Location-based admission control for differentiated services in 3G cellular networks

Rudesindo Núñez-Queija*
Department of Mathematics and Computer Science, Eindhoven University of Technology
5600 MB Eindhoven
The Netherlands
sindo@cwi.nl

Hwee-Pink Tan
EURANDOM
P.O. Box 513, 5600 MB Eindhoven
The Netherlands
tan@eurandom.tue.nl

ABSTRACT

Third generation wireless systems can simultaneously accommodate flow transmissions of users with widely heterogeneous applications. As resources are limited (particularly in the air interface), admission control is necessary to ensure that all active users are accommodated with sufficient capacity to meet their specific Quality of Service requirements. Our admission control rule protects users with stringent capacity requirements ("streaming traffic") while offering sufficient capacity over longer time intervals to delay-tolerant users ("elastic traffic"). Performance evaluation of wireline differentiated-services platforms is already difficult due to the inherently large dimensionality of models to capture the diversity of user applications. In wireless systems, this is further exemplified as the location of users adds to the dimensionality problem. Using time-scale decomposition, we develop approximations to evaluate the performance of a differentiated admission control strategy to support integrated services with capacity requirements in a realistic downlink transmission scenario for a single radio cell.

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 are commonly grouped into two broad categories: **Elastic traffic** correspond 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 [1, 9, 14]; alternatively, *adaptive* streaming traffic (that are TCP-friendly and mimic elastic traffic) are considered in [4, 12, 11]. Markovian models have been developed for the exact analysis of these systems [14, 13]. 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 [1, 11], 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. In our previous work [7], we disregard the location of the users in the admission control model by assuming that all users are located at the cell border, consuming *more* resource than they actually do. As a result, fewer users can be admitted, giving rise to a *conservative* admission control model. Here, we generalize the admission control model by taking into account the distance of each user from the base station to achieve a more *realistic* representation of the actual scenario. We describe the model in Section 2 and develop an approximation based on time-scale decomposition in Section 3. Numerical results are presented in Section 4. Some concluding remarks are outlined in Section 5.

2. 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

*At the time of this research, this author is also affiliated with CWI (The Netherlands). He is currently also affiliated with TNO Information and Communication Technology (The Netherlands).

MSWiM '06 Malaga, Spain

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 $P_u^r = P_u \Gamma_u$, 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

$$\left(\frac{E_b}{N_0}\right)_u = \frac{W}{R_u} \frac{P_u^r}{\eta + I_u^a + I_u^r},$$

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_u^a, I_u^r) is the intr*a* / int*er*-cell interference at user u respectively. As the name suggests, intra (*inter*)-cell interference is caused by simultaneous *interfering* transmissions received at user u from the base station in the serving cell (*neighboring* cells). For linear and hexagonal networks [3], it can be shown that I_u^r increases as δ_u increases.

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:

$$R_u \leq \frac{W P_u \Gamma_u}{\epsilon_u(\eta + I_u^a + I_u^r)}. \tag{1}$$

2.1 Resource Sharing

According to Eq. (1), in addition to its location and transmission power, the feasible data transmission rate at user u depends on the resource sharing mode through the intra-cell interference, I_u^a. This is due to simultaneous transmissions from the serving base station of user u using non-orthogonal codes (with total power P_u^a) to other users in the *same* cell received at user u. Quantitatively, we can write $I_u^a = \alpha P_u^a \Gamma_u$, where α is the code non-orthogonality factor. Depending on the resource sharing mode, we have the following expressions for P_u^a:

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

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. From Eq. (1), R_u can be maximized by minimizing P_u^a, i.e., by applying 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; hence, while the received signal for a streaming user is interfered by simultaneous transmissions to all other users, the received signal for an elastic user is only interfered by simultaneous transmissions to streaming users only.

2.2 Cell Partitioning

According to Eq. (1), the transmission power, P_u, required 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 the exact location of each user u, a maximum number of users can be admitted if the base station allocates *exactly* $\tilde{P}_u$ to each user u. However, for our analysis to be tractable, it is necessary to quantize the location of each user in the cell. We do so 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 ($J{=}\infty$).

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 [10], and will be denoted by $\epsilon_{j,e}$ and $\epsilon_{j,s}$ for elastic and streaming users in segment j, respectively.

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.

According to our 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

$$r_{j,s}\epsilon_{j,s}[\alpha(P - P_{j,s})\Gamma_j + \eta + I_j^r] \leq W P_{j,s}\Gamma_j,$$

so that

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

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

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

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 at least

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

Note that β_j is *insufficient* to guarantee capacity $r_{j,e}$ if streaming traffic consumes more than P_s. In contrast, γ_j is always sufficient to achieve rate $r_{j,s}$.

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

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

It is worth noting that the functions $\beta_j(\mathbf{N}_s)$ and $P_s(\mathbf{N}_s)$ depend only on $N_{j,s}$ through the weighted sum $\sum_{j=0}^{J} N_{j,s}\gamma_j$.

Conditions (3) and (4) completely determine the admission policy: a newly-arrived user will be accepted only if the resulting system state, (N_e,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:

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

Notice 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. We define

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

and

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

Taking the maximum of β_j and $\beta_j(\mathbf{N}_s)$ ensures that if streaming traffic uses less than the reserved capacity, i.e., $P_s(\mathbf{N}_s) < P_s$, the minimum capacity requirement for elastic users in segment j can be guaranteed, even if streaming traffic claims the full reserved power at a later stage. Hence, a newly-arrived elastic user in segment i will be admitted if

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

While the admission control proposed in [1] 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.

REMARK 1. *For a single radio cell, let us define each segment j as the annulus between concentric rings of radius δ_{j-1} and δ_j. In this case, Γ_j decreases and I_j^r increases (assuming linear or hexagonal networks) with δ_j. If we further assume that ($r_{j,e}$, $r_{j,s}$) = (r_e, r_s) and ($\epsilon_{j,e}$, $\epsilon_{j,s}$) = (ϵ_e, ϵ_s), and disable time-sharing amongst elastic users, then we have the following:*

$$\beta_j \leq \beta_J = \frac{r_e\epsilon_e[\alpha P + \frac{\eta + I_J^r}{\Gamma_J}]}{W + \alpha r_e\epsilon_e},$$

and

$$\gamma_j \leq \gamma_J = \frac{r_s\epsilon_s[\alpha P + \frac{\eta + I_J^r}{\Gamma_J}]}{W + \alpha r_s\epsilon_s}.$$

Replacing β_j $\mathbf{1}_{N_{j,e}>0}$ by β_J and γ_j by γ_J, our model simplifies to the conservative model defined in [7].

2.4 Rate allocation

As mentioned above, streaming users are accommodated with exactly their required capacities, i.e., $r_{j,s}$ in segment j. For elastic users, the rates 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. Using (2), an elastic user in segment j attains a data rate

$$r_{j,e}(N_e, \mathbf{N}_s) = \frac{1}{N_e} f_{j,e}\mu_{j,e}(\mathbf{N}_s),$$

where

$$\mu_{j,e}(\mathbf{N}_s) = \frac{1}{f_{j,e}} \times \frac{W(P - P_s(\mathbf{N}_s))\Gamma_j}{\epsilon_{j,e}[\alpha P_s(\mathbf{N}_s)\Gamma_j + \eta + I_j^r]}$$

can be interpreted as the *total departure rate* of elastic users if all elastic users are in segment j.

Remark 2. *When serving users in time-sharing mode, significant performance improvements can be obtained through opportunistic scheduling [5]. The above can be adapted to include such gains by redefining* $r_{j,e}(N_e, \boldsymbol{N}_s)$ *as follows:*

$$r_{j,e}(N_e, \boldsymbol{N}_s) = \frac{G(N_e)}{N_e} f_{j,e}\mu_{j,e}(\boldsymbol{N}_s),$$

for some gain function $G(\cdot)$. *Then, the interpretation of* $\mu_{j,e}(\boldsymbol{N}_s)$ *given above only applies if there is only one elastic user in segment* j. *Since the analysis is not affected, we will include the factor* $G(\cdot)$, *but in our numerical results, we will assume* $G(\cdot) \equiv 1$.

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 [14, 13], we develop an approximation based on time-scale decomposition to evaluate the cell performance and assess the accuracy through comparison with simulation. In our previous work [7], an incoming request is only distinguished based on its type, and hence, if there are only elastic (streaming) users in the cell, the number of on-going transmissions can be modeled by an egalitarian processor-sharing queue (Erlang-loss queue). Here, users within each type are further distinguished according to their distance from the base station (segment index j).

3.1 Quasi-stationary Approximation

We develop a quasi-stationary approximation for elastic flows, to be denoted $\mathbf{A}(\mathbf{Q})$, 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 [8]. 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:

$$\begin{aligned}\mathbb{P}^Q(\mathbf{n}_e|\mathbf{n}_s) &\equiv \mathbb{P}(\mathbf{N}_e^Q = \mathbf{n}_e \mid \mathbf{N}_s^Q = \mathbf{n}_s) \\ &= \frac{c_e^Q(\mathbf{n}_s)n_e!}{\prod_{k=0}^{n_e} G(k)} \prod_{j=1}^{J} \frac{\rho_{j,e}(\mathbf{n}_s)^{n_{j,e}}}{n_{j,e}!}, \end{aligned} \quad (7)$$

where we have defined $G(0) = 1$ (for notational convenience), $\rho_{j,e}(\mathbf{n}_s) = \frac{\lambda_{j,e}}{\mu_{j,e}(\mathbf{n}_s)}$ and the normalization constant $c_e^Q(\mathbf{n}_s)$ is such that adding (7) 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, equals

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

From (7) we can also obtain the distribution of the total number of active elastic users by summing over all admitted combinations of $n_{j,e}$ with $\sum_j n_{j,e} = n_e$. For the special case where $\beta_i \equiv \beta$ for all i – we call this *uniform admission control*[1] –, the distribution for the total number of elastic users takes a very appealing form:

$$\mathbb{P}(N_e^Q = n_e \mid \mathbf{N}_s^Q = \mathbf{n}_s) = \frac{c_e^Q(\mathbf{n}_s)\rho_e(\mathbf{n}_s)^{n_e}}{\prod_{k=0}^{n_e} G(k)}. \quad (8)$$

This also leads to the following simple expression for the normalization constant:

$$c_e^Q(\mathbf{n}_s) = \left(\sum_{k=0}^{n_e^{Q,\max}(\mathbf{n}_s)} \frac{\rho_e(\mathbf{n}_s)^k}{\prod_{k=0}^{n_e} G(k)} \right)^{-1},$$

where $n_e^{Q,\max}(\mathbf{n}_s) = \lfloor (P - \overline{P}_s(\mathbf{n}_s)/\beta) \rfloor$. Note that, without opportunistic scheduling, $G(\cdot) \equiv 1$ and (8) reduces further to a simple truncated geometric distribution.

We emphasize that, assuming quasi-stationarity, (7) and (8) are valid for general distributions of elastic requests [8]. 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* [9].

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

[1]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).

control, it is either accepted or blocked. Under our approximation assumptions, the probability of acceptance in segment i, $A^Q_{i,s}(\mathbf{n}_s)$, is given by:

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

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

$$\Lambda^Q_{i,s}(\mathbf{n}_s) \quad = \quad \lambda_{i,s} A^Q_{i,s}(\mathbf{n}_s).$$

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

In general, there is no closed-form expression for the equilibrium distribution of $\mathbf{N}^Q_s$ 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 $\mathbf{N}^Q_s$ is much smaller than 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 1], then $A^Q_{i,s}(\mathbf{n}_s) \equiv A^Q_s(n_s)$ is independent of i and depends on $\mathbf{n}_s$ only through the total number of streaming flows. $\mathbf{N}^Q_s$ can then be shown to be *balanced* [2] and can be reduced to the framework of [8]. It follows that, for *arbitrary* holding time distributions of streaming flows, and $0 \le n_s \le n_s^{\max} = \lfloor \frac{P}{\gamma} \rfloor$:

$$\mathbb{P}(\mathbf{N}^Q_s = \mathbf{n}_s) = c^Q_s \prod_{k=0}^{n_s-1} A^Q_s(k) \prod_{j=1}^{J} \frac{(\rho_{j,s})^{n_{j,s}}}{n_{j,s}!}, \qquad (9)$$

with $\rho_{j,s} = \lambda_{j,s}/\mu_{j,s}$ and $c^Q_s = P(N^Q_s = 0)$ can be determined by normalizing (9) 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):

$$\mathbb{P}(N^Q_s = n_s) = c^Q_s \frac{(\rho_s)^{n_s}}{n_s!} \prod_{k=0}^{n_s-1} A^Q_s(k), \qquad (10)$$

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

$$c^Q_s = \left(\sum_{n_s=0}^{n_s^{\max}} \frac{(\rho_s)^{n_s}}{n_s!} \prod_{k=0}^{n_s-1} A^Q_s(k) \right)^{-1}.$$

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

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

The unconditional blocking probabilities in segment i are

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

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

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

3.2 Fluid Approximation

The fluid approximation (from the perspective of elastic flows), denoted by $\mathbf{A(F)}$, 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 $\mathbf{A(Q)}$, 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:

$$\begin{aligned} \mathbb{P}^F(\mathbf{n}_s \mid \mathbf{n}_e) &\equiv \mathbb{P}(\mathbf{N}^F_s = \mathbf{n}_s \mid \mathbf{N}^F_e = \mathbf{n}_e) \\ &= c^F_s(\mathbf{n}_e) \prod_{j=1}^{J} \frac{\rho_{j,s}^{n_{j,s}}}{n_{j,s}!}, \end{aligned} \qquad (11)$$

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 (11) adds to 1 when summing (for fixed $\mathbf{n}_e$) over all $\mathbf{n}_s$ such that $P_e(\mathbf{n}_e, \mathbf{n}_s) + P_s(\mathbf{n}_s) \le P$. 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:

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

and

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

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

3.2.2 *Unconditional marginal distributions*

Next, we consider the dynamics of elastic flows. When $\mathbf{N}^F_e = \mathbf{n}_e > 0$, 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 $\mathbf{n}_e$):

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

where the summation is taken over all $\mathbf{n}_s$ for which $P_e(\mathbf{n}_e, \mathbf{n}_s) + P_s(\mathbf{n}_s) \le P$. 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:

$$A^F_{i,e}(\mathbf{n}_e) \equiv \mathbb{P}(\overline{P}_s(\mathbf{N}^F_s) + \overline{P}_e(\mathbf{n}_e + \mathbf{e}_i, \mathbf{N}^F_s) \le P \mid \mathbf{N}^F_e = \mathbf{n}_e),$$

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

$$\Lambda^F_{i,e}(\mathbf{n}_e) \quad\equiv\quad \lambda_{i,e} A^F_{i,e}(\mathbf{n}_e).$$

As for the quasi-stationary approximation, in general, there is no closed-form expression for the distribution of $\mathbf{N}^F_e$. However, under additional assumptions, $\mathbf{N}^Q_s$ is *balanced* [2]. 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. Then, we can write

$$\mu_{j,e}(\mathbf{n}_s) = \frac{\nu_j}{f_{j,e}}\, h(\mathbf{n}_s),$$

with

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

and

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

Moreover, because of the uniform admission control for elastic users, the dynamics of $\mathbf{N}^F_s$ depends on $\mathbf{N}^F_e$ only through the total number of elastic users N_e, so that

$$\mathbb{E}[h(\mathbf{N}^F_s) \mid \mathbf{N}^F_e = \mathbf{n}_e] = \mathbb{E}[h(\mathbf{N}^F_s) \mid N^F_e = n_e] \equiv g(n_e)$$

where n_e is the total number of elastic users in state $\mathbf{n}_e$. Consequently, we have

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

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

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

$$\mathbb{P}(\mathbf{N}^F_e = \mathbf{n}_e) = c^F_e \prod_{k=1}^{n_e} \frac{k\, A^F_e(k-1)}{G(k)\, g(k)} \prod_{j=1}^{J} \left(\frac{\rho_{j,e}}{\nu_j}\right)^{n_{j,e}}, \quad (12)$$

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

$$\mathbb{P}(N^F_e = n_e) = c^F_e \left(\sum_j \frac{\rho_{j,e}}{\nu_j}\right)^{n_e} \prod_{k=1}^{n_e} \frac{k\, A^F_e(k-1)}{G(k)\, g(k)}, \quad (13)$$

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

$$c^F_e = \left(\sum_{n_e=0}^{n^{\max}_e} \left(\sum_j \frac{\rho_{j,e}}{\nu_j}\right)^{n_e} \prod_{k=1}^{n_e} \frac{k\, A^F_e(k-1)}{G(k)\, g(k)}\right)^{-1}.$$

REMARK 3. *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 ($\eta_j + I^r_j << \alpha P_s(\boldsymbol{n}_s)\Gamma_j$) by choosing*

$$\nu_j = \frac{W\Gamma_j}{\epsilon_{j,e}},$$

and

$$h(\boldsymbol{n}_s) = \frac{P - P_s(\boldsymbol{n}_s)}{\alpha P_s(\boldsymbol{n}_s)}.$$

We conclude this section with the following unconditional performance measures:

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

The unconditional blocking probabilities in segment i are

$$p^F_{i,e} = \sum_{\mathbf{n}_e}(1 - A^F_{i,e}(\mathbf{n}_e))\mathbb{P}(\mathbf{N}^F_e = \mathbf{n}_e),$$

and

$$p^F_{i,s} \quad=\quad \sum_{\mathbf{n}_e}(1 - A^F_{i,s}(\mathbf{n}_e))\mathbb{P}(\mathbf{N}^F_e = \mathbf{n}_e).$$

4. PERFORMANCE EVALUATION

We consider a single UMTS cell whose radius, δ_J, is computed using the reference link budget given in Table 8.3 [10] and the Okumura-Haka propagation model [16] for an urban macro cell. The inter-cell interference at each location within the cell is computed based on the conservative approximation for a hexagonal network [3].

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) and mean file size, f_e (holding time, $\frac{1}{\mu_s}$). The base station performs admission control according to the type and location of each user, 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. While representative numerical results for the quasi-stationary approximation, $\mathbf{A(Q)}$, are presented here, more extensive results (that include $\mathbf{A(F)}$) are found in [6].

4.1 Performance Insensitivity with Traffic Parameter Distribution

We develop a simulation program for our model by considering arrival / departure events of traffic requests (elastic or streaming), assuming that the dynamics of streaming flows take place on a much slower time scale than those of elastic flows (i.e., quasi-stationary traffic regime).

We select the distribution of the traffic parameters according to the following cases: **I** with exponentially distributed (d_s, s_e) , **II** with exponentially distributed d_s and hyper-exponentially distributed s_e with parameter a_e (cf.[15], p. 359), where $\forall s \ge 0$, P($s_e > s$) $= \frac{a_e e^{\frac{-a_e s}{f_e}} + e^{\frac{-s}{a_e f_e}}}{a_e+1}$, Var[$s_e$] $= (a_e +$

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-Haka Model [16]
Inter-cell Interference Model	Hexagonal network with maximum tx. power [3]
Link budget	Table 8.3 [10]
r_e (kbps)	128
r_s (kbps)	128

Table 1: UMTS cell and traffic parameters for performance evaluation.

$\frac{1}{a_e} - 1)f_e^2$ and **III** with exponentially distributed s_e and Erlang-k distributed d_s, where $\forall d \geq 0$ and $k > 0$, $f_s(d) = \frac{k\mu_s(k\mu_s d)^{k-1}}{(k-1)!}e^{-k\mu_s d}$, $\text{Var}[d_s] = \frac{1}{k\mu_s^2}$.

Hence, each simulation scenario is defined as follows:

1. Select the traffic parameter distribution according to Case **I**, **II** or **III**;
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 $u_s = \frac{\lambda_s r_s}{\mu_s}$;
2. For each l, fix the traffic *mix*, $\frac{u_e}{lc}$, by choosing u_e, $0 \leq u_e \leq l\ c$;
3. For each traffic mix, select $(\lambda_e,\ \lambda_s)$ according to a quasi-stationary traffic regime.

We generate 5 sets of simulation results for each scenario, and compute the sample mean for $(p_e,\ p_s,\ S_e)$. The results are tabulated in Table 2 for $a_e = 100$ (**II**) and $k = 2$ (**III**) for a *fully-loaded* cell (i.e., $l = 1$). We observe that the performance measures obtained for Cases **II** and **III** are within 10% of those obtained for Case **I**. Hence, the performance is *almost* insensitive to the traffic parameter distributions, thus justifying the insensitive approximations proposed here.

4.2 Accuracy of Quasi-stationary Approximation

To apply the quasi-stationary approximation to estimate the cell performance analytically, we need to partition the cell into J rings such that $\delta_j = \frac{j}{J}\delta_J$, $1 \leq j \leq J$, where the arrival rate of users in each ring j is $\lambda_j = \frac{\delta_j^2 - \delta_{j-1}^2}{\delta_J^2}\lambda$, where $\delta_0 = 0$, due to the assumption of uniformly distributed arrivals. We investigate the accuracy of the approximation for various values of J (denoted by $\mathbf{A(Q,J)}$) by bench-marking against simulation results (Case **I**) obtained in Section 4.1. We plot $(p_e,\ p_s)$ and $(\text{E}[N_e],\ S_e)$ as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$, for $\mathbf{A(Q,J)}$ in Fig. 1 and 2 respectively.

We observe that, for P = 20W, the cell performance obtained with simulation is well approximated by $\mathbf{A(Q,J{=}1)}$, and that $\mathbf{A(Q,J)}$ is almost invariant with the value of J. Although cell partitioning (with increasing J) was intended to improve the accuracy of the approximations by reducing the quantization error of estimating each user's location, for the given base station transmission power, the cell performance can be well approximated using the conservative admission control in [7], which does not exploit user location.

In order to investigate the performance gain with exploiting user location, we repeat the simulations for the case of P = 0.2W, and plot $(\text{E}[N_e],\ \text{E}[N_s])$ and $(p_e,\ S_e)$ as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$, for $\mathbf{A(Q,J)}$ in Fig. 3 and 4 respectively. In this case, we note that as cell partitioning becomes finer (increasing J), the performance obtained with $\mathbf{A(Q,J)}$ approaches the simulation performance. We expect the accuracy of $\mathbf{A(Q,J)}$ to be further improved as J is further increased.

4.3 Performance sensitivity in different traffic regimes

In the last two sections, we obtained the cell performance through simulations for a quasi-stationary traffic regime, where the dynamics of streaming flows take place on a much slower time scale than those of elastic flows. In particular, we showed in Section 4.2 that $\mathbf{A(Q,J{=}1)}$ accurately approximates the cell performance with P=20W, and when the base station power is reduced to 0.2W, the accuracy of $\mathbf{A(Q,J)}$ improves as J is increased.

Here, we define two other traffic regimes: fluid (neutral) traffic regimes, where the dynamics of streaming flows take place on *a much faster* (similar) time scales than those of elastic flows. Our objective is to investigate if $\mathbf{A(Q,J)}$ can be applied to approximate non quasi-stationary traffic regimes.

We generate 5 sets of simulation results for each simulation scenario, and compute the sample mean for $(p_e,\ p_s)$ and $(\text{E}[N_e],\ S_e)$. For Case **I** (P=20W), we plot these metrics as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$, alongside $\mathbf{A(Q,J)}$ in Fig. 5 and 6. We observe that under heavy load condition ($\frac{u_e}{c} \geq 0.5$), as the load increases, the performance metrics become invariant with respect to the traffic regime. In addition, as expected, the accuracy of $\mathbf{A(Q,J)}$ is degraded as we move from the quasi-stationary to the neutral regime, and further with the transition into the fluid regime. In this case, $\mathbf{A(F)}$ is necessary to approximate the performance in the latter regime.

We repeat the simulations for Case **II** (P=0.2W) under moderate loading condition ($\alpha = 0.6$), and the sample means of $(\text{E}[N_e],\ S_e)$ as a function of the traffic mix in Fig. 7. Accordingly, under reduced power constraints, the performance metrics are *almost* invariant in the various traffic regimes, and hence, if $\mathbf{A(Q,J)}$ is sufficiently accurate for the quasi-stationary regime, it will also be a good approximation for the other traffic regimes.

5. 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 differentiated admission control strategy that protects users with stringent capacity requirements ("streaming traffic") while offering sufficient capacity over longer time intervals to delay-tolerant users ("elastic traffic").

u_e/c		0,1			0,3			0,5			0,7			0,9	
Case	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
$E[N_e]$	0,868	0,892	0,856	2,350	2,358	2,345	1,980	1,938	1,990	0,877	0,884	0,852	0,243	0,250	0,241
$E[N_s]$	30,906	31,176	30,845	25,349	25,377	25,388	18,870	18,846	18,936	11,558	11,586	11,398	3,809	3,875	3,815
S_e	1,925	1,999	1,898	1,660	1,671	1,651	0,806	0,789	0,813	0,254	0,256	0,246	0,055	0,056	0,055

Table 2: Impact of traffic parameter distribution on ($\mathbf{E}[N_e]$, $\mathbf{E}[N_s]$, S_e) for P = 20W with various elastic load compositions (I: exponentially distributed (s_e, d_s), II: exponentially distributed d_s and hyper-exponentially distributed s_e (a_e=100) and III: exponentially distributed s_e and Erlang-2 distributed d_s).

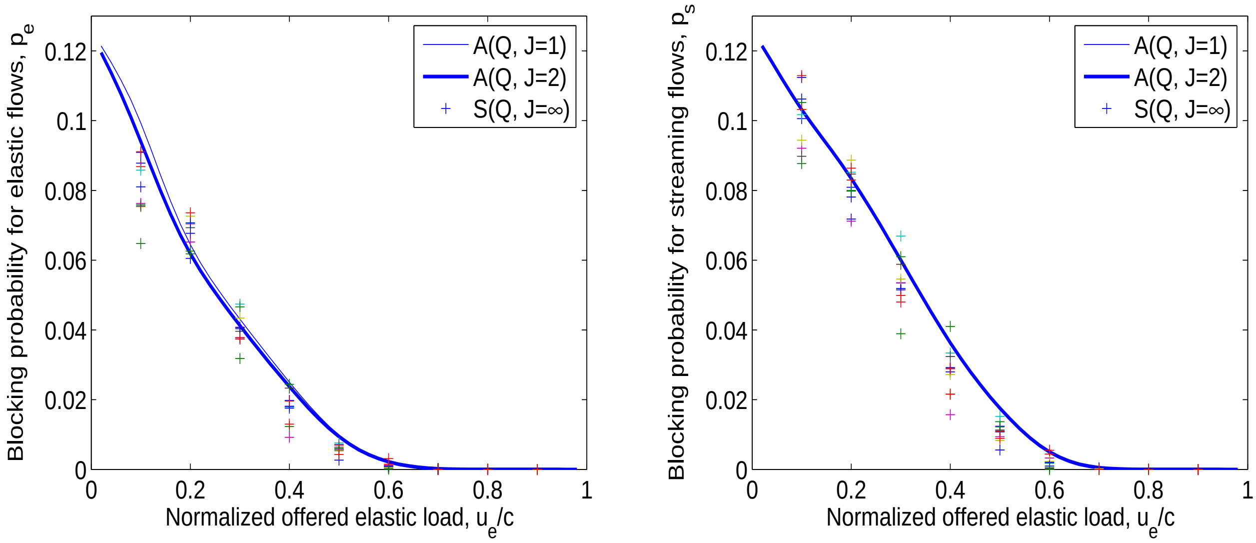

Figure 1: Blocking probability for elastic (left) and streaming requests (right) vs normalized offered elastic load obtained with approximation and simulation for Case I (P=20W).

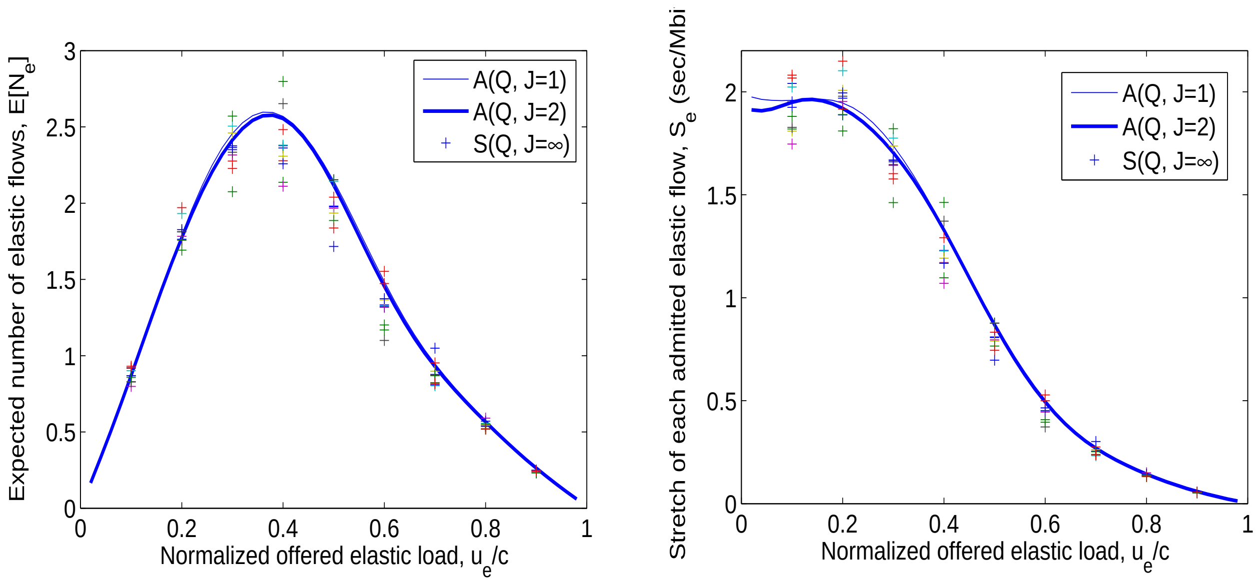

Figure 2: Number of active elastic requests (left) and stretch of each admitted elastic request vs normalized offered elastic load obtained with approximation and simulation for Case I (P=20W).

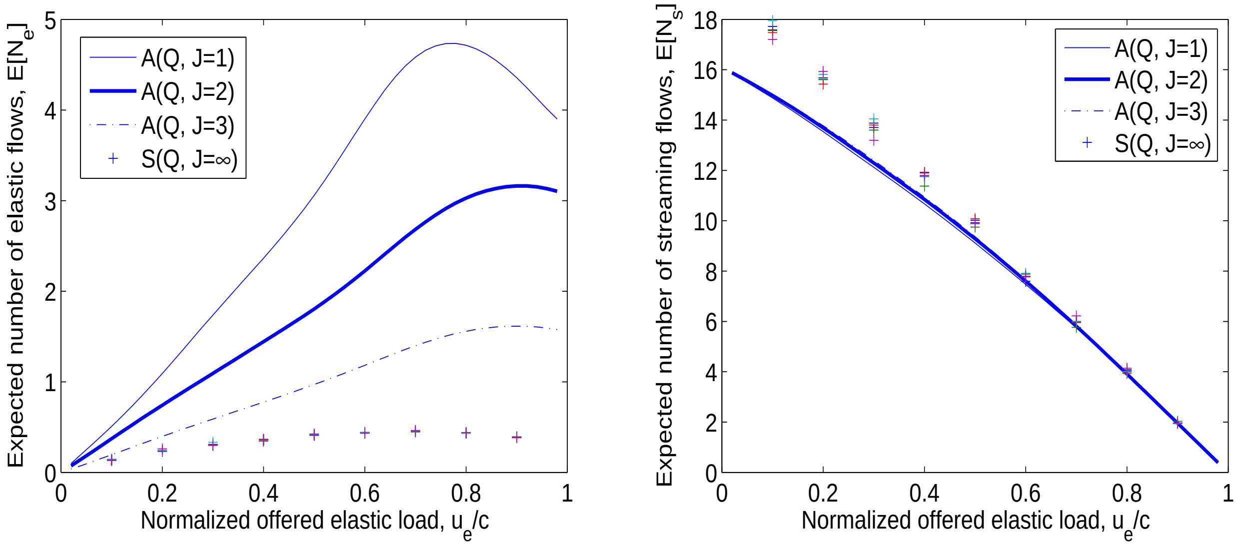

Figure 3: Number of active elastic (left) and streaming (right) requests vs normalized offered elastic load obtained with approximation and simulation for Case I (P=0.2W).

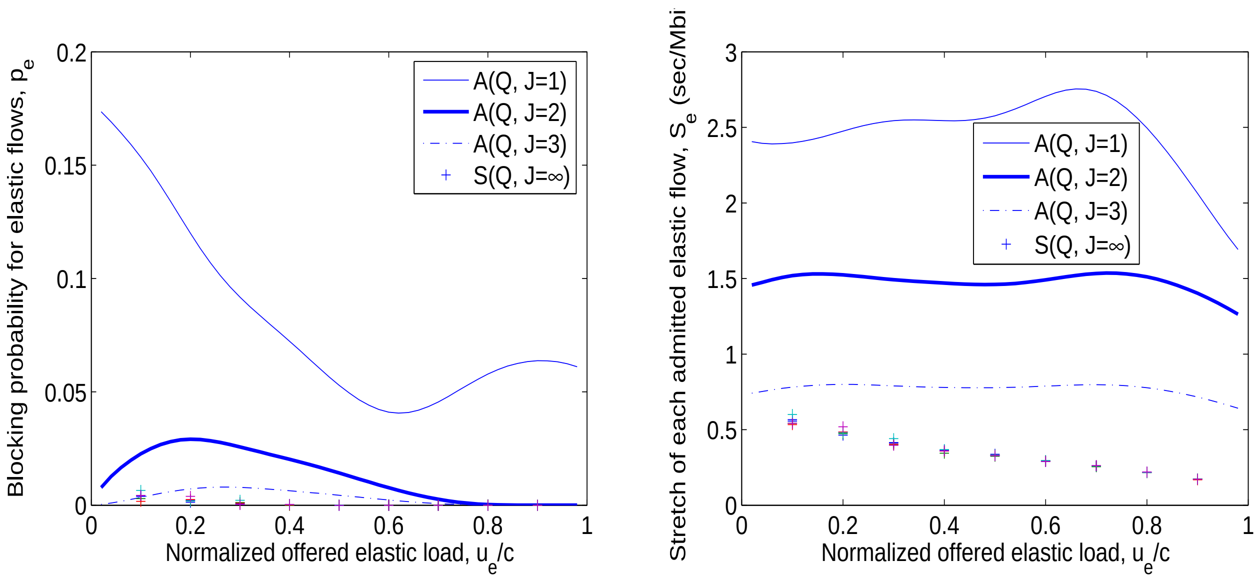

Figure 4: Blocking probability (left) and stretch (right) of elastic requests vs normalized offered elastic load obtained with approximation and simulation for Case I (P=0.2W).

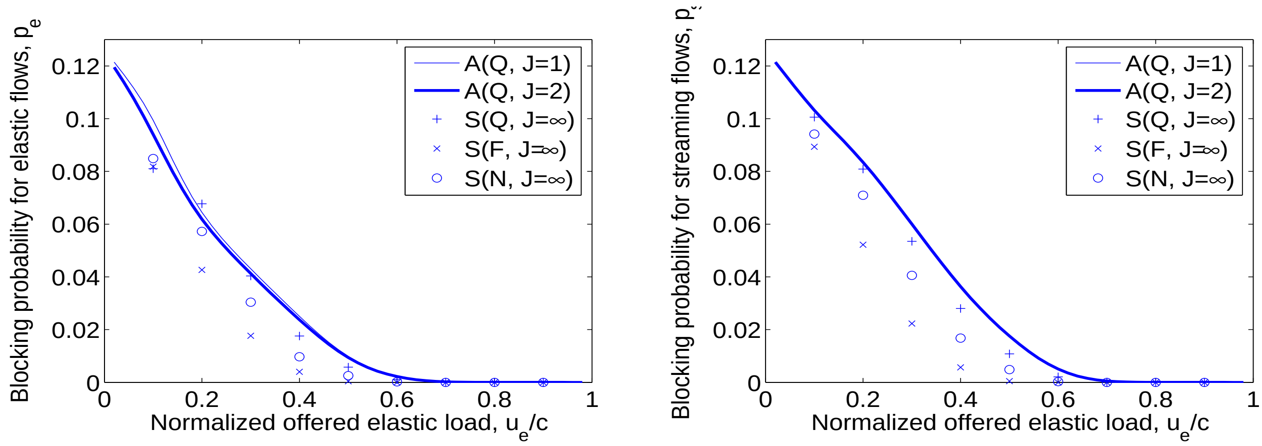

Figure 5: Blocking probability for elastic (left) and streaming requests (right) vs normalized offered elastic load obtained with approximation and simulation in different traffic regimes for Case I (P=20W).

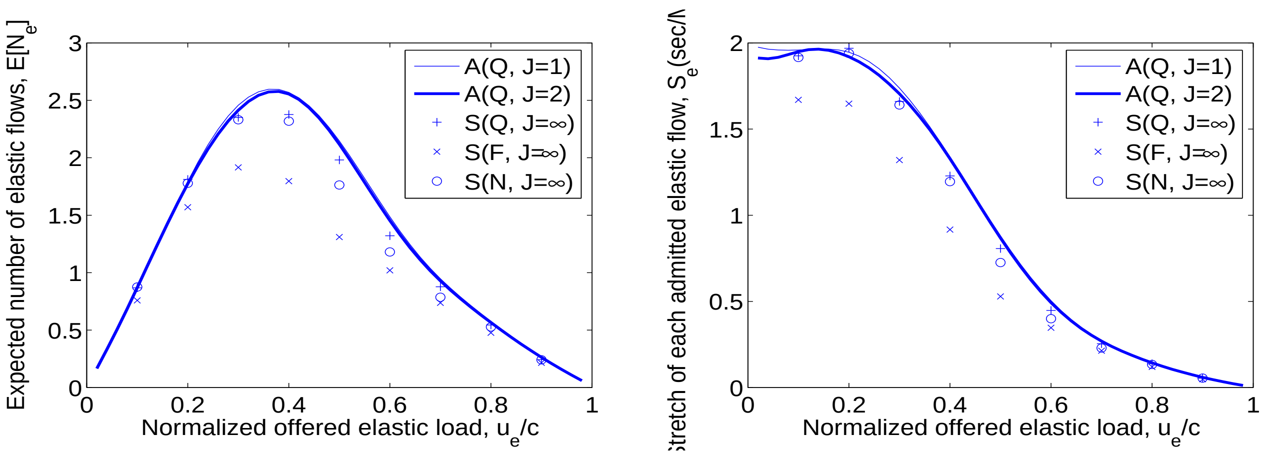

Figure 6: Number of active elastic requests (left) and stretch of each admitted elastic request vs normalized offered elastic load obtained with approximation and simulation in different traffic regimes for Case I (P=20W).

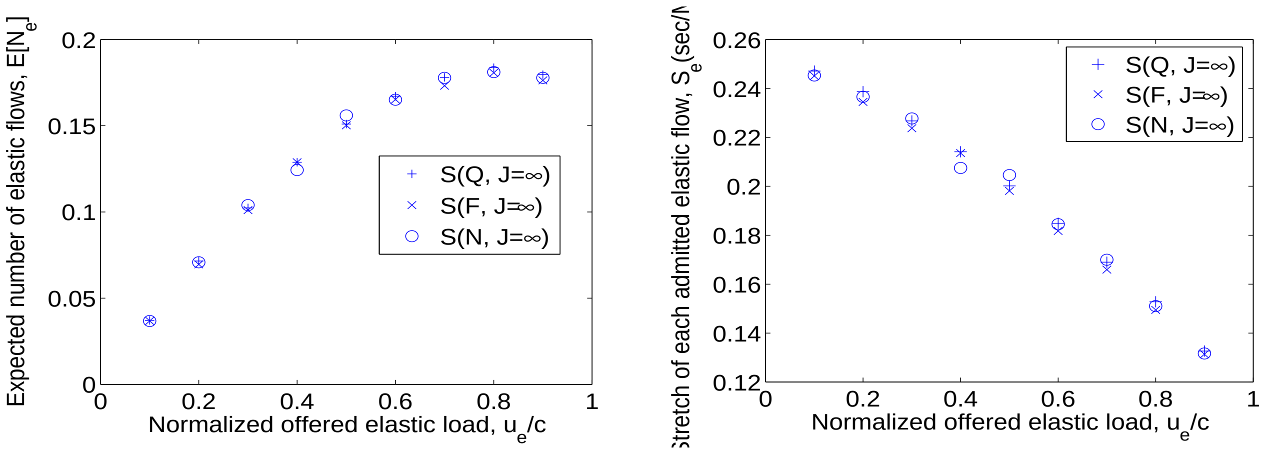

Figure 7: Number of active elastic requests (left) and stretch of each admitted elastic request vs normalized offered elastic load obtained with approximation and simulation in different traffic regimes for Case II (P=0.2W, $l = 0.6$).

Since the exact analysis to evaluate the performance of such an integrated services system is non-tractable in general, we apply time-scale decomposition to develop approximations for the cell performance for a single cell scenario. In our previous work [7], we developed a conservative model for the admission control strategy, where each incoming request is only distinguished based on its type (streaming or elastic). We generalize the model in this paper by (a) further distinguishing users within each type according to their distance from the base station by partitioning the cell into segments and (b) introducing a time-sharing resource sharing mechanism to improve the rate allocation to elastic traffic while guaranteeing the capacity requirements of all users.

For the quasi-stationary traffic regime (where traffic parameters are selected such that the dynamics of elastic requests take place at a much finer time scale than that of streaming requests), simulation results suggest that the performance is almost insensitive to traffic parameter distributions. In addition, we demonstrate that the generalized model approximates the cell performance better than the conservative model, and the accuracy improves as the cell partitioning becomes finer.

6. 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.

7. REFERENCES

[1] 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*, 2156:69–81, September 2001.

[2] T. Bonald and A. Proutière. Insensitive bandwidth sharing in data networks. *Queueing Systems*, 44:69–100, 2003.

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

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

[5] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. *IEEE/ACM Transanctions on Networking*, pages 636–647, June 2005.

[6] O. J. Boxma, A. F. Gabor, R. Núñez-Queija, and H. P. Tan. Integration of Streaming and Elastic Traffic in a Single UMTS Cell: Modeling and Performance Analysis. Technical report, 2006. [Online]. Available: http://euridice.tue.nl/∼hptan/publications/umts_single_cell.pdf.

[7] 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. *NGI 2006*, April 2006.

[8] J. W. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12:245–284, 1979.

[9] F. Delcoigne, A. Proutière, and G. Regnie. Modeling integration of streaming and data traffic. *Performance Evaluation*, 55(3-4):185–209, February 2004.

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

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

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

[13] R. Núñez-Queija. *Processor-Sharing Models for Integrated-Services Networks.* PhD thesis, Eindhoven University of Technology, 2000.

[14] 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*, pages 1039–1050, 1999. Eds. D. Smith and P. Key. Elsevier, Amsterdam.

[15] H. C. Tijms. *Stochastic Models — An Algorithmic Approach.* John-Wiley and Sons, 1994.

[16] Y. Wang and T. Ottosson. Cell search in W-CDMA. *IEEE Jounnal on Selected Areas in Communications*, 18(8):1470–1482, August 2000.