

Simple Approximations of Performance Metrics for a link integrating elastic and streaming traffic

Henda Ben Cheikh, Olivier Brun, Jean-Marie Garcia

▶ **To cite this version:**

Henda Ben Cheikh, Olivier Brun, Jean-Marie Garcia. Simple Approximations of Performance Metrics for a link integrating elastic and streaming traffic. The International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (PEMWN 2014), Nov 2014, Sousse, Tunisia. hal-02062271

HAL Id: hal-02062271

https://laas.hal.science/hal-02062271

Submitted on 8 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Simple Approximations of Performance Metrics for a link integrating elastic and streaming traffic

H. Ben Cheikh, O. Brun, J.M. Garcia

LAAS-CNRS
Université de Toulouse
7 Avenue du Colonel Roche
31077 Toulouse, France.
email: {hbenchei,brun,jmg }@laas.fr

Abstract. We consider a flow-level model of a single-bottleneck link where elastic flows compete with fixed-rate streaming flows. We assume that a call-admission control mechanism is used to limit the overall volume of streaming traffic. We further assume that streaming flows are served with priority, the remaining bandwidth being fairly shared by elastic flows according to Balanced Fairness. We derive exact performance results for streaming flows, and propose two different approximations for elastic flows. The first approximation provides insensitive performance bounds, whereas the second one is based on a quasi-stationary assumption. Simulation results show that the performances are well estimated with the proposed approximations.

Keywords: Performance estimation, streaming traffic, elastic traffic, balanced fairness.

1 Introduction

Internet traffic consists of data sent by different kinds of applications. Given the large variety of applications, it is useful to distinguish between two broad classes of flows: elastic and streaming flows. Elastic flows correspond to the transfer of digital documents (e.g., web page, file, emails) and adapt their transmission rate to the available bandwidth by means of a transport protocol typically Transmission Control Protocol (TCP). They are characterized by a fixed size and variable duration which determine user perceived performance. The time required to transfer a document or equivalently the throughput achieved for a document transfer constitute the main performance criterion for an elastic flow. Streaming flows are produced by audio and video applications such as video streaming and voice over IP services. This kind of traffic is often associated with applications, which use User Datagram Protocol (UDP). They are characterized by a fixed rate and variable size which determine user perceived performance. The amount of data received within the admissible delay or equivalently data loss constitute the main performance criterion for a streaming flow. Since streaming flows usually have stringent quality of service requirements, it makes sense to give them some form of priority over elastic ones. If nothing is done to prevent streaming flows from grabbing the whole network capacity, this can however lead to a severe performance degradation for elastic flows [9].

In this paper, we investigate the performance of elastic and fixed-rate streaming flows when call admission control is used to limit the number of simultaneous streaming flows. We consider an idealized flow-level model where the number of ongoing flows randomly varies as new flows are initiated, and where the flow rate adaptation is perfect and instantaneous. Flow-level models have been introduced by Massoulié and Roberts for evaluating Internet performance [11]. Most studies on performance evaluation of internet traffic concern elastic traffic. In [11], Massoulié and Roberts provide the first analytical performance results for networks with multiple resources. An extension to homogeneous hypercubes networks was proposed by Bonald and Proutière [3]. In [6], [2], Bonald et al. derived performance results for a general network topology under balanced fairness. The performance under balanced fairness is much easier to analyse than under other sharing schemes [1]. This is due to the

reversibility of the underlying Markov process, which allows one to derive the equilibrium distribution by a simple recursion. Moreover, the performance of any network whose resources are shared according to balanced fairness is insensitive to traffic characteristics and depends on traffic characteristics through the traffic intensity only. The integration of streaming and elastic traffic has been considered only in few works. Most studies on the integration of streaming and elastic flows where priority is given to streaming flows over elastic flows [9] relies on the assumption that the flows sharing the link are homogeneous, however. In practice, flows have different bandwidth requirements or constraints.

The main contribution of the present paper is to provide simple and explicit performance approximations for heterogeneous elastic flows with differing maximum bit rates so called multi-rate systems [5].

The rest of this paper is organized as follows. In the next section, we describe the basic model of a single bottleneck link shared by elastic and fixed-rate streaming flows. We extend the results to a multi-class system in section 3 and to the multi-rate case in section 4. We give some concluding remarks in Section 5.

2 Basic model

The basic model consists of a single link of capacity C shared by elastic and streaming flows without any rate limit. In the following, elastic flows are called flows of class 1, while streaming flows are considered to be of class 2. Elastic flows arrive according to a Poisson process at rate λ_1 and have a fixed amount of data to transmit of mean $1/\mu_1$ Mbps. Streaming flows arrive according to a Poisson process at rate λ_2, stay for an arbitrarily distributed random duration of mean $\frac{1}{\mu_2}$ during which they send data at a constant bit rate d Mbps. We define $\rho_i = \lambda_i/\mu_i$ as the traffic intensity of class-i. We denote by x_i the number of class-i flows and $\mathbf{x} = (x_1, x_2)$ the network state. We let $\phi_i(\mathbf{x})$ be the rate allocated to class-i flows at a state $\mathbf{x}$. We assume that streaming flows have the priority. Since streaming flows have a constant bit rate, we have $\phi_2(\mathbf{x}) = x_2 d$. In order to limit the bandwidth allocated to streaming traffic to at most $C_s < C$, an admission control mechanism is used: Denoting by $N_s = \frac{C_s}{d}$, an arriving streaming flow is admitted in state $\mathbf{x}$ if and only if $x_2 < N_s$. As well, we assume that at a state $\mathbf{x}$ elastic flows fairly share the remaining bandwidth unused by streaming flows. The bandwidth allocated to flows of the same class are shared equally. A necessary and sufficient condition for stability is,

$$\rho_1 < C - C_s \tag{1}$$

This stability condition is assumed to be satisfied in the sequel.

2.1 Performance metrics for streaming traffic

The main performance metric for streaming flows is the blocking probability. Focusing on the number of accepted streaming flows, the set of allowed states is $\mathcal{X}_s = \{x_2 \ : x_2 \leq N_s\}$, and a streaming flow call is blocked whenever $x_2 \in \mathcal{X}_s^{'} = \{x_2 \ : \ C_s - d \leq x_2\, d \leq C_s\}$. The marginal distribution of x_2 is easily obtained using standard results from the theory of multi-rate loss networks [10]

$$\pi_S(x_2) = \pi_S(\mathbf{0})\, \frac{\rho_2^{x_2}}{x_2!}, \quad x_2 \in \mathcal{X}^s, \tag{2}$$

from which the blocking probability follows

$$B = \sum_{x_2 \in \mathcal{X}_s^{'}} \pi_S(x_2). \tag{3}$$

2.2 Approximate performance metrics for elastic traffic

In this section, we evaluate the performance of elastic flows through their throughput γ_1 defined as the ratio $\rho_1/E[x_1]$ where $E[x_1]$ denotes the expected number of elastic flows in progress. We recall results in absence of streaming traffic. We then extend the results to account for streaming traffic.

Absence of elastic flow In absence of streaming flows, this system reduces to a single processor sharing queue, the model originally defined by Massoulié and Roberts [11]. The stationary distribution of the number of ongoing elastic flows is

$$\pi(\mathbf{x}) = (1-\rho_1)\rho_1. \tag{4}$$

Corresponding to the mean number of elastic flows

$$E[x_1] = \frac{\rho_1}{C-\rho_1}. \tag{5}$$

From which the elastic flow throughput follows

$$\gamma_1 = C - \rho_1. \tag{6}$$

Presence of streaming flow In presence of streaming flows, we propose two performance approximations for elastic flows. We first provide insensitive performance bounds. We then propose performance approximations based on a quasi-stationary assumption.

Insensitive bounds We note that the system can be represented as a network of two processor sharing nodes. At a state $\mathbf{x}$, the corresponding service rate at each node is given by

$$\phi_2(\mathbf{x}) = x_2 d, \qquad 0 \leq x_2 \leq N_s, \tag{7}$$

and,

$$\phi_1(\mathbf{x}) = (C - x_2 d) \qquad 0 \leq x_2 \leq N_s. \tag{8}$$

If the system is insensitive to traffic characteristics (e.g., flow size distribution), one can easily evaluate the stationary distribution of the network state and thus derive all performance metrics. As explained in [2], a necessary condition for insensitivity is the following balance property

$$\frac{\phi_1(\mathbf{x}-e_2)}{\phi_1(\mathbf{x})} = \frac{\phi_2(\mathbf{x}-e_1)}{\phi_2(\mathbf{x})}, \tag{9}$$

for all states $\mathbf{x}$ such that $x_1 > 0$ and $x_2 > 0$. where $e_1 = (1,0)$ and $e_2 = (0,1)$.
Since

$$\frac{\phi_1(\mathbf{x}-e_2)}{\phi_1(\mathbf{x})} > 1 = \frac{\phi_2(\mathbf{x}-e_1)}{\phi_2(\mathbf{x})}. \tag{10}$$

We deduce that the stationary distribution of the network state is sensitive to traffic characteristics. In such a situation, it usually extremely difficult to derive explicit expressions for performance measures without making some specific assumptions about traffic characteristics. However, following the approach described in [4], insensitive stochastic bounds on the system $\mathbf{x}(t)$ at time t, valid for any traffic characteristics, can be derived, provided the monotonicity property holds e.g. removing a customer from any node does not decrease the service rate of any other customer. Note that this monotonicity property is satisfied by many real systems. More precisely, consider a network of k processor-sharing nodes. Formally, the monotonicity property can be stated as follows

$$\phi_i(\mathbf{x} - e_j) \geq \phi_i(\mathbf{x}), \tag{11}$$

for all i, j where i, j corresponds to any processor sharing node in the system, and for all vector states $\mathbf{x} = (x_1, ...x_k)$ such that $x_i > 0$ and $x_j > 0$ where x_i represents the number of flows in node i. It easy to check from equation (10), that this condition is satisfied. The network state $x(t)$ at any time t satisfies

$$x^-(t) \leq x(t) \leq x^+(t); \tag{12}$$

Where $x^-(t)$ and $x^+(t)$ are the states of two virtual insensitive processor sharing systems at time t. The lower and upper bounds $x^-(t)$ and $x^+(t)$ are entirely characterized by the so-called balance function Φ^- and Φ^+, respectively (see [4]). It remain difficult in general to find an explicit expression for these functions. A notable exception is the case of biased networks. The so-called "bias property" holds if nodes can be numbered in such a way for all pairs of nodes i, j such that $i \leq j$:

$$\frac{\phi_i(\mathbf{x} - e_j)}{\phi_i(\mathbf{x})} \leq \frac{\phi_j(\mathbf{x} - e_i)}{\phi_j(\mathbf{x})}. \tag{13}$$

For biased networks, we have

$$\Phi^-(\mathbf{x}) = \left(\prod_{i=1}^{x_k} \phi_k(ie_k) ... \times \prod_{i=1}^{x_1} \phi_1(ie_1 + \sum_j x_j e_j) \right)^{-1}, \tag{14}$$

and,

$$\Phi^+(\mathbf{x}) = \left(\prod_{i=1}^{x_1} \phi_1(ie_1) \times ... \times \prod_{i=1}^{x_k} \phi_k(ie_k + \sum_{j=1}^{k-1} x_j e_j) \right)^{-1}. \tag{15}$$

The stationary distribution of the system $x^-(t)$ and $x^+(t)$, provided they exist, are then given by

$$\pi^-(\mathbf{x}) = \pi^-(0)\Phi^-(\mathbf{x}) \prod_{i=1}^{k} \rho_i^{x_i}, \tag{16}$$

Where $\pi^-(0) = \left(\sum_{\mathbf{x}} \Phi^-(\mathbf{x}) \prod_{i=1}^{k} \rho_i^{x_i} \right)^{-1}$

and,

$$\pi^+(\mathbf{x}) = \pi^+(0)\Phi^+(\mathbf{x}) \prod_{i=1}^{k} \rho_i^{x_i}, \tag{17}$$

Where $\pi^+(0) = \left(\sum_{\mathbf{x}} \Phi^+(\mathbf{x}) \prod_{i=1}^{k} \rho_i^{x_i} \right)^{-1}$.

One easily sees that the bias property holds in this case, yielding the following result.

If $\rho_1 \leq C - C_s$, it follows from (16) and (14) that $\gamma_1^- < \gamma_1$, where γ_1^- denotes the elastic flow throughput for the lower bound

$$\gamma_1^- = \rho_1 / E^-[\mathbf{x}_1] = C - \rho_1 \tag{18}$$

where $E^-[\mathbf{x}_1] = \frac{\rho_1}{C - \rho_1}$.

Similarly it follows from (17) and (15) that $\gamma_1^+ > \gamma_1$, where γ_1^+ denotes the elastic flow throughput for the upper bound,

$$\gamma_1^+ = \rho_1 / E^+[\mathbf{x}_1], \tag{19}$$

where $E^+[\mathbf{x}_1] = \frac{\sum_{x_2=0}^{N_s} \frac{\rho_1}{C-x_2 d-\rho_1} \alpha(x_2)}{\sum_{x_2=0}^{N_s} \alpha(x_2)}$ and $\alpha(x) = \frac{1}{1-\frac{\rho_1}{C-xd}} \frac{\rho_2^x}{d^x x!}$.

Quasi-stationary analysis The basic idea of the Quasi-stationary (QS) assumption is to assume that the number of elastic flows evolves rapidly with respect to the number of streaming flows and thus reach a statistical equilibrium before the number of streaming flows has evolved. We can use the QS assumption as follows. Assuming that x_2 is fixed, we can obtain the mean number $E[x_1|x_2]$ of elastic flows in progress by replacing C with $C(x_2) = C - x_2 d$ in equation 5. It follows, the mean number of class-i elastic flows in progress can be approximated,

$$E[x_1] = \sum_{x_2 \in \mathcal{X}_s} E[x_1|x_2]\, \pi_s(x_2). \tag{20}$$

Equation (20) immediately yields the following approximation for elastic flow throughput

$$\gamma_1 = \rho_1 / \sum_{x_2 \in \mathcal{X}_s} E[x_1|x_2]\, \pi_s(x_2). \tag{21}$$

Example 1. Pour tudier la prcision des approximations, on considre dans un premier temps To study the accuracy of the proposed approximation, we consider as a first example a single link of capacity $C = 30Mbps$ shared by 2 classes of flows. Class-1 correspond to elastic flow whereas class-2 corresponds to streaming flows. Consider that $c = 4Mbps$ and $d = 2Mbps$. Consider that elastic flows represent a proportion 90 % of the total traffic and streaming flows represent 10 % of the total traffic. Define $\alpha = \frac{\lambda_2 \mu_2}{\lambda_1 \mu_1}$.

Figure 1 shows the evolution of the expected number of elastic flows as a function of the offered load for different value of α.

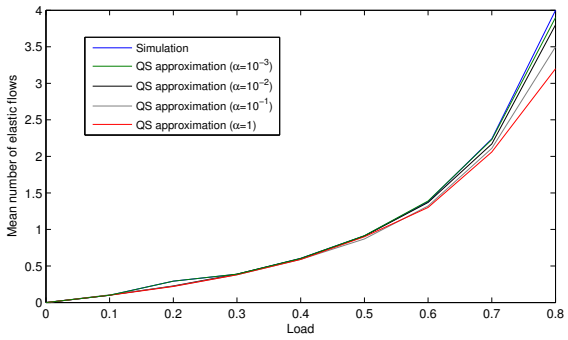


Fig. 1: Mean number of elastic flows in progress

We observe that the results obtained under the QS assumption are close to the numerical results obtained using discrete-event simulation for different value of α. We also note that the relative error decreases when the QS assumption is satisfied i.e., α is small enough.

Figure 2 compare the proposed approximations (QS approximation and insensitive bounds) when α is small enough ($\alpha = 10^{-3}$) with the results obtained with discrete-event simulations for class-1 elastic flows with respect to the traffic load. We note that the QS approximations provide more accurate results then the insensitive bounds. The relative error of the QS approximation is below 3 % in all traffic regimes, whereas the accuracy of the upper and lower insensitive bound decreases as the total link utilization increases.

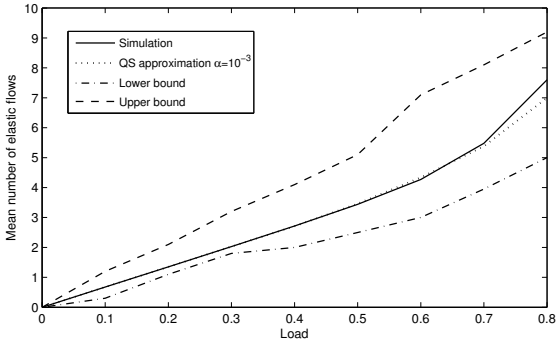


Fig. 2: Mean number of elastic flows in progress

3 Multi-class extension

In this section, we extend the results obtained in section 2 to multi-class systems. To simplify we will use the following notation. We denote by $\mathbb{R}_+$ the set of non-negative real numbers, and $\mathbb{N}$ the set of natural numbers. We let $\mathbf{e}_i$ be the vector with 1 in position i, and 0 elsewhere (the size of the vectors will be clear from the context). Given two vectors $\mathbf{y} = (y_1, \ldots, y_M) \in \mathbb{R}_+^M$ and $\mathbf{x} = (x_1, \ldots, x_M) \in \mathbb{N}^M$, $|\mathbf{x}|$ denotes the sum $\sum_{m=1}^M x_m$, $\mathbf{x}!$ denotes the product $\prod_{m=1}^M x_m!$, and $\mathbf{y}^{\mathbf{x}}$ is the product $\prod_{m=1}^M y_m^{x_m}$.

3.1 Model description

We consider again an isolated bottleneck link of finite capacity C shared by streaming and elastic flows. We let $\mathcal{S}$ and $\mathcal{E}$ be the set of streaming and elastic flow classes, respectively. Streaming flows of class i arrive according to a Poisson process at rate λ_i and have an arbitrarily distributed random duration of mean $1/\mu_i$ during which they send data at a constant bit rate d_i Mbps. Similarly, elastic flows of class i also arrive according to a Poisson process at rate λ_i and have a fixed amount of data to transmit of mean $1/\mu_i$ Mbps. Each class-i elastic flow has a maximum bit rate. To simplify the analysis of the model, we shall assume that this maximum bit rate is identical for all classes. Thus, we shall assume that the maximum bandwidth allocation of an individual flow is c Mbps whatever its class. We denote by x_i the number of ongoing flows of class i. We let $\mathbf{x} = (\mathbf{x}^s, \mathbf{x}^e)$ be the state of the system, where $\mathbf{x}^s = (x_i)_{i\in\mathcal{S}}$ and $\mathbf{x}^e = (x_i)_{i\in\mathcal{E}}$. We denote by $\phi_i(\mathbf{x})$ the bandwidth allocated to class-i flows in state $\mathbf{x}$. Again, we assume that streaming flows have priority. Since streaming flows have a constant bit rate, we have $\phi_i(\mathbf{x}) = x_i d_i$ for $i \in \mathcal{S}$. In order to limit the bandwidth allocated to streaming traffic to at most $C_s < C$, an admission control mechanism is used. Thus, an arriving streaming flow of class k is admitted in state $\mathbf{x}$ if and only if $\sum_{i\in\mathcal{S}} \phi_i(\mathbf{x}) \leq C_s - d_k$. We assume that

elastic flows share the remaining capacity unused by streaming flows according to balanced fairness (BF)[2]. Flows of the same class equally share the allocated bandwidth. In the following, we refer to $\rho = (\rho_i)_{i\in\mathcal{E}\cup\mathcal{S}}$ as the traffic intensity vector where $\rho_i = \lambda_i/\mu_i$ corresponds to the traffic intensity of class-i flows and denote by $\theta = \sum_{i\in\mathcal{E}} a_{i,l}\rho_i$ the total offered elastic traffic. A necessary and sufficient condition for stability is therefore

$$\theta < C - C_s. \tag{22}$$

For the remainder of the paper we will assume that the stability condition is satisfied.

3.2 Performance metrics for streaming traffic

As in the previous section, we evaluate the performance of streaming flows through their blocking probability. We let $p_s(n) = \sum_{\mathbf{x}_s d=n} \pi_s(\mathbf{x}_s)$ where $\pi_s(\mathbf{x}_s)$ denotes the marginal distribution of $\mathbf{x}_s$. We assume that C and d_i are integers for all $i \in \mathcal{S}$ and evaluate $p_s(n)$ for $n \leq C_s$ in a similar way as in the Kaufman-Roberts recursive formula [10] from which we deduce lemma 1,

Lemma 1. *The blocking probability of class $i \in \mathcal{S}$ is given by,*

$$B_i = \sum_{n>C-c_i} p_s(n), \tag{23}$$

where

$$p_s(n) = p_s(0)\sum_{i\in\mathcal{S}} \frac{\rho_i}{n} p_s(n-d_i), \tag{24}$$

with $p_s(0) = \left(\sum_{k=0}^{C_s}\sum_{i\in\mathcal{S}} \frac{\rho_i}{k} p_s(k-d_i)\right)^{-1}$, $p_s(0) = 1$ *and* $p_s(n) = 0$ *for* $n < 0$.

3.3 Approximate performance metrics for elastic traffic

As in section 2, we evaluate the performance of class-i elastic flows through their throughput γ_i defined as the ratio $\rho_i/E[x_i]$ where $E[x_i]$ denotes the expected number of elastic flows in progress. We first provide insensitive performance bounds for elastic flows. We then propose simple performance approximations based on the QS assumption.

Insensitive bounds We observe that this case can be modelled as a network of processor sharing nodes where customers in nodes i correspond to class-i flows. The corresponding service rate at each node is given by

$$\phi_i(\mathbf{x}) = \begin{cases} x_i d_i & if\, i \in \mathcal{S} \\ x_i min\left(c, \frac{C-\sum_{k\in\mathcal{S}}\phi_k(\mathbf{x})}{|\mathbf{x}^e|}\right) & otherwise \end{cases}$$

One can easily verify that the balance property is violated by the service rates. Indeed, we have

$$\frac{\phi_j(\mathbf{x}-e_i)}{\phi_j(\mathbf{x})} = \frac{\phi_i(\mathbf{x}-e_j)}{\phi_i(\mathbf{x})}, \tag{25}$$

$$\frac{\phi_{j'}(\mathbf{x}-e_{i'})}{\phi_{j'}(\mathbf{x})} = \frac{\phi_{i'}(\mathbf{x}-e_{j'})}{\phi_{i'}(\mathbf{x})}, \tag{26}$$

and,

$$\frac{\phi_i(\mathbf{x}-e_{i'})}{\phi_i(\mathbf{x})} > 1 = \frac{\phi_{i'}(\mathbf{x}-e_i)}{\phi_{i'}(\mathbf{x})}, \tag{27}$$

for all $i, j \in \mathcal{E}$, $i^{'}, j^{'} \in \mathcal{S}$ and for all state $\mathbf{x}$ such that $x_i > 0$, $x_j > 0$, $x_{i'} > 0$ and $x_{j'} > 0$.

From (27), we deduce that the stationary of the network state is sensitive to traffic characteristics. However, following the same approach described in section 2, we can obtain insensitive performance bounds for elastic flows.

From (25), (26) and (27), we deduce that the monotonicity property and the bias property holds. The corresponding balance functions may then be written as

$$\Phi^{-}(\mathbf{x}) = \left(\prod_{i \in \mathcal{E}} \prod_{j=1}^{x_i} \phi_i(\sum_{k=1}^{i-1} x_k e_k + j e_i) \times \prod_{i \in \mathcal{S}} \prod_{j=1}^{x_i} \phi_i(\sum_{m \in \mathcal{E}} e_m x_m + \sum_{k=1}^{i-1} x_k e_k + j e_i) \right)^{-1}. \tag{28}$$

and,

$$\Phi^{+}(\mathbf{x}) = \left(\prod_{i \in \mathcal{S}} \prod_{j=1}^{x_i} \phi_i(\sum_{k=1}^{i-1} x_k e_k + j e_i) \times \times \prod_{i \in \mathcal{E}} \prod_{j=1}^{x_i} \phi_i(\sum_{m \in \mathcal{S}} e_m x_m + \sum_{k=1}^{i-1} x_k e_k + j e_i) \right)^{-1}. \tag{29}$$

(28)-(29) immediately yields the following lower and upper performance bounds for class-i throughput

$$\gamma_i^{-} = \rho_i / \sum_{\mathbf{x}} x_i \pi^{-}(0) \Phi^{-}(\mathbf{x}) \rho^{\mathbf{x}}, \tag{30}$$

where $\pi^{-}(0) = \left(\sum_{\mathbf{x}} \Phi^{-}(\mathbf{x}) \rho^{\mathbf{x}}\right)^{-1}$.

$$\gamma_i^{+} = \rho_i / \sum_{\mathbf{x}} x_i \pi^{+}(0) \Phi^{+}(\mathbf{x}) \rho^{\mathbf{x}}, \tag{31}$$

where $\pi^{+}(0) = \left(\sum_{\mathbf{x}} \Phi^{+}(\mathbf{x}) \rho^{\mathbf{x}}\right)^{-1}$.

Note that for the multi-class case the insensitive bounds are quite loose. We propose in the next subsection simple performance approximations for elastic flows based on the QS assumption.

Quasi-stationary analysis We can easily extend the QS approach to the multi-class case as follows: Assume that $\mathbf{x}^s$ is kept fixed. By letting $n = \sum_{i \in \mathcal{S}} \phi_i(\mathbf{x}^s)$, elastic flows share the remaining bandwidth $C - n$ according to BF, which is equivalent to the ordinary *Processor Sharing* (PS) discipline as long as $\sum_{i \in \mathcal{E}} x_i^e \leq N = (C - n)/c$. The resulting system can be analysed using the theory of *Generalized Processor Sharing* (GPS) queues [8]. Provided that the total offered elastic traffic $\theta = \sum_{k \in \mathcal{E}} \rho_k < C - n$, it yields the following simple expression of the mean number of class-i elastic flows in progress conditioned on n (see [7])

$$E\left[x_i^e | n\right] = \frac{\rho_i}{c} \; + \; B(n) \, \frac{\rho_i}{C - n - \theta}, \tag{32}$$

where $B(n)$ represents the congestion probability of an equivalent link of capacity $C - n$ and is given by the well-known Erlang delay formula, i.e.,

$$B(n) = \frac{\frac{1}{N!} \left(\frac{\theta}{c}\right)^N \frac{C_n}{C - n - \theta}}{\sum_{i=0}^{N-1} \frac{1}{i!} \left(\frac{\theta}{c}\right)^i + \frac{1}{N!} \left(\frac{\theta}{c}\right)^N \frac{C - n}{C - n - \theta}}. \tag{33}$$

The above QS approximation immediately yields the following approximation for class-i throughput:

$$\gamma_i = \rho_i / \sum_{n \leq C_s} E[x_i^e|n]\, p_S(n). \tag{34}$$

Although (32) was derived in the case of a common rate limit for all elastic classes, a similar expression (although slightly more complex) can be obtained in section 4 for the multi-rate case.

Example 2. To illustrate the result obtained in this section, we consider as a second example that a link of capacity 30 Mbps is shared by 4 traffic classes, the first two classes corresponding to elastic traffic while the other ones correspond to streaming traffic. It is assumed that $c_1 = c_2 = 4Mbps$, $d_3 = 3$ and $d_4 = 4$ Mbps. We further assume that the total offered traffic is composed of 90 % of elastic traffic.

Figure 3 compare the proposed approximations with the results obtained with discrete-event simulations for both class-1 and class-2 elastic flows. The relative error of the QS approximation is below 5% in all traffic regimes, whereas the accuracy of the upper and lower insensitive bound decreases as the total link utilization increases.

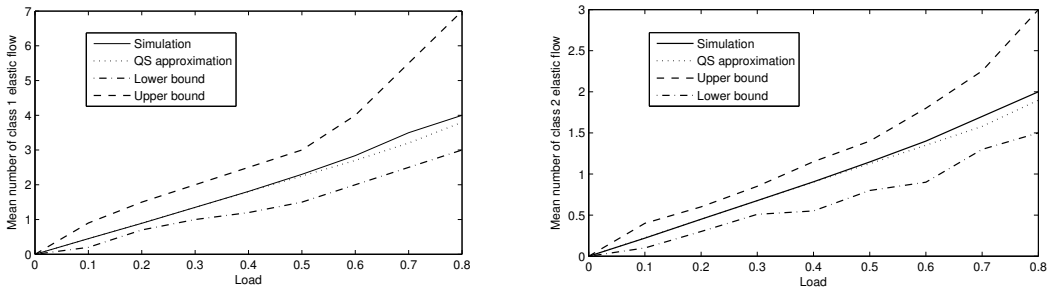


Fig. 3: Mean number of elastic flows in progress

4 Multi-rate extension

In this section, we propose an extension of the QS approach to the multi-rate case. We assume that each elastic flow of class i has its own rate limit c_i and denote by $c = (c_i)_{i \in \mathcal{E}}$ the vector of per-flow rate limit.

4.1 Absence of streaming traffic

Observe that in absence of streaming traffic, as shown in [2], one can compute the probability of each state $\mathbf{x}$ from

$$\pi(\mathbf{x}) = \pi(\mathbf{0})\, \Phi(\mathbf{x}) \rho^{\mathbf{x}}, \tag{35}$$

Where Φ refers to the so-called Balance function, recursively defined by

$$\Phi(x) = \max \left\{ \frac{1}{C} \sum_{i \in \mathcal{E}} \Phi(\mathbf{x} - e_i), \max_{i \in \mathcal{E}} \frac{\Phi(\mathbf{x} - e_i)}{c_i\, x_i} \right\}, \tag{36}$$

with $\Phi(\mathbf{0}) = 1$, and $\Phi(\mathbf{x}) = 0$ for any state $\mathbf{x}$ such that $x_i < 0$ for some i.

In theory, all performance metrics of interest can be derived from (36)-(35). In practice however, the approach based on the computation of state probabilities suffers from the curse of dimensionality. If truncation of the state space $\mathbb{N}^M$ is feasible in light traffic regimes, the direct computation of the above performance metrics cannot be done when either the number of flow classes gets large, or when traffic intensities are not enough small. In order to evaluate the number of expected class-i elastic flows in progress in an explicit and simple way, we propose the following proposition.

Proposition 1. *The mean number of class-i elastic flows in progress is given by*

$$E[x_e^i] = \frac{\rho_i}{c_i} \sum_{\mathbf{x}c \leq C} \pi(\mathbf{x}) + \frac{\rho_i}{C-\theta}(1 - \sum_{\mathbf{x}c \leq C-c_i} \pi(\mathbf{x})) + \frac{1}{C-\theta} \sum_k \rho_k \sum_{C-c_k < \mathbf{x}c < C} x_i \pi(\mathbf{x}), \tag{37}$$

where $\pi(\mathbf{x}) = \pi(0)\frac{\rho^{\mathbf{x}}}{\mathbf{x}!c^{\mathbf{x}}}$ *for all* $\mathbf{x}$ *such that* $|c\mathbf{x}| \leq C$, *and*

$$\pi(0) = \left(\sum_{|\mathbf{x}c| \leq C} \frac{\rho^{\mathbf{x}}}{\mathbf{x}!c^{\mathbf{x}}} + \frac{1}{C-\theta} \sum_{k \in \mathcal{E}} \rho_k \sum_{C-c_k \leq |\mathbf{x}c| \leq C} \frac{\rho^{\mathbf{x}}}{\mathbf{x}!c^{\mathbf{x}}} \right)^{-1}.$$

Proof. see Appendix.

4.2 Presence of streaming traffic

The QS approach can be easily extended to the multi-rate case. In presence of streaming traffic, we can obtain the mean number $E[\mathbf{x}_i^e|n]$ of class i elastic flows in progress given the total streaming flow throughput n=$\sum_{i \in \mathcal{S}} \phi_s(\mathbf{x}_s)$ by replacing C with $= C - n$ in equation 37. We obtain the following approximation for class-i throughput:

$$\gamma_i = \rho_i / \sum_{n \leq C_s} E[x_i^e|n]\, p_S(n). \tag{38}$$

Example 3. As a final example, we consider again Example 2 but suppose now that $c_1 = 2Mbps$ and $c_2 = 3Mbps$. Figure 4 shows the evolution of the mean number of class 1 elastic flows obtained using the QS approach and discrete-event simulation results as a function of the traffic load. We observe that the result obtained under the QS assumption are very close to the simulation results. The relative error is below 3% for all traffic regime.

5 Conclusion

We have investigated the performance of streaming and elastic flows when call admission control is used. Exact results for streaming flows and performance approximations for elastic flows were presented. Future work include the extension of our results to an arbitrary network topology.

References

1. T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Syst. Theory Appl.*, 53(1-2):65–84, June 2006.
2. T. Bonald, A. Penttinen, and J. T. Virtamo. On light and heavy traffic approximations of balanced fairness. In *Proceedings of ACM SIGMETRICS/Performance*, pages 109–120, 2006.
3. T. Bonald and A. Proutière. Insensitive bandwidth sharing in data networks. *Queueing Syst.*, 44(1):69–100, 2003.

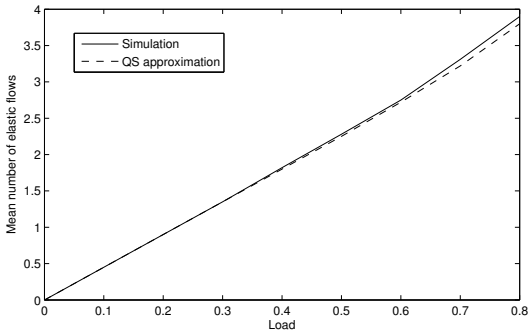


Fig. 4: Mean number of class 1 elastic flows in progress

4. T. Bonald and A. Proutière. On stochastic bounds for monotonic processor sharing networks. *Queueing Syst.*, 47(1-2):81–106, 2004.
5. T. Bonald and J. Virtamo. A recursive formula for multirate systems with elastic traffic. *IEEE Communications Letters*, 9(8):753–755, Aug 2005.
6. T. Bonald and J. T. Virtamo. Calculating the flow level performance of balanced fairness in tree networks. *Perform. Eval.*, 58(1):1–14, 2004.
7. O. Brun, A. Al Sheikh, and J. M. Garcia. Flow-level modelling of tcp traffic using gps queueing networks. In *International Teletraffic Congress*, pages 1–8. IEEE, 2009.
8. J. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12(3):245–284, 1979.
9. F. Delcoigne, A. Proutière, and G. Régnié. Modeling integration of streaming and data traffic. *CPerformance evaluation*, 2004.
10. J. Roberts, U. Mocci, and J. virtamo. Broadband network teletraffic. *Springer*, 1996.
11. J. W. Roberts and L. Massoulié. Bandwidth sharing and admission control for elastic traffic. In *Telecommunication Systems*, pages 185–201, 1998.

Appendix

Proof of proposition 1

In this case, the balance function can be simplified to

$$\Phi(\mathbf{x}) = \begin{cases} \prod_k \frac{1}{x_k! c_k^{x_k}} & \text{if} \quad |\mathbf{x}c| \leq C \\ \frac{1}{C} \sum_k \Phi(x - e_k) & otherwise \end{cases}$$

It follows that $\forall i \in \mathcal{E}$ $\phi_i(\mathbf{x}) = x_i c_i$ in the absence of congestion: $|\mathbf{x}c| \leq C$. We first note that

$$\begin{aligned} \pi(\mathbf{x} - e_i) &= \frac{1}{\rho_i} \pi(0) \phi(\mathbf{x} - e_i) \rho_i^{\mathbf{x}}, \\ &= \frac{\phi_i(\mathbf{x})}{\rho_i} \pi(0) \phi(\mathbf{x}) \rho_i^{\mathbf{x}}, \\ &= \frac{\phi_i(\mathbf{x})}{\rho_i} \pi(\mathbf{x}) \end{aligned} \tag{39}$$

it yields

$$\begin{aligned}\sum_{|\mathbf{x}c|\leq C} x_i\pi(\mathbf{x}) &= \sum_{|\mathbf{x}c|\leq C} x_i \frac{\rho_i \pi(\mathbf{x}-e_i)}{x_i c_i} \\ &= \frac{\rho_i}{c_i}\sum_{|\mathbf{x}c|\leq C} \pi(\mathbf{x}-e_i) \\ &= \frac{\rho_i}{c_i}\sum_{|\mathbf{x}c|\leq C-c_i} \pi(\mathbf{x})\end{aligned} \tag{40}$$

We have,

$$\begin{aligned}\sum_{[\mathbf{x}c|>C} x_i\pi(\mathbf{x}) &= \frac{1}{C}\sum_{|\mathbf{x}c|>C}\sum_k x_i\rho_k\pi(\mathbf{x}-e_k) \\ &= \frac{\rho_i}{C}\sum_{|\mathbf{x}c|>C-c_i}\pi(\mathbf{x}) + \frac{1}{C}\sum_k \rho_k \sum_{|\mathbf{x}c|>C-c_k} x_i\pi(\mathbf{x}) \\ &= \frac{\rho_i}{C}\sum_{|\mathbf{x}c|>C-c_i}\pi(\mathbf{x}) + \frac{1}{C}(\sum_k \rho_k \sum_{|\mathbf{x}c|>C} x_i\pi(\mathbf{x}) + \sum_k \rho_k \sum_{C-c_k<|\mathbf{x}c|<C} x_i\pi(\mathbf{x})) \\ &= \frac{\rho_i}{C-\theta}(1-\sum_{|\mathbf{x}c|\leq C-c_i}\pi(\mathbf{x})) + \frac{1}{C-\theta}\sum_k \rho_k \sum_{C-c_k<|\mathbf{x}c|<C} x_i\pi(\mathbf{x})\end{aligned} \tag{41}$$

from which the result follows since $E[x_e^i]$ can be written as

$$\begin{aligned}E[x_e^i] &= \sum_{\mathbf{x}} x_i\pi(\mathbf{x}) \\ &= \sum_{|\mathbf{x}c|\leq C} x_i\pi(\mathbf{x}) + \sum_{|\mathbf{x}c|>C} x_i\pi(\mathbf{x})\end{aligned} \tag{42}$$

where $\pi(\mathbf{x}) = \pi(0)\Phi(\mathbf{x})\rho^{\mathbf{x}}$ and,

$$\begin{aligned}\pi(0)^{-1} &= \sum_{\mathbf{x}} \Phi(\mathbf{x})\rho^{\mathbf{x}} \\ &= \sum_{|\mathbf{x}c|\leq C} \Phi(\mathbf{x})\rho^{\mathbf{x}} + \sum_{|\mathbf{x}c|>C} \Phi(\mathbf{x})\rho^{\mathbf{x}}\end{aligned} \tag{43}$$

Since

$$\begin{aligned}\sum_{|\mathbf{x}c|>C} \Phi(\mathbf{x})\rho^{\mathbf{x}} &= \frac{1}{C}\sum_{|\mathbf{x}c|>C}\sum_{k\in\mathcal{E}} \rho_k\Phi(\mathbf{x}-e_k)\rho^{\mathbf{x}-e_k}, \\ &= \frac{1}{C}\sum_{k\in\mathcal{E}}\rho_k \sum_{|\mathbf{x}c|>C-c_k} \Phi(\mathbf{x})\rho^{\mathbf{x}}, \\ &= \frac{\theta}{C}\sum_{|\mathbf{x}c|>C} \Phi(\mathbf{x})\rho^{\mathbf{x}} + \frac{1}{C}\sum_{k\in\mathcal{E}}\rho_k \sum_{C-c_k\leq|\mathbf{x}c|\leq C} \Phi(\mathbf{x})\rho^{\mathbf{x}}.\end{aligned}$$

It yields

$$\sum_{|\mathbf{x}c|>C} \Phi(\mathbf{x})\rho^{\mathbf{x}} = \frac{1}{C-\theta} \sum_{k\in\mathcal{E}} \rho_k \sum_{C-c_k\leq|\mathbf{x}c|\leq C} \Phi(\mathbf{x})\rho^{\mathbf{x}}, \tag{44}$$

We deduce,

$$\pi(0)^{-1} = \sum_{|\mathbf{x}c|\leq C} \Phi(\mathbf{x})\rho^{\mathbf{x}} + \frac{1}{C-\theta} \sum_{k\in\mathcal{E}} \rho_k \sum_{C-c_k\leq|\mathbf{x}c|\leq C} \Phi(\mathbf{x})\rho^{\mathbf{x}} \tag{45}$$