Integration of Streaming and Elastic Traffic in a Single UMTS Cell: Modeling and Performance Analysis

Onno J. Boxma[*,†,‡], Adriana F. Gabor[*,†], Rudesindo Núñez-Queija[*,‡] and Hwee Pink Tan[†]

[*]Department of Mathematics and Computer Science, Technical University of Eindhoven
5600 MB Eindhoven, (The Netherlands)

[†]EURANDOM, P.O. Box 513, 5600 MB Eindhoven (The Netherlands)

[‡]CWI, P.O. Box 94079, 1090 GB Amsterdam (The Netherlands)

Abstract— **Using time-scale decomposition, we develop approximations to evaluate the performance of an admission control strategy for integrated services in a single UMTS radio cell. Simulation results suggest that the performance is almost insensitive to traffic parameter distributions, and is well estimated by our proposed approximations.**

I. Introduction

UMTS is a 3G cellular network that is expected to support a large variety of applications which are commonly grouped into two broad categories:

Elastic flows correspond to the transfer of digital documents (e.g., Web pages, emails, stored audio / videos). They are characterized by their size, i.e., the volume of the document to be transferred. These flows are flexible, or "elastic", towards rate fluctuations, the total transfer time being a typical performance measure.

Streaming flows correspond to the real-time transfer of various signals (e.g., voice, streaming audio / video). They are characterized by their duration as well as the transmission rate. For "streaming" applications, stringent transmission rate guarantees are necessary to ensure real-time communication.

Various papers that study the integration of elastic and streaming traffic have been published recently [1], [2], [3], [4], [5], [6]. In terms of bandwidth sharing policy, the classical approach is to give head-of-line priority to packets of streaming flows in order to offer packet delay and loss guarantees [1], [2], [4]; alternatively, *adaptive* streaming flows (that are TCP-friendly and mimic elastic flows) are considered in [3], [5], [6].

In terms of modeling approach, while Markovian models have been developed for the exact analysis of the integrated-services system, they can be numerically cumbersome. Hence, a fluid model is proposed in [2], [3], [4], [5], [6] to provide closed form limit results and approximations. These results can serve as performance bounds, and hence yield useful insight.

In this study, we define a model (Section II) for a single UMTS cell that supports integrated services and gives priority to streaming flows through bandwidth reservation and ensures stability through admission control on both types of flows. To evaluate its performance, we develop approximations based on time-scale decomposition in Section III and numerical results are presented in Section IV. Some concluding remarks and future directions are outlined in Section V.

II. Model

We consider a UMTS cell with a single downlink channel whose limited *resource* (e.g., transmission power at the base station) is shared amongst streaming and elastic requests, that arrive as independent Poisson processes with rate λ_s and λ_e respectively. Let us denote by c the total amount of resource available.

We assume that a part of the total resource, $c_s \leq c$, is reserved for streaming requests, and each *admitted* streaming request is allocated (statically) an average rate r_s (Kbps) during its entire lifetime, d_s, which is generally distributed with mean $\frac{1}{\mu_s}$ (sec). The resource not claimed by streaming traffic is equally shared amongst all elastic requests, where each requires a minimum average rate of r_e

(Kbps) at all times and the corresponding size, s_e, is generally distributed with mean f_e (bits). Note that although the resource that can be maximally ensured for on-going elastic transfers is c_e = c-c_s, they are permitted to use more than c_e. However, the surplus is immediately allocated to streaming traffic if a new streaming request arrives. The model is illustrated in Fig. 1.

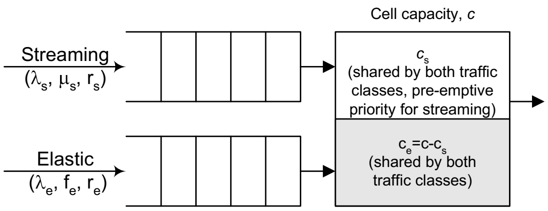
Cell capacity, c
Streaming
(λs, μs, rs)
Elastic
(λe, fe, re)
cs
(shared by both traffic classes, pre-emptive priority for streaming)
ce=c-cs
(shared by both traffic classes)

Fig. 1. Model of a single UMTS cell with streaming and elastic traffic.

A. Resource Sharing

Let P be the total power available at the base station, and P_u be the power transmitted to user *u*, where $P_u \leq P$. The power received by a user u is $P_u^r = P_u \Gamma_u$, where Γ_u denotes the attenuation due to path-loss. As a measure of the quality of service for 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 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* / inte*r*-cell interference at user *u* respectively. The intra-cell interference arises due to *simultaneous* transmissions to the other users in the *same* cell as user *u*; on the other hand, the inter-cell interference is due to the base stations' transmissions in *neighboring* cells.

Given a target error probability ϵ, in order to ensure a good quality of service, it is necessary that for each active user u, $\left(\frac{E_b}{N_0}\right)_u \geq \epsilon$, for some threshold ϵ, which is assumed to be the same for all users. A necessary condition for satisfying the quality of service criterion is that the rate R_u of each admitted user u satisfies:

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

To formulate the criterion for admission control, we consider two types of resource sharing schemes:

Time-sharing : When the resource is time-shared, I_u^a = 0 and P_u = P, since the base station transmits *all* its power to one user at any time. If ϕ_u denotes the fraction of time the base station transmits to user u, where $\sum_u \phi_u = 1$, then $\phi_u R_u$ corresponds to the *average* rate of user *u* and Eq. (1) can be written as follows:

$$\begin{aligned} \phi_u R_u &\leq \frac{\phi_u W P \Gamma_u}{\epsilon(\eta + I_u^r)} \\ &\leq \frac{\phi_u W P}{\epsilon} \min_u \frac{\Gamma_u}{(\eta + I_u^r)} \\ &= \frac{\phi_u W P \Gamma_{min}}{\epsilon(\eta + I_{max}^r)}, \end{aligned} \tag{2}$$

where (Γ_{min}, I_{max}^r) correspond to the maximum attenuation due to path-loss and maximum inter-cell interference in the cell. For typical propagation models, the attenuation due to path-loss for a user at distance δ from the base station is proportional to $\frac{1}{\delta^\gamma}$, where γ is a positive path-loss component. Hence, the maximum attenuation occurs when the user is at the edge of the cell. For linear and hexagonal networks [7], it can be shown that the total inter-cell interference is maximum when the user is at the edge of the cell.

To implement our resource reservation scheme, we assume that a fixed fraction of time, ϕ_s, is reserved for streaming traffic, such that ϕ_s = $\frac{c_s}{c}$. If ($\mathbb{E}$, $\mathbb{S}$) denote the set of on-going elastic and streaming requests respectively, where ($|\mathbb{E}|$, $|\mathbb{S}|$) = (N_e, N_s), then we have the following condition:

$$\sum_{u \in \mathbb{E}} \phi_u \leq 1 - \phi_s.$$

For an admitted elastic request, its average rate, $\phi_u R_u$, must satisfy $r_e \leq \phi_u R_u \leq \frac{\phi_u W P \Gamma_{min}}{\epsilon(\eta + I_{max}^r)}$. Summing Eq. (2) over $\mathbb{E}$, we obtain the following:

$$\begin{aligned} N_e r_e &\leq \frac{P W \Gamma_{min}}{\epsilon(\eta + I_{max}^r)} \sum_{u \in \mathbb{E}} \phi_u \\ &\leq \frac{W(1 - \phi_s) P \Gamma_{min}}{\epsilon(\eta + I_{max}^r)}. \end{aligned} \tag{3}$$

Finally, summing Eq. (2) over $\mathbb{E} \cup \mathbb{S}$, and noting that $\phi_u R_u$ = r_s, $u \in \mathbb{S}$, we obtain the following:

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

Power-sharing : With power-sharing, unlike in a time-sharing system, the base station transmits to all active users at the same time, and the received signals of different users are distinguished through codes. As a result, interference is also generated by users in the same cell (intra-cell interference) and is given by I_r^a = $\alpha(P - P_u)\Gamma_u$, where α is the code non-orthogonality factor. Then, Eq. (1) can be written as follows:

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

which can be re-written as follows:

$$\begin{aligned} \frac{R_u}{W + \alpha\epsilon R_u} &\leq \frac{P_u}{\epsilon(\frac{\eta + I_u^r}{\Gamma_u} + \alpha P)} \\ &\leq \frac{P_u}{\epsilon} \min_u \frac{1}{\frac{\eta + I_u^r}{\Gamma_u} + \alpha P} \\ &= \frac{P_u}{\epsilon(\frac{\eta + I_{max}^r}{\Gamma_{min}} + \alpha P)}. \end{aligned} \qquad (4)$$

Since the function $\frac{R_u}{W + \alpha\epsilon R_u}$ is an increasing function of R_u, it follows that for every elastic request u, the following should hold:

$$\frac{r_e}{W + \alpha\epsilon r_e} \leq \frac{P_u}{\epsilon(\frac{\eta + I_{max}^r}{\Gamma_{min}} + \alpha P)}.$$

To implement our resource reservation scheme, we assume that a fixed portion of P, P_s, is reserved for streaming traffic. Then, we have the following condition:

$$\sum_{u \in \mathbb{E}} P_u \leq P - P_s = P_e.$$

By summing over $\mathbb{E}$, we obtain the following:

$$\frac{N_e r_e}{W + \alpha\epsilon r_e} \leq \frac{P_e}{\epsilon(\frac{\eta + I_{max}^r}{\Gamma_{min}} + \alpha P)}. \qquad (5)$$

Finally, summing Eq. (4) over $\mathbb{E} \cup \mathbb{S}$, and noting that $R_u = r_s$, $u \in \mathbb{S}$, we obtain the following:

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

from which we obtain the following:

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

where r = max(r_e, r_s).

B. Admission Control

Hence, with time-sharing and power-sharing, the UMTS cell can be modeled as a link with capacity c = $\frac{W P \Gamma_{min}}{\epsilon(\eta + I_{max}^r)}$ and $\frac{P(W + \alpha\epsilon r)}{\epsilon(\frac{\eta + I_{max}^r}{\Gamma_{min}} + \alpha P)}$ respectively. We note that the capacity c is *independent* of the allocation of resources to *individual* users, as long as the total allocation satisfies the bandwidth reservation.

The admission control can be implemented based on the number of ongoing requests, (N_e, N_s), where a new elastic request will be accepted only if the following conditions hold:

$$\begin{aligned} N_s r_s + (N_e + 1) r_e &\leq c \\ (N_e + 1) r_e &\leq c - c_s. \end{aligned}$$

On the other hand, a new streaming request will be admitted as long as

$$(N_s + 1) r_s + N_e r_e \leq c.$$

Note that our admission control model is conservative, since it implicitly assumes that all users are located at the edge of the cell. In addition, while the admission control proposed in [2] is similar, it results in equal blocking probabilities for both types of traffic, which is not the case with our strategy due to bandwidth reservation.

For the convenience of the analysis that follows, we define $K_e(n_s)$ = $\lfloor \frac{\min\{c - n_s r_s,\ c_e\}}{r_e} \rfloor$ and $K_s(n_e)$ = $\lfloor \frac{c - n_e r_e}{r_s} \rfloor$, where $K_i^{n_j}$ is the maximum number of type-i flows when n_j type-j flows are present. In addition, we denote the conditional probability of event $\mathcal{B}$, given event $\mathcal{A}$, $P(\mathcal{B} \mid \mathcal{A})$ as $\mathbb{P}_{\mathcal{A}}^{\mathcal{B}}$.

III. Analysis

Since exact analysis of our model is non-tractable in general and computationally involved when assuming exponentially distributed holding times and file sizes, we develop various approximation techniques and assess their accuracy through comparison with simulation.

A. Quasi-stationary Approximation for Elastic Flows

For the quasi-stationary approximation, to be denoted $\mathbf{A}(\mathbf{Q})$, 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. The corresponding condition is that

$$\mu_s E[N_s] + \lambda_s << \frac{c - r_s E[N_s]}{f_e} + \lambda_e, \qquad (6)$$

where the expression on the LHS (RHS) corresponds to the average rate at which the number of streaming (elastic) flows changes. Although the above condition cannot be easily checked (due to the dependence on $E[N_s]$), it is ensured to be satisfied if

$$\mu_s \frac{c}{r_s} + \lambda_s << \lambda_e.$$

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 considering a fixed number of streaming flows, i.e., $N_s = n_s$. We construct an approximation assuming that the number of active elastic flows *instantaneously* reaches a new statistical equilibrium whenever the number of streaming flows changes. To avoid any confusion we will mark all quantities (such as queue lengths and performance measures) resulting from this approximation approach by adding a superscript ^{Q} to the notation.

From the capacity constraint and the reservation policy, it follows that $n_e r_e \leq \min\{c - n_s r_s,\ c_e\}$. In this case, elastic traffic behaves like an $M/G/1/K_e(n_s)$ processor-sharing (PS) queue with $K_e(n_s)$ service positions, capacity $c - n_s r_s$ and average departure rate $\mu_e(n_s) = \frac{c - n_s r_s}{f_e}$. Hence, from [8],

$$\begin{aligned} \mathbb{P}^{N_e^Q=n_e}_{N_s^Q=n_s} &\equiv P(N_e^Q = n_e \mid N_s^Q = n_s) \\ &= \frac{\rho_e(n_s)^{n_e}(1-\rho_e(n_s))}{1-\rho_e(n_s)^{K_e(n_s)+1}}, \qquad (7) \end{aligned}$$

where $\rho_e(n_s) = \frac{\lambda_e}{\mu_e(n_s)} = \frac{\lambda_e f_e}{c - n_s r_s}$. Notice [8] that this expression is insensitive to the file size distribution, other than through its mean. As a further remark, we observe that whether or not $\rho_e(n_s) < 1$ is of no concern, since N_e^Q is limited due to the assumption that $r_e > 0$. Often, when applying a time-scale decomposition, this matter is of importance, giving rise to an additional assumption commonly referred to as *uniform stability* [4].

Next, we consider the dynamics of streaming flows. When N_s^Q=n_s, streaming flows depart at a rate $n_s\mu_s$. When a new streaming flow arrives, due to admission control, we have two possible scenarios: either the newly arrived streaming flow is accepted or it is blocked. Under our approximation assumptions, the probability of acceptance is P($N_e^Q r_e$+(n_s+1)$r_s \leq c \mid N_s^Q$=n_s). Notice that the admission probability of streaming flows equals 1 if $(n_s+1)r_s \leq c_s$. Substituting Eq. (7) into this expression and noting that $N_e^Q r_e \leq c_e$, the *effective* arrival rate of streaming flows, $\Lambda_s(n_s)$, is given as follows:

$$\begin{aligned} \Lambda_s(n_s) &= \lambda_s P(N_e^Q \leq K_e(n_s+1) \mid N_s^Q = n_s) \\ &= \lambda_s \frac{1-\rho_e(n_s)^{K_e(n_s+1)+1}}{1-\rho_e(n_s)^{K_e(n_s)+1}}. \end{aligned}$$

Hence, it follows that, for 0≤ $n_s \leq \lfloor \frac{c}{r_s} \rfloor$:

$$P(N_s^Q = n_s) = \frac{\prod_{i=0}^{n_s-1}\Lambda_s(i)}{n_s!\mu_s^{n_s}} P(N_s^Q = 0),$$

where P(N_s^Q=0) can be computed using $\sum_{n_s=0}^{\lfloor \frac{c}{r_s} \rfloor} P(N_s^Q = n_s)$=1. Consequently, it follows that:

$$P(N_e^Q = n_e) = \sum_{n_s=0}^{\lfloor \frac{c}{r_s} \rfloor} \mathbb{P}^{N_e^Q=n_e}_{N_s^Q=n_s} P(N_s^Q = n_s).$$

In particular, the *conditional* blocking probability of newly-arrived streaming flows is $P(N_e^Q r_e + (n_s+1)r_s > c \mid N_s^Q = n_s)$. Un-conditioning on N_s^Q, and noting that blocking can occur only for $\lfloor \frac{c_s}{r_s} \rfloor \leq n_s \leq \lfloor \frac{c}{r_s} \rfloor$, the blocking probability for streaming flows, p_s^Q, is given as follows:

$$\begin{aligned} p_s^Q &= \sum_{n_s=\lfloor \frac{c_s}{r_s} \rfloor}^{\lfloor \frac{c}{r_s} \rfloor} \mathbb{P}^{N_e^Q > K_e(n_s+1)}_{N_s^Q=n_s} P(N_s^Q = n_s) \\ &= 1 - \frac{1}{\lambda_s} \sum_{n_s=\lfloor \frac{c_s}{r_s} \rfloor}^{\lfloor \frac{c}{r_s} \rfloor} \Lambda_s(n_s) P(N_s^Q = n_s). \end{aligned}$$

The corresponding blocking probability for elastic flows, p_e^Q, is given as follows:

$$p_e^Q = \sum_{n_s=0}^{\lfloor \frac{c}{r_s} \rfloor} \mathbb{P}^{N_e^Q \geq K_e(n_s)}_{N_s^Q=n_s} P(N_s^Q = n_s).$$

B. Fluid Approximation for Elastic Flows

For the fluid approximation, denoted by $\mathbf{A(F)}$, we assume that the dynamics of elastic flows are much slower than those of streaming flows, i.e.,

$$\frac{c - r_s E[N_s]}{f_e} + \lambda_e << \mu_s E[N_s] + \lambda_s, \qquad (8)$$

which is certainly true if $\frac{c}{f_e} + \lambda_e << \lambda_s$. This assumption is valid when we consider the combination of voice calls (streaming) and large file transfer (elastic) applications. Under this assumption, the dynamics of streaming flows can be studied by considering a fixed number of elastic flows. Similar to **A(Q)**, we will construct an approximating two-dimensional process under the assumption that N_s immediately reach steady state, whenever N_e changes. This approximation will be reflected in the notation by adding a superscript ^{F} whenever not doing so might give rise to confusion.

From the capacity constraint, it follows that $n_s r_s \leq c - n_e r_e$. By modeling the streaming flows as an Erlang-loss queue with finite capacity $K_s(n_e)$, it follows that:

$$\begin{aligned} \mathbb{P}^{N_s^F=n_s}_{N_e^F=n_e} &\equiv P(N_s^F = n_s \mid N_e^F = n_e) \\ &= \frac{\frac{\rho_s^{n_s}}{n_s!}}{\sum_{i=0}^{K_s(n_e)} \frac{\rho_s^i}{i!}}, \qquad (9) \end{aligned}$$

where $\rho_s = \frac{\lambda_s}{\mu_s}$. As before, we emphasize that the above expression depends on the holding time distribution only through its mean.

Next, we consider the dynamics of elastic flows. When N_e^F=$n_e > 0$, elastic flows depart at a rate

$\mu_e(n_e)$ given as follows:

$$\begin{aligned}\mu_e(n_e) &= \frac{E[c - N_s^F r_s \mid N_e^F = n_e]}{f_e} \\ &= \sum_{n_s=0}^{K_s(n_e)} \frac{c - n_s r_s}{f_e} \mathbb{P}_{N_e^F=n_e}^{N_s^F=n_s}.\end{aligned}$$

Hence, from the admission control conditions and Eq. (9), the *effective* arrival rate of elastic flows, $\Lambda_e(n_e)$, is given as follows:

$$\begin{aligned}\Lambda_e(n_e) &= \lambda_e \cdot \mathbb{P}_{N_e^F=n_e}^{N_s^F \le K_s(n_e+1)} \\ &= \lambda_e \sum_{l=0}^{K_s(n_e+1)} \frac{\frac{\rho_s^l}{l!}}{\sum_{i=0}^{K_s(n_e)} \frac{\rho_s^i}{i!}}.\end{aligned}$$

Using Cohen's results for his generalized PS model [8] for general service times with service rate $\mu_e(n_e)$ and arrival rate $\Lambda_e(n_e)$, it follows that, for 0$\le n_e \le \lfloor \frac{c_e}{r_e} \rfloor$:

$$P(N_e^F = n_e) = \prod_{i=0}^{n_e-1} \frac{\Lambda_e(i)}{\mu_e(i+1)} P(N_e^F = 0), \tag{10}$$

where P(N_e^F=0) can be computed using $\sum_{n_e=0}^{K_s(n_e+1)} P(N_e^F = n_e)$=1. Consequently, $P(N_s^F = n_s)$ is obtained by substituting Eq. (9) and Eq. (10) into the following expression:

$$P(N_s^F = n_s) = \sum_{n_e=0}^{\lfloor \frac{c_e}{r_e} \rfloor} \mathbb{P}_{N_e^F=n_e}^{N_s^F=n_s} \cdot P(N_e^F = n_e).$$

On the other hand, the newly-arrived elastic flow is blocked if n_e = $\lfloor \frac{c_e}{r_e} \rfloor$ or if $N_s^F r_s + (n_e+1) r_e > c$. Hence, the blocking probability for elastic flows, p_e^F, is given by:

$$\begin{aligned}p_e^F &= \sum_{n_e=0}^{\lfloor \frac{c_e}{r_e} \rfloor - 1} \mathbb{P}_{N_e^F=n_e}^{N_s^F > K_s(n_e+1)} \cdot P(N_e^F = n_e) \\ &+ P(N_e^F = \left\lfloor \frac{c_e}{r_e} \right\rfloor).\end{aligned}$$

The corresponding blocking probability for streaming flows, p_s^F, is given as follows:

$$p_s^F = \sum_{n_e=0}^{\lfloor \frac{c_e}{r_e} \rfloor} \mathbb{P}_{N_e^F=n_e}^{N_s^F \ge K_s(n_e} P(N_e^F = n_e).$$

IV. Performance Evaluation

In this section, we evaluate the performance of a single UMTS cell with elastic and streaming requests (Fig. 1) through simulation for the following cell parameters: *c* = 1000 Kbps, c_s = 500 Kbps, r_s = 60 Kbps and r_e = 40 Kbps. While the arrival processes are assumed to be Poisson, we consider the following distributions for (d_s, s_e), given that E[d_s] = $\frac{1}{\mu_s}$ and E[s_e] = f_e:

Hyper-exponential distribution : A common distribution that can be used to characterize the behavior of (d_s, s_e) is the hyper-exponential distribution *with balanced means*, which is defined as follows (cf.[9], p. 359):

$$\forall d \ge 0, P(d_s > d) = \frac{a_s e^{-a_s d \mu_s} + e^{\frac{-d\mu_s}{a_s}}}{a_s + 1},$$

$$\forall s \ge 0, P(s_e > s) = \frac{a_e e^{\frac{-a_e s}{f_e}} + e^{\frac{-s}{a_e f_e}}}{a_e + 1}.$$

The parameters (a_s, a_e) completely characterizes the behavior of (d_s, s_e) respectively and can be interpreted as follows: A fraction $\frac{a_s}{a_s+1}$ ($\frac{a_e}{a_e+1}$) of small streaming (elastic) requests of mean duration (size) $\frac{1}{a_s \mu_s}$ ($\frac{f_e}{a_e}$)and a fraction $\frac{1}{a_s+1}$ ($\frac{1}{a_e+1}$) of large streaming (elastic) requests of mean duration (size) $\frac{a_s}{\mu_s}$ ($a_e f_e$). Increasing the parameter a_s (a_e) increases the variance of d_s (s_e). If a_s (or a_e) = 1, the hyper-exponential distribution reduces to an exponential distribution.

Erlang distribution : Another common distribution that is useful for characterizing the behavior of d_s is the Erlang distribution, which has the following density:

$$\forall d \ge 0 \text{ and } k > 0, f_s(d) = \frac{k\mu_s (k\mu_s d)^{k-1}}{(k-1)!} e^{-k\mu_s d},$$

where Var[d_s]=$\frac{1}{k\mu_s^2}$. Hence, a larger value of k implies a smaller variance for d_s. If k=1, the Erlang distribution reduces to an exponential distribution.

Once the distribution of (d_s, s_e) is selected, we characterize each simulation run according to the following procedure:

1. Fix the total offered traffic by choosing the *loading factor*, $\alpha > 0$, where u_e + u_s = α c,
u_e = $\lambda_e f_e$ and u_s = $\frac{\lambda_s r_s}{\mu_s}$;
2. For each α, fix the traffic *mix*, $\frac{u_e}{\alpha c}$, by choosing u_e, $0 \le u_e \le \alpha$ c;
3. For each traffic mix, select (λ_e, λ_s) to fit one of the following traffic *regimes*:
a. Quasi-stationary Regime (denoted by **S(Q)**), where Eq. (6) is satisfied;
b. Fluid Regime (denoted by **S(F)**), where Eq. (8) is satisfied;
c. Neutral Regime (denoted by **S(N)**), where neither Eq. (6) nor (8) is satisfied.

We note that in Step 3 of the above procedure, f_e and μ_s can be computed once (α, ρ_e, λ_e, λ_s) are specified. The simulation duration, *T*, is selected such that

$$\min\{\lambda_e,\ \lambda_s\} \cdot T \geq N_c,$$

where N_c is chosen (default value = 10000) such that $N_c \cdot \min\{p_e,\ p_s\}$ is not too small.

From the simulations, we compute various performance metrics which are of interest to our integrated-services model. In addition to the queue length as well as blocking probability for each class of traffic, the expected residence time for each admitted elastic request, $E[R_e]$, can be computed in terms of ($E[N_e]$, p_e) using Little' Theorem as follows:

$$E[R_e] = \frac{E[N_e]}{\lambda_e(1-p_e)}.$$

We define the *stretch*, S_e, for each admitted elastic flow by normalizing $E[R_e]$ by f_e as follows:

$$S_e = \frac{E[R_e]}{f_e}.$$

Applying Little's Theorem to each admitted streaming flow, we have the following:

$$E[R_s] = \frac{E[N_s]}{\lambda_s(1-p_s)}.$$

Since $E[R_s]$ = $\frac{1}{\mu_s}$, p_s can be expressed in terms of (λ_s, μ_s) and $E[N_s]$ as follows:

$$p_s = 1 - E[N_s]\frac{\mu_s}{\lambda_s}. \qquad (11)$$

Eq. (11) is used to verify the corresponding expression for p_s obtained for each approximation.

A. Approximation Techniques : Limiting behavior and Accuracy

We apply each approximation technique, **A** $\in$ $\{$**A(Q)**, **A(F)**$\}$, developed in Section III to estimate the performance of our model. We first analyze the limiting behavior of each approximation technique for the following extreme regimes: (i) $u_s \to 0$ (ii) $u_e \to 0$, where u_e + u_s = α c, $\alpha > 0$. The limiting values of (p_e, p_s, S_e) for each approximation technique can be expressed in terms of the cell parameters and α as shown in Table I.

In regime (i), where the offered load comprises almost entirely of elastic requests, the limiting performance for each approximation technique is the same. Almost all streaming requests are admitted ($p_s \to 0$) since $u_s << c_s$, the bandwidth reserved for this class of flows. The performance for elastic requests can be well-approximated by a processor sharing queue with limited capacity, K_e=$\lfloor \frac{c_e}{r_e} \rfloor$, where the blocking probability, $p_e \to \frac{\alpha^{K_e}-\alpha^{K_e+1}}{1-\alpha^{K_e+1}}$ as $u_e \to \alpha\ c$ ($u_s \to 0$). As the capacity K_e increases, the likelihood of blocking is reduced and hence, more elastic requests are admitted (i.e., $E[N_e]$ increases). Since S_e = $\frac{E[N_e]}{u_e(1-p_e)}$, the behavior of S_e with increased K_e is not immediately obvious.

In regime (ii), the offered load comprises almost entirely of streaming requests, and hence, the performance with respect to these requests can be well-approximated by an Erlang-loss queue with limited capacity, K_s = $\lfloor \frac{c}{r_s} \rfloor$, where the blocking probability, p_s, is given in Table I. Almost all elastic requests are admitted (i.e., $p_e \to 0$) since c-$K_s r_s \geq r_e$ for the given set of cell parameters, i.e., even with the maximum number of admitted streaming requests, there is *sufficient* residual capacity to serve an elastic request. However, in contrast to regime (i), the stretch of each admitted elastic request depends on the traffic regime, as shown in Table I.

For a few special cases of α, the expressions for $\lim\limits_{u_s \to 0} p_e$ and $\lim\limits_{u_s \to 0} S_e$ can be further simplified as follows:

$$\lim_{u_s \to 0} p_e \quad = \quad \begin{cases} 0, & \alpha \to 0; \\ \frac{1}{K_e+1}, & \alpha = 1; \\ 1, & \alpha \to \infty. \end{cases}$$

$$\lim_{u_s \to 0} S_e \quad = \quad \begin{cases} \frac{1}{c}, & \alpha \to 0; \\ \frac{K_e+1}{2c}, & \alpha = 1; \\ \frac{K_e}{c}, & \alpha \to \infty. \end{cases}$$

Next, for a fully-loaded cell (i.e., α = 1), we plot (p_e, p_s) as a function of the traffic mix, $\frac{u_e}{c}$, 0$\leq$ u_e $\leq$ c, for each approximation technique in Fig. 2 and Fig. 3 respectively. For each traffic mix, we generate simulation results with 5 sets of traffic parameters for each traffic regime and plot them together with the approximations in Fig. 2-3, assuming that (d_s, s_e) are *exponentially* distributed, i.e., a_e = 1 and k = a_s = 1. Qualitatively, we note that **A(Q)** (**A(F)**) is accurate in the quasi-stationary (fluid) regime in terms of each metric.

A Weighted Approximation for Blocking Probabilities : For the neutral traffic regime, Fig. 2-3 suggest that the blocking probabilities obtained (with simulation) are typically in between **A(Q)** and **A(F)**. Hence, it seems worthwhile to try and obtain a good estimate of the performance for such a regime by *weighing* the performance obtained with **A(Q)** and **A(F)** (denoted **A(W)**).

According to Eq. (6) and Eq. (8), the criteria used to define the traffic regime is the relative dynamics of streaming and elastic flows, given by $\mu_s E[N_s] + \lambda_s$ and $\frac{c - r_s E[N_s]}{f_e} + \lambda_e$ respectively. Hence, a natural approach to define the *weight* allocated to the quasi-

Performance metric, x	p_s	p_e	S_e	
Approximation, **A**	**A(Q), A(F)**	**A(Q), A(F)**	**A(Q)**	**A(F)**
$\lim_{u_s\to 0} \mathrm{x}_{\mathrm{A}}$; $K_{\mathrm{e}} = \left\lfloor c_{\mathrm{e}} / r_{\mathrm{e}} \right\rfloor$	0	$\dfrac{\alpha^{K_e} - \alpha^{K_e+1}}{1-\alpha^{K_e+1}}$	$\dfrac{1-(K_e+1)\alpha^{K_e} + K_e\alpha^{K_e+1}}{c(1-\alpha)(1-\alpha^{K_e})}$	
$\lim_{u_e\to 0} \mathrm{x}_{\mathrm{A}}$; $K_{\mathrm{s}} = \left\lfloor c / r_{\mathrm{s}} \right\rfloor, \hat{u}_s = \alpha c / \mathrm{r}_{\mathrm{s}}$	$\dfrac{\hat{u}_s^{K_s} / K_s!}{\sum_{i=0}^{K_s} \hat{u}_s^{\,i} / i!}$	0	$\dfrac{\sum_{i=0}^{K_s} \hat{u}_s^{\,i} / i!(c-ir_s)}{\sum_{i=0}^{K_s} \hat{u}_s^{\,i} / i!}$	$\dfrac{\sum_{i=0}^{K_s} \hat{u}_s^{\,i} / i!}{\sum_{i=0}^{K_s} \hat{u}_s^{\,i} (c-ir_s) / i!}$

TABLE I

LIMITING PROPERTIES FOR EXTREME REGIMES (I) $u_s \to 0$ AND (II) $u_e \to 0$.

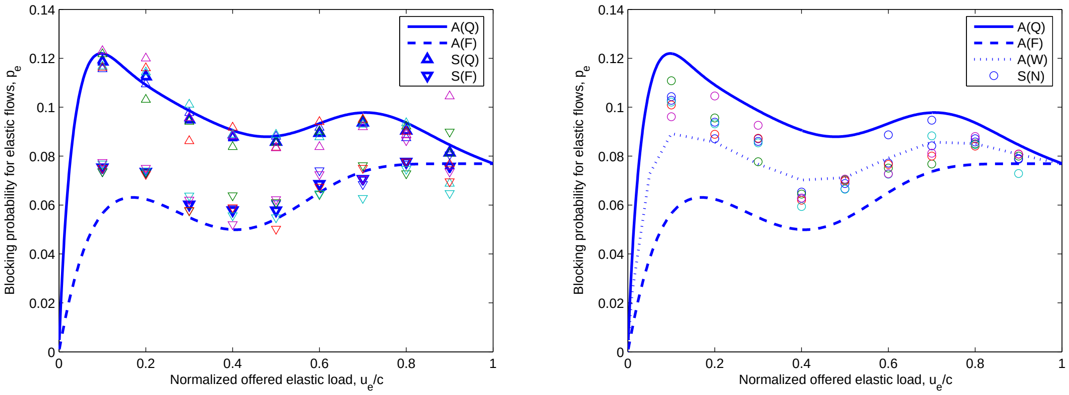

Fig. 2. Blocking probability for elastic requests vs normalized offered elastic load obtained for the 5 cases in quasi-stationary and fluid regimes (left) and neutral regime (right).

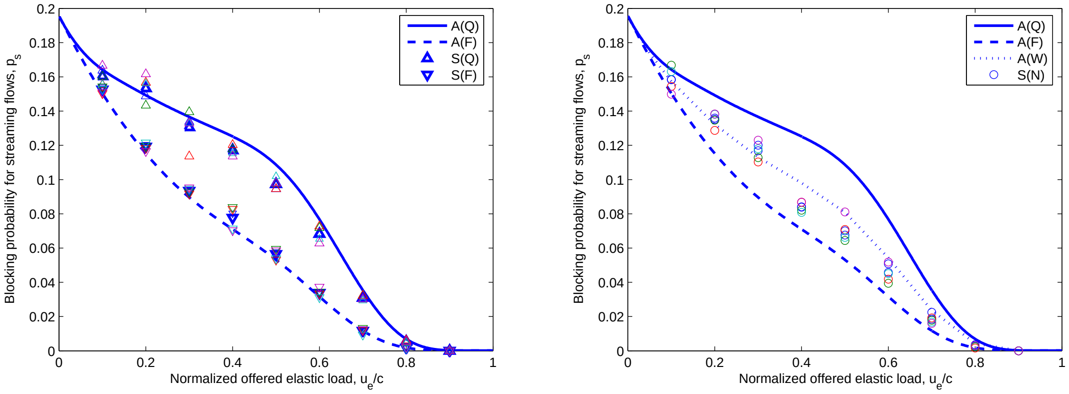

Fig. 3. Blocking probability for streaming requests vs normalized offered elastic load obtained for the 5 cases in quasi-stationary and fluid regimes (left) and neutral regime (right).

stationary approximation, w_Q, is as follows:

$$w_Q = \frac{\frac{c-r_sE[N_s]}{f_e} + \lambda_e}{\frac{c-r_sE[N_s]}{f_e} + \lambda_e + \mu_s E[N_s] + \lambda_s}, \qquad (12)$$

where the corresponding weight allocated to the fluid approximation, w_F = 1- w_Q. In this way, when the dynamics of elastic flows occur at a *faster* rate than that of streaming flows (towards quasi-stationary regime), $w_Q > w_F$ and vice versa. As a result, if $x_{\mathbf{A}}$ denotes the performance metric x obtained with approximation $\mathbf{A}$, then we can define the performance obtained with the weighted approximation as follows:

$$x_{\mathbf{A(W)}} = w_Q x_{\mathbf{A(Q)}} + w_F x_{\mathbf{A(F)}}.$$

Accordingly, $E[N_s] = w_Q E[N_s]_{\mathbf{A(Q)}}$ + (1-w_Q)$E[N_s]_{\mathbf{A(F)}}$, and together with Eq. (12), w_Q can be computed by solving the following quadratic equation:

$$Aw_Q^2 + Bw_Q = C,$$

where

$$\begin{aligned} A &= (E[N_s]_{\mathbf{A(Q)}} - E[N_s]_{\mathbf{A(F)}})(\mu_s f_e - r_s) \\ B &= c + (\lambda_e + \lambda_s) f_e + E[N_s]_{\mathbf{A(F)}}(\mu_s f_e - r_s) \\ &- r_s(E[N_s]_{\mathbf{A(F)}} - E[N_s]_{\mathbf{A(Q)}}) \\ C &= c + \lambda_e f_e - r_s E[N_s]_{\mathbf{A(F)}} \end{aligned}$$

We demonstrate the accuracy of **A(W)** for the case of balanced traffic mix. We consider $w_Q \in$ {0.1, 0.2, $\cdots$, 0.9}, and for each w_Q, we generate simulation runs by selecting 9 sets of traffic parameters. We plot the blocking probabilities obtained alongside the corresponding estimates with **A(Q)**, **A(F)** and **A(W)** in Fig. 4. We observe that the blocking probabilities for both types of requests are well-estimated by **A(W)**.

To quantify the accuracy of the approximations, for a given traffic mix, let $\beta_{S,A}$ denote the proportion of simulation runs such that $|\frac{x_S - x_A}{x_A}| \leq \gamma$, where x_S, x_A is the value of metric *x* obtained with simulation run **S** and approximation **A** respectively, where **S** $\in$ {**S(Q)**, **S(F)**, **S(N)**} and **A** $\in$ {**A(Q)**, **A(F)**}. We say that **A** is an *accurate* approximation for traffic regime **S** if $\beta_{S,A}$ exceeds some threshold β_0. In addition, if $\beta_{S,A(Q)} > \beta_{S,A(F)}$, then we say that approximation **A(Q)** is a better fit in traffic regime **S** than **A(F)**. Some results for the quantitative assessment of the accuracy of each approximation technique are tabulated in Table II. According to Fig. 2-3 and Table II, **A(W)** seems to be a promising approximation for the blocking probabilities in all traffic regimes.

B. Performance Insensitivity with Traffic Parameter Distribution

Next, we evaluate the impact of the distribution of (d_s, s_e) on the performance of the integrated-services system in different traffic regimes. Since each performance measure obtained with the approximation techniques converges to the same value for $u_e \to c$, we focus on the following cases: (i) $\frac{u_e}{c}$ = 0.5 (Balanced traffic mix) and (ii) $\frac{u_e}{c}$ = 0.1 (Dominant composition of streaming traffic).

1) Hyper-exponential distribution for (d_s, s_e): While the simulations have been conducted for [a_e, a_s] = [1,1] in Section IV-A, we repeat the simulations for (a) [a_e, a_s] = [1,100] and (b) [a_e, a_s] = [100,1], where a larger value of a_e (or a_s) implies a larger variance for s_e (d_s). We compute the sample mean for (p_e, p_s, S_e) over all the simulation runs for each traffic mix, and the results are tabulated in Table III. We observe that the performance measures are *almost* insensitive to the variance of the traffic parameters. Quantitatively, in terms of S_e, the performance measures obtained for cases (a) and (b) are within 2% of the corresponding measures obtained assuming exponentially distributed (s_e, d_s).

2) Erlang-k distribution for d_s with exponentially distributed s_e: We repeat the simulations in the previous section for scenarios where s_e is exponentially distributed (i.e., a_e = 1) and d_s is Erlang-*k* distributed, for *k* $\in$ {1,2,3}. The corresponding results are shown in Table IV. We observe that the performance measures are *almost* insensitive to the value of *k* Quantitatively, in terms of S_e, the performance measures obtained for *k* = 2 and *k* = 3 are within 4% and 10% of the corresponding measures obtained assuming exponentially distributed (s_e, d_s) for $\frac{u_e}{c}$ = 0.5 and 0.1 respectively.

C. Impact of traffic load on performance

The results from the previous sections have been obtained assuming a fully-loaded UMTS cell, i.e., α = 1, where u_e + u_s = α c. Here, we plot (p_e, p_s) as a function of α for each approximation technique in Fig. 5, assuming a balanced traffic mix and exponentially distributed (d_s, s_e). For $\alpha \in$ {0.2, 0.4, $\cdots$, 1.8}, we run the simulation for 9 sets of traffic parameters for each traffic regime and plot a representative set of results together with the approximations in Fig. 5. Some results for the quantitative assessment of the accuracy of each approximation technique are tabulated in Table V.

Regarding the performance of elastic flows, we observe that a cross-over point, $\alpha_0 \approx$ 1.3, exists such that for α < (>) α_0, **A(Q)** achieves a better (worse) performance than **A(F)**. On the other hand, the **A(Q)** always results in higher blocking for streaming flows than **A(F)**. According to Table I, the limiting behavior of **A(Q)** and **A(F)** coincides and is given as follows:

$$\lim_{\alpha \to y} p_e = \lim_{\alpha \to y} p_s = \begin{cases} 0, & y = 0; \\ 1, & y = \infty. \end{cases} \quad (13)$$

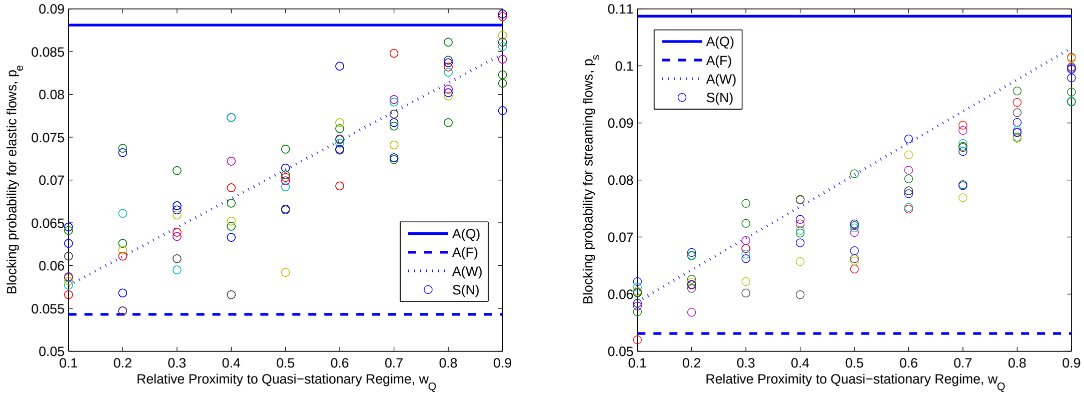

Fig. 4. Blocking probability for elastic (left) and streaming (right) requests for neutral traffic regime, assuming balanced traffic mix, fully-loaded cell and exponentially-distributed $(d_s,\, s_e)$.

S	S(F)			S(N)												S(Q)		
w_Q	~0			0,2			0,4			0,6			0,8			~1		
A	A(Q)	A(F)	A(W)	A(Q)	A(F)	A(W)	A(Q)	A(F)	A(W)	A(Q)	A(F)	A(W)	A(Q)	A(F)	A(W)	A(Q)	A(F)	A(W)
p_e	0,00	0,67	0,89	0,00	0,33	0,56	0,00	0,11	0,67	0,11	0,00	0,89	0,89	0,00	1,00	1,00	0,00	1,00
p_s	0,00	0,67	1,00	0,00	0,11	0,89	0,00	0,00	0,78	0,00	0,00	0,67	0,00	0,00	0,78	0,22	0,00	1,00

TABLE II

Accuracy of approximation techniques in terms of $\beta_{S,A}$ for a fully-loaded UMTS cell, exponentially distributed $(d_s,\, s_e)$, $\gamma = 0.1$ and balanced traffic mix in various traffic regimes.

u_e/c	0,5									0,1								
S	S(Q)			S(F)			S(N)			S(Q)			S(F)			S(N)		
$[a_e,a_s]$	[1,1]	[1,100]	[100,1]	[1,1]	[1,100]	[100,1]	[1,1]	[1,100]	[100,1]	[1,1]	[1,100]	[100,1]	[1,1]	[1,100]	[100,1]	[1,1]	[1,100]	[100,1]
p_e	0,086	0,087	0,086	0,057	0,059	0,055	0,069	0,072	0,069	0,122	0,117	0,119	0,075	0,075	0,072	0,103	0,105	0,101
p_s	0,097	0,100	0,100	0,057	0,057	0,055	0,070	0,075	0,071	0,164	0,161	0,164	0,152	0,153	0,153	0,157	0,160	0,157
S_e	10,159	10,154	10,163	10,088	10,152	10,018	10,114	10,291	10,228	7,916	7,817	7,892	5,996	6,092	5,927	7,606	7,720	7,509

TABLE III

Impact of distribution of traffic parameters $(s_e,\, d_s)$ on $(p_e,\, p_s,\, S_e)$ for $\frac{u_e}{c}$=0.5 and 0.1.

u_e/c	0,5									0,1								
S	S(Q)			S(F)			S(N)			S(Q)			S(F)			S(N)		
k	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
p_e	0,086	0,087	0,085	0,057	0,056	0,058	0,069	0,072	0,071	0,117	0,123	0,120	0,075	0,072	0,072	0,105	0,104	0,110
p_s	0,097	0,099	0,097	0,057	0,057	0,056	0,070	0,075	0,074	0,161	0,165	0,162	0,153	0,152	0,152	0,160	0,161	0,162
S_e	10,088	10,144	10,065	10,088	10,152	10,018	10,114	10,447	10,384	7,817	7,975	7,903	6,092	6,651	6,610	7,720	7,658	7,765

TABLE IV

Impact of distribution of traffic parameters $(s_e,\, d_s)$ on $(p_e,\, p_s,\, S_e)$ for $\frac{u_e}{c}$=0.5 and 0.1.

$$\lim_{\alpha \to y} S_e \quad \approx \quad \begin{cases} \frac{1}{c}, & y = 0; \\ \frac{1}{r_e}, & y = \infty. \end{cases} \tag{14}$$

In fact, according to Fig. 5, for α = 0.2 (1.8), the overall load is sufficiently low (high) such that the performance converges to the limiting performance given in Eq. (13) and (14). Hence, under very low or high loading conditions, the performance can be accurately estimated by **A(Q)** or **A(F)**. For intermediate values of traffic load, according to Table V, **A(Q)** (**A(F)**) is accurate in the quasi-stationary (fluid) traffic regime.

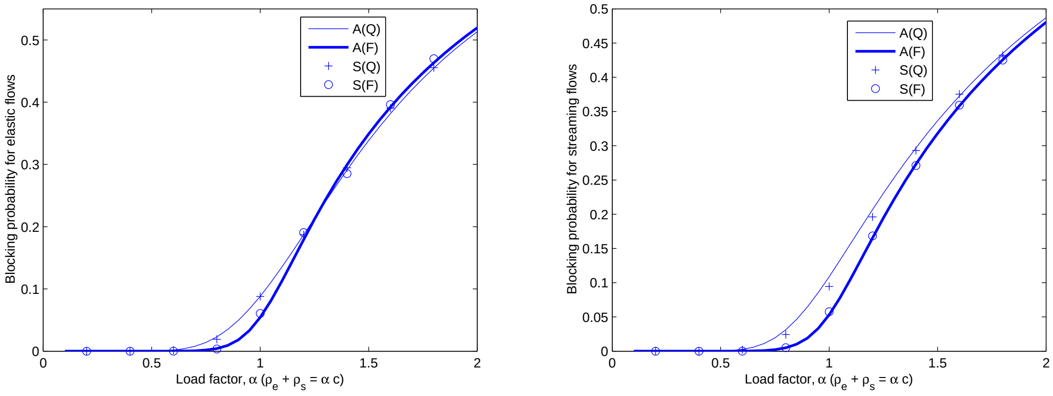

Fig. 5. Blocking probability for elastic (left) and streaming (right) traffic vs traffic load obtained with simulation and various approximation techniques (Balanced traffic mix).

α		0,2			0,8			1,0			1,2			1,8		
A		A(Q)	A(F)	A(L)	A(Q)	A(F)	A(L)	A(Q)	A(F)	A(L)	A(Q)	A(F)	A(L)	A(Q)	A(F)	A(L)
S	S(Q)	1,00	1,00	1,00	1,00	0,00	0,00	1,00	1,00	0,00	1,00	0,00	0,00	1,00	1,00	1,00
	S(F)	1,00	1,00	1,00	0,00	0,89	1,00	1,00	1,00	0,00	0,00	1,00	0,00	1,00	1,00	1,00
	S(N)	1,00	1,00	1,00	0,00	0,00	0,00	0,89	0,78	0,00	0,22	1,00	0,00	1,00	1,00	1,00

TABLE V

TABULATION OF $\beta_{S,A}$ TO ASSESS ROBUSTNESS OF APPROXIMATION TECHNIQUES, $\mathbf{A}$, IN VARIOUS TRAFFIC REGIMES, $\mathbf{S}$ ([a_e, a_s]=[1,1], $\frac{u_e}{\alpha c}$=0.5 AND γ = 0.05).

V. CONCLUSIONS AND FUTURE WORK

In this study, we evaluate the performance of an admission control strategy for integrated services in a single UMTS cell. The integrated services comprise elastic and streaming requests, and priority is given to streaming requests through resource reservation. We model the UMTS cell as a link with fixed capacity that is independent of the actual resource allocation to individual requests.

We develop a quasi-stationary (fluid) approximation to estimate the cell performance in traffic regimes where the dynamics of streaming requests take place on a much slower (faster) time scale than those of elastic requests. In addition, we propose a weighted version of the quasi-stationary and fluid approximations for traffic regimes where the relative dynamics of both request types are similar (neutral traffic regime). Simulation results suggest that the cell performance is almost insensitive to traffic parameter distributions, and is accurately estimated by the proposed approximations.

We currently explore other approximation techniques for the neutral traffic regime. In addition, our existing model for admission control is conservative since it assumes that all users are at the edge of the cell. We consider extensions to the model by relaxing the assumptions and explicitly taking into account the location of the users.

REFERENCES

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

[2] 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, pp. 69–81, September 2001.

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

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

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

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

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

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

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