Performance Analysis of Admission Control for Integrated Services with Minimum Rate Guarantees

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

[*]Eindhoven University of Technology,
Department of Mathematics and Computer Science,
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— **In this paper, we focus on an admission control strategy for streaming and elastic users that enforces a minimum rate guarantee for each elastic user through pre-emptive capacity reservation. We propose approximations to estimate the performance of this strategy. We apply time-scale decomposition for the limiting regimes, and for non-limiting regimes we propose a novel weighted approximation. Simulation results suggest that the performance is almost insensitive to traffic parameter distributions, and is well estimated by our proposed approximations. Our work is motivated by the integration of services in 3rd generation wireless systems such as UMTS and CDMA 2000.**

I. Introduction

Future generation broadband networks are expected to support a large variety of applications, typically 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, with the transfer time as 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 performance of elastic and streaming traffic integration have been published [1], [2], [3], [4], [5], [6], [7]. In terms of resource sharing policy, the classical approach is to give *absolute* priority to streaming flows (through head-of-line packet marking) 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 results and approximations for *limiting* traffic regimes based on time scale decomposition.

The above works do not consider the provision of minimum rate guarantees. However, such guarantees are important to maintain satisfactory perceived quality of service, especially for elastic users. Hence, in this paper, we propose such a strategy to enforce minimum rate guarantees through pre-emptive capacity reservation. To model a link that supports such a strategy, we apply both processor sharing and Erlang-loss models and develop approximations for limiting regimes based on time-scale decomposition. Comparison with numerical simulations suggests that these approximations are accurate and form performance bounds. This led us to propose a novel weighted approximation for non-limiting regimes.

We describe our model and admission control strategy in detail in Section II. We present the time-scale decomposition analysis in Section III and show representative numerical results in Section IV. Some concluding remarks and future directions are

outlined in Section V. We illustrate the applicability of our model for evaluating the downlink performance of cellular systems in the Appendix.

II. Model

We consider a link whose limited *resource* (c Kbps in total) is shared amongst streaming and elastic requests with fixed and minimum rate requirements of r_s (Kbps) and r_e (Kbps) respectively. We assume that streaming and elastic requests arrive as independent Poisson processes with rate λ_s and λ_e respectively. The duration of each *admitted* streaming request, d_s, is generally distributed with mean $\frac{1}{\mu_s}$ (sec). The corresponding size of each admitted elastic request, s_e, is generally distributed with mean f_e (bits).

A part of the total resource, $c_s < c$, is *reserved* for streaming requests: at any time, $\frac{c_s}{r_s}$ streaming requests will be guaranteed admission; however, if there are at least $\frac{c_s}{r_s}$ ongoing streaming requests, then a new streaming request will be admitted as long as the minimum rate r_e can be guaranteed for ongoing elastic requests. On the other hand, an elastic request will be admitted as long as ongoing requests can maintain their rate requirements and the number of ongoing elastic requests does not exceed $\frac{c-c_s}{r_e}$. Note that the capacity unused by streaming requests may be equally shared amongst elastic requests; however, this surplus is immediately re-allocated to streaming requests when a new streaming request arrives.

The above admission control strategy can be quantified as follows: let (N_s, N_e) be the number of on-going streaming and elastic requests respectively. Due to the total resource constraint and resource reservation for streaming traffic, $N_s r_s$ + $N_e r_e \leq c$ and $N_e r_e \leq c$-c_s. The second inequality ensures that all active elastic flows can be accommodated with their minimum rate by the ensured capacity $c_e = c - c_s$. Consequently, a new elastic request will be accepted only if $N_s r_s$ + (N_e+1)r_e $\leq c$ and (N_e+1)$r_e \leq c$-c_s. On the other hand, a new streaming request will be admitted as long as (N_s+1)r_s + $N_e r_e \leq c$. The model is illustrated in Fig. 1.

Note that a different integrated admission control scheme was proposed in [2]. While this scheme aims to ensure equal blocking probabilities, our scheme enforces rate requirements for both types of traffic.

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.

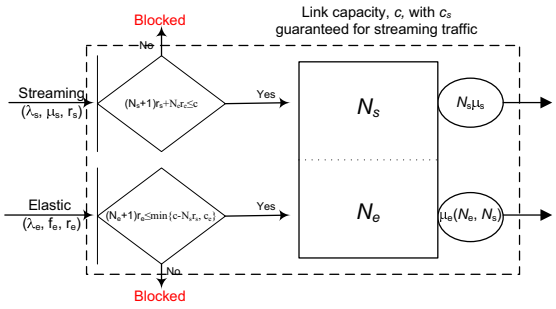

Fig. 1. Model of single link with capacity reservation and minimum rate guarantee for integrated services.

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 (see [1], [5] for similar models), 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

$$\mu_s E[N_s] + \lambda_s << \frac{c - r_s E[N_s]}{f_e} + \lambda_e, \quad (1)$$

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],

$$\begin{aligned} \mathbb{P}^{N_e^Q=n_e}_{N_s^Q=n_s} &\equiv P(N_e^Q = n_e \mid N_s^Q = n_s) \\ &= \frac{\rho_e(n_s)^{n_e}(1-\rho_e(n_s))}{1-\rho_e(n_s)^{K_e(n_s)+1}}, \end{aligned} \quad (2)$$

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. (2) 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:

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

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

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

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:

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

B. Fluid Approximation for Elastic Flows

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

$$\frac{c - r_s E[N_s]}{f_e} + \lambda_e << \mu_s E[N_s] + \lambda_s, \quad (3)$$

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 $\mathbf{A(Q)}$, we will construct an approximating two-dimensional process under the assumption that N_s immediately reaches 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:

$$\mathbb{P}^{N_s^F=n_s}_{N_e^F=n_e} = \frac{\frac{\rho_s^{n_s}}{n_s!}}{\sum_{i=0}^{K_s(n_e)} \frac{\rho_s^i}{i!}}, \quad (4)$$

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:

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

Hence, from the admission control conditions and Eq. (4), the *effective* arrival rate of elastic flows, $\Lambda_e(n_e)$, is given as follows:

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

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 $0 \leq n_e \leq \lfloor \frac{c_e}{r_e} \rfloor$:

$$P(N_e^F = n_e) = \prod_{i=0}^{n_e-1} \frac{\Lambda_e(i)}{\mu_e(i+1)} P(N_e^F = 0), \quad (5)$$

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. (4) and Eq. (5) into the following expression:

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

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:

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

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

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

IV. Performance Evaluation

In this section, we evaluate the performance of an isolated link with elastic and streaming requests (Fig. 1) through simulation for the following parameters: c = 1000 Kbps, c_s = 500 Kbps, r_s = 60 Kbps and r_e = 40 Kbps. We consider the following distributions for (s_e, d_s), given that E[s_e] = f_e and E[d_s] = $\frac{1}{\mu_s}$:

Hyper-exponential distribution : A common distribution used to characterize the behavior of s_e is the hyper-exponential distribution *with balanced means*, which is defined as follows (cf.[9], p. 359): $\forall s \geq 0, P(s_e > s) = \frac{a_e e^{\frac{-a_e s}{f_e}} + e^{\frac{-s}{a_e f_e}}}{a_e+1}$. The parameter a_e completely characterizes the behavior of s_e and can be interpreted as follows: A fraction $\frac{a_e}{a_e+1}$ of small elastic requests of mean size $\frac{f_e}{a_e}$ and a fraction $\frac{1}{a_e+1}$ of large elastic requests of mean size $a_e f_e$. Increasing a_e increases the variance of s_e and the special case of a_e = 1 corresponds to the exponential distribution.

Erlang distribution : A common distribution to characterize the behavior of d_s is the Erlang distribution, which has the following density:

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

It reduces to an exponential distribution for k=1; increasing k reduces the variance for d_s.

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*, $\tau > 0$, where $u_e + u_s = \tau\ c$, $u_e = \lambda_e f_e$ and $u_s = \frac{\lambda_s r_s}{\mu_s}$;
2. For each τ, fix the traffic *mix*, $\frac{u_e}{\tau c}$, by choosing u_e, $0 \leq u_e \leq \tau\ c$;
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. (1) is satisfied;
   b. Fluid Regime (denoted by **S(F)**), where Eq. (3) is satisfied;
   c. Neutral Regime (denoted by **S(N)**), where neither Eq. (1) nor (3) is satisfied.

We note that in Step 3 of the above procedure, f_e and μ_s can be computed once (τ, 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 the blocking probability for each type of request. 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's Law as follows:

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

We define the *stretch*, S_e, by normalizing $E[R_e]$ with f_e as follows:

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

A. Performance Insensitivity with Traffic Parameter Distribution

First, we evaluate the impact of the distribution of (d_s, s_e) on the performance of a fully-loaded link (τ = 1) in different traffic regimes. We consider the following traffic mix: (i) $\frac{u_e}{c}$ = 0.5 (Balanced traffic mix) and (ii) $\frac{u_e}{c}$ = 0.1 (Dominant composition of streaming traffic).

For each case, we define 3 sets of simulations as follows: **Case I** [a_e, k] = [1, 1], **Case II** [a_e, k] = [100, 1] and **Case III** [a_e, k] = [1, 3]. For each set, we compute the sample mean for (p_e, p_s, S_e) over all the simulation runs at each traffic mix, and the results are tabulated in Table I. We observe that the performance measures are *almost* insensitive to the traffic parameter distributions, thus justifying the insensitive approximations proposed here.

B. A Weighted Approximation for Blocking Probabilities

For each scenario in **Case I**, 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, alongside the corresponding simulation results. Qualitatively, we note that **A(Q)** (**A(F)**) is accurate in the quasi-stationary (fluid) regime for each metric. The non-monotonicity of the blocking probability for elastic flows is due to the counter-acting effects of reservation and change of traffic composition. Notice that this does not affect the blocking probability for streaming flows.

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 estimate the performance metric x in such a regime by *weighing* the corresponding metrics obtained with **A(Q)** and **A(F)** (denoted **A(W)**) as follows:

$x_{\mathbf{A(W)}} = w_Q x_{\mathbf{A(Q)}} + (1 - w_Q) x_{\mathbf{A(F)}}$,

where w_Q is the weight allocated to **A(Q)** and $x_{\mathbf{A}}$ is the value of x obtained with approximation **A**.

According to Eq. (1) and Eq. (3), 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 w_Q is as follows:

$$w_Q = \frac{\frac{c - r_s E[N_s]}{f_e} + \lambda_e}{\frac{c - r_s E[N_s]}{f_e} + \lambda_e + \mu_s E[N_s] + \lambda_s}. \quad (6)$$

In this way, when the dynamics of elastic flows occur at a *faster* rate than that of streaming flows (towards quasi-stationary regime), $w_Q > 0.5$ and vice versa. Accordingly, $E[N_s] = w_Q E[N_s]_{\mathbf{A(Q)}} + (1\text{-}w_Q) E[N_s]_{\mathbf{A(F)}}$, and together with Eq. (6), w_Q can be computed by solving the following quadratic equation:

$$A w_Q^2 + B w_Q = C,$$

where

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

We demonstrate the accuracy of **A(W)** for the case of a 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)**.

V. Conclusions and Future Work

In this study, we evaluate the performance of an admission control strategy that ensures minimum and fixed rates for elastic and streaming requests respectively in an isolated link. We develop approximations to estimate the performance in limiting traffic regimes where the dynamics of both types of requests take place at significantly different time scales. Based on the former, we propose a weighted approximation for non-limiting traffic regimes. Simulation results suggest that the performance is almost insensitive to traffic parameter distributions, and is accurately estimated by the proposed approximations. Our model can, see e.g. [3], be used to assess the performance of downlink communication in 3^{rd} generation cellular systems when the difference in distances to the base station within the cell can be neglected. We are currently investigating extensions to include signal attenuation due to path loss. We also plan to investigate more enhanced weighing schemes of the proposed approximations.

Appendix

Let us consider downlink transmissions in UMTS (with W-CDMA in FDD mode) to elastic and streaming users, where the base station can transmit to all active users (denoted by $\mathbb{A}$) simultaneously. 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*

u_e/c	0,5									0,1								
S	**S(Q)**			**S(F)**			**S(N)**			**S(Q)**			**S(F)**			**S(N)**		
Case	**I**	**II**	**III**	**I**	**II**	**III**	**I**	**II**	**III**	**I**	**II**	**III**	**I**	**II**	**III**	**I**	**II**	**III**
p_e	0,086	0,086	0,085	0,057	0,055	0,058	0,069	0,069	0,071	0,122	0,119	0,120	0,075	0,072	0,072	0,103	0,101	0,110
p_s	0,097	0,100	0,097	0,057	0,055	0,056	0,070	0,071	0,074	0,164	0,164	0,162	0,152	0,153	0,152	0,157	0,157	0,162
S_e	10,159	10,163	10,065	10,088	10,018	10,018	10,114	10,228	10,384	7,916	7,892	7,903	5,996	5,927	6,610	7,606	7,509	7,765

TABLE I

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.

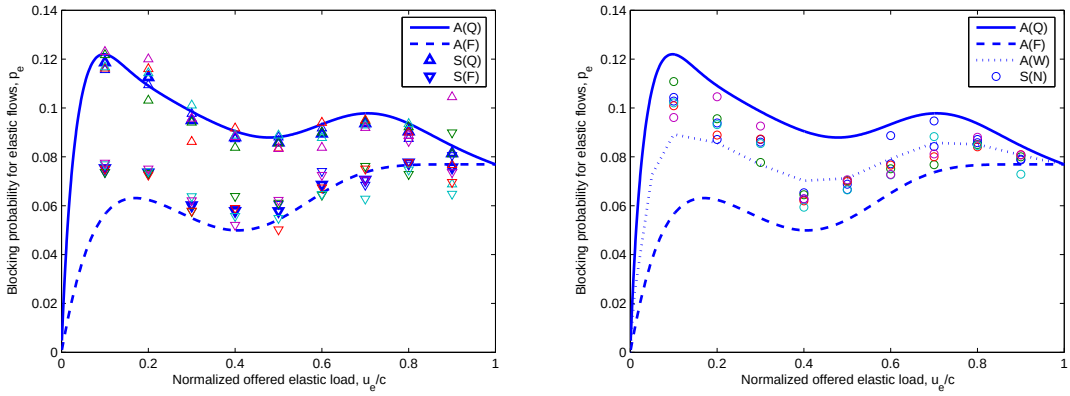

Fig. 2. Blocking probability for elastic requests vs normalized offered elastic load obtained for the 5 cases in quasi-stationary and fluid regimes (left) and neutral regime (right).

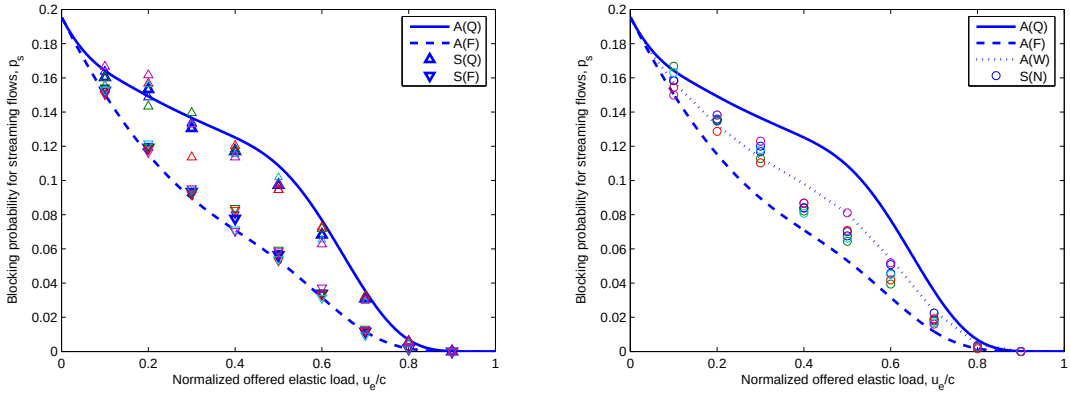

Fig. 3. Blocking probability for streaming requests vs normalized offered elastic load obtained for the 5 cases in quasi-stationary and fluid regimes (left) and neutral regime (right).

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 intra / inter-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, and is given by $I_r^a = \alpha \sum_{j\in\mathbb{A}, j\neq u} P_j \Gamma_u \leq \alpha(P - P_u)\Gamma_u$, where α is the code non-orthogonality factor. On the other hand, the inter-cell interference is due to the base stations' transmissions in *neighboring* cells.

To achieve a given target error probability (assumed to be zero here), it is necessary that for each active user u, $\left(\frac{E_b}{N_0}\right)_u \geq \epsilon$, for some threshold

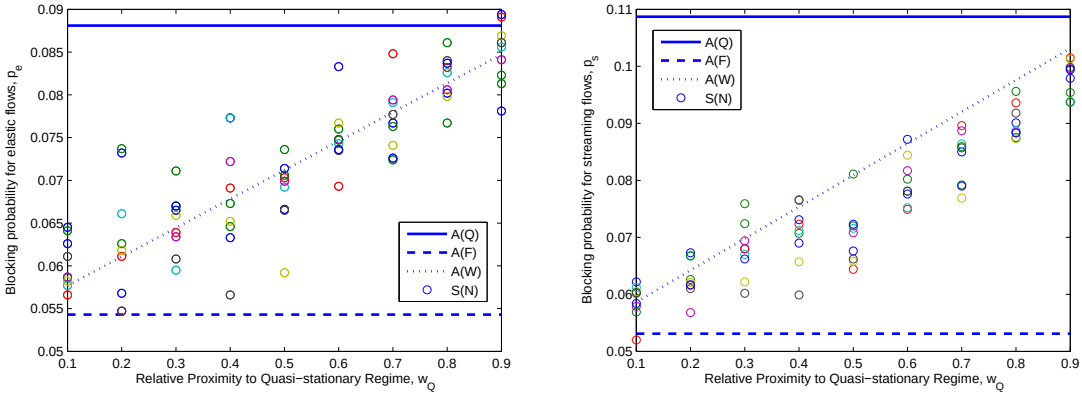

Fig. 4. Blocking probability for elastic (left) and streaming (right) requests for neutral traffic regime, assuming balanced traffic mix, fully loaded cell and exponentially distributed (d_s, s_e).

ϵ, which is assumed to be the same for all users. Hence, for each active user u, we have the following:

$$R_u \leq \frac{WP_u}{\epsilon(\alpha(P-P_u)+\frac{\eta+I_u^r}{\Gamma_u})}. \quad (7)$$

which can be re-written as follows:

$$\frac{R_u}{W+\alpha\epsilon R_u} \leq \frac{P_u}{\epsilon(\frac{\eta+I_{max}^r}{\Gamma_{min}}+\alpha P)}. \quad (8)$$

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

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

Next, let's assume that a fixed portion of P, P_s, is reserved for streaming traffic. Then, we have that $\sum_{u\in\mathbb{E}} P_u \leq P - P_s = P_e$.

By summing over the set of N_e *active* elastic users, we obtain the following:

$$\frac{N_e r_e}{W+\alpha\epsilon r_e} \leq \frac{P_e}{\epsilon(\frac{\eta+I_{max}^r}{\Gamma_{min}}+\alpha P)}. \quad (9)$$

Finally, summing Eq. (8) over $\mathbb{A}$, and noting that $R_u = r_s$ for streaming users, we have:

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

where $r = \max(r_e, r_s)$.

Hence, our model can be applied to the downlink transmission scenario in UMTS by defining $c = \frac{P(W+\alpha\epsilon r)}{\epsilon(\frac{\eta+I_{max}^r}{\Gamma_{min}}+\alpha P)}$ and choosing $c_s = \frac{P_s}{P}$. We note that the resulting cell capacity c is conservative, since the admission control criteria is defined based on users at the edge of the cell (where $I_u^r = I_{max}^r$ and $\Gamma_u = \Gamma_{min}$).

ACKNOWLEDGMENT

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.

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] M. Fischer and T. Harris, "A model for evaluating the performance of an integrated circuit- and packet- switched multiplex structure," *IEEE Transactions on Communications*, vol. 24, pp. 195–202, February 1976.

[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.