On Performance Bounds for the Integration of Elastic and Adaptive Streaming Flows

Thomas Bonald, Alexandre Proutière

On Performance Bounds for the Integration of Elastic and Adaptive Streaming Flows

Thomas Bonald and Alexandre Proutière
France Telecom R&D
38-40 rue du Général Leclerc, 92794 Issy-les-Moulineaux, France
{thomas.bonald,alexandre.proutiere}@francetelecom.com

ABSTRACT

We consider a network model where bandwidth is fairly shared by a dynamic number of elastic and adaptive streaming flows. Elastic flows correspond to data transfers while adaptive streaming flows correspond to audio/video applications with variable rate codecs. In particular, the former are characterized by a fixed size (in bits) while the latter are characterized by a fixed duration. This flow-level model turns out to be intractable in general. In this paper, we give performance bounds for both elastic and streaming traffic by means of sample-path arguments. These bounds present the practical interest of being insensitive to traffic characteristics like the distributions of elastic flow size and streaming flow duration.

Categories and Subject Descriptors

C.4 [**Computer System Organization**]: Performance of Systems

General Terms

Performance

Keywords

Multi-service network, elastic traffic, adaptive streaming traffic, flow-level analysis, insensitive bounds

1. INTRODUCTION

Despite considerable research and standardization efforts to define network architectures offering quality of service (QoS) guarantees [14, 15], IP networks remain largely operated on a "best effort" basis. This may notably be explained by the difficulty of defining appropriate service classes, marking the corresponding packets, controlling the end-to-end QoS over a multi-provider path, policing traffic and pricing services accordingly.

SIGMETRICS/Performance'04, June 12–16, 2004, New York, NY, USA.

Given the large variety of applications, it is useful to distinguish between two broad classes of flow:

Elastic flows correspond to the transfer of digital documents (e.g., Web pages, emails, stored audio/videos). They are characterized by a fixed size (the volume of the transferred document) and a variable duration, which determines user perceived performance. The duration of an elastic flow depends on its rate which varies with respect to the level of congestion in the network, typically under the control of TCP.

Streaming flows correspond to the real-time transfer of various signals (e.g., voice, streaming audio/video). They are characterized by a fixed duration (inherent to the original signal) and a possibly variable size, which corresponds to the amount of data received within the admissible delay and therefore determines user perceived performance. The actually transferred size of a streaming flow depends on packet delays and losses as well as on the sending rate which may vary with respect to the level of congestion in the network when variable rate codecs are used.

A classical approach to offering packet delay and loss guarantees to streaming flows is to give them some form of priority over elastic flows. However, even this simple form of service differentiation seems difficult to introduce in practice. It is necessary to mark each packet as belonging to an elastic or a streaming flow and this marking requires tight control. Another major problem concerns the performance of elastic flows that may suffer from long starvation periods.

An alternative approach consists in relying on the self-adaptation of streaming flows. Streaming flows are then expected to be "TCP-friendly", i.e., to mimic elastic flows so as to get the same rate *as if* they were themselves elastic [12]. Assuming fair sharing of network resources between elastic and streaming flows, both types of flow are then similarly affected by congestion periods: while elastic flows last longer, the quality of streaming applications suffers from reduced data rates [1, 2, 11].

A great advantage of this approach is that there is no need for explicit differentiation between elastic and streaming flows. The service model remains based on the initial "best effort" paradigm of the Internet. To offer commercially viable services, it remains necessary to predict the performance resulting from an expected demand in elastic and adaptive streaming traffic. The question is: how much capacity is necessary for the network to be transparent to both elastic and streaming flows? This question is far from

obvious given the inherently random nature of traffic, i.e., the fact that the number of ongoing flows randomly varies as new flows are initiated and other cease.

In this paper, we partially answer this question by analyzing the stochastic flow-level dynamics of a network integrating elastic and adaptive streaming flows. Specifically, we provide performance bounds for both types of flow using an idealized model where the flow rate adaptation is perfect and instantaneous. These bounds have the great practical interest of being insensitive to traffic characteristics like the distributions of elastic flow size and streaming flow duration.

Flow-level models have only recently been introduced for evaluating Internet performance. Massoulié and Roberts used the processor sharing queue to represent a single link whose capacity is fairly shared by a dynamic number of elastic flows [19]. Various extensions of this model have been proposed to account for the way TCP actually shares bandwidth [3, 18], the presence of rate limits and multiple bottlenecks [4, 6, 16] and the impact of user behaviour [8, 13].

Most studies on the integration of elastic and streaming flows concern systems where priority is given to streaming traffic. Núñez Queija [20] and Delcoigne et. al. [10] have shown the extreme sensitivity of this system to traffic characteristics and the severe degradation of elastic traffic performance in the absence of any mechanism like admission control that prevents streaming traffic from grabbing the whole link capacity. To our knowledge, the issue of the integration of elastic and adaptive streaming flows in realistic dynamical scenarios with random flow arrivals and departures has only recently been addressed by Key et. al. [17]. They notably proved for exponentially distributed elastic flow sizes and streaming flow durations that streaming traffic does not affect the stability condition, i.e., the number of ongoing flows remains finite provided the *elastic* traffic load is less than 1. In this paper we focus rather on performance bounds, valid for any traffic characteristics. The stability of the system follows directly from that of the lower bound in all considered scenarios.

In the next section, we present the basic model of a single link shared fairly by elastic and adaptive streaming flows in the absence of rate limit. The impact of a common rate limit is evaluated in the following section. The extension of these results to several rate limits and network scenarios is presented in Section 4. Section 5 concludes the paper.

2. A SINGLE BOTTLENECK LINK

The basic model consists of a single unit capacity link shared by elastic and adaptive streaming flows without any rate limit. Specifically, we assume that the link capacity is fairly shared between ongoing flows so that each flow has a rate:

$$\gamma(x) = \frac{1}{x_e + x_s},$$

where x_e and x_s are the number of ongoing elastic and streaming flows, respectively, and $x = (x_e, x_s)$ denotes the network state. We first present the traffic assumptions and performance metrics and then show that the analysis of this system reduces to that of a processor sharing network with state-dependent service rates.

2.1 Traffic assumptions

Elastic and streaming flows arrive as independent Poisson processes of respective intensities λ_e and λ_s. Elastic flows have i.i.d. sizes of mean $1/\mu_e$. Streaming flows have i.i.d. durations of mean $1/\mu_s$. As the link is of unit capacity, $1/\mu_s$ may be viewed as the mean potential size of a streaming flow, equal to its actual size in the absence of any other flow. Thus we define the elastic and streaming traffic intensities as:

$$\rho_e = \frac{\lambda_e}{\mu_e} \quad \text{and} \quad \rho_s = \frac{\lambda_s}{\mu_s}.$$

The overall traffic intensity is denoted by $\rho = \rho_e + \rho_s$.

2.2 Performance metrics

Users of data transfer applications experience quality of service through the time necessary to transfer a document. Thus we evaluate the performance of elastic traffic through the elastic flow throughput γ_e, defined as the ratio of the mean flow size $1/\mu_e$ to the mean flow duration τ_e:

$$\gamma_e = \frac{1}{\mu_e \tau_e}.$$

By Little's law, we have $\mathrm{E}[x_e] = \lambda_e \tau_e$ so that:

$$\gamma_e = \frac{\rho_e}{\mathrm{E}[x_e]}. \tag{1}$$

Users of audio and video applications, on the other hand, experience quality of service through their instantaneous rate. Thus we evaluate the performance of streaming traffic through the streaming flow throughput γ_s, defined as the mean instantaneous rate conditioned on the fact that there is at least one ongoing streaming flow:

$$\gamma_s = \mathrm{E}[\gamma(x) | x_s > 0]. \tag{2}$$

Note that we choose mean performance metrics for the sake of simplicity. The bounds derived in this paper also apply to finer metrics like the probability that the instantaneous rate is less than a given threshold, see the example in §2.5.

2.3 A processor sharing network

The system can be represented as a network of two coupled processor sharing nodes. Customers in one node, node "e", correspond to the elastic flows, while customers in the other node, node "s", correspond to the streaming flows. The arrival rates at nodes e and s are λ_e and λ_s, respectively. Service requirements are i.i.d. of unit mean at each node. The service rate of node e is given by:

$$\phi_e(x) = \frac{x_e}{x_e + x_s} \mu_e. \tag{3}$$

This corresponds to the "departure rate" of elastic flows in state x. Since the duration of a streaming flow is independent of its bit rate, the service rate of node s is simply given by:

$$\phi_s(x) = x_s \mu_s. \tag{4}$$

This corresponds to the departure rate of streaming flows in state x.

Absence of streaming traffic. In the absence of streaming traffic, this queueing system reduces to a single processor sharing queue, the model originally considered by Massoulié and Roberts [19]. Under the stability condition $\rho_e < 1$, the stationary distribution of the number of ongoing elastic flows is then:

$$\pi(x_e) = (1 - \rho_e)\rho_e^{x_e},$$

corresponding to the elastic flow throughput:

$$\gamma_e = 1 - \rho_e. \tag{5}$$

This system is *insensitive* to the elastic flow size distribution.

Presence of streaming traffic. If this insensitivity property were preserved with streaming traffic, one could easily evaluate the stationary distribution of the network state. Unfortunately, the presence of streaming traffic breaks down insensitivity. As explained in Appendix A, a necessary condition for insensitivity is that:

$$\frac{\phi_e(x - f_s)}{\phi_e(x)} = \frac{\phi_s(x - f_e)}{\phi_s(x)}, \quad \forall x: \ x_e > 0, \ x_s > 0,$$

where $f_e \equiv (1,0)$ and $f_s \equiv (0,1)$. One can easily verify that this "balance" property is violated by the service rates (3), (4). We deduce that the stationary distribution of the network state x does depend on the distributions of elastic flow size and streaming flow duration. Thus we cannot expect explicit results without making specific assumptions about traffic characteristics. This is why we prefer to derive performance bounds, valid for any traffic characteristics.

2.4 Insensitive bounds

As explained in Appendix A, insensitive bounds can be derived for any processor sharing network provided the following monotonicity property holds: removing a customer from any node does not decrease the service rate of any other customer. The monotonicity property clearly holds for the service rates (3), (4). We deduce that the network state $x(t)$ at any time t satisfies:

$$\hat{x}(t) \leq x(t) \leq \check{x}(t), \tag{6}$$

where $x \leq y$ means $x_e \leq y_e$, $x_s \leq y_s$ and $\hat{x}(t)$, $\check{x}(t)$ are the states of two "virtual" insensitive processor sharing networks at time t. Denoting by $\hat{\Phi}$ and $\check{\Phi}$ the corresponding balance functions, the stationary distributions of the network states $\hat{x}(t)$, $\check{x}(t)$ are respectively given by:

$$\hat{\pi}(x) = \hat{\pi}(0)\hat{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s}, \quad \check{\pi}(x) = \check{\pi}(0)\check{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s}, \tag{7}$$

provided the following stability condition holds:

$$\sum_x \check{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s} < \infty. \tag{8}$$

We shall see that this condition is satisfied if and only if $\rho_e < 1$. In view of (6), we deduce that the original system is stable if and only if $\rho_e < 1$: streaming traffic does not affect the stability condition. This is due to the adaptive nature of the considered streaming flows.

We now derive the balance functions $\hat{\Phi}$, $\check{\Phi}$ and the corresponding performance metrics using the fact that the following "bias" property holds (refer to Appendix A):

$$\frac{\phi_e(x - f_s)}{\phi_e(x)} \geq 1 = \frac{\phi_s(x - f_e)}{\phi_s(x)}, \quad \forall x: \ x_e > 0, \ x_s > 0.$$

We refer to the network state $\hat{x}(t)$ as the "upper" bound since it leads to a smaller number of ongoing elastic and streaming flows, corresponding to higher values of the performance metrics. Similarly, we refer to $\check{x}(t)$ as the lower bound.

Upper bound. Using the bias property, we obtain:

$$\hat{\Phi}(x) = \frac{1}{x_s!} \frac{1}{\mu_e^{x_e}\mu_s^{x_s}}. \tag{9}$$

This corresponds to "virtual" service rates:

$$\hat{\phi}_e(x) = \mu_e \geq \phi_e(x) \quad \text{and} \quad \hat{\phi}_s(x) = \phi_s(x).$$

Thus the system behaves as if each type of traffic were in isolation in the sense that elastic traffic is not affected by streaming traffic and streaming traffic is not affected by elastic traffic.

If $\rho_e < 1$, we deduce from (1), (7), (9) that $\gamma_e \leq \hat{\gamma}_e$, where $\hat{\gamma}_e$ denotes the elastic flow throughput for the upper bound:

$$\hat{\gamma}_e = 1 - \rho_e,$$

and from (2), (7) and (9) that $\gamma_s \leq \hat{\gamma}_s$, where $\hat{\gamma}_s$ denotes the streaming flow throughput for the upper bound:

$$\hat{\gamma}_s = \frac{1 - \rho_e}{e^{\rho_s} - 1} \sum_{k \geq 0} \rho_e^k \sum_{l \geq 1} \frac{\rho_s^l}{(l+k) \times l!}.$$

Lower bound. Similarly, it follows from the bias property that:

$$\check{\Phi}(x) = \binom{x_e + x_s}{x_e} \frac{1}{x_s!} \frac{1}{\mu_e^{x_e}\mu_s^{x_s}}. \tag{10}$$

This corresponds to "virtual" service rates:

$$\check{\phi}_e(x) = \phi_e(x) \quad \text{and} \quad \check{\phi}_s(x) = \frac{x_s}{x_e + x_s} \times x_s\mu_s \leq \phi_s(x).$$

We deduce that the stability condition (8) holds if and only if $\rho_e < 1$, in which case

$$\sum_x \check{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s} = \frac{1}{1 - \rho_e} e^{\frac{\rho_s}{1 - \rho_e}}.$$

If $\rho_e < 1$, it follows from (1), (7), (10) that $\gamma_e \geq \check{\gamma}_e$, where $\check{\gamma}_e$ denotes the elastic flow throughput for the lower bound:

$$\check{\gamma}_e = \frac{(1 - \rho_e)^2}{1 - \rho_e + \rho_s}.$$

Similarly, we obtain $\gamma_s \geq \check{\gamma}_s$, where $\check{\gamma}_s$ denotes the streaming flow throughput for the lower bound:

$$\check{\gamma}_s = \frac{1 - \rho_e}{e^{\frac{\rho_s}{1 - \rho_e}} - 1} \sum_{k \geq 0} \rho_e^k \sum_{l \geq 1} \binom{k+l}{l} \frac{\rho_s^l}{(l+k) \times l!}.$$

All traffic elastic. Another simple lower bound can be obtained by considering that all traffic is elastic. This corresponds to a single processor sharing queue with two customer classes of respective mean service requirements $1/\mu_e$ and $1/\mu_s$. The "virtual" service rates are:

$$\tilde{\phi}_e(x) = \phi_e(x) \quad \text{and} \quad \tilde{\phi}_s(x) = \frac{x_s}{x_e + x_s}\mu_s \leq \phi_s(x).$$

In particular, the bound is less tight than the above lower bound. It is stable if and only if $\rho < 1$, in which case the stationary distribution is:

$$\tilde{\pi}(x) = \tilde{\pi}(0)\tilde{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s},$$

with

$$\tilde{\Phi}(x) = \binom{x_e + x_s}{x_e} \frac{1}{\mu_e^{x_e} \mu_s^{x_s}}.$$

We obtain $\gamma_e \geq \tilde{\gamma}_e$ and $\gamma_s \geq \tilde{\gamma}_s$, with:

$$\tilde{\gamma}_e = 1 - \rho, \quad \tilde{\gamma}_s = \frac{(1-\rho)(1-\rho_e)}{\rho_s} \ln\left(\frac{1-\rho_e}{1-\rho}\right).$$

2.5 Numerical example

Figure 1 compares the bounds of §2.4 to simulations of the model described in §2.1 with exponential elastic flow sizes ($\mu_e = 1$) and exponential streaming flow durations ($\mu_s = 1$), when streaming traffic represents a fraction $\rho_s/\rho = 0.2$ of the overall traffic. We observe that elastic traffic is only slightly penalized by the presence of adaptive streaming traffic. This is a significant difference with the scenario considered in [10, 20] where streaming traffic is given priority and strongly penalizes elastic traffic.

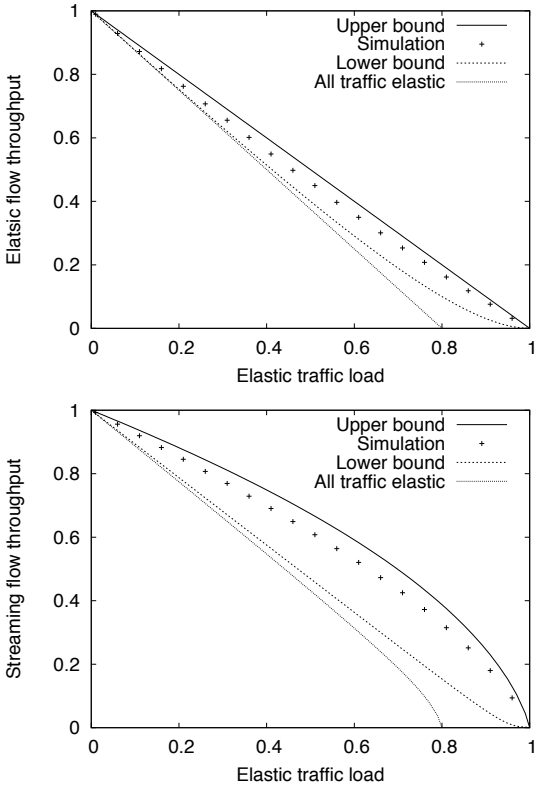

Figure 1: Performance of elastic traffic (upper graph) and streaming traffic (lower graph) for a unit capacity link with 20% of streaming traffic.

Figure 2 gives the results obtained for the same scenario but a different performance metric for streaming traffic, namely the probability that the instantaneous bit rate is less than 0.1. We observe that when overall traffic intensity ρ is less than 1, corresponding to an elastic traffic intensity $\rho_e < 0.8$, the instantaneous bit rate is rarely less than 0.1.

3. ACCOUNTING FOR RATE LIMITS

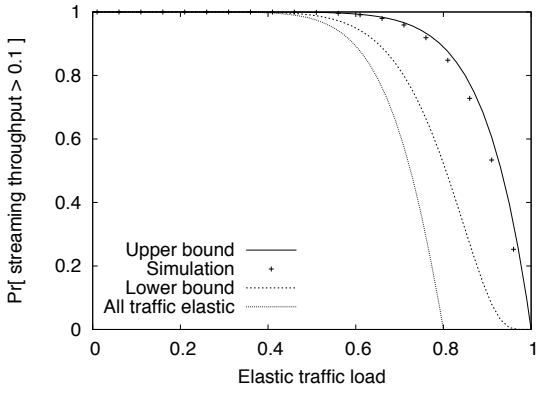

Figure 2: Performance of streaming traffic for a unit capacity link with 20% of streaming traffic.

We now extend the model of Section 2 to account for a common rate limit $a < 1$ for both elastic and streaming flows, due to the speed of the user's access line for instance. For streaming flows, the rate limit may also represent the maximum rate of the audio/video codec. Thus the rate of each flow is given by:

$$\gamma(x) = \min\left(a, \frac{1}{x_e + x_s}\right).$$

3.1 Traffic assumptions

The traffic assumptions are the same as in §2.1. In view of the flow rate limit, the mean potential size of a streaming flow is now equal to a/μ_s. Thus we define the elastic and streaming traffic intensities as:

$$\rho_e = \frac{\lambda_e}{\mu_e} \quad \text{and} \quad \rho_s = \frac{\lambda_s}{\mu_s} a.$$

The overall traffic intensity is still denoted by $\rho = \rho_e + \rho_s$.

3.2 A processor sharing network

As in §2.3, the system can be represented as a network of two coupled processor sharing nodes. The corresponding service rates are:

$$\phi_e(x) = x_e \mu_e \times \min\left(a, \frac{1}{x_e + x_s}\right). \tag{11}$$

and

$$\phi_s(x) = x_s \mu_s. \tag{12}$$

Again, the balance property is violated so that network performance is sensitive to the distributions of elastic flow size and streaming flow duration.

3.3 Insensitive bounds

As the monotonicity property and the bias property still hold for the service rates (11), (12), we can derive insensitive bounds characterized by balance functions $\hat{\Phi}$ and $\check{\Phi}$ as above. Let n be the largest integer such that $n \times a \leq 1$.

Upper bound. We get:

$$\hat{\Phi}(x) = \frac{1}{x_s!} \frac{1}{\mu_e^{x_e} \mu_s^{x_s}} \times \begin{cases} \frac{(1/a)^{x_e}}{x_e!} & \text{if } x_e \leq n, \\ \frac{(1/a)^n}{n!} & \text{otherwise.} \end{cases}$$

This corresponds to "virtual" service rates:

$$\hat{\phi}_e(x) = x_e\mu_e \times \min(a, \frac{1}{x_e}) \geq \phi_e(x) \quad \text{and} \quad \hat{\phi}_s(x) = \phi_s(x).$$

Again, the upper bound corresponds to a system where each type of traffic is in isolation.

Lower bound. Similarly, it follows from the bias property that:

$$\check{\Phi}(x) = \binom{x_e + x_s}{x_e} \frac{1}{\mu_e^{x_e}\mu_s^{x_s}} \times \begin{cases} \frac{(1/a)^{x_e}}{(x_e+x_s)!} & \text{if } x_e + x_s \leq n, \\ \frac{1}{x_s!} & \text{if } x_s > n, \\ \frac{(1/a)^{n-x_s}}{n!} & \text{otherwise.} \end{cases}$$

This corresponds to "virtual" service rates $\check{\phi}_e(x) = \phi_e(x)$ and $\check{\phi}_s(x) \leq \phi_s(x)$, with:

$$\check{\phi}_s(x) = x_s\mu_s \times \begin{cases} 1 & \text{if } x_e + x_s \leq n, \\ \frac{x_s}{x_e+x_s} & \text{if } x_s > n, \\ \frac{1}{a(x_e+x_s)} & \text{otherwise.} \end{cases}$$

As in the absence of rate limits, the stability condition (8) holds if and only if $\rho_e < 1$.

All traffic elastic. Another lower bound is obtained by considering that all traffic is elastic. As the mean potential size of streaming flows is equal to a/μ_s, we deduce that the "virtual" service rates are

$$\tilde{\phi}_e(x) = \phi_e(x)$$

and

$$\tilde{\phi}_s(x) = x_s\frac{\mu_s}{a} \times \min\left(a, \frac{1}{x_e + x_s}\right) \leq \phi_s(x).$$

It is stable if and only if $\rho < 1$, in which case the stationary distribution is:

$$\tilde{\pi}(x) = \tilde{\pi}(0)\tilde{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s},$$

with

$$\tilde{\Phi}(x) = \binom{x_e + x_s}{x_e} \frac{1}{\mu_e^{x_e}(\mu_s/a)^{x_s}} \times \begin{cases} \frac{(1/a)^{x_e+x_s}}{(x_e+x_s)!} & \text{if } x_e + x_s \leq n, \\ \frac{(1/a)^n}{n!} & \text{otherwise.} \end{cases}$$

3.4 Numerical example

Figure 3 illustrates the tightness of the bounds in case of a single link with a common rate limit equal to 0.1 for all flows. Again the simulation results are obtained for exponential elastic flow sizes and streaming flow durations with the same mean ($\mu_e = \mu_s$). Observe that the bounds are relatively tight when the overall traffic intensity ρ is less than 1, corresponding to an elastic traffic intensity $\rho_e < 0.8$. We deduce that performance is approximately insensitive for this range of traffic loads.

4. MULTICLASS EXTENSION

Finally, we extend the results to account for the fact that flows may have different rate limits or may be limited by the capacity of several links. We consider a network of L links of respective capacities $C_1, \ldots, C_L$ shared by N classes of flow. Class-i flows are characterized by a rate limit a_i and a fixed route r_i consisting of a subset of the links $\{1, \ldots, L\}$. We assume that network resources are shared according to balanced fairness [6]. This allocation has been shown to provide insensitive performance and to constitute a good approximation of usual "fair" allocations like max-min fairness and proportional fairness [22]. We first recall the definition and the main properties of balanced fairness and then derive performance bounds from the analysis of a processor sharing network.

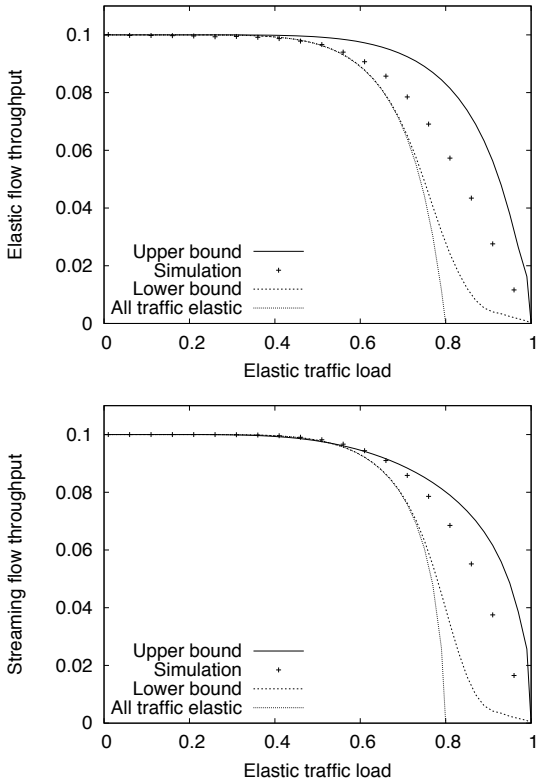

Figure 3: Performance of elastic traffic (upper graph) and streaming traffic (lower graph) for a unit capacity link with a flow rate limit $a = 0.1$ and 20% of streaming traffic.

Notation. We denote by e_i the unit vector with 1 in component i and 0 elsewhere, for $i = 1, \ldots, N$. For any $y \leq z \in \mathbb{N}^N$ and any $a \in \mathbb{R}_+^N$, we use the vectorial notation:

$$\binom{z}{y} \equiv \prod_{i=1}^{N} \binom{z_i}{y_i}, \quad y! \equiv \prod_{i=1}^{N} y_i! \quad \text{and} \quad a^y \equiv \prod_{i=1}^{N} a_i^{y_i}.$$

4.1 Balanced fairness

Let $y = (y_1, \ldots, y_N)$ be the network state, where y_i is the number of class-i flows. We denote by $\psi_i(y)$ the overall rate of class-i flows in state y. The rate constraints are:

$$\forall i = 1, \ldots, N, \quad \psi_i(y) \leq y_i a_i, \tag{13}$$

and

$$\forall l = 1, \ldots, L, \quad \sum_{i:l\in r_i} \psi_i(y) \leq C_l. \tag{14}$$

Balanced fairness refers to the allocation:

$$\psi_i(y) = \frac{\Psi(y - e_i)}{\Psi(y)}, \tag{15}$$

where Ψ is the function recursively defined by $\Psi(0) = 1$ and:

$$\Psi(y) = \max\left\{ \max_{i:y_i>0} \frac{\Psi(y-e_i)}{y_i a_i}, \max_{l=1,\ldots,L} \frac{1}{C_l} \sum_{i:x_i>0, l\in r_i} \Psi(y-e_i) \right\}.$$

This is the unique insensitive allocation such that at least one rate constraint is attained in any state $y \neq 0$. Such an allocation has been considered in the context of elastic traffic only [6].

In the following, class-i flows may be either elastic or streaming. We denote by $x_{e,i}$ the number of class-i elastic flows, $x_{s,i}$ the number of class-i streaming flows. The network state is now $x = (x_e, x_s)$, with $x_e = (x_{e,1}, \ldots, x_{e,N})$ and $x_s = (x_{s,1}, \ldots, x_{s,N})$. The rate of an elastic class-i flow is the same as that of a streaming class-i flow and is given by:

$$\gamma_i(x) = \frac{\psi_i(x_e + x_s)}{x_{e,i} + x_{s,i}}. \tag{16}$$

A particular example is a single unit capacity link with a common rate limit a as considered in Section 3. This implies the state-dependent rate:

$$\psi(y) = \min(y \times a, 1).$$

The corresponding balance function Ψ is:

$$\Psi(y) = \frac{1}{y!\, a^y} \text{ if } y \leq n, \ \Psi(y) = \frac{1}{n!\, a^n} \text{ otherwise,}$$

where n is the largest integer such that $n \times a \leq 1$.

4.2 Traffic assumptions

Class-i elastic and streaming flows arrive as independent Poisson processes of respective intensities $\lambda_{e,i}$ and $\lambda_{s,i}$. Class-i elastic flows have i.i.d. sizes of mean $1/\mu_{e,i}$. Class-i streaming flows have i.i.d. durations of mean $1/\mu_{s,i}$. The mean potential size of a class-i streaming flow is equal to $a_i/\mu_{s,i}$. Thus we define the class-i elastic and streaming traffic intensities as:

$$\rho_{e,i} = \frac{\lambda_{e,i}}{\mu_{e,i}} \quad \text{and} \quad \rho_{s,i} = \frac{\lambda_{s,i}}{\mu_{s,i}}\, a_i.$$

The overall class-i traffic intensity is $\rho_i = \rho_{e,i} + \rho_{s,i}$.

4.3 A processor sharing network

The system can be viewed as a network of $2N$ coupled processor sharing nodes. Customers in nodes (e,i) correspond to class-i elastic flows, while customers in nodes (s,i) correspond to class-i streaming flows. The respective service rates are:

$$\phi_{e,i}(x) = x_{e,i}\mu_{e,i} \times \gamma_i(x), \ \phi_{s,i}(x) = x_{s,i}\mu_{s,i}. \tag{17}$$

The balance property holds per type of traffic in the sense that, in view of (15), (16) and (17),

$$\frac{\phi_{e,i}(x - f_{e,j})}{\phi_{e,i}(x)} = \frac{\phi_{e,j}(x - f_{e,i})}{\phi_{e,j}(x)}, \quad \forall x : x_{e,i} > 0, \ x_{e,j} > 0,$$

and

$$\frac{\phi_{s,i}(x - f_{s,j})}{\phi_{s,i}(x)} = \frac{\phi_{s,j}(x - f_{s,i})}{\phi_{s,j}(x)}, \quad \forall x : x_{s,i} > 0, \ x_{s,j} > 0,$$

where $f_{e,i} \equiv (e_i, 0)$ and $f_{s,i} \equiv (0, e_i)$, but in general,

$$\frac{\phi_{e,i}(x - f_{s,j})}{\phi_{e,i}(x)} \neq 1 = \frac{\phi_{s,j}(x - f_{e,i})}{\phi_{s,j}(x)}.$$

We conclude that network performance is sensitive to the distributions of elastic flow size and streaming flow duration.

4.4 Insensitive bounds

In the following, we assume that the monotonicity property holds. We give examples in §4.5 where this property is indeed satisfied. Note that the bias property then follows, as:

$$\frac{\phi_{e,i}(x - f_{s,j})}{\phi_{e,i}(x)} \geq 1 = \frac{\phi_{s,j}(x - f_{e,i})}{\phi_{s,j}(x)}, \quad \forall x : x_{e,i} > 0, \ x_{s,j} > 0.$$

The balance functions $\hat{\Phi}$ and $\check{\Phi}$ that characterize the upper and lower bounds can then easily be derived. We have:

$$\hat{\pi}(x) = \hat{\pi}(0)\hat{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s}, \quad \check{\pi}(x) = \check{\pi}(0)\check{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s}, \tag{18}$$

provided the stability condition holds:

$$\sum_x \check{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s} < \infty. \tag{19}$$

Upper bound. Using the bias property, we get:

$$\hat{\Phi}(x) = \frac{\Psi(x_e)}{x_s!}\ \frac{1}{\mu_e^{x_e}\mu_s^{x_s}}.$$

This corresponds to "virtual" service rates:

$$\hat{\phi}_{e,i}(x) = \mu_{e,i}\psi_i(x_e) \quad \text{and} \quad \hat{\phi}_{s,i}(x) = \phi_{s,i}(x).$$

The upper bound corresponds to a system where each type of traffic is in isolation.

Lower bound. Similarly, it follows from the bias property that:

$$\check{\Phi}(x) = \binom{x_e + x_s}{x_e} \frac{\Psi(x_e + x_s)}{\Psi(x_s)} \frac{1}{x_s!}\ \frac{1}{\mu_e^{x_e}\mu_s^{x_s}}.$$

This corresponds to "virtual" service rates $\check{\phi}_{e,i}(x) = \phi_{e,i}(x)$ and:

$$\check{\phi}_{s,i}(x) = \frac{x_{s,i}}{x_{e,i} + x_{s,i}}\ \frac{\psi_i(x_e + x_s)}{\psi_i(x_s)}\ x_{s,i}\mu_{s,i}.$$

A key result is that the stability condition does not depend on streaming traffic. Let $\bar{\rho}_{e,l} = \sum_{i:l\in r_i} \rho_{e,i}$ be the overall elastic traffic intensity on link l. The proof of Theorem 1 is given in Appendix B.

THEOREM 1. *The stability condition (19) holds if and only if $\bar{\rho}_{e,l} < C_l$ for all links l.*

All traffic elastic. Using the monotonicity property, we obtain another lower bound by considering that all traffic is elastic. As the mean potential size of class-i streaming flows is equal to $a_i/\mu_{s,i}$, we deduce that the "virtual" service rates are $\tilde{\phi}_{e,i}(x) = \phi_{e,i}(x)$ and:

$$\tilde{\phi}_{s,i}(x) = x_{s,i}\frac{\mu_{s,i}}{a_i} \times \gamma_i(x) \leq \phi_{s,i}(x).$$

It is stable if and only if $\bar{\rho}_l \equiv \sum_{i:l\in r_i} \rho_i < C_l$ for all links l, in which case the stationary distribution is:

$$\tilde{\pi}(x) = \tilde{\pi}(0)\tilde{\Phi}(x)\lambda_e^{x_e}\lambda_s^{x_s},$$

with

$$\tilde{\Phi}(x) = \binom{x_e + x_s}{x_e} \Psi(x_e + x_s) \frac{1}{\mu_e^{x_e} (\mu_s/a)^{x_s}}.$$

4.5 Numerical examples

Tree network. We first consider a 2-branch tree as depicted Figure 4, in the absence of rate limits. The network consists of three links, a trunk of normalized capacity 1 and two branches of respective capacities C_1 and C_2, with $C_1, C_2 \leq 1$ and $C_1 + C_2 > 1$, and two routes, each route containing the trunk and one branch. The results apply more generally to any tree network, with an arbitrary number of multiplexing stages. Such topologies are practically interesting as they can represent access networks where bandwidth might be a scarce resource.

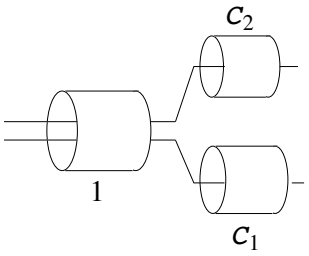

Figure 4: A 2-branch tree.

For this simple network topology, the function Ψ characterizing balanced fairness is explicit [6]: if y_1 and y_2 denote the number of flows on route 1 and 2, respectively, we get

$$\Psi(y) = \frac{1}{C_1^{y_1}}, \quad \text{if } y_2 = 0, \quad \Psi(y) = \frac{1}{C_2^{y_2}}, \quad \text{if } y_1 = 0,$$

and

$$\Psi(y) = \sum_{z=1}^{y_1} \binom{y_1 - z + y_2 - 1}{y_1 - z} \frac{1}{C_1^z} + \sum_{z=1}^{y_2} \binom{y_1 - 1 + y_2 - z}{y_2 - z} \frac{1}{C_2^z},$$

otherwise.

To apply the bounds derived in §4.4, we need the monotonicity property, i.e.,

$$\psi_i(y + e_j) \leq \psi_i(y), \quad \forall i \neq j, \ \forall y : y_i > 0,$$

and

$$\frac{\psi_i(y + e_i)}{y_i + 1} \leq \frac{\psi_i(y)}{y_i}, \quad \forall i, \ \forall y : y_i > 0.$$

In view of (15), this is equivalent to the inequalities:

$$\Psi(y+e_i)\Psi(y+e_j) \leq \Psi(y)\Psi(y+e_i+e_j), \quad \forall i \neq j, \ \forall y, \quad (20)$$

and

$$y_i \Psi(y)^2 \leq (y_i+1)\Psi(y-e_i)\Psi(y+e_i), \quad \forall i, \ \forall y : y_i > 0. \quad (21)$$

We prove the following result in Appendix C.

PROPOSITION 1. *The inequalities (20) and (21) hold.*

Figures 5 and 6 illustrate the tightness of the bounds for a 2-branch tree of capacities $C_1 = 1$ and $C_2 = 0.5$. The branches are equally loaded and, as in previous examples, streaming traffic represents 20% of the overall traffic intensity.

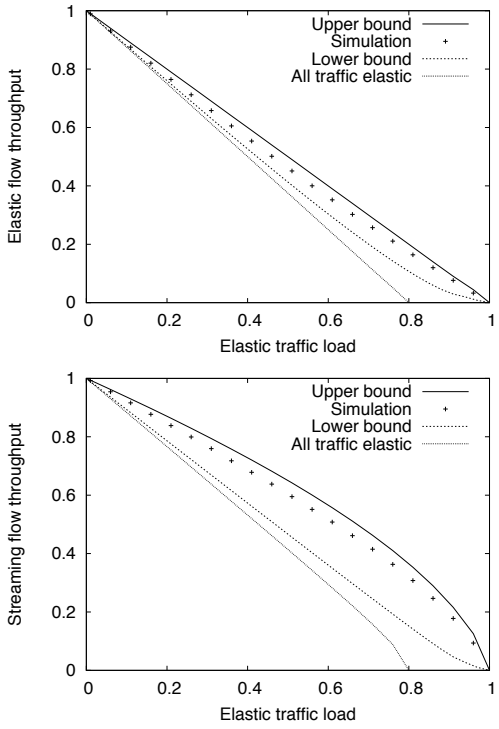

Figure 5: Performance of elastic traffic (upper graph) and streaming traffic (lower graph) on branch 1 of a tree network ($C_1 = 1$, $C_2 = 0.5$).

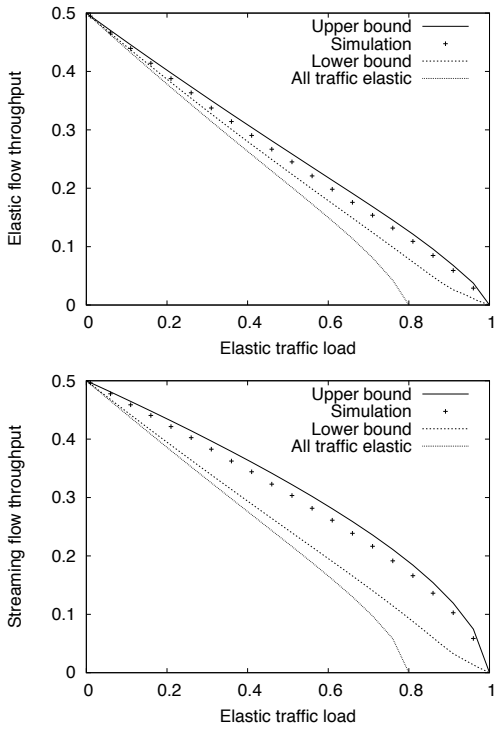

Figure 6: Performance of elastic traffic (upper graph) and streaming traffic (lower graph) on branch 2 of a tree network ($C_1 = 1$, $C_2 = 0.5$)

Multirate system. We now consider the case of a single unit capacity link with two different rate limits a_1 and a_2, with $a_1, a_2 \leq 1$, as depicted in Figure 7.

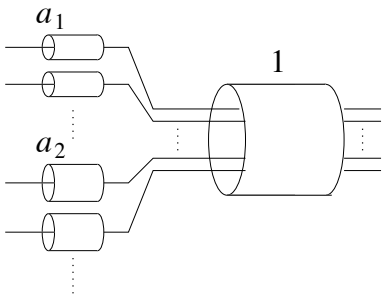

Figure 7: A 2-class multirate system.

Again, the function Ψ characterizing balanced fairness is explicit in this case [6]. The monotonicity property holds if and only if the corresponding inequalities (20) and (21) are satisfied, which may be proved in a similar way as Proposition 1. Figures 8 and 9 give the results obtained for the rate limits $a_1 = 0.5$ and $a_2 = 0.2$. The traffic intensity is the same for each class and streaming traffic still represents 20% of the overall traffic.

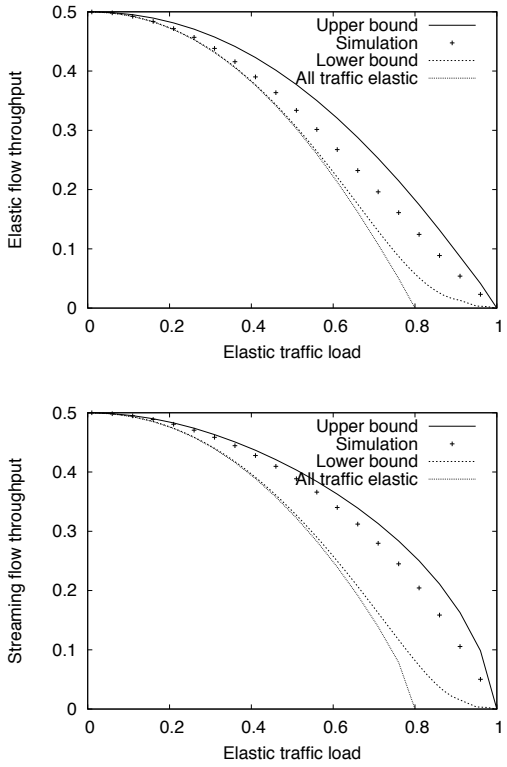

Figure 8: Performance of elastic traffic (upper graph) and streaming traffic (lower graph) of class-1 users for a multirate system ($a_1 = 0.5$, $a_2 = 0.2$).

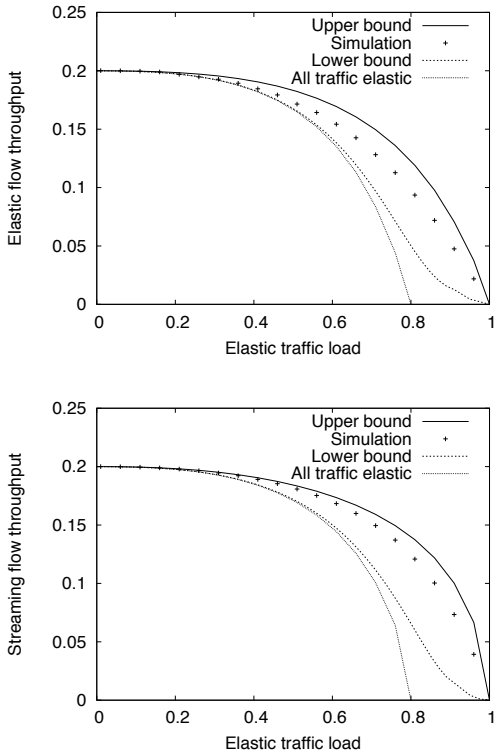

Figure 9: Performance of elastic traffic (upper graph) and streaming traffic (lower graph) of class-2 users for a multirate system ($a_1 = 0.5$, $a_2 = 0.2$).

5. CONCLUSION

We have derived performance bounds for a network integrating elastic and adaptive streaming flows. These bounds have the great practical interest of being insensitive to traffic characteristics like the distributions of elastic flow size and streaming flow duration. Provisioning rules can then be developed based on forecasts of elastic and streaming traffic demands only, independently of the complex traffic structure which is continually evolving as new applications emerge. This requires that both elastic and streaming flows are truly adaptive, which could be enforced by the implementation of packet-level mechanisms like per-flow fair queueing in routers.

APPENDIX

We first provide the necessary material on insensitivity results for networks of processor sharing nodes, then give the proof of Theorem 1 and Proposition 1.

A. INSENSITIVITY RESULTS

We consider a network of N processor sharing nodes. Customers arrive as a Poisson process of intensity λ_i at node i, require i.i.d. services of unit mean and leave the network once served. Nodes are coupled through their state-dependent service rates. We denote by $\phi_i(x)$ the service rate of node i in state $x = (x_1, \ldots, x_N)$, where x_i is the number of customers at node i. Let e_i be the unit vector with 1 in component i and 0 elsewhere, for $i = 1, \ldots, N$.

Balanced networks. A particular class of processor sharing networks is characterized by the following balance property:

$$\frac{\phi_i(x-e_j)}{\phi_i(x)} = \frac{\phi_j(x-e_i)}{\phi_j(x)}, \quad \forall i,j,\ \forall x:\ x_i>0,\ x_j>0.$$

Such networks are known as Whittle networks [21]. They are characterized by a balance function Φ recursively defined by $\Phi(0)=1$ and:

$$\Phi(x) = \frac{\Phi(x-e_i)}{\phi_i(x)}, \quad \forall i, \forall x:\ x_i>0.$$

Note that this definition is unique in view of the balance property. For any x, $\Phi(x)$ may be viewed as the weight of any direct path from state x to state 0, where a direct path is a set of consecutive states $x(0)\equiv x, x(1), x(2), \ldots, x(n)\equiv 0$ such that $x(m) = x(m-1) - e_{i(m)}$ for some $i(m)$, $m = 1,\ldots,n$, with $n = |x|$, and the weight of such a path is the inverse of the product of $\phi_{i(m)}(x(m))$ for $m=1,\ldots,n$ (refer to Figure 10).

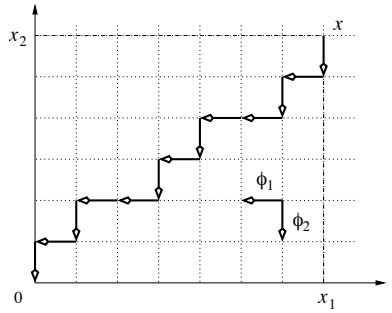

Figure 10: The balance function $\Phi(x)$ is equal to the weight of any direct path from state x to state 0.

A Whittle network is stable if and only if:

$$\sum_x \Phi(x) \prod_{i=1}^N \lambda_i^{x_i} < \infty,$$

in which case the stationary distribution is:

$$\pi(x) = \pi(0)\Phi(x) \prod_{i=1}^N \lambda_i^{x_i}.$$

In particular, this stationary distribution is *insensitive* to the distribution of service requirements at any node.

Non-balanced networks. It has recently been shown that the balance property is in fact a necessary condition for insensitivity [5]. Thus for a non-balanced network, the stationary distribution is sensitive and it proves extremely difficult to derive explicit expressions. Stochastic bounds on the network state $x(t)$ can be derived, however, provided the network is monotonic in the following sense:

$$\forall x, \forall i,j: x_i>0,\ x_j>0, \quad \xi_i(x-e_j) \geq \xi_i(x),$$

where $\xi_i(x) \equiv \phi_i(x)/x_i$ denotes the per-customer service rate at node i in state x. The proof is based on sample-path comparisons with two balanced networks [7].

Let $\hat{\Phi}$ and $\check{\Phi}$ be the balance functions recursively defined by $\hat{\Phi}(0) = \check{\Phi}(0) = 1$ and:

$$\hat{\Phi}(x) = \min_{i:\ x_i>0} \frac{\hat{\Phi}(x-e_i)}{\phi_i(x)}.$$

$$\check{\Phi}(x) = \max_{i:\ x_i>0} \frac{\check{\Phi}(x-e_i)}{\phi_i(x)},$$

For any x, $\hat{\Phi}(x)$ and $\check{\Phi}(x)$ thus correspond to the minimum weight and the maximum weight of all direct paths from state x to state 0 (cf. Figure 10). The service rates in these networks are respectively larger and smaller than in the original network:

$$\hat{\phi}_i(x) \geq \phi_i(x) \quad \text{and} \quad \check{\phi}_i(x) \leq \phi_i(x), \quad \forall i, \forall x: x_i>0.$$

We deduce from the monotonicity property that the number of customers at each node of the networks characterized by the balance functions $\hat{\Phi}$ and $\check{\Phi}$ is respectively smaller and larger than in the original network. Denoting by $\hat{x}(t)$ and $\check{x}(t)$ the corresponding network states at time t, we deduce that:

$$\hat{x}(t) \leq x(t) \leq \check{x}(t), \tag{22}$$

where $x \leq y$ means $x_i \leq y_i$ for all i. The stationary distributions of the network states $\hat{x}(t)$, $\check{x}(t)$ are respectively given by:

$$\hat{\pi}(x) = \hat{\pi}(0)\hat{\Phi}(x) \prod_{i=1}^N \lambda_i^{x_i}, \quad \check{\pi}(x) = \check{\pi}(0)\check{\Phi}(x) \prod_{i=1}^N \lambda_i^{x_i},$$

provided the stability condition holds:

$$\sum_x \check{\Phi}(x) \prod_{i=1}^N \lambda_i^{x_i} < \infty.$$

Note that, in view of (22), this condition implies that the original network and that characterized by the balance function $\hat{\Phi}$ are stable.

It proves difficult to derive explicit expressions for the balance functions $\hat{\Phi}$ and $\check{\Phi}$ in general. A particular case is when the nodes can be numbered in such a way that the following bias property holds for the original network:

$$\frac{\phi_i(x-e_j)}{\phi_i(x)} \leq \frac{\phi_j(x-e_i)}{\phi_j(x)}, \quad \forall i \leq j,\ \forall x:\ x_i>0,\ x_j>0.$$

In this case, we simply have:

$$\hat{\Phi}(x) = \frac{\hat{\Phi}(x-e_i)}{\phi_i(x)}, \quad \check{\Phi}(x) = \frac{\check{\Phi}(x-e_i)}{\phi_i(x)},$$

where i is the maximum index and the minimum index such that $x_i>0$, respectively. Thus $\hat{\Phi}(x)$ and $\check{\Phi}(x)$ correspond to weights of straight paths from state x to state 0. For $N=2$ classes, we obtain:

$$\hat{\Phi}(x) = \frac{1}{\phi_2(x)\ldots\phi_2(x_1+e_2)\phi_1(x_1)\ldots\phi_1(e_1)}.$$

$$\check{\Phi}(x) = \frac{1}{\phi_1(x)\ldots\phi_1(x_2+e_1)\phi_2(x_2)\ldots\phi_2(e_2)},$$

B. PROOF OF THEOREM 1

Necessary stability condition. Assume that $\bar{\rho}_{e,l} \geq C_l$ for some link l. It follows from (14) and (15) that:

$$\forall y, \quad \Psi(y) \geq \frac{1}{C_l} \sum_{i: l \in r_i, y_i > 0} \Psi(y - e_i).$$

Let $\mathcal{X}$ be the set of states x such that $x_s = 0$ and $x_{e,i} = 0$ for all i such that $l \notin r_i$. We deduce from previous inequality that:

$$\forall x \in \mathcal{X}, \quad \Psi(x_e) \geq \binom{x_e}{x_{e,i}, i : l \in r_i} \frac{1}{C_l^{|x_e|}},$$

where the first term follows from the number of direct paths from state x to state 0. In particular,

$$\sum_{x \in \mathcal{X}} \check{\Phi}(x) \lambda_e^{x_e} \lambda_s^{x_s} = \sum_{x \in \mathcal{X}} \Psi(x_e) \rho_e^{x_e} \geq \sum_{n=0}^{\infty} \left(\frac{\bar{\rho}_{e,l}}{C_l} \right)^n = \infty.$$

Sufficient stability condition. Assume that $\bar{\rho}_{e,l} < C_l$ for all links l. As in [6, Theorem 2], we use the fact that:

$$\forall y, \quad \Psi(y) \leq \Psi'(y), \tag{23}$$

whereΨ' is the balance function associated with the so-called "store-and-forward" allocation:

$$\Psi'(y) = \sum_{z : z_i + \sum_{l \in r_i} z_{l,i} = y_i} \prod_{i=1}^{N} \frac{1}{a_i^{z_i} \, z_i!} \prod_{l=1}^{L} \binom{\sum_{i: l \in r_i} z_{l,i}}{z_{l,i}, i : l \in r_i} \frac{1}{C_l^{\bar{z}_l}},$$

with $\bar{z}_l \equiv \sum_{i: l \in r_i} z_{l,i}$. This function satisfies for any $\varrho \in \mathbb{R}_+^N$ such that $\bar{\varrho}_l \equiv \sum_{i: l \in r_i} \varrho_i < C_l$ for all l:

$$\sum_y \Psi'(y) \rho^y = \prod_{i=1}^{N} e^{\frac{\varrho_i}{a_i}} \prod_{l=1}^{L} \left(\frac{1}{C_l - \bar{\varrho}_l} \right). \tag{24}$$

Denoting by $C' = \max_l C_l$ the maximum link capacity, we also use the inequality:

$$\forall y, \quad \Psi(y) \geq \frac{1}{C'^{|y|}}, \tag{25}$$

which simply follows from the fact that, in view of (14),

$$\forall y, \ \forall i : y_i > 0, \quad \Psi(y) \geq \frac{1}{C'} \Psi(y - e_i).$$

We deduce from (23) and (25) that:

$$\sum_x \check{\Phi}(x) \lambda_e^{x_e} \lambda_s^{x_s} \leq \sum_x \binom{x_e + x_s}{x_e} \Psi'(x_e + x_s) C'^{|x_s|} \rho_e^{x_e} \frac{\rho_s^{x_s}}{x_s!}.$$

Now let $\varepsilon \in \mathbb{R}_+^N$ with $\varepsilon_i > 0$ for all $i = 1, \ldots, N$ be such that $\bar{\rho}_{e,l} + \bar{\varepsilon}_l < C_l$, where $\bar{\varepsilon}_l \equiv \sum_{i: l \in r_i} \varepsilon_i$. We have:

$$\sum_{x_e} \binom{x_e + x_s}{x_e} \Psi'(x_e + x_s) \rho_e^{x_e} \varepsilon^{x_s} \leq \sum_y \Psi'(y) (\rho_e + \varepsilon)^y.$$

Using (24), we deduce:

$$\sum_x \check{\Phi}(x) \lambda_e^{x_e} \lambda_s^{x_s} \leq \prod_{i=1}^{N} e^{\frac{\rho_i + \varepsilon_i}{a_i}} \prod_{l=1}^{L} \left(\frac{1}{C_l - \bar{\rho}_{e,l} - \bar{\varepsilon}_l} \right) \times \sum_{x_s} \frac{\rho_s'^{x_s}}{x_s!}.$$

where ρ_s' is the vector defined by $\rho'_{s,i} = \rho_{s,i} C' / \varepsilon_i$ for all i. We conclude that:

$$\sum_x \check{\Phi}(x) \lambda_e^{x_e} \lambda_s^{x_s} \leq \prod_{i=1}^{N} e^{\rho'_{s,i} + \frac{\rho_i + \varepsilon_i}{a_i}} \prod_{l=1}^{L} \left(\frac{1}{C_l - \bar{\rho}_{e,l} - \bar{\varepsilon}_l} \right).$$

The stability condition (19) holds.

C. PROOF OF PROPOSITION 1

Proof of inequality (20). One can readily verify that the inequality holds when either $y_1 = 0$ or $y_2 = 0$. Now it follows from the Pareto-efficiency of tree networks [9] that the trunk is saturated in any state y such that $y_1 > 0$ and $y_2 > 0$, so that $\Psi(y) = \Psi(y - e_1) + \Psi(y - e_2)$. Thus inequality (20) is equivalent to:

$$\Psi(y)^2 \geq \Psi(y + e_1 - e_2) \Psi(y + e_2 - e_1). \tag{26}$$

We prove (26) by induction on $y_1 + y_2$. Note that the result holds for $y = (1, 1)$, since $1/C_1 + 1/C_2 \geq 1/(C_1 C_2)$. It is not difficult to extend the result to any states y such that $y_1 = 1$ or $y_2 = 1$. Assume the result holds for all y such that $y_1 + y_2 \leq n$, and let y be any state such that $y_1 + y_2 = n + 1$. We can assume that $y_1 > 1$ and $y_2 > 1$. Define $a = \Psi(y + e_1 - 2e_2)$, $b = \Psi(y - e_2)$, $c = \Psi(y - e_1)$ and $d = \Psi(y + e_2 - 2e_1)$. We know that $b^2 \geq ac$ and $c^2 \geq bd$. Using the fact that the trunk is saturated in states y, $y - e_1 + e_2$ and $y + e_1 - e_2$, one can easily verify that:

$$\begin{aligned} \Psi(y)^2 - \Psi(y + e_1 - e_2) \Psi(y + e_2 - e_1) &= (b^2 - ac) + (c^2 - bd) + bc - ad, \\ &\geq bc - ad. \end{aligned}$$

The proof then follows from the fact that:

$$\frac{bc}{ad} = \frac{b^2}{ac} \times \frac{c^2}{bd} \geq 1.$$

Proof of inequality (21). We prove the inequality for $i = 2$. As before, we can easily check that the result holds when $y_1 \leq 1$ or $y_2 \leq 1$. If $y_1 > 1$ and $y_2 > 1$, inequality (21) is equivalent to:

$$\begin{aligned} \Gamma(y) \equiv & \Big[(y_2 + 1) \Psi(y + e_1 - e_2) \Psi(y + e_2 - e_1) \\ & + \Psi(y + e_1 - e_2)^2 + 2 \Psi(y) \Psi(y + e_1 - e_2) \Big] \\ & - (y_2 \Psi(y)^2) \geq 0. \end{aligned} \tag{27}$$

We prove (27) by induction on $y_1 + y_2$. Note that the result holds for $y = (1, 1)$. One can also readily extend the result to any state y such that $y_1 = 1$ or $y_2 = 1$. Assume the result holds for all y such that $y_1 + y_2 \leq n$, and let y bet any state such that $y_1 + y_2 = n + 1$. We can assume that $y_1 > 1$ and $y_2 > 1$. The trunk is saturated in states y, $y + e_1 - e_2$ and $y + e_2 - e_1$ so that, applying the result for $y + e_1 - e_2$ and $y + e_2 - e_1$, we get:

$$\Gamma(y) \geq (y_2 + 1) ad + 3ac + 2ab + b^2 - (y_2 - 1) bc.$$

Applying $\Gamma(y-e_2)\geq 0$ and $\Gamma(y-e_1)\geq 0$ successively, we get:

$$\begin{aligned}
&\frac{(y_2+1)ad+3ac+2ab+b^2}{(y_2-1)bc}\\
=\ &\frac{a((y_2+1)d+b+2c)+ac+ab+b^2}{(y_2-1)bc}\\
\geq\ &\frac{y_2ac^2/b+ac+ab+b^2}{(y_2-1)bc}\\
=\ &\frac{\frac{ac}{b}(y_2c+2b+a)+b(a+b)-ac(1+a/b)}{(y_2-1)bc}\\
\geq\ &\frac{(y_2-1)bc+b(a+b)-\frac{ac}{b}(a+b)}{(y_2-1)bc}.
\end{aligned}$$

Now using $b^2\geq ac$, we obtain $\Gamma(y)\geq 0$.

REFERENCES

[1] N. Argiriou, L. Georgiadis, Channel Sharing by Rate-Adaptive Streaming Applications, in: *Proc. of IEEE Infocom*, 2002.

[2] D. Bansal, H. Balakrishnan, S. Floyd and S. Shenker, Dynamic Behaviour of Slowly-Responsive Congestion Control Algorithms, in: *Proc. of ACM SIGCOMM*, 2001.

[3] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié and J. Roberts, Statistical bandwidth sharing: A study of congestion at flow level, in: *Proc. of ACM SIGCOMM*, 2001.

[4] T. Bonald and L. Massoulié, Impact of Fairness on Internet Performance, in: *Proc. of ACM Sigmetrics*, 2001.

[5] T. Bonald and A. Proutière, Insensitivity in processor-sharing networks, Performance Evaluation 49 (2002) 193–209.

[6] T. Bonald and A. Proutière, Insensitive bandwidth sharing in data networks, Queueing Systems 44 (2003) 69–100.

[7] T. Bonald and A. Proutière, On stochastic bounds for monotonic processor sharing networks, to appear in Queueing Systems, 2004.

[8] T. Bonald and J. Roberts, Congestion at flow level and the impact of user behaviour, Computer Networks 42 (2003) 521–536.

[9] T. Bonald, J. Virtamo, Calculating flow level performance under balanced fairness, to appear in Performance Evaluation, 2004.

[10] F. Delcoigne, A. Proutière and G. Régnié, Modeling integration of streaming and data traffic, Performance Evaluation 55 (2004) 185–209.

[11] R. Doshi, P. Cao, Streaming Traffic Fairness Over Low Bandwidth WAN Links, in: *Proc. of the third IEEE Workshop on Internet Applications*, San Jose, 2003.

[12] S. Floyd and K. Fall, Promoting the use of end-to-end congestion control in the Internet, IEEE/ACM Transactions on Networking 7 (1999) 458–472.

[13] F. Guillemin, P. Robert and B. Zwart, Heavy tailed M/G/1-PS queues with impatience and admission control in packet networks, in: *Proc. of IEEE Infocom*, 2003.

[14] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski, Assured Forwarding PHB, IETF RFC 2597, 1999.

[15] V. Jacobson, K. Nichols and K. Poduri, An Expedited Forwarding PHB, IETF RFC 2598, 1999.

[16] F. Kelly, R. Williams, Fluid model for a network operating under a fair bandwidth-sharing policy, to appear in Annals of Applied Probability, 2003.

[17] P. Key, L. Massoulié, A. Bain and F. Kelly, A network flow model for mixtures of file transfers and streaming traffic, in: *Proc. of ITC 18*, 2003.

[18] A.A. Kherani and A. Kumar, Stochastic Models for Throughput Analysis of Randomly Arriving Elastic Flows in the Internet, in: *Proc. of IEEE Infocom*, 2002.

[19] L. Massoulié and J.W. Roberts, Bandwidth sharing and admission control for elastic traffic, Telecommunication Systems 15 (2000) 185–201.

[20] R. Núñez Queija, Sojourn times in a processor sharing queue with service interruptions, Queueing Systems 34 (1-4) (2000) 351–386.

[21] R.F. Serfozo, *Introduction to Stochastic Networks*, Springer Verlag, 1999.

[22] V. Timonen, Simulation studies on performance of balanced fairness, Research Report 6/2003, Helsinki University of Technology, Networking Laboratory, 2003.