Admission control for differentiated services in 3G cellular networks

Onno J. Boxma [a,b], Rudesindo Núñez-Queija [c,b,1],
Adriana F. Gabor [a,b,2], Hwee-Pink Tan [a,*,3]

[a] *EURANDOM, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*

[b] *Eindhoven University of Technology, Department of Mathematics and Computer Science, 5600 MB Eindhoven, The Netherlands*

[c] *CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*

Abstract

Third generation wireless systems can simultaneously accommodate flow transmissions of users with widely heterogeneous applications. As radio resources are limited, we propose an admission control rule that protects users with stringent capacity requirements ("streaming traffic") while offering sufficient capacity over longer time intervals to delay-tolerant users ("elastic traffic"). Using time-scale decomposition, we develop approximations to evaluate the performance of our differentiated admission control strategy to support integrated services with capacity requirements in a realistic downlink transmission scenario for a single radio cell.

Key words: Third generation wireless systems, differentiated admission control, time-scale decomposition.

[*] Corresponding author. Authorship listed in alphabetical order.
Email addresses: boxma@win.tue.nl (Onno J. Boxma), sindo@cwi.nl (Rudesindo Núñez-Queija), a.f.gabor@tue.nl (Adriana F. Gabor), tanhp@tcd.ie (Hwee-Pink Tan).

[1] Presently also afiliated with TNO Information and Communication Technology, The Netherlands. No longer affiliated with Eindhoven University of Technology.

[2] No longer affiliated with EURANDOM.

[3] Present address: Centre for Telecommunications Value-chain Research, Lloyd Institute, Trinity College Dublin, Dublin 2, Ireland.

Preprint submitted to Performance Evaluation *3 April 2007*

1 Introduction

Third Generation (3G) cellular networks such as UMTS and CDMA2000 are expected to support a large variety of applications, where the traffic they carry is commonly grouped into two broad categories: **Elastic traffic** corresponds to the transfer of digital documents (e.g., Web pages, emails, stored audio / videos) characterized by their size, i.e., the volume to be transferred. Applications carrying elastic traffic are flexible, or "elastic", towards capacity fluctuations, the total transfer time being a typical performance measure. **Streaming traffic** corresponds to the real-time transfer of various signals (e.g., voice, streaming audio / video) characterized by their duration as well as their transmission rate. Stringent capacity guarantees are necessary to ensure real-time communication to support applications carrying streaming traffic.

Various papers have been published recently that study *wired* links carrying integrated (elastic and streaming) traffic. In terms of resource sharing, the classical approach is to give head-of-line priority to packets of streaming traffic in order to offer packet delay and loss guarantees [7,5,11]; alternatively, *adaptive* streaming traffic (that is TCP-friendly and mimics elastic traffic) is considered in [9,3,13]. Markovian models have been developed for the exact analysis of these systems [11,14]. However, they can be numerically cumbersome due to the inherently large dimensionality required to capture the diversity of user applications. Hence, various approximations have been proposed [7,13], where closed-form limit results were obtained that can serve as performance bounds, and hence yield useful insight.

In this study, we consider downlink transmissions of integrated traffic in a single 3G radio cell and propose an admission control strategy that allocates priority to streaming traffic through resource reservation and guarantees the capacity requirements of all users while maximizing the data rate of each elastic user. The location-dependence of the wireless link capacity adds to the dimensionality problem already inherent in the performance analysis of corresponding *wireline* integrated services platforms. We describe the system model in Section 2 and develop an approximation based on time-scale decomposition in Section 3 to evaluate the user-level performance. We define two base station models based on abstractions of the system model in Section 4 and present numerical results comparing both models in Section 5. Some concluding remarks are outlined in Section 6.

2 System model

We consider a 3G radio cell (e.g., UMTS/W-CDMA) with a single downlink channel whose transmission power at the base station (resource) is shared amongst users carrying streaming and elastic traffic. We assume that the base station transmits at full power, denoted by P, whenever there is at least one user in the cell. In addition, a part of the total power, $P_s \leq P$, is *statically* reserved for streaming traffic, where unclaimed power is *equally* shared amongst all elastic users. Note that although the resource that can be maximally *guaranteed* for on-going elastic traffic is $P_e = P - P_s$, they are permitted to use more than P_e. However, the surplus is immediately allocated to streaming traffic when a new streaming user arrives.

With W-CDMA technology, the base station can transmit to *multiple* users simultaneously using orthogonal code sequences. Let $P_u \leq P$ be the power transmitted to user u. The power received by user u is $P^r_u = P_u \Gamma_u$, where Γ_u denotes the attenuation due to path-loss. For typical radio propagation models, Γ_u for user u at distance δ_u from its serving base station is proportional to $(\delta_u)^{-\gamma}$, where γ is a positive path-loss exponent.

As a measure of the quality of the received signal at user u, we consider *the energy-per-bit to noise-density ratio*, $\left(\frac{E_b}{N_0}\right)_u$, given by

$$\left(\frac{E_b}{N_0}\right)_u = \frac{W}{R_u} \frac{P^r_u}{\eta + I^a_u + I^r_u},$$

where W is the CDMA chip rate, R_u is the *instantaneous* data rate of user u, η is the background noise (assumed to be constant throughout the cell) and I^r_u is the inte$r$-cell interference at user u caused by simultaneous *interfering* transmissions received at user u from the base station in the serving cell. For linear and hexagonal networks, it can be shown [12] that I^r_u increases as δ_u increases. On the other hand, intr$a$-cell interference, I^a_u, is due to simultaneous transmissions from the serving base station of user u using non-orthogonal codes (with total power P^a_u) to other users in the *same* cell received at user u. Quantitatively, we can write $I^a_u = \alpha P^a_u \Gamma_u$, where α is the code non-orthogonality factor.

To achieve a target error probability corresponding to a given Quality of Service (QoS), it is necessary that $\left(\frac{E_b}{N_0}\right)_u \geq \epsilon_u$, for some threshold ϵ_u. Equivalently, the data rate R_u of each admitted user u is upper-bounded as follows:

$$R_u \leq \frac{W P_u \Gamma_u}{\epsilon_u(\eta + \alpha P^a_u \Gamma_u + I^r_u)}. \tag{1}$$

Accordingly, for a given P_u, α and user-type, the feasible data rate of user u depends on its location (through Γ_u and I_u^r) and the intra-cell interference power, P_u^a.

2.1 Location-dependence

According to Eq. (1), the transmission power, P_u, needed to support the capacity requirement, r_u, of user u is given by:

$$P_u \geq \frac{r_u \epsilon_u [\alpha P_u^a \Gamma_u + \eta + I_u^r]}{W \Gamma_u} \equiv \tilde{P}_u. \tag{2}$$

Ideally, given perfect knowledge of the location of each user u at the base station, a maximum number of users can be admitted by allocating *exactly* $\tilde{P}_u$ to each user u. However, in reality, only the *quantized* location of each user in the cell is known. This is obtained, e.g., by dividing the cell into J disjoint segments, where we assume that the path-loss, intra-cell and inter-cell interference are the same for any user in segment $j = 1, \ldots, J$, denoted by (Γ_j, I_j^a, I_j^r), respectively. As J increases, the location quantization becomes finer and approaches the ideal case (J=∞); on the other hand, the special case of J=1 corresponds to the case where user-location is unknown.

Accordingly, we assume that elastic and streaming users arrive at segment j as independent Poisson processes at rates $\lambda_{j,e}$ and $\lambda_{j,s}$, with capacity requirements of $r_{j,e} > 0$ and $r_{j,s} > 0$ respectively. Elastic users in segment j have a general file size (or service requirement) distribution with mean $f_{j,e}$ (bits) and, similarly, the holding times of streaming users may be taken to have mean $1/\mu_{j,s}$ (secs). The total arrival rates of elastic and streaming users to the cell are denoted by $\lambda_e = \sum_{j=1}^{J} \lambda_{j,e}$ and $\lambda_s = \sum_{j=1}^{J} \lambda_{j,s}$. The minimum energy-to-noise ratio, ϵ_u, may depend on the user type and location [1], and will be denoted by $\epsilon_{j,e}$ and $\epsilon_{j,s}$ for elastic and streaming users in segment j, respectively.

2.2 Resource Sharing

Given the transmission power, P_u, the mechanism via which the total power, P, is shared amongst all users (resource sharing) determines the total intra-cell interference power experienced at user u, P_u^a. When the base station transmits to all users in the cell simultaneously, each user u experiences the maximum intra-cell interference power, given by P-P_u; on the other hand, if time is slotted and the base station transmits only to one user in each time slot (*time* *sharing*), then there will be no interference power. Accordingly, we have the following expressions for P_u^a:

$$P_u^a \begin{cases} = P - P_u, & \text{simultaneous transmission to } \textit{all} \text{ users in the cell;} \\ < P - P_u, & \text{simultaneous transmission to } \textit{some} \text{ users in the cell;} \\ = 0, & \textit{no} \text{ simultaneous transmission (}\textit{time-sharing}\text{).} \end{cases}$$

2.3 Admission Control

We propose an admission control strategy that ensures the required capacity r_u of each admitted user u is satisfied. Let $N_{j,e}$ and $N_{j,s}$ denote the number of elastic and streaming users in segment j respectively, and define $N_j = N_{j,e} + N_{j,s}$. We further define the vectors $\mathbf{N}_e = (N_{1,e}, \ldots, N_{J,e})$ and $\mathbf{N}_s = (N_{1,s}, \ldots, N_{J,s})$ and let N_e and N_s be the total number of elastic and streaming users in the cell respectively. Let (β_j, γ_j) be the *minimum* transmission power required by an (elastic, streaming) user in segment j to sustain a capacity requirement of $(r_{j,e}, r_{j,s})$, respectively. Depending on the resource sharing mechanism employed, (β_j, γ_j) can be evaluated using Eq. (2).

Streaming users are always accommodated with exactly their required capacity, consuming a total power of

$$P_s(\mathbf{N}_s) = \sum_{j=1}^{J} N_{j,s}\gamma_j.$$

Hence, the capacity of elastic users must be achievable with power $P_e = P - P_s$. Since all elastic users receive an equal portion of the available power, we conclude that

$$N_e\beta_j \leq P_e,$$

must hold for all j with $N_{j,e} > 0$, or equivalently,

$$N_e\beta_j\mathbf{1}_{(N_{j,e}>0)} \leq P_e, \qquad \forall j. \tag{3}$$

The indicator function $\mathbf{1}_E$ equals 1 if expression E holds and is 0 otherwise. Note that the J conditions in (3) only limit the *total* number of elastic users N_e, but that the maximum number of users does depend on the entire vector

$\mathbf{N}_e$. Similarly, the fact that elastic users share power equally, together with the minimum power restrictions of both elastic and streaming users, imply that

$$N_e \beta_j \mathbf{1}_{(N_{j,e}>0)} + P_s(\mathbf{N}_s) \leq P, \quad \forall j. \tag{4}$$

Conditions (3) and (4) completely determine the admission policy: a newly-arrived user will be accepted only if the resulting system state, ($\mathbf{N}_e$, $\mathbf{N}_s$), satisfies all $2J$ conditions. Alternatively, these conditions may be formulated in terms of the *required power* for each user type. Similar to $P_s(\mathbf{N}_s)$, we determine the transmission power required by elastic requests:

$$P_e(\mathbf{N}_e, \mathbf{N}_s) \equiv N_e \times \max_{j:N_{j,e}>0} \{\beta_j\} .$$

Note that this expression depends on the system state, ($\mathbf{N}_e$, $\mathbf{N}_s$).

Our admission control policy for streaming users can now be formulated as follows: a newly-arrived streaming user in segment i will be admitted if

$$P_e(\mathbf{N}_e, \mathbf{N}_s + \mathbf{e}_i) + P_s(\mathbf{N}_s + \mathbf{e}_i) \leq P, \tag{5}$$

where the vector $\mathbf{e}_i$ has its i^{th} component equal to 1 and all other components are 0.

For elastic users, we must incorporate the power reservation restrictions as well. If we define

$$\overline{P}_s(\mathbf{N}_s) \equiv \max \{P_s, P_s(\mathbf{N}_s)\} ,$$

then a newly-arrived elastic user in segment i will be admitted if

$$P_e(\mathbf{N}_e + \mathbf{e}_i, \mathbf{N}_s) + \overline{P}_s(\mathbf{N}_s) \leq P \tag{6}$$

While the admission control proposed in [7] is similar, it results in equal blocking probabilities for both types of traffic. Due to resource reservation in our case, the blocking probabilities will depend on both the user type and location.

2.4 Rate allocation

While streaming users are accommodated with exactly their required capacities, i.e., $r_{j,s}$ in segment j, the date rates allocated to elastic users depend on the number, type and location of other users. The available transmission power for elastic flows is $P - P_s(\mathbf{N}_s)$, of which all active elastic users receive an equal portion regardless of their location. Using Eq. (1), an elastic user in segment j attains a data rate

$$r_{j,e}(N_e, \mathbf{N}_s) = \frac{W \frac{P - P_s(\mathbf{N}_s)}{N_e}}{\epsilon_e[\alpha P^a_{j,e} + \frac{\eta + I^r_j}{\Gamma_j}]}, \tag{7}$$

where $P^a_{j,e}$ is the total intra-cell interference experienced by that user, which depends on the resource sharing mechanism. Accordingly, the *departure rate* of elastic users in segment j is given by:

$$\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s) = \frac{N_{j,e} r_{j,e}(N_e, \mathbf{N}_s)}{f_{j,e}}. \tag{8}$$

3 Analysis

Since exact analysis of our model is non-tractable in general and computationally involved when assuming exponentially distributed holding times and file sizes [11,14], we develop an approximation based on time-scale decomposition to evaluate the cell performance and assess the accuracy through comparison with simulation.

3.1 Quasi-stationary Approximation

We develop a quasi-stationary approximation for elastic flows, to be denoted $\mathbf{A}(\mathbf{Q}, \mathbf{J})$, where we assume that the dynamics of streaming flows take place on a much slower time scale than those of elastic flows. More specifically, we assume that elastic traffic practically reaches statistical equilibrium while the number of active streaming calls remains unchanged, i.e., we assume that all $\mu_{j,s}$ and $\lambda_{j,s}$ are much smaller than any of the quantities $1/f_{j,e}$ and $\lambda_{j,e}$. This assumption is reasonable when we consider the combination of voice calls (streaming) and web-browsing or email (elastic) applications. Under this assumption, the dynamics of elastic flows can be studied by fixing the number of streaming flows in each segment, i.e., we fix the vector $\mathbf{N}_s \equiv \mathbf{n}_s$.

3.1.1 Conditional distribution for elastic traffic

We construct an approximation assuming that the number of active elastic flows *instantaneously* reaches a new statistical equilibrium whenever $\mathbf{N}_s$

changes. For fixed $\mathbf{N}_s \equiv \mathbf{n}_s$, the elastic traffic behaves like a J-class $M/G/1$ processor-sharing (PS) queue with admission control dictated by both (3) and (4). To avoid any confusion, we will append a superscript ^{Q} to all quantities (such as queue lengths and performance measures) resulting from this approximation.

For general service requirement distributions of elastic users and an admission region of the type $\sum_j N_{j,e}^Q \leq M$, the steady state distribution of the numbers of jobs in each segment was shown to be a multivariate geometric distribution [6]. This can be shown to imply the same stationary distribution (up to a multiplicative constant) for the elastic users under the quasi-stationary assumption. For phase-type distributions, this can be proved formally by taking M large enough so that the set of allowable states (3) and (4) can be included. The joint process of queue lengths and service phases is reversible, so that state-space truncation does not destroy detailed balance and one can obtain the stationary distribution of the restricted process by re-normalization of the steady-state measure:

$$\begin{aligned}\mathbb{P}^Q(\mathbf{n}_e|\mathbf{n}_s) \equiv& \mathbb{P}(\mathbf{N}_e^Q = \mathbf{n}_e \mid \mathbf{N}_s^Q = \mathbf{n}_s) \\ =& c_e^Q(\mathbf{n}_s) n_e! \prod_{j=1}^{J} \frac{\rho_{j,e}(\mathbf{n}_s)^{n_{j,e}}}{n_{j,e}!},\end{aligned} \tag{9}$$

where we have defined $\rho_{j,e}(\mathbf{n}_s) = \frac{\lambda_{j,e}}{\mu_{j,e}(\mathbf{n}_s)}$ and the normalization constant $c_e^Q(\mathbf{n}_s)$ is such that adding (9) over all $\mathbf{n}_e$ that satisfy (3) and (4) gives a total of 1, for each fixed $\mathbf{n}_s$. We finally recall that $n_e = \sum_{j=1}^J n_{j,e}$.

The conditional acceptance probability of newly-arrived elastic flows in segment i is

$$A_{i,e}^Q(\mathbf{n}_s) \equiv \mathbb{P}(P_e(\mathbf{N}_e^Q + \mathbf{e}_i, \mathbf{n}_s) \leq P - \overline{P}_s(\mathbf{n}_s) \mid \mathbf{N}_s^Q = \mathbf{n}_s).$$

From (9), we can also obtain the distribution of n_e by summing over all admitted combinations of $n_{j,e}$ such that $\sum_j n_{j,e} = n_e$. For the special case where $\beta_i \equiv \beta$ for all i – we call this *uniform admission control*[4] –, the distribution

[4] With uniform admission control, the minimum required power is the same for all users, irrespective of their locations. As a consequence, the minimum rates are determined by the locations: users further away from the base station or with larger inter-cell interference must compromise for a lower rate. Thus, although the admission policy is the same, users in different segments are distinguished by the achievable rates (as well as their own traffic distributions).

of n_e reduces to a simple truncated geometric distribution:

$$\mathbb{P}(N_e^Q = n_e \mid \mathbf{N}_s^Q = \mathbf{n}_s) = \frac{\rho_e(n_s)^{n_e}(1-\rho_e(\mathbf{n}_s))}{1-\rho_e(\mathbf{n}_s)^{n_e^{Q,\max}(\mathbf{n}_s)}}, \tag{10}$$

where $n_e^{Q,\max}(\mathbf{n}_s) = \lfloor \left(P - \overline{P}_s(\mathbf{n}_s)\right)/\beta \rfloor$ and $\rho_e(\mathbf{n}_s) = \frac{\lambda_e}{\mu_e(\mathbf{n}_s)}$ is the total departure rate of elastic requests from the cell.

We emphasize that, assuming quasi-stationarity, (9) and (10) are valid for general distributions of elastic requests [6]. Note that these expressions are insensitive to the file size distributions, other than through their means. As a further remark, we observe that stability is of no concern in our model, since $\mathbf{N}_e^Q$ is bounded due to the assumption that $r_{j,e} > 0$. Often, when applying a time-scale decomposition, the issue of stability is of considerable importance, giving rise to an additional assumption commonly referred to as *uniform stability* [5].

Remark 1 *According to Eq. (8), the departure rate of elastic requests depends on the system state, ($\boldsymbol{n}_e$, $\boldsymbol{n}_s$). However, to apply Eq. (9) and (10), the departure rate can depend on the system state through $\boldsymbol{n}_s$ only. We illustrate how this can be achieved with various resource sharing mechanisms in Section 4.*

3.1.2 Unconditional marginal distributions

Next, we consider the dynamics of streaming flows. When $\mathbf{N}_s^Q$=$\mathbf{n}_s$, streaming flows depart at a rate $\sum_j n_{j,s}\mu_{j,s}$. When a new streaming flow arrives in segment i, due to admission control, it is either accepted or blocked. Under our approximation assumptions, the probability of acceptance in segment i, $A_{i,s}^Q(\mathbf{n}_s)$, is given by:

$$\mathbb{P}\left(P_e(\mathbf{N}_e^Q, \mathbf{n}_s + \mathbf{e}_i) \leq P - P_s(\mathbf{n}_s + \mathbf{e}_i) \mid \mathbf{N}_s^Q = \mathbf{n}_s\right).$$

Hence, the effective arrival rate of streaming flows in segment i, $\Lambda_{i,s}^Q(\mathbf{n}_s)$, is given as follows:

$$\Lambda_{i,s}^Q(\mathbf{n}_s) = \lambda_{i,s} A_{i,s}^Q(\mathbf{n}_s).$$

As a side remark, note that $A_{i,s}^Q(\mathbf{n}_s) = 1$ if $P_s(\mathbf{n}_s + \mathbf{e}_i) \leq P_s$, since the admission control on elastic flows ensures that $N_e^Q \beta_j \mathbf{1}_{(N_{j,e}>0)} \leq P - P_s$ for all j.

In general, there is no closed-form expression for the equilibrium distribution of $\mathbf{N}_s^Q$ and we must assume exponential or phase-type holding time distributions and resort to standard methods to (numerically) solve the equilibrium distribution of a finite-state Markov process. Note that the dimension of this process $\mathbf{N}_s^Q$ is much smaller than that of the original process $(\mathbf{N}_e, \mathbf{N}_s)$: the component $\mathbf{N}_e$ is "eliminated" in the approximation. However, if we apply *uniform* admission control for streaming traffic by taking $\gamma_j \equiv \gamma$ independent of j [see the earlier Footnote 4], then $A_{i,s}^Q(\mathbf{n}_s) \equiv A_s^Q(n_s)$ is independent of i and depends on $\mathbf{n}_s$ only through the total number of streaming flows. $\mathbf{N}_s^Q$ can then be shown to be *balanced* [4] and can be reduced to the framework of [6]. It follows that, for *arbitrary* holding time distributions of streaming flows, and $0 \leq n_s \leq n_s^{\max} = \lfloor \frac{P}{\gamma} \rfloor$:

$$\mathbb{P}(\mathbf{N}_s^Q = \mathbf{n}_s) = c_s^Q \prod_{k=0}^{n_s - 1} A_s^Q(k) \prod_{j=1}^{J} \frac{(\rho_{j,s})^{n_{j,s}}}{n_{j,s}!}, \tag{11}$$

with $\rho_{j,s} = \lambda_{j,s}/\mu_{j,s}$ and $c_s^Q = P(N_s^Q = 0)$ can be determined by normalizing (11) to a probability distribution. Letting $\rho_s = \sum_j \rho_{j,s}$, we further obtain the distribution of the total number of active streaming flows (still for uniform admission control):

$$\mathbb{P}(N_s^Q = n_s) = c_s^Q \frac{(\rho_s)^{n_s}}{n_s!} \prod_{k=0}^{n_s - 1} A_s^Q(k), \tag{12}$$

which in this case results again in a simple expression for the normalizing constant:

$$c_s^Q = \left(\sum_{n_s=0}^{n_s^{\max}} \frac{(\rho_s)^{n_s}}{n_s!} \prod_{k=0}^{n_s - 1} A_s^Q(k) \right)^{-1}.$$

To conclude this section, we now calculate several relevant performance measures (not restricting anymore to uniform admission control) by un-conditioning on $\mathbf{N}_s^Q$. In general, the unconditional distribution for the number of elastic users is

$$\mathbb{P}(\mathbf{N}_e^Q = \mathbf{n}_e) = \sum_{\mathbf{n}_s} \mathbb{P}^Q(\mathbf{n}_e \mid \mathbf{n}_s) \mathbb{P}(\mathbf{N}_s^Q = \mathbf{n}_s). \tag{13}$$

The unconditional blocking probabilities in segment i are

$$p_{i,s}^Q = \sum_{\mathbf{n}_s} (1 - A_{i,s}^Q(\mathbf{n}_s)) \mathbb{P}(\mathbf{N}_s^Q = \mathbf{n}_s), \tag{14}$$

for streaming flows; similarly, for elastic flows, we have:

$$p^Q_{i,e} = \sum_{\mathbf{n}_s}(1 - A^Q_{i,e}(\mathbf{n}_s))\mathbb{P}(\mathbf{N}^Q_s = \mathbf{n}_s).$$

3.2 Fluid Approximation

The fluid approximation (from the perspective of elastic flows), denoted by $\mathbf{A(F, J)}$, complements the quasi-stationary approximation: We now assume that the dynamics of elastic flows are much slower than those of streaming flows, i.e., the $\lambda_{j,s}$ and $\mu_{j,s}$ are much larger than the $\lambda_{j,e}$ and $1/f_{j,e}$. This assumption is valid when we consider the combination of voice calls (streaming) and large file transfer (elastic) applications. The dynamics of streaming flows can then be studied by fixing the number of elastic flows in each segment. This approximation will be reflected in the notations by adding a superscript ^{F}. Similar to $\mathbf{A(Q,\ J)}$, we will construct an approximating $2J$-dimensional process under the assumption that $\mathbf{N}^F_s$ immediately reaches steady state, whenever $\mathbf{N}^F_e$ changes.

3.2.1 Conditional distribution of streaming traffic

We fix the number of elastic flows in each segment: $\mathbf{N}^F_e = \mathbf{n}_e$. Under the "fluid" approximation assumption, we can model the streaming flows as a J-class Erlang-loss queue with finite capacity:

$$\begin{aligned}\mathbb{P}^F(\mathbf{n}_s \mid \mathbf{n}_e) &\equiv \mathbb{P}(\mathbf{N}^F_s = \mathbf{n}_s \mid \mathbf{N}^F_e = \mathbf{n}_e)\\ &= c^F_s(\mathbf{n}_e)\prod_{j=1}^{J}\frac{\rho_{j,s}^{n_{j,s}}}{n_{j,s}!},\end{aligned} \tag{15}$$

where $\rho_{j,s} = \frac{\lambda_{j,s}}{\mu_{j,s}}$. As before, we emphasize that the above expression depends on the holding time distribution only through its mean. The constant $c^F_s(\mathbf{n}_e)$ can again be determined by requiring that (15) adds to 1 when summing (for fixed $\mathbf{n}_e$) over all $\mathbf{n}_s$ such that $P_e(\mathbf{n}_e, \mathbf{n}_s) + P_s(\mathbf{n}_s) \leq P$. For uniform admission control, i.e., $\gamma_i \equiv \gamma$ independent of i, this results in an elegant form of the distribution for the *total* number of streaming users (a truncated Poisson distribution), as well as for the normalization constant:

$$\mathbb{P}(N^F_s = n_s \mid \mathbf{N}^F_e = \mathbf{n}_e) = c^F_s(\mathbf{n}_e)\frac{(\rho_s)^{n_s}}{n_s!},$$

and

$$c_s^F(\mathbf{n}_e) = (\sum_{k=0}^{n_s^{F,\max}(\mathbf{n}_e)} \frac{(\rho_s)^k}{k!})^{-1},$$

where $n_s^{F,\max}(\mathbf{n}_e)$ is the maximum number of streaming users for which $P_e(\mathbf{n}_e, \mathbf{n}_s) + P_s(\mathbf{n}_s) \leq P$.

3.2.2 Unconditional marginal distributions

Next, we consider the dynamics of elastic flows. When $\mathbf{N}_e^F = \mathbf{n}_e > 0$, elastic flows in segment j (if any) experience an average data rate (recall that n_e is the sum over all components of the vector $\mathbf{n}_e$):

$$\begin{aligned} \overline{r}_{j,e}(\mathbf{n}_e) &\equiv \mathbb{E}[r_{j,e}(n_e, \mathbf{N}_s^F) \mid \mathbf{N}_e^F = \mathbf{n}_e] \\ &= \sum_{\mathbf{n}_s} r_{j,e}(n_e, \mathbf{n}_s)\, \mathbb{P}^F(\mathbf{n}_s \mid \mathbf{n}_e), \end{aligned} \tag{16}$$

where the summation is taken over all $\mathbf{n}_s$ for which $P_e(\mathbf{n}_e, \mathbf{n}_s) + P_s(\mathbf{n}_s) \leq P$. The (state-dependent) departure rate of elastic flows from segment j is

$$n_{j,e}\overline{r}_{j,e}(\mathbf{n}_e)/f_{j,e}.$$

In order to fully describe the dynamics of the elastic flows, we now determine the arrival rate, which also depends on the state $\mathbf{n}_e$ because of the employed admission control. Under our approximation assumptions, the probability of acceptance in segment i is given by:

$$A_{i,e}^F(\mathbf{n}_e) \equiv \mathbb{P}(\overline{P}_s(\mathbf{N}_s^F) + \overline{P}_e(\mathbf{n}_e + \mathbf{e}_i, \mathbf{N}_s^F) \leq P \mid \mathbf{N}_e^F = \mathbf{n}_e),$$

and, consequently, the effective arrival rate of elastic flows in segment i is

$$\Lambda_{i,e}^F(\mathbf{n}_e) \equiv \lambda_{i,e} A_{i,e}^F(\mathbf{n}_e).$$

As for the quasi-stationary approximation, in general, there is no closed-form expression for the distribution of $\mathbf{N}_e^F$. However, under additional assumptions, $\mathbf{N}_e^F$ is *balanced* [4]. This is the case, for example, if we assume perfectly orthogonal codes ($\alpha = 0$) **and** apply *uniform* admission control for elastic traffic by taking $\beta_j \equiv \beta$ independent of j. In this case, the dynamics of $\mathbf{N}_s^F$ depends on $\mathbf{N}_e^F$ only through the total number of elastic users N_e, so if we define

$$h(\mathbf{n}_s) = P - P_s(\mathbf{n}_s),$$

then, we can write

$$\mathbb{E}[h(\mathbf{N}_s^F) \mid \mathbf{N}_e^F = \mathbf{n}_e] = \mathbb{E}[h(\mathbf{N}_s^F) \mid N_e^F = n_e] \equiv g(n_e).$$

If we further define

$$\nu_j = \frac{W\Gamma_j}{\epsilon_{j,e}[\eta + I_j^r]},$$

then, from Eq. (7) and (16), we obtain

$$\overline{r}_{j,e}(\mathbf{n}_e) \equiv \overline{r}_{j,e}(n_e) = \frac{\nu_j\, g(n_e)}{n_e}.$$

Furthermore, $A_{i,e}^F(\mathbf{n}_e)$ is independent of i and depends on $\mathbf{n}_e$ only through the total number of elastic flows, i.e., $A_{i,e}^F(\mathbf{n}_e) \equiv A_e^F(n_e)$.

It follows that, for *arbitrary* file size distributions, and $0 \leq n_e \leq n_e^{\max} = \lfloor \frac{P_e}{\beta} \rfloor$:

$$\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e) = c_e^F \prod_{k=1}^{n_e} \frac{k\, A_e^F(k-1)}{g(k)} \prod_{j=1}^{J} \left(\frac{\rho_{j,e}}{\nu_j}\right)^{n_{j,e}}, \tag{17}$$

with $\rho_{j,e} = \lambda_{j,e}\, f_{j,e}$ and $c_e^F = P(N_e^F = 0)$ can be determined after normalization. We further obtain the distribution of the *total* number of file transmissions (still for uniform admission control and $\alpha = 0$):

$$\mathbb{P}(N_e^F = n_e) = c_e^F \left(\sum_j \frac{\rho_{j,e}}{\nu_j}\right)^{n_e} \prod_{k=1}^{n_e} \frac{k\, A_e^F(k-1)}{g(k)}, \tag{18}$$

leading to a simple expression for the normalizing constant as before:

$$c_e^F = \left(\sum_{n_e=0}^{n_e^{\max}} \left(\sum_j \frac{\rho_{j,e}}{\nu_j}\right)^{n_e} \prod_{k=1}^{n_e} \frac{k\, A_e^F(k-1)}{g(k)}\right)^{-1}.$$

Remark 2 *If the codes are not perfectly orthogonal ($\alpha > 0$), we can still apply the above analysis in case the background noise and inter-cell interference are negligible ($\eta_j + I_j^r << \alpha P_{j,e}^a \Gamma_j$) by choosing*

$$\nu_j = \frac{W}{\alpha \epsilon_{j,e} P_{j,e}^a}.$$

We conclude this section with the following unconditional performance measures:

$$\mathbb{P}(\mathbf{N}_s^F = \mathbf{n}_s) = \sum_{\mathbf{n}_e} \mathbb{P}^F(\mathbf{n}_s \mid \mathbf{n}_e)\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e). \tag{19}$$

The unconditional blocking probabilities in segment i are

$$p_{i,e}^F = \sum_{\mathbf{n}_e}(1 - A_{i,e}^F(\mathbf{n}_e))\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e), \tag{20}$$

and

$$p_{i,s}^F = \sum_{\mathbf{n}_e}(1 - A_{i,s}^F(\mathbf{n}_e))\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e).$$

4 Specific models

Based on the abstraction of the model in terms of (a) location-dependence and (b) resource sharing mechanism, we can define variants of the base station model, where our objective is to evaluate the performance gain achieved with location-awareness and time-sharing so as to ascertain if the added processing complexity at the base station is justified.

4.1 CDMA model

In this model, the base station does not maintain location information of each user (i.e., J=1), and transmits *simultaneously* to all users. As such, each user u is distinguished only in terms of its type (i.e., streaming ($u \equiv s$) vs elastic ($u \equiv e$)). Hence, the system state ($\mathbf{N}_e$, $\mathbf{N}_s$) reduces to (N_e, N_s) and we can drop the subscript j from the notations. In addition, due to simultaneous transmission to *all* users, $P_u^a = P$-P_u, and Eq. (1) can be written as follows:

$$R_u \leq \frac{WP_u}{\epsilon_u[\alpha(P - P_u) + \frac{\eta + I_u^r}{\Gamma_u}]},$$

which can be re-written as follows:

$$P_u \geq \frac{R_u\epsilon(\frac{\eta + I_{max}^r}{\Gamma_{min}} + \alpha P)}{W + \alpha\epsilon R_u}. \tag{21}$$

For linear and hexagonal networks and typical propagation models, $\Gamma_u = \Gamma_{min}$ and $I_u^r = I_{max}^r$ when user u is located at the *edge* of the cell.

Accordingly, we can define the minimum power required by an (elastic, streaming) user to sustain capacity requirements of $(r_e,\, r_s)$ as follows:

$$\beta = \frac{r_e \epsilon(\frac{\eta + I_{max}^r}{\Gamma_{min}} + \alpha P)}{W + \alpha \epsilon r_e},$$
$$\gamma = \frac{r_s \epsilon(\frac{\eta + I_{max}^r}{\Gamma_{min}} + \alpha P)}{W + \alpha \epsilon r_s}. \tag{22}$$

Eq. (3) and (4) can be written as follows:

$$N_e \beta \leq P_e,$$
$$N_e \beta + N_s \gamma \leq P. \tag{23}$$

By substituting Eq. (22) into Eq. (23) and defining $r = \max(r_e,\, r_s)$, we obtain the following:

$$N_e r_e + N_s r_s \leq \frac{P(W + \alpha \epsilon r)}{\epsilon(\frac{\eta + I_{max}^r}{\Gamma_{min}} + \alpha P)}. \tag{24}$$

4.1.1 Equivalent wired link analysis

According to Eq. (24), if we define $c \equiv \frac{P(W+\alpha\epsilon r)}{\epsilon(\frac{\eta+I_{max}^r}{\Gamma_{min}}+\alpha P)}$, then the downlink transmission scenario in the CDMA model can be approximated by a wired link with capacity c shared amongst streaming and elastic requests, where $c_s = \frac{P_s}{P} c$ is reserved for streaming requests. Details of the analysis of this model based on the quasi-stationary and fluid approximations (denoted by $\mathbf{A(Q)}$ and $\mathbf{A(F)}$ respectively) can be found in [10].

4.1.2 General analysis

Referring to Remark 1, to apply the quasi-stationary approximation, the departure rate of elastic users from the cell should only depend on the system state $(\mathbf{N}_s,\, \mathbf{N}_e)$ through N_s. From Eq. (8), we have the following expression:

$$\mu_e(N_e, N_s) = \frac{W[P - P_s(N_s)]}{\epsilon_e[\alpha(P - \frac{P - P_s(N_s)}{N_e}) + \frac{\eta + I_{max}^r}{\Gamma_{min}}]}.$$

It is not straightforward to obtain an approximation, $\mu_e(N_s)$, for $\mu_e(N_e,\, N_s)$. On the other hand, the fluid approximation developed in Section 3.2 can be applied for this model.

4.2 HSDPA model

Based on our definition in Section 1, each streaming (elastic) user u has a fixed (minimum) capacity requirement, denoted by r_u. According to our resource reservation policy, while each streaming user transmits at *fixed* rate r_u, the transmission rate of an elastic user u, R_u ($\geq r_u$), depends on the resource unclaimed by streaming traffic, given by P-$P_s(\mathbf{N}_s)$. From Eq. (1), R_u can be maximized by minimizing P_u^a. One approach to do so is to apply time-sharing amongst elastic users.

If we aggregate all elastic users, the resource sharing mechanism is such that the base station transmits using (almost-)orthogonal codes to all users, where the aggregate elastic user may be assigned several codes. Within the aggregate user, elastic users sharing the same code are served in a time-slotted fashion so that they do not interfere with one another, but only with elastic users using different codes and streaming traffic. This resource sharing mode is similar to UMTS / HSDPA, where up to $N_c = 4$ codes can be shared amongst data/elastic users. We assume that $N_c = 1$ in our study.

For the HSDPA model defined here, the base station has the capability to maintain *quantized* location information (at different levels of granularity) and also supports time-sharing resource sharing amongst elastic requests as described above.

4.2.1 Impact on admission control

According to the above resource sharing policy, the received signal at each streaming user u in segment j is interfered by simultaneous transmissions to all other users, i.e., $P_u^a = P$-P_u and from (2) we obtain

$$\gamma_j = \frac{r_{j,s}\epsilon_{j,s}[\alpha P\Gamma_j + \eta + I_j^r]}{(W + \alpha r_{j,s}\epsilon_{j,s})\Gamma_j}.$$

For an elastic user u in segment j, we have $P_u^a = P_s(\mathbf{N}_s)$ since its received signal is only interfered by streaming users. Hence, the power required by an elastic user in segment j to sustain its capacity requirement, $r_{j,e}$, depends on the number and location of streaming users as follows:

$$\beta_j(\mathbf{N}_s) = \frac{r_{j,e}\epsilon_{j,e}[\alpha P_s(\mathbf{N}_s)\Gamma_j + \eta + I_j^r]}{W\Gamma_j}.$$

The admission control scheme is such that a newly-arrived user is blocked only if accepting it would violate either the static reservation policy or the minimum power requirement of any user. At any time, streaming traffic can claim a portion P_s of the total power P. Therefore, the power required by an elastic user in segment j is given by:

$$\beta_j \equiv \beta_j(\mathbf{N}_s) = \frac{r_{j,e}\epsilon_{j,e}[\alpha P_s(\mathbf{N}_s)\Gamma_j + \eta + I_j^r]}{W\Gamma_j}.$$

4.2.2 Impact on rate allocation

Using Eq. (7) and (8), with time-sharing amongst elastic users, the departure rate of elastic users in segment j is given by:

$$\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s) = \frac{N_{j,e}W[P - P_s(\mathbf{N}_s)]}{f_{j,e}N_e\epsilon_e[\alpha P_s(\mathbf{N}_s) + \frac{\eta+I_j^r}{\Gamma_j}]}. \qquad (25)$$

Since $\frac{N_{j,e}}{N_e} \leq 1$, we have the following:

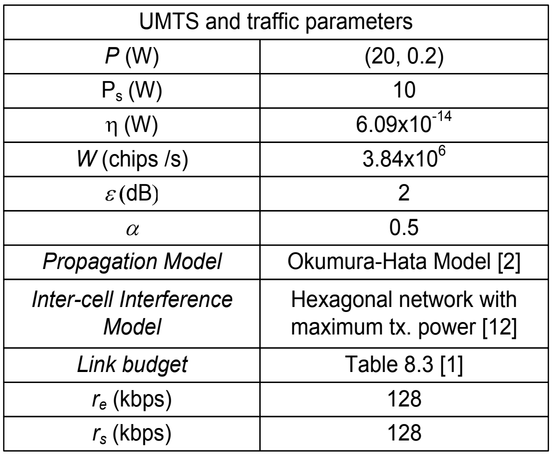
$$\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s) \leq \frac{W[P - P_s(\mathbf{N}_s)]}{f_{j,e}\epsilon_e[\alpha P_s(\mathbf{N}_s) + \frac{\eta+I_j^r}{\Gamma_j}]} \equiv \mu_{j,e}(\mathbf{N}_s). \qquad (26)$$

Referring to Remark 1, to apply the quasi-stationary approximation, it is necessary to remove the dependence of $\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s)$ on $\mathbf{N}_e$ in Eq. (25). This can be achieved by approximating $\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s)$ with $\mu_{j,e}(\mathbf{N}_s)$; this approximation is exact when location information is unknown (i.e., J=1).

As with the basic CDMA model, the fluid approximation can be applied for this model. Further details on the analysis of this model can be found in [8].

5 Performance Evaluation

We consider a single UMTS cell whose radius, δ_J, is computed using the reference link budget given in Table 8.3 of [1] and the Okumura-Hata propagation model [2] for an urban macro cell. The inter-cell interference at each location



UMTS and traffic parameters	
P (W)	(20, 0.2)
P_s (W)	10
η (W)	6.09×10^{-14}
W (chips /s)	3.84×10^{6}
ε (dB)	2
α	0.5
Propagation Model	Okumura-Hata Model [2]
Inter-cell Interference Model	Hexagonal network with maximum tx. power [12]
Link budget	Table 8.3 [1]
r_e (kbps)	128
r_s (kbps)	128

Table 1
UMTS cell and traffic parameters for performance evaluation.

within the cell is computed based on the conservative approximation for a hexagonal network [12].

Elastic (streaming) users arrive at the cell according to a Poisson process at rates λ_e (λ_s), capacity requirement r_e (r_s), target energy-to-noise ratio ϵ_e (ϵ_s), mean file size f_e (holding time $\frac{1}{\mu_s}$) and are assumed to be uniformly distributed over the cell. In addition to the mean number of users, (E[N_e], E[N_s]), and blocking probabilities, (p_e, p_s), for each class of traffic, we define the *stretch*, S_e, for each admitted elastic user by normalizing the expected residence time, $E[R_e]$, by the mean file size, f_e, i.e., $S_e = \frac{E[R_e]}{f_e} = \frac{E[N_e]}{\lambda_e(1-p_e)}$ (cf. Little's Theorem). A summary of the cell and traffic parameters is given in Table 1.

In [10] and [8], through simulations, we have demonstrated that the user-performance obtained with the basic CDMA and HSDPA model is almost insensitive to the actual distribution of the traffic parameters. This justifies the application of the approximation techniques we develop, which depend on the traffic parameter distribution only through the mean values. In addition, we also demonstrated the accuracy of the approximations, particularly for the extreme (quasi-stationary and fluid) traffic regimes.

Here, we focus on the comparison of the basic CDMA model and the HSDPA model for the base station based on simulation as well as the approximations. Unless otherwise stated, we assume that (d_s, s_e) are exponentially distributed with mean $\frac{1}{\mu_s}$ and f_e respectively.

5.1 Simulation Procedure

We develop a simulation program for our model by considering arrival / departure events of traffic requests (elastic or streaming). Each simulation scenario is defined according to the following procedure:

1. Fix the level of location quantization, J:
   J=1 : no location information;
   J>1 : location information available.
2. Fix the total offered traffic by choosing the *loading factor*, $l > 0$, where $u_e + u_s = l\ c$,
   $u_e = \lambda_e f_e$ and $u_s = \frac{\lambda_s r_s}{\mu_s}$;
3. For each l, fix the traffic *mix*, $\frac{u_e}{lc}$, by choosing u_e,
   $0 \leq u_e \leq l\ c$;
4. For each traffic mix, select (λ_e, λ_s) to fit one of the following traffic regimes:
   a. Quasi-stationary Regime (**S(Q,J)**, cf. Section 3.1);
   b. Fluid Regime (**S(F,J)**, cf. Section 3.2);
   c. Neutral Regime (**S(N,J)**, fits neither a. nor b.)

We generate 5 sets of simulation results for each scenario, for which the sample mean for each performance metric is computed and used for performance comparison.

5.2 Impact of time-sharing (J=1)

We begin by investigating the performance gain achieved with time-sharing by comparing the performance obtained for the basic CDMA model and HSDPA model without location-awareness for various traffic regimes.

5.2.1 Quasi-stationary regime

We plot (p_e, p_s) and (E[N_e], S_e) as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$ in Fig. 1 and 2 respectively. We note that, since it is not straightforward to apply the quasi-stationary approximation to the CDMA model (cf. Section 4.1.2), we utilize the equivalent wired link analysis to obtain the corresponding quasi-stationary approximation.

Based on the simulation results, we observe a performance gain achieved as a result of time-sharing in terms of reduced blocking probabilities, queue length and sojourn time. This gain is expected since, for a given number of streaming requests, time-sharing amongst elastic flows reduces the intra-cell interference power experienced by each elastic user, thereby increasing the data rate per elastic user. This gain is marginal when elastic load is low, since the additional interference experienced by an elastic user due to other elastic users (without time-sharing) is insignificant.

In terms of the accuracy of approximations, we observe that the performance obtained with the HSDPA model is well-tracked by the corresponding approximation; on the other hand, the equivalent wired link analysis results in overly conservative estimates of the performance for the CDMA model.

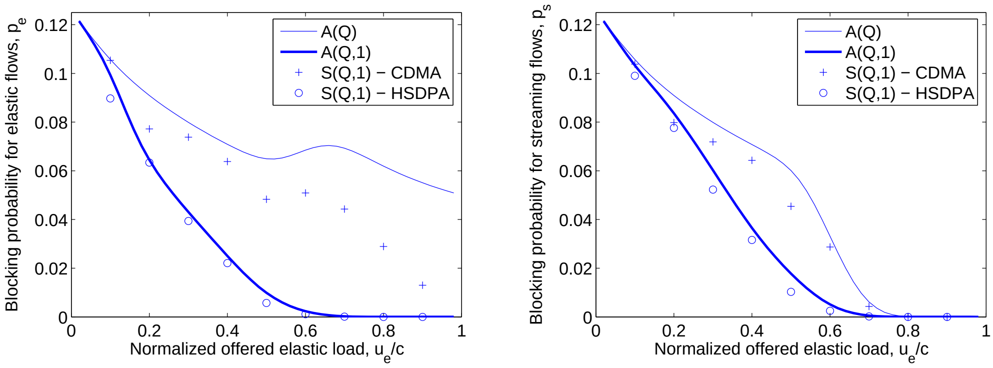


Fig. 1. Blocking probability for elastic (left) and streaming requests (right) vs normalized offered elastic load obtained for quasi-stationary regime (J=1).

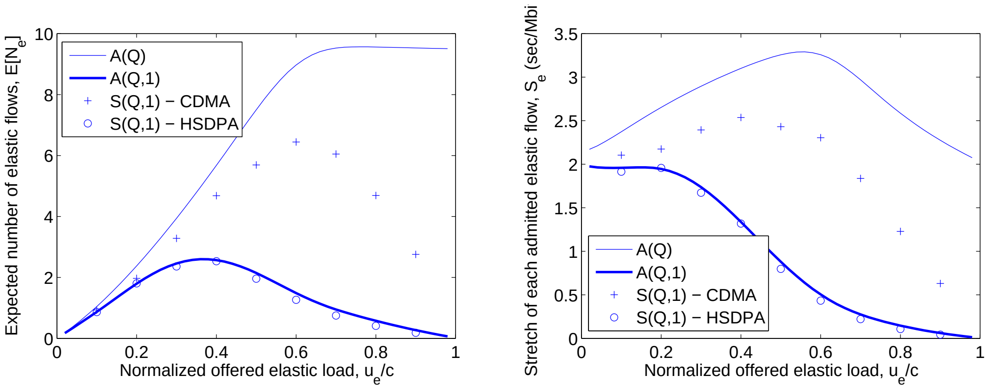


Fig. 2. Number of active elastic requests (left) and stretch of each admitted elastic request vs normalized offered elastic load obtained for quasi-stationary regime (J=1).

5.2.2 Fluid regime

We plot (p_e, p_s) and (E[N_e], S_e) as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$ in Fig. 3 and 4 respectively.

As with the quasi-stationary regime, we observe a performance gain achieved as a result of time-sharing in terms of reduced blocking probabilities, queue length and sojourn time. In terms of the accuracy of approximations, the blocking performance obtained with both models is well-tracked by the corresponding approximations. However, the approximations achieved more optimistic estimates of the queue length and sojourn time of elastic users.

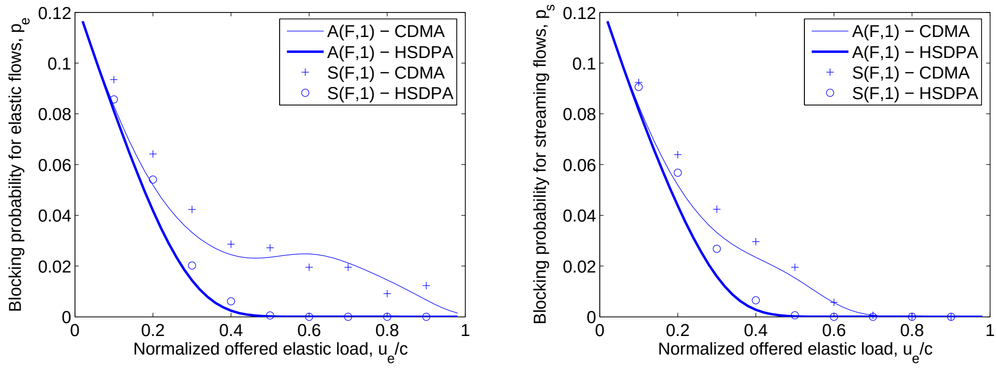


Fig. 3. Blocking probability for elastic (left) and streaming requests (right) vs normalized offered elastic load obtained for fluid regime (J=1).

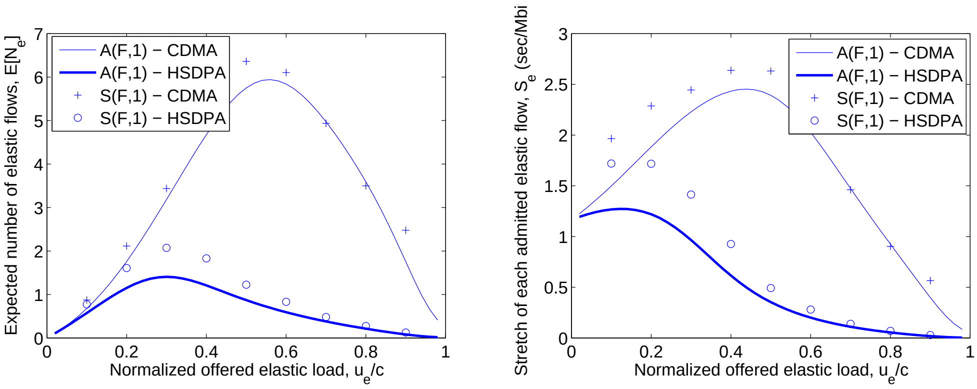


Fig. 4. Number of active elastic requests (left) and stretch of each admitted elastic request vs normalized offered elastic load obtained for fluid regime (J=1).

5.2.3 Neutral regime

We plot (p_e, p_s) and (E[N_e], S_e) as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$ in Fig. 5 and 6 respectively.

As with the extreme traffic regime, we observe a performance gain achieved as a result of time-sharing in terms of reduced blocking probabilities, queue length and sojourn time. For each performance metric, we note that the quasi-stationary (fluid) approximation upper (lower) bounds the performance obtained in the neutral traffic regime, where a tighter bound is obtained with the HSDPA model.

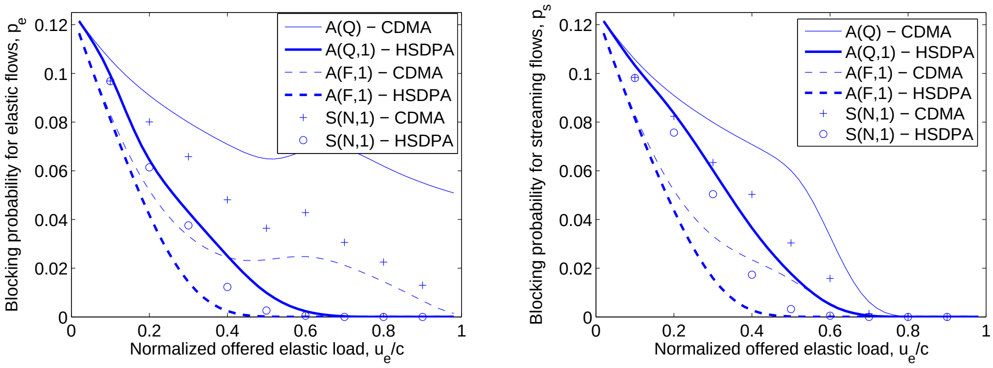


Fig. 5. Blocking probability for elastic (left) and streaming requests (right) vs normalized offered elastic load obtained for neutral regime (J=1).

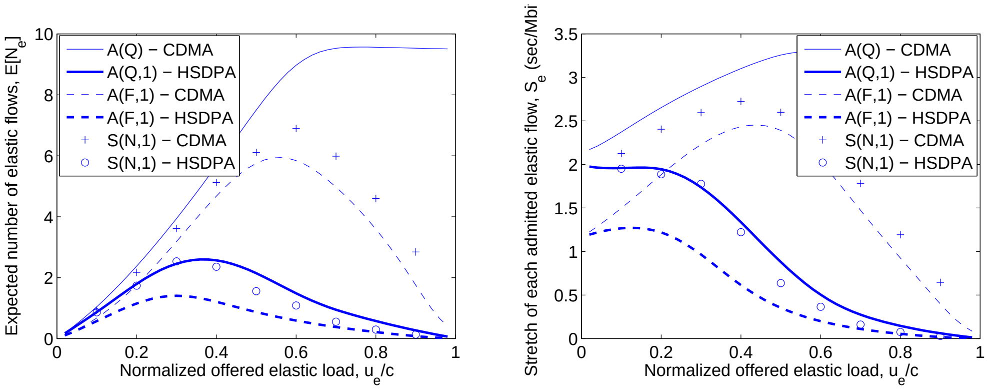


Fig. 6. Number of active elastic requests (left) and stretch of each admitted elastic request vs normalized offered elastic load obtained for neutral regime (J=1).

5.3 Impact of location-awareness (HSDPA model)

Next, we investigate the performance gain achieved with location-awareness by comparing the performance obtained for the HSDPA model with various degrees of location quantization, J. We define each segment j as the annulus between concentric rings of radius δ_{j-1} and δ_j such that $\delta_j = \frac{j}{J}\delta_J$, 1$\leq j \leq J$. Since user arrivals are uniformly distributed over the cell, their arrival rate in each ring j is $\lambda_j = \frac{\delta_j^2 - \delta_{j-1}^2}{\delta_J^2}\lambda$, where $\delta_0 = 0$.

5.3.1 P=20W

We plot (p_e, p_s) and (E[N_e], S_e) as a function of the traffic mix, $\frac{u_e}{c}$, 0$\leq u_e \leq c$, for **S(N,J)** in Fig. 7 and 8 respectively for J=1, 2 and ∞ (corresponding to the case of exact location information) for a neutral traffic regime. We observe that the cell performance obtained with simulation is lower bounded (well approximated) by **A(F,J=1)** (**A(Q,J=1)**), and that **S(N,J)** is almost invariant with the value of J. Hence, no significant performance gain is achieved through exploiting more accurate location information in this case, and therefore, the performance can be approximated using location-unaware approximations (J=1).

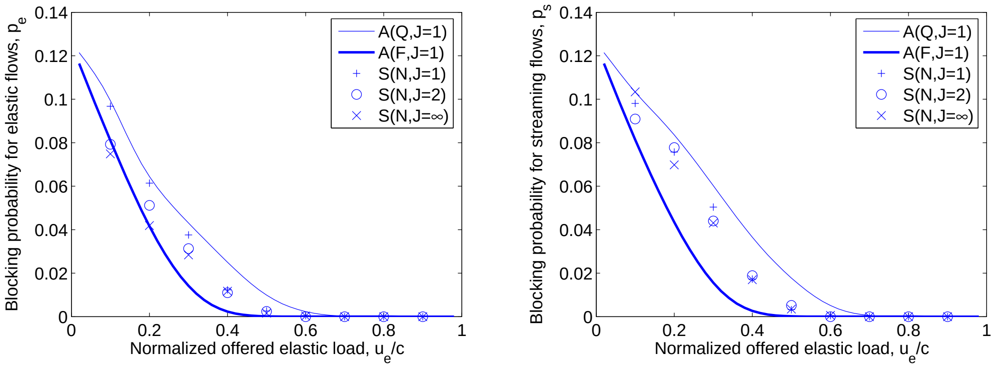


Fig. 7. Blocking probability for elastic (left) and streaming requests (right) vs normalized offered elastic load obtained with approximation and simulation for HSDPA model (J=1,2,∞, P=20W, neutral regime).

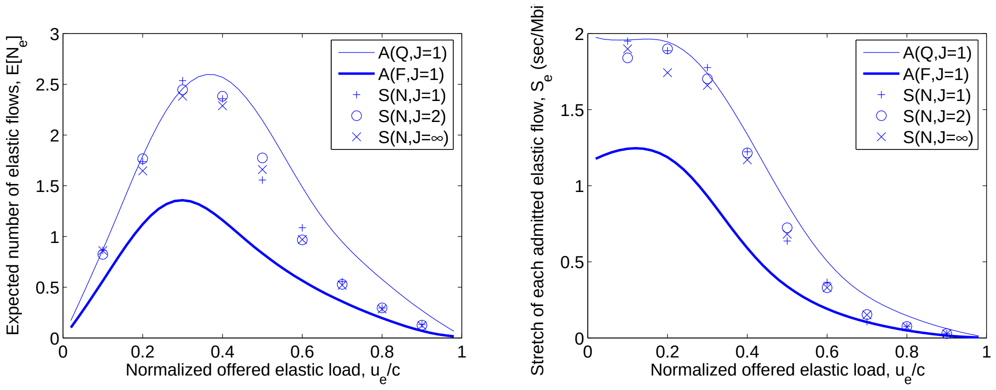


Fig. 8. Number of active elastic requests (left) and stretch of each admitted elastic request vs normalized offered elastic load obtained with approximation and simulation for HSDPA model (J=1,2,∞, P=20W, neutral regime).

5.3.2 P=0.2W

In order to demonstrate the performance gain with exploiting user location, we repeat the simulations for the case of P = 0.2W, and plot (p_e, p_s) and (E[N_e], S_e) as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$, for $\mathbf{S(F,J)}$ in Fig. 9 and 10 respectively. In this case, we note that as cell partitioning becomes finer (increasing J), the performance obtained with $\mathbf{S(F,J)}$ is improved significantly (e.g., reduced blocking and sojourn time).

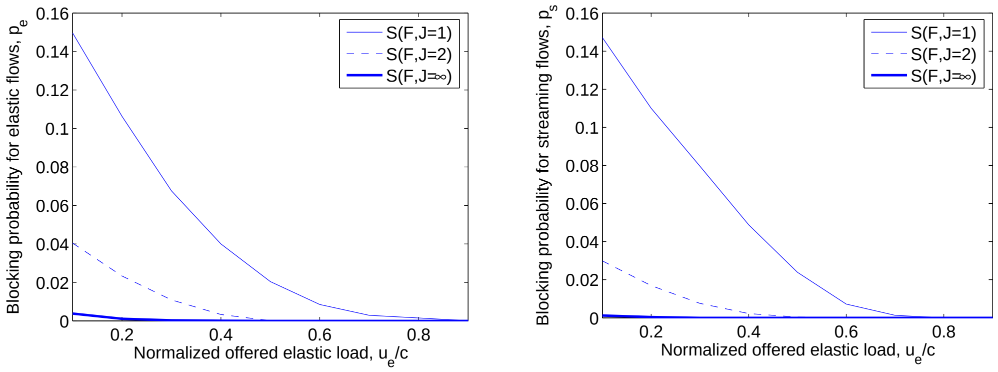


Fig. 9. Blocking probability for elastic (left) and streaming requests (right) vs normalized offered elastic load obtained with simulation for HSDPA model (J=1,2,∞, P=0.2W, fluid regime).

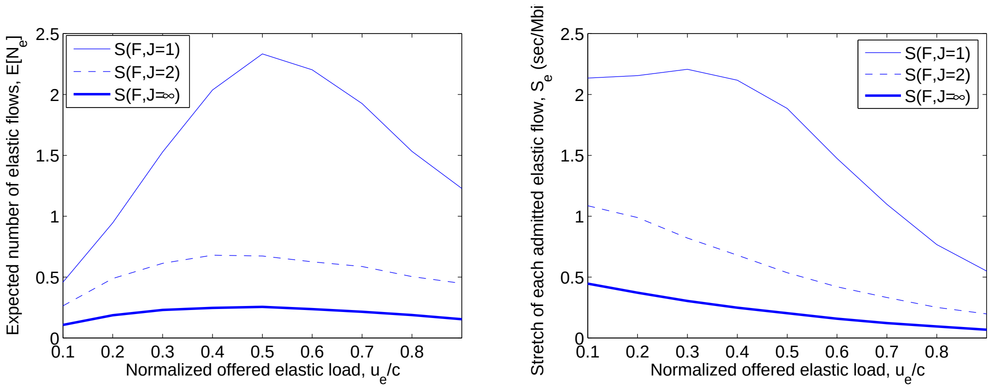


Fig. 10. Number of active elastic requests (left) and stretch of each admitted elastic request vs normalized offered elastic load obtained with simulation for HSDPA model (J=1,2,∞, P=0.2W, fluid regime).

6 Conclusions

Third generation wireless systems can simultaneously accommodate users carrying widely heterogeneous applications. Since resources are limited, particularly in the air interface, admission control is necessary to ensure that all active users are accommodated with sufficient bandwidth to meet their specific Quality of Service requirements. We propose a general traffic management framework that supports differentiated admission control, resource sharing and rate allocation strategies, such that users with stringent capacity requirements ("streaming traffic") are protected while sufficient capacity over longer time intervals to delay-tolerant users ("elastic traffic") is offered. This framework permits users within each type to be distinguished according to their distance from the base station through cell partitioning, and also supports a time-sharing resource sharing mode to improve rate allocation to elastic traffic while guaranteeing the capacity requirements of all users.

Since the exact analysis to evaluate the performance of such an integrated services system is non-tractable in general, we define extreme traffic regimes (quasi-stationary and fluid) for which time-scale decomposition can be applied to isolate the traffic streams, from which known results from fluid queueing models are used to approximate the performance for each user type. For the extreme traffic regimes, simulation results suggest that the performance is almost insensitive to traffic parameter distributions, and is well approximated by our proposed approximations. In addition, we also demonstrate the performance gain achieved by exploiting location information about each user, as well as applying time-sharing amongst elastic users to improve their rate allocation.

7 Acknowledgments

The support of Vodafone is gratefully acknowledged. This research is partially supported by the Dutch Bsik/BRICKS project and is performed within the framework of the European Network of Excellence Euro-NGI. Much of the research was performed while Adriana F. Gabor and Hwee-Pink Tan were affiliated with EURANDOM.

References

[1] H. Holma and A. Toskala, *WCDMA for UMTS, Radio access for third generation mobile communications.* John-Wiley and Sons, 2001.

[2] Y. Wang and T. Ottosson, “Cell search in W-CDMA,” *IEEE Journal on Selected Areas in Communications*, vol. 18, 2000, pp. 1470–1482.

[3] P. Key, L. Massoulié, A. Bain, and F. Kelly, “Fair internet traffic integration: network flow models and analysis,” *Annales des Telecommunications*, vol. 59, 2004, pp. 1338–1352.

[4] T. Bonald and A. Proutière, “Insensitive bandwidth sharing in data networks,” *Queueing Systems*, vol. 44, 2003, pp. 69–100.

[5] F. Delcoigne, A. Proutière, and G. Regnie, “Modeling integration of streaming and data traffic,” *Performance Evaluation*, vol. 55, 2004, pp. 185–209.

[6] J. W. Cohen, “The multiple phase service network with generalized processor sharing,” *Acta Informatica*, vol. 12, 1979, pp. 245–284.

[7] N. Benameur, S. B. Fredj, F. Delcoigne, S. Oueslati-Boulahia, and J. W. Roberts, “Integrated admission control for streaming and elastic traffic,” *Lecture Notes in Computer Science*, vol. 2156, 2001, pp. 69–81.

[8] R. Núñez-Queija and H. P. Tan, "Location-based admission control for differentiated services in 3G cellular networks," *Proc. of the 9th ACM-IEEE MSWiM*, 2006, pp. 322–329.

[9] T. Bonald and A. Proutière, "On performance bounds for the integration of elastic and adaptive streaming flows," *Proceedings of the ACM SIGMETRICS / Performance*, 2004, pp. 235–245.

[10] O. J. Boxma, A. F. Gabor, R. Núñez-Queija, and H. P. Tan, "Performance analysis of admission control for integrated services with minimum rate guarantees," *Proc. of 2nd NGI*, 2006, pp. 41–47.

[11] R. Núñez-Queija, J. L. van den Berg, and M. R. H. Mandjes, "Performance evaluation of strategies for integration of elastic and stream traffic," *Proc. ITC 16*, 1999. Eds. D. Smith and P. Key. Elsevier, Amsterdam, pp. 1039–1050.

[12] T. Bonald and A. Proutière, "Wireless downlink data channels: User performance and cell dimensioning," *Proc. of the ACM MOBICOM*, 2003, pp. 339–352.

[13] P. Key and L. Massoulié, "Fluid Limits and Diffusion Approximations for Integrated Traffic Models," Technical Report MSR-TR-2005-83, Microsoft Research, June 2005.

[14] R. Núñez-Queija, *Processor-Sharing Models for Integrated-Services Networks.* PhD thesis, Department of Mathematics and Computer Science, Eindhoven University of Technology, 2000.