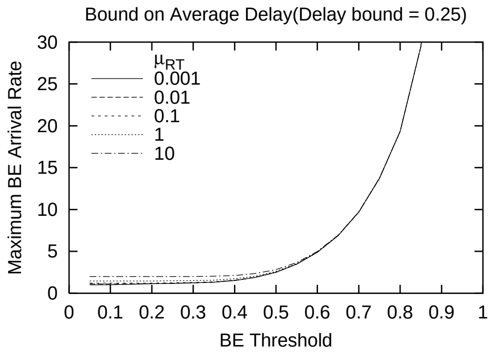

Fig. 4. BE Delay Aware Capacity

VI. Conclusion

We have modelled resource sharing of BE applications with RT applications in WCDMA networks. Both type of traffic have flexibility to adapt to the available bandwidth but unlike BE traffic, RT traffic requires strict minimum bounds on the throughput. We studied the performance of both BE and RT traffic and examined the impact of reservation of some portion of the bandwidth for the BE applications. We introduced a novel capacity definition related to the delay of BE traffic and showed how to compute it.

References

[1] Eitan Altman. Capacity of multi-service cdma cellular networks with best-effort applications. In *Proceedings of ACM MOBICOM*, September 2002.

[2] Eitan Altman, Damien Artiges, and Karim Traore. On the integration of best-effort and guaranteed performance services. *European Transactions on Telecommunications , Special Issue on Architectures, Protocols and Quality of Service for the Internet of the Future*, 2, February-March 1999.

[3] Harri Holma and Antti Toskala, editors. *WCDMA for UMTS, Radio Access For Third Generation Mobile Communications*. John Wiley & Sons, Ltd., 2001.

[4] Insoo Koo, JeeHwan Ahn, Jeong-A Lee, and Kiseon Kim. Analysis of erland capacity for the multimedia DS-CDMA systems. *IEICE Transactions of Fundamentals*, E82-A(5):849–55, May 1999.

[5] Jaana Laiho and Achim Wacker. Radio network planning process and methods for WCDMA. *Annales des Télécommunications*, 56(5-6):317–31, 2001.

[6] R. Leelahakriengkrai and R. Agrawal. Scheduling in multimedia CDMA wireless networks. Technical Report ECE-99-3, ECE Dept., University of Wisconsin-Madison, July 1999.

[7] N. Mandayam, J. Holtzman, and S. Barberis. Performance and capacity of a voice/data CDMA system with variable bit rate sources. In *Special Issue on Insights into Mobile Multimedia Communications*. Academic Press Inc., January 1997.

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

[9] R.Nú nez Queija and O.J. Boxma. Analysis of a multi-server queueing model of ABR. *J. Appl. Math. Stoch. Anal.*, 11(3), 1998.

[10] S. Ramakrishna and Jack M. Holtzman. A scheme for throughput maximization in a dual-class CDMA system. *IEEE Journal Selected Areas in Comm.*, 16:830–44, 1998.

[11] Audrey M. Viterbi and Andrew J. Viterbi. Erlang capacity of a power controlled CDMA system. *IEEE Journal on Selected Areas in Communications*, 11(6):892–900, August 1993.

[12] Qiang Wu, Wei-Ling Wu, and Jiong-Pan Zhou. Effects of slow fading SIR errors on CDMA capacity. In *Proceedings of IEEE VTC*, pages 2215–17, 1997.