# Integration of streaming and elastic traffic: Modeling and Performance Analysis

Henda Ben Cheikh
Esprit School of Engineering, Tunis, Tunisia

Email: henda.bencheikh@esprit.tn

## ABSTRACT

We study a flow-level model where elastic flows compete with fixed-rate streaming flows. We assume that streaming flows are served with priority, the remaining bandwidth being shared by elastic flows according to Balanced Fairness. We derive exact performance results for streaming flows, and propose performance approximations for elastic flows. Simulation results show that the relevant performance metrics are well estimated with the proposed approximations.

## Categories and Subject Descriptors

D.4.8 [Performance]: Queueing Theory.

## Keywords

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

## 1. INTRODUCTION

Although elastic traffic is still predominant on the Internet, we are witnessing a drastic increase of the share of streaming traffic. In this context, network operators increasingly need simple and accurate methods for predicting the performance resulting from an expected demand in elastic and streaming traffic. In [4], insensitive performance bounds for both types of flow have been obtained assuming that streaming flows are "TCP-friendly", that is that they share network ressources fairly with elastic flows as if they were themselves elastic.

However, since streaming flows usually have stringent quality of service requirements, it makes sense to give them some form of priority over best-effort traffic. 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 [7]. The use of a call admission control mechanism to limit the bandwidth taken by streaming traffic has been advocated in [1], where its impact on performance has been investigated using event-driven simulations.

In this paper, we investigate the performance of streaming and elastic flows when a 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.

For a single link, we obtain in Section 2 exact performance results for streaming flows, and compare two different approximation schemes for the computation of the throughput of elastic flows. Preliminary results on how to extend our approach to an arbitrary network topology are also presented in Section 3.

## 2. THE SINGLE LINK CASE

We consider a single link of capacity $C$ that is 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 $\lambda_i$, stay for 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 $\lambda_i$, but instead of having a fixed duration, they have a fixed amount of data to transmit of mean $1/\mu_i$ Mbits. In addition, the sending rate of class-$i$ elastic flows is limited to $c_i$ Mbps. In the following, we refer to $\rho_i = \lambda_i/\mu_i$ as the traffic intensity of class-$i$. 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 $\mathbf{x}$. 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 $\sum_{i\in\mathcal{S}} \phi_i(\mathbf{x}) \leq C_s - d_k$ (lost calls are cleared). We assume that elastic flows share the remaining capacity $C - \sum_{i\in\mathcal{S}} \phi_i(\mathbf{x})$ according to balanced fairness (BF) [2], [8].

A necessary and sufficient condition for stability is therefore

$$\sum_{i\in\mathcal{E}} \rho_i < C - C_s. \tag{1}$$

### 2.1 Performance metrics for streaming traffic

The main performance metric for streaming flows is the

VALUETOOLS 2015, December 14-16, Berlin, Germany

DOI 10.4108/eai.14-12-2015.2262692

blocking probability. Focusing on the number of accepted streaming flows, the set of allowed states is $\mathcal{X}^s = \{\mathbf{x}^s : \sum_{j\in\mathcal{S}} d_j\, x_j^s \le C_s\}$, and a class-$i$ call is blocked whenever $\mathbf{x}^s \in \mathcal{X}_i^s = \{\mathbf{x}^s : C_s - d_i \le \sum_{j\in\mathcal{S}} d_j\, x_j^s \le C_s\}$. The marginal distribution of $\mathbf{x}^s$ is easily obtained using standard results from the theory of multi-rate loss networks [9]

$$\pi_S(\mathbf{x}^s) = \pi_S(\mathbf{0}) \prod_{i\in\mathcal{S}} \frac{\rho_i^{x_i^s}}{x_i^s!}, \quad \mathbf{x}^s \in \mathcal{X}^s, \tag{2}$$

from which the blocking probability of class $i \in \mathcal{S}$ follows: $B_i = \sum_{\mathbf{x}^s\in\mathcal{X}_i^s} \pi_S(\mathbf{x}^s)$.

The well-known Kaufman-Roberts algorithm can be used to efficiently compute $B_i$ [9].

## 2.2 Approximate performance metrics for elastic traffic

The main performance metric for class-$i$ elastic flows is the throughput $\gamma_i$ defined as the ratio $\rho_i/E[x_i^e]$, where $E[x_i^e]$ denotes the expected number of elastic flows in progress.

It is readily verified that in our model the stationary distribution of the network state $\mathbf{x}$ does depend on the distributions of elastic flow size and streaming flow durations. However, following the approach described in [4], insensitive performance bounds can be easily derived. These bounds are however quite loose in general. Due to the lack of space, we do not describe this approach.

We propose below a different approach. This approach is based on a quasistationarity (QS) assumption. The basic idea 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. To simplify the presentation, we shall assume that all elastic flows have a common rate limit $c$, i.e., $c_i = c$ for all $i \in \mathcal{E}$.

Assume that $\mathbf{x}^s$ is kept fixed. Then elastic flows share the remaining bandwidth $C(\mathbf{x}^s) = C - \sum_{i\in\mathcal{S}} \phi_i(\mathbf{x}^s)$ according to BF, which is equivalent to the ordinary PS as long as $\sum_{i\in\mathcal{E}} x_i^e \le N = C(\mathbf{x}^s)/c$. The resulting system can be analyzed using the theory of *Generalized Processor Sharing* (GPS) queues [6].

Provided that the total offered elastic traffic $\theta = \sum_{k\in\mathcal{E}} \rho_k < C(\mathbf{x}^s)$, it yields the following simple expression of the mean number of class-$i$ elastic flows in progress conditioned on $\mathbf{x}^s$

$$E\left[x_i^e|\mathbf{x}^s\right] = \frac{\rho_i}{c} + B(\mathbf{x}^s)\,\frac{\rho_i}{C(\mathbf{x}^s)-\theta}, \tag{3}$$

where $B(\mathbf{x}^s)$ represents the congestion probability of an equivalent link of capacity $C(\mathbf{x}^s)$ and is given by the well-known Erlang delay formula, i.e.,

$$B(\mathbf{x}^s) = \frac{\frac{1}{N!}\left(\frac{\theta}{c}\right)^N \frac{C}{C-\theta}}{\sum_{i=0}^{N-1} \frac{1}{i!}\left(\frac{\theta}{c}\right)^i + \frac{1}{N!}\left(\frac{\theta}{c}\right)^N \frac{C}{C-\theta}}. \tag{4}$$

We note that the total mean number of elastic flows in progress is given by the expected number of customers in the corresponding $M/M/N/\infty$ queue. The above QS approximation immediately yields the following approximation for class-$i$ throughput:

$$\gamma_i = \rho_i / \sum_{\mathbf{x}^s\in\mathcal{X}} E[x_i^e|\mathbf{x}^s]\,\pi_S(\mathbf{x}^s). \tag{5}$$

Although (3) 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 the multi-rate case.

## 2.3 Validity of the QS approximation

To illustrate the results, we consider that a link of capacity $C = 30Mbps$ is shared by 4 traffic classes, the first two corresponding to elastic traffic while the other ones correspond to streaming traffic. It is assumed that $c_1 = c_2 = 4$ Mbps, $d_3 = 3$ and $d_4 = 4$ Mbps. We also assume that the total offered traffic is composed of 90 % of elastic traffic.

Figure 1 compares the proposed approximations with the results obtained with discrete-event simulations for class-1 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.

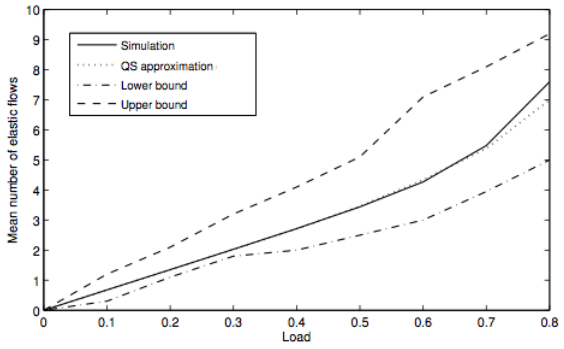


Figure 1: Mean number of class 1 flow in progress.

# 3. ARBITRARY NETWORK TOPOLOGY

We now extend our model to an arbitrary network topology. We denote by $\mathcal{L} = \{1, ..., K\}$ the set of links and assume that link $l$ has capacity $C_l$ Mbps. We define the routing matrix $\mathbf{A}$ as follows: $a_{i,l} = 1$ if class-$i$ flows are routed through link $l$, and 0 otherwise. In the following, $C_s^l$ and $\theta_l = \sum_{i\in\mathcal{E}} a_{i,l}\rho_i$ denotes the maximimum bandwidth that can be used for streaming traffic on link $l$ and the total offered elastic traffic on that link, respectively. We assume that elastic flows share the remaining link capacities according to balanced fairness.

A necessary and sufficient condition for stability is

$$\theta_l < C_l - C_s^l, \forall l \in \mathcal{L}. \tag{6}$$

## 3.1 Performance metrics for streaming traffic

From the conceptual point of view, the problem of calculating end-to-end blocking in a multi-rate loss network can be handled in a nice and simple way, since the product form solution in Section 2.1 still applies (the allowed set is now limited by several linear capacity constraints). However, from a practical point of view, the evaluation of block-

ing probabilities using the product form solution becomes numericaly intractable. One has then to resort to numerical approximation, e.g., using an extension of Kelly's fixed point algorithm to the multi-rate case (see [9] for details).

## 3.2 Performance metrics for elastic traffic

We first recall performance results in absence of streaming traffic. We then extend the results to account for streaming traffic.

### 3.2.1 *Absence of streaming traffic*

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

$$\pi(\mathbf{x}) = \pi(0)\Phi(\mathbf{x})\prod_{i\in\mathcal{E}}\rho_i^{x_i}, \tag{7}$$

where $\Phi$ correspond to the balance function recursively defined by $\Phi(0) = 1$ and

$$\Phi(\mathbf{x}) = \max\{\max_{l\in\mathcal{L}}\frac{1}{C_l}\sum_{i\in\mathcal{E}}\Phi(\mathbf{x}-\mathbf{e}_i)a_{i,l}, \max_{i\in\mathcal{E}}\frac{\Phi(\mathbf{x}-\mathbf{e}_i)}{c_i x_i}\} \tag{8}$$

where $e_i$ represents a unit vector with 1 in position i, and 0 elsewhere. In theory, all performance metrics of interest can be derived from (7). In practice however, the approach based on the computation of state probabilities suffers from the curse of dimensionality. If truncation of the state space is feasible in light traffic regimes, the direct computation of the performance metrics cannot be done when either the number of flow classes gets large, or when traffic intensities are not small enough.

In order to evaluate the number of expected class-$i$ elastic flows in progress in an explicit and simple way, we propose the following approximation.

$$E[x_i] = \frac{\rho_i}{c} + \sum_{l\in\mathcal{L}} a_{i,l} b_l \frac{\rho_i}{C_l - \theta_l}, \tag{9}$$

where $b_l$ is the congestion probability of link $l$ as given in (4). The form of expression (9) is first motivated by numerical observations. We also note that it coincides with the exact result (3) in the case of a single link, and that it is in agreement with the light and heavy traffic approximations obtained in [3].

Remark that in some pathological cases (e.g., a single traffic class passing through multiple links of the same capacity in series) the above approximation is highly inaccurate. Thus, some form of independence of link congestion probabilities is required. As we will show in the following example, this independance assumption is often satisfied in practice, and the approximation provides fairly accurate results.

EXAMPLE 1. *We consider the line lot network of Figure 2 where $C_1 = 25$ Mbps, $C_2 = 30$ Mbps, $C_3 = 35$ Mbps and $c = 4$ Mbps. Defining $p_i = \rho_i / \sum_k \rho_k$ as the proportion of class-$i$ elastic traffic, we consider the cases $p_0 \in \{0.01, 0.1, 0.3\}$ and $p_1 = 2p_2 = 0.5p_3$.*

*Figure 3 shows the evolution of the relative error of the approximation obtained with (9) for the expected number of class-1 flows in progress as a function of the link load. For all scenarios, the relative error is always smaller than 3 %.*

EXAMPLE 2. *Consider the more complex network topology depicted in Figure 4. The network has 10 links and 10 flow classes. The load distribution is given by: $p_1 \in \{0.01, 0.1, 0.2, 0.3\}$ and $p_i = p_j$ for $\forall i, j \neq 1$.*

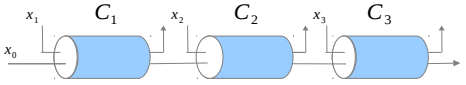


Figure 2: Line lot network

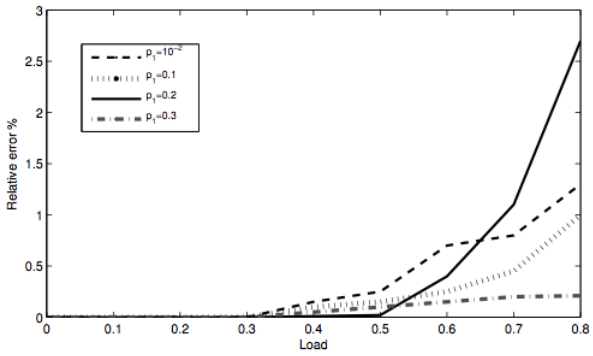


Figure 3: Relative error for class-0 in the network of Figure (cf. 2)

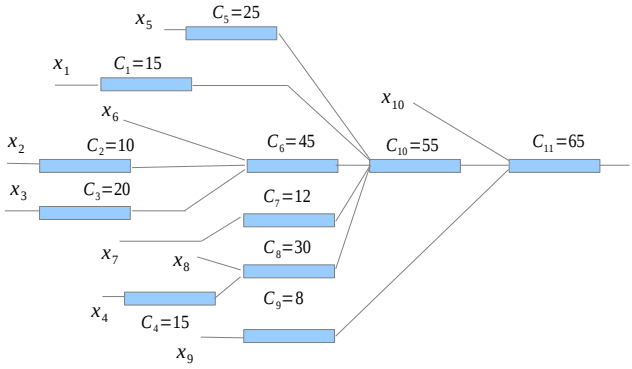


Figure 4: Tree network.

*Figure 5 shows the relative error obtained on the expected number of class-1 flows with respect to the load of the most loaded link. For all considered scenarios, the relative error is always less than 6.5 %.*

### 3.2.2 *Presence of streaming traffic*

In presence of streaming flows, we can use the QS assumption as follows. Given the vector $\mathbf{x}_s$ of accepted streaming flows, we can obtain the mean number $E[x_i^e|\mathbf{x}^s]$ of class-$i$ elastic flows in progress by replacing $C_l$ with $C_l(\mathbf{x}^s) =$

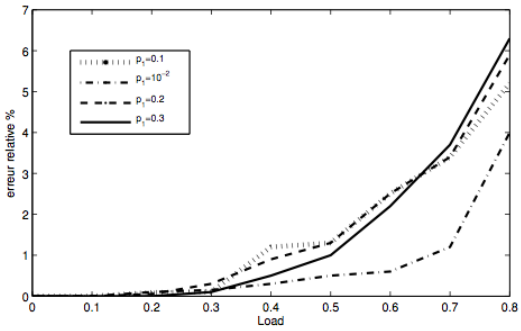


Figure 5: Relative error for class-1 in the Tree network of Figure 4.

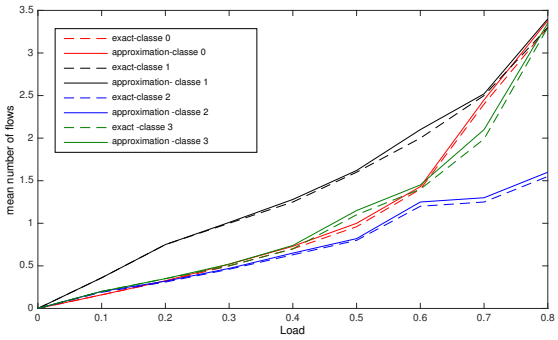


Figure 7: Mean number of classe 1, 2 and 3 flows.

$C_l - \sum_{k\in\mathcal{S}} a_{k,l}\phi_k(\mathbf{x}_s)$ in (9). As for the single link case, we obtain an approximation of class-$i$ throughput:

$$\gamma_i = \rho_i / \sum_{\mathbf{x}^s\in\mathcal{X}} E[x_i^e|\mathbf{x}^s]\,\pi_S(\mathbf{x}^s). \qquad (10)$$

We emphazise that this approach requires however the prior computation of the marginal distribution $\pi_S(\mathbf{x}^s)$ of the number of streaming flows, which is clearly feasible only for a small network due to the curse of dimensionality.

EXAMPLE 3. *Consider two scenarios: (i) in the first, we consider the same exemple as in Figure 2 but with two streaming traffic classes of rate $d_4 = 3$ and $d_5 = 4$ Mbps such that class 4 flows (resp. 5) go through link 1 (resp. 3) (ii) in the second one, we consider the same exemple as in Figure 2 but with three streaming traffic classes of rate $d_4 = 2$, $d_5 = 3$ and $d_6 = 4$ such that flows of class 4 go through link 0, flows of class 5 use link 2 and class 6 flows go through link 3. For both scenarios, we consider that elastic trafic represents 80% of total load.*

*Figure 6 and 7 shows the evolution of the mean number of class-0, 1, 2 and 3 elastic flows obtained with simulations and QS approximation (cf. equation 10)) with respect to the load of the most loaded link. For all considered scenarios, the relative error is always less than 4% in all traffic regimes.*

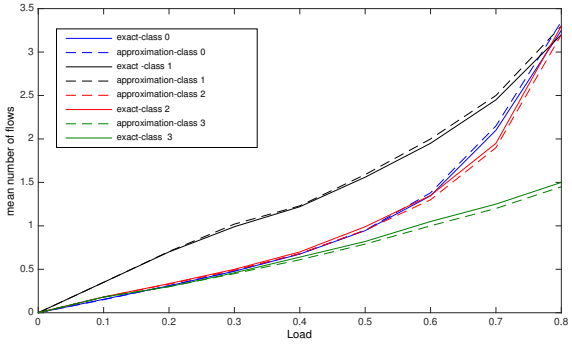


Figure 6: Mean number of classe 1, 2 and 3 flows.

# 4. CONCLUSION

We have investigated the performance of streaming and elastic flows when call admission control is used. For the single link case, exact results for streaming flows and an approximation based on a quasistationarity assumption for elastic flows were presented. The extension of this approach to an arbitrary network topology was also discussed. Future work includes the developement of an approximation scheme that do not require the prior computation of the marginal distribution of the number of streaming flows.

# 5. REFERENCES

[1] N. Benameur, S.Fredj, F.Delcoigne, S.Oueslati, and J.W.Roberts. Integrated admission control for streaming and elastic traffic. *COST 263 International workshop*, Springer, Sep 2001.

[2] 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.

[3] T. Bonald, A. Penttinen, and J. Virtamo. On light and heavy traffic approximations of balanced fairness. In *Proceedings of SIGMETRICS '06/Performance '06*, pages 109–120, New York, NY, USA, 2006. ACM.

[4] T. Bonald and A. Proutière. On performance bounds for the integration of elastic and adaptive streaming flows. In *Proceedings of ACM SIGMETRICS '04/Performance '04*, pages 235–245, New York, NY, USA, 2004.

[5] L. Breslau, E. W. Knightly, S. Shenker, I. Stoica, and H. Zhang. Endpoint admission control: Architectural issues and performance. *In Proceedings of SIGCOMM*, 2000.

[6] J. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12(3):245–284, 1979.

[7] F. Delcoigne, A. Proutière, and G. Régnié. Modeling integration of streaming and data traffic. *Perform. Eval.*, 55(3-4):185–209, Feb. 2004.

[8] L. Massoulié. Statistical bandwidth sharing: A study of congestion at flow level. *Ann. Appl. Probab*, 2007.

[9] J. Roberts, U. Mocci, and J. virtamo. Broadband network teletraffic. *Springer*, 1996.

[10] I. Verloop and R. Núñez-Queija. Asymptotically optimal parallel resource assignment with interference. *Queueing Systems*, 65:43–92, 2010.