Alcom-FT Technical Report Series
ALCOMFT-TR-01-178

Effective bandwidths in a multiclass priority queueing system

Vincent Dumas
IMB, Université Bordeaux I
351, Cours de la Libération
33405 Talence Cedex, France

Fabrice Guillemin
France Telecom R&D
2, Avenue Pierre Marzin
22300 Lannion, France

Philippe Robert
INRIA, Domaine de Voluceau
Rocquencourt, B.P. 105
78153 Le Chesnay Cedex, France

Abstract—**A new admission control algorithm is proposed in this paper for traffic streams regulated by a dual leaky bucket enforcing the peak rate, the mean rate and the maximum burst size. The proposed algorithm relies on a model, which stochastically dominates the theoretical system composed of deterministic worst case traffic sources multiplexed in a finite capacity queue. This dominating model is the $M^{[X]}/M/1/K$ system, where the batch size depends on the maximum burst size of the individual sources, the Poisson input rate is the sum of the mean rates and the mean service time is the service time of one packet. We derive effective bandwidths for both the multiclass $M^{[X]}/M/1/K$ system and the $M^{[X]}/M/1/K$ system including a head of line priority level. In both cases, the remarkable property is that the admissibility regions are delimited by hyperplanes.**

Keywords— **Effective bandwidths, priority queueing systems, statistical multiplexing, asymptotic estimates.**

I. Introduction

In order to offer quality of service to a user in broadband packet networks, a common approach followed by standardisation bodies consists of a negotiation of traffic parameters along with a certain quality of service level between the user and the network before transmitting data. Thus, when traffic delivered by a user is conforming to the negotiated parameters, the quality of service objectives (information loss rate, information transfer delay, etc.) are met within the network. To prevent from a degradation of the quality of service due to malicious usage, the negotiated traffic parameters are enforced at network access by the so-called policing function. This approach has prevailed in the standardisation of ATM networks via the specification of the VBR (variable bit rate) service, and more recently in the evolution of the Internet (e.g., IntServ and CR-LDP for MPLS).

To meet the negotiated quality of service objectives, the network operator has to provision sufficient resources and has to decide whether a new connection can be accepted or not depending on the new parameters and on those of the already established traffic streams. Thus, admission control plays a critical role in the sense that this function has the difficult task of accepting a sufficient amount of traffic in order to otpimize network utilisation while limiting the load so that quality of service requirements are met. The admission control algorithm is thus an essential tool for a network operator.

Admission control has been studied in great detail, especially in connection with the VBR service in ATM network. A basic requirement for this function is that only a limited computational effort should be needed to accept or reject a connection. This requirement comes from the fact that admission control should be performed by switches with limited computational capacities. Moreover, acceptance or rejection of a connection should be done very rapidly in order to limit the latency of the network; ideally the network should respond in real time to a request.

These requirements for the admission control function have led to the development of the concept of effective bandwidth. Theoretically, when connections are multiplexed in a buffer of finite capacity, all the traffic characteristics of the connections should be taken into account to determine the ideal number of connections, which can be accepted. This is however too complicated and the idea of effective bandwidth is to capture the characteristics of a connection through a single parameter in order to easily decide acceptance or rejection.

A huge amount of work has been devoted to the study of effective bandwidths (see for instance [1] and references therein). Since the fundamental papers by Guerin *et al* [2], and Kelly [3], different refinements have been introduced in the concept of effective bandwidths. In particular, these quantities naturally appear as coefficients in the exponential decay term of the survivor probability function of the occupation of a queue fed with Markovian sources (see [4], [5]).

All the theoretical investigations on effective bandwidths have led to the design of several admission con-

Partially supported by the contract 00-1B320 with France Telecom and the IST Programme of the E U under contract number IST-1999-14186 (ALCOM-FT)

trol algorithms. Maybe, the most popular algorithm is that developed by Elwalid, Mitra and Wentworth [6], which allows the computation of the effective bandwidth of a source described by the peak rate, the mean rate and the maximum burst size. This algorithm relies in fact on the reduction of a buffered system to a loss model (i.e., with no buffer), which is then used to estimate the effective bandwidth of a source via refined Chernoff bounds. In this algorithm, special effort has been devoted to bounding the region of admissibility, which is in general a concave domain, by an optimal hyper-plane. This is intended to greatly simplify the admission decision; see [7] for the problem and the impact of linearizing admissibility regions.

While most admission control algorithms have focused on FIFO queues, less attention has been paid to queues with priority schemes. However, most ATM switches or even routers supporting DiffServ, which are available today on the market, implement such schemes. In particular, a salient problem is to develop an admission algorithm, which takes into account a head of line priority level. Such a time priority is in general given to real time traffic, which must pass through network nodes as fast as possible.

The problem of admission control in the presence of priorities has been investigated, for instance, by Berger and Whitt [8], [9]. Most of the time, the higher priority level is taken into account by reducing the service rate seen by the lower priority level. This is the so-called reduced service rate (RSR) approximation. However, this approximation may lead, see for instance [10], to severely underestimate the workload in the lower priority level queue and then to accept too many connections at the risk of not meeting the quality objectives.

In this paper, we investigate the concept of effective bandwidth for a queue with two priority levels. In contrary to earlier studies on the same topic, we do not search an "optimal" effective bandwidth, i.e., an effective bandwidth which results in an optimal use of the transmission link capacity. Instead, we derive a conservative upper bound for the effective bandwidth of a traffic source described by its mean rate and maximum burst size. In particular, we relax the constraint on the peak rate, which is assumed to be infinite. It is worth noting that the amount of resources, which is not used by traffic subject to admission control, can nevertheless be consumed by best effort or elastic traffic, supported by the so called UBR and ABR services in ATM networks.

For this purpose, we introduce a system, which stochastically dominates the system composed of worst case traffic sources multiplexed in a finite capacity queue. This system is an $M^{[X]}/M/1/K$ queue, where the batch size depends on the maximum burst size of the individual sources, the Poisson input rate is the sum of the mean rates and the mean service time is the service time of one packet. The crucial advantage of this system is that the admissibility region is delimited by a hyper-plane. Moreover, a head of line priority (for real time traffic) can easily be introduced and effective bandwidths can still be derived.

The organization of this paper is as follows: The motivation for introducing the $M^{[X]}/M/1/K$ system is presented in Section II. The analysis of this system is carried out in Section III. The extension to a multiclass system is performed in Section IV. The analysis of the system with a priority level is carried out in Section V. Further extensions of the system are presented as concluding remarks in Section VI. Throughout this paper, we pay special attention to the case of VBR traffic in ATM networks, which is the principal target of all the calculations developed hereafter.

II. PROBLEM FORMULATION

We consider in this paper statistical multiplexing of traffic sources, which are regulated by a dual leaky bucket. The first leaky bucket controls the peak rate p of a source with a stringent tolerance. The second leaky bucket enforces the mean rate m of a source with a given tolerance so that the maximum burst size at the peak rate at the output of the regulator is upper bounded by some parameter b. This maximum burst size in number of packets is related to the tolerance τ of the leaky bucket as $b = \lfloor 1 + \tau/(p-m) \rfloor$.

When multiplexing such sources in a FIFO queue with finite capacity, On/Off worst-case traffic is in general considered to compute the volume of traffic admissible for a given loss probability objective. An On/Off worst case traffic pattern is composed of bursts of size b at the peak rate such that the distance between the starting times of two consecutive bursts is b/m. The duration of a burst is b/p and the silence period between two bursts is $b(p-m)/mp$.

Note that On/Off traffic may not be the worst case [11]. This however does not impact the analysis performed in this paper since we consider a model, which asymptotically dominates the worst case of [11].

The complete analysis of a queue fed by the superposition of On/Off worst case traffic sources has been carried out, for instance, in [12], [13]. However, the exact evaluation of performance measures of interest, such as the loss probability, relies on intricate computations, which cannot be carried out by simple processors implemented in network elements. This is why numerous studies have tried to find out approximations, which can be used in practice to develop admission control algorithms (see for instance [6], [14], [1]).

While most of the admission algorithms proposed so far intend to find out optimal parameters so that the volume of bandwidth unused due to connection rejection is as small as possible, we develop in this paper an algorithm, which takes into account an extremal worst case behavior. This is motivated by the fact that even if more accurate admission algorithms can be designed, their practical use remains delicate, in particular due to the concavity of the admissibility region. Moreover, we simplify as far as possible the model in order to take into account a priority level. This last point is motivated by the fact that the reduced service rate approximation does not completely capture the impact of the high priority queue on the low priority queue.

For this purpose, we note that the loss probability when multiplexing On/Off worst case traffic as described above is smaller than the loss probability obtained when data of a connection arrive in batches of size b. Moreover, the loss probability with periodic batches is smaller than the loss probability when batches arrive according to a Poisson process. This point can be rigorously proved by invoking stochastic ordering arguments (see [15]). Finally, to further simplify the system, we suppose that packet service times are exponentially distributed instead of being deterministic (in the case of ATM cells). We are thus led to study the $M^{[X]}/M/1/K$ system. Note furthermore that the Poisson batch arrival process stochastically dominates the worst case traffic pattern described in [11].

The $M^{[X]}/M/1/K$ system has already been considered by Gravey *et al* [14] to determine the maximal burst size in order to allow a bandwidth allocation on the basis of the mean rate. In this paper, we allow for heterogeneous burst sizes. The counterpart is that we have an over-allocation factor on the mean rate.

The real motivation for introducing the $M^{[X]}/M/1/K$ system is in that we can easily account for a priority level and carry out exact computations. When considering a priority level, we assume that the high priority packets have preemptive priority over the low priority packets. This is clearly a distortion of the real system, since the service of a packet cannot be interrupted. Nevertheless, if the size of packets is small (and this is precisely the case for ATM cells), the difference between the preemptive resume priority and the real system will not be significant. This is why we neglect this border effect in the following.

In the case of ATM VBR traffic, let us mention that the mean rate m is equal to the sustainable cell rate (SCR) of the connection, the peak rate p is equal to the peak cell rate (PCR) and b is equal to the maximum burst size (MBS).

III. THE $M^{[X]}/M/1/K$ QUEUE

The first two sub-sections are essentially reminders of the results obtained by Tijms and Van Ommeren in [16]. The last sub-section presents new material, which will be exploited in the following.

A. *Overflow probability*

Consider a finite capacity single-server queue, at which batches of packets arrive according to a Poisson process with rate λ. The batch size has a general probability distribution $\{\beta_j,\ j = 1, 2, \ldots\}$ with generating function $\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j$ and finite mean $\bar{\beta}$.

The buffer capacity is denoted by K. The system can contain only K packets, including the one in service. A batch whose size exceeds the remaining capacity in the buffer is partially lost due to overflow of the excess. The channel can handle only one packet at a time and the service times of the packets are independent random variables having a common probability distribution, which we here assume exponential, with finite mean σ. Letting $\rho = \lambda\bar{\beta}\sigma$, it is assumed that the offered load ρ is less than one.

Furthermore, let

$\pi_{\text{loss}}(K) =$ the long-run fraction of packets that are lost.

The main result of this section is the following theorem.

Theorem 1 (Tijms and Van Ommeren [16]) For any K,

$$\pi_{\text{loss}}(K) = \frac{(1-\rho)\sum_{j \geq K} q_j(\infty)}{1 - \rho \sum_{j \geq K} q_j(\infty)}, \tag{1}$$

where $\{q_j(\infty),\, j = 0, 1, \ldots\}$ is the distribution of generating function

$$Q(z) = \frac{(1-\rho)\beta^*(z)}{1 - \rho z \beta^*(z)}, \quad |z| \leq 1, \tag{2}$$

and $\beta^*(z) = [1-\beta(z)]/[(1-z)\bar{\beta}]$ is the generating function of the distribution $\{\beta_j^* = \sum_{i>j} \beta_i/\beta,\, j = 0, 1, \ldots\}$.

Proof: Equation (1) is a rewriting of equation (1) in [16]. Equation (2) is a rewriting of equation (5) and equation (4) in [16] in the case of exponential service times. ■

B. *Asymptotics*

By Theorem 1, an asymptotically exponential expansion for the overflow probability $\pi_{\text{loss}}(K)$ is obtained when the tail probabilities $q_j(\infty)$ exhibit a geometric behavior for large enough j. These tail probabilities will decrease geometrically fast only when the batch-size distribution has no extremely long tails. Therefore, we make the following assumption.

Assumption 1: The radius of convergence R of the power series $\beta(z) = \sum_{j=1}^{\infty} \beta_j z^j$ is larger than one, and $\rho R \beta^*(R) > 1$.

Note that the convergence radius of the power series $\beta^*(z) = \sum_{j=1}^{\infty} \beta_j^* z^j$ is equal to R since $\beta^*(z) = [1 - \beta(z)]/[(1-z)\bar{\beta}]$. (The point $z = 1$ is a removable singularity.)

Under the above assumption there exists a unique number $z_0 \in (1, R)$ such that

$$\rho z_0 \beta^*(z_0) = 1. \tag{3}$$

This is easily checked by using the fact that the function $z \to \rho z \beta^*(z)$ is a (strictly) increasing continuous function, which takes the value $\rho < 1$ at $z = 1$ and which is such that $\rho R \beta^*(R) > 1$.

We can then establish the following result:

Theorem 2 (Tijms and Van Ommeren [16]) As j tends to ∞,

$$q_j(\infty) \sim \frac{(1-\rho)\beta^*(z_0)}{\rho\left(\beta^*(z_0) + z_0 \beta^{*\prime}(z_0)\right)} z_0^{-j-1}, \tag{4}$$

where z_0 is defined by equation (3).

Proof: Since the power series $\beta^*(z) = \sum_{n=0}^{\infty} \beta_n^* z^n$ is convergent for $|z| < R$, it follows that the right hand side of equation (2) is analytic in the domain $|z| < R$, possibly except at the zeros of the denominator $1 - \rho z \beta^*(z)$.

It is easily checked that $1 - \rho z \beta^*(z)$ has no zeros in the domain $\{z : |z| < z_0\}$ since $|\rho z \beta^*(z)| \leq \rho |z| \beta^*(|z|) < 1$ for $|z| < z_0$. Moreover, the real number z_0 is the unique zero of the function $z \to 1 - \rho z \beta^*(z)$ on the circle $|z| = z_0$, since the real part of $\rho z_0 e^{i\theta} \beta^*(z_0 e^{i\theta})$ is

$$\rho \sum_{j=0}^{\infty} \beta_j^* z_0^{j+1} \cos((j+1)\theta) \leq \rho \sum_{j=0}^{\infty} \beta_j^* z_0^{j+1} = 1,$$

where the inequality is strict unless $\cos(\theta) = 1$ (because $\beta_0^* > 0$). Finally, z_0 is a zero with multiplicity one since $\rho\left(\beta^*(z_0) + z_0 \beta^{*\prime}(z_0)\right) > 0$.

From the above analysis, there exists a number R_1 with $z_0 < R_1 < R$ such that $1 - \rho z \beta^*(z)$ has no zeros in the domain $|z| \leq R_1$ except at the point $z = z_0$. Consequently, we can write equation (2) as $Q(z) = H(z)/(z - z_0)$ for some analytic function $H(z)$ in $|z| \leq R_1$ with $H(z_0) \neq 0$. Using the Taylor expansion $H(z) = H(z_0) + (z - z_0)U(z)$, we next find that $Q(z)$ is of the form

$$Q(z) = \frac{a_{-1}}{z_0 - z} + U(z) \tag{5}$$

for $|z| \leq R_1$, $z \neq z_0$, where $U(z)$ is some analytic function in the domain $|z| \leq R_1$ and the residue a_{-1} is given by

$$a_{-1} = \lim_{z \to z_0} (z_0 - z) Q(z) = \frac{(1-\rho)\beta^*(z_0)}{\rho[\beta^*(z_0) + z_0 \beta^{*\prime}(z_0)]}.$$

Since $U(z)$ is analytic for $|z| \leq R_1$, a Taylor series expansion $U(z) = \sum_{j=0}^{\infty} u_j z^j$ holds for $|z| \leq R_1$. The power series $\sum_{j=0}^{\infty} u_j z^j$ is convergent for $z = R_1$ and so $u_j R_1^j$ is bounded in j. Since $Q(z) = \sum_{j=0}^{\infty} q_j(\infty) z^j$ for $|z| < z_0$, we obtain from the series expansion of the right hand side of equation (5) that

$$q_j(\infty) = \frac{(1-\rho)\beta^*(z_0)}{\rho\left(\beta^*(z_0) + z_0 \beta^{*\prime}(z_0)\right)} z_0^{-j-1} + O(R_1^{-j})$$

for all $j \geq 0$.

Using the fact that $R_1 > z_0$, we finally get the desired result. This completes the proof. ■

As an immediate consequence of the Theorems 1 and 2, we have the following asymptotic representation for the loss probability $\pi_{\text{loss}}(K)$.

Theorem 3 (Tijms and Van Ommeren [16]) As K tends to ∞,

$$\pi_{\text{loss}}(K) \sim \frac{(1-\rho)^2}{\rho^2 (z_0 - 1)\left(\beta^*(z_0) + z_0 \beta^{*\prime}(z_0)\right)} z_0^{-K-1}.$$

Proof: Use the fact that

$$z_0 \beta^*(z_0) = 1/\rho$$

in equation (4) and the definition of $\pi_{\text{loss}}(K)$ given by equation (1). This completes the proof. ■

C. Upper bounds

In this section our goal is two-fold. First, we want to simplify the equivalent of Theorem 3 as a function of z_0. More precisely, we derive an upper bound that is a power function of z_0. Second, we want to find an accurate upper bound on $\pi_{\text{loss}}(K)$ in the case when $\rho R \beta^*(R) < 1$. This point, which at first glance might seem groundless for practical applications, will appear crucial when a higher priority queue will be added to our model.

Let us first consider the case when $\rho R \beta^*(R) > 1$. From equation (3), we may consider z_0 as a decreasing function of ρ, with $z_0(1) = 1$. Let us differentiate equation (3) with respect to ρ (in what follows, z_0 and z_0' are understood as $z_0(\rho)$ and $z_0'(\rho)$). We obtain: $z_0 \beta^{*\prime}(z_0) z_0' + z_0' \beta^*(z_0) = -1/\rho^2$, or equivalently,

$$z_0' = -\frac{1}{\rho^2[\beta^*(z_0) + z_0 \beta^{*\prime}(z_0)]}.$$

Since there exists a $\rho^* \in (\rho, 1)$ such that $z_0 - 1 = (\rho - 1)z_0'(\rho^*)$, we deduce that (with $z_0^* = z_0(\rho^*)$):

$$(z_0 - 1)[\beta^*(z_0) + z_0\beta^{*\prime}(z_0)] = \frac{1-\rho}{\rho^{*2}} \frac{\beta^*(z_0) + z_0\beta^{*\prime}(z_0)}{\beta^*(z_0^*) + z_0^*\beta^{*\prime}(z_0^*)} > 1 - \rho,$$

where the inequality stems from the fact that $z_0 > z_0^* > 1$ and the functions β^* and $\beta^{*\prime}$ are increasing over the interval $(1, R)$.

The above arguments along with Theorem 3 yield the following upper bound for $\pi_{\text{loss}}(K)$.

Proposition 1: If $\rho R \beta^*(R) > 1$, then for large enough K,

$$\pi_{\text{loss}}(K) \le \frac{1-\rho}{\rho^2} z_0^{-K}.$$

Now, assume that $\rho R \beta^*(R) < 1$ (with still $R > 1$), hence $R < \infty$ and $\beta^*(R) < \infty$. It follows that the right side of equation (2) is analytic in the domain $|z| < R$, with finite limits on $|z| = R$, and diverges for $|z| > R$. Hence the power series $Q(z) = \sum_{j=0}^{\infty} q_j(\infty) z^j$ is convergent for $|z| \le R$, and divergent for $|z| > R$, and so $q_j(\infty) = o(R^{-j})$ but $\limsup_{j\to\infty} q_j(\infty) z^j = \infty$ for all $z > R$. As a consequence of Theorem 1, we thus have

Proposition 2: If $\rho R \beta^*(R) < 1$, then for large enough K, $\pi_{\text{loss}}(K) \le R^{-K}$.

This bound is accurate in the sense that for all $z > R$,

$$\limsup_{K\to\infty} \pi_{\text{loss}}(K) z^K = \infty.$$

IV. The multiclass $M^{[X]}/M/1/K$ queue

A. *The equivalent single-source model*

Let us now modify the $M^{[X]}/M/1/K$ model studied in the previous section in order to introduce different *source classes*. Assume that there are N source classes and $x_i \in \mathbb{N}$ sources of class i, $1 \le i \le N$. Each source of class i generates batches of packets according to a Poisson process with rate λ_i; the batch size has a general probability distribution with generating function $\beta_i(z)$ and finite mean $\bar{\beta}_i$.

In the following, we shall focus on the case when the batch size is deterministic, but for the moment it is sufficient to assume that the convergence radius of the power series $\beta_i(z)$ is infinite. All sources are assumed to be independent from one another. The service times of the packets are independent random variables having a common exponential distribution with finite mean σ.

Let $x = (x_i)_{1\le i\le N}$. Obviously our new model is equivalent to the single-source model of Section III, where batches arrive at rate

$$\lambda = \lambda(x) = \sum_{i=1}^{N} x_i \lambda_i,$$

have the generating function

$$\beta(z) = \beta(z, x) = \sum_{i=1}^{N} \frac{x_i \lambda_i}{\lambda(x)} \beta_i(z),$$

and mean

$$\bar{\beta} = \bar{\beta}(x) = \sum_{i=1}^{N} \frac{x_i \lambda_i}{\lambda(x)} \bar{\beta}_i.$$

Let $\rho_i = \lambda_i \bar{\beta}_i \sigma$, $1 \le i \le N$, and

$$\rho = \rho(x) = \lambda(x)\beta(x)\sigma = \sum_{i=1}^{N} x_i \rho_i,$$

We assume that the offered load $\rho(x)$ is less than one.

Denoting by $\pi_{\text{loss}}(K, x)$ the overflow probability, we want to characterize $\{x \in \mathbb{N}^N / \pi_{\text{loss}}(K, x) \le \varepsilon\}$ for some given ε (say, $\varepsilon = 10^{-7}$). In applications, the buffer capacity is $K \ge 1000$, so that we consider that we can confidently use the upper bounds of section III (namely that of Proposition 1 here, since $R = \infty$).

An important point is that the equivalent single-source model makes sense for all (not necessarily integer-valued) $x \in \mathbb{R}_+^N$, and that $\pi_{\text{loss}}(K, x)$ is non-decreasing along radii, that is, if $\theta > 1$, $\pi_{\text{loss}}(K, \theta x) \ge \pi_{\text{loss}}(K, x)$, since the single-source model associated with θx can be obtained from the single-source model associated with x by just increasing the arrival rate from $\lambda(x)$ to $\theta\lambda(x)$.

B. *Effective bandwidths*

According to Proposition 1, it is guaranteed that the loss probability $\pi_{\text{loss}}(K, x) \le \varepsilon$ provided that

$$[1 - \rho(x)]/\rho(x)^2 \le 1$$

and $z_0 = z_0(x) \ge (1/\varepsilon)^{1/K}$, or equivalently provided $\rho(x) \ge (-1 + \sqrt{5})/2$ and (in view of equation (3))

$$\rho(x)(1/\varepsilon)^{1/K} \beta^*((1/\varepsilon)^{1/K}, x) \le 1.$$

Notice that

$$\begin{aligned} \rho(x) z \beta^*(z, x) &= \lambda(x)\bar{\beta}(x)\sigma z \frac{1 - \beta(z, x)}{(1 - z)\bar{\beta}(x)} \\ &= \sum_{i=1}^{N} x_i \rho_i z \beta_i^*(z), \end{aligned}$$

hence both conditions are linear in x.

In view of the two above conditions, one has to wonder whether they are compatible and, moreover, what can be said for all x such that $\rho(x) < (-1+\sqrt{5})/2$. For such x it is guaranteed that $\pi_{\text{loss}}(K,x) \leq \varepsilon$ provided we can find a $\theta > 1$ such that $\rho(\theta x) \geq (-1+\sqrt{5})/2$ and

$$\sum_{i=1}^{N} \theta x_i \rho_i (1/\varepsilon)^{1/K} \beta_i^*((1/\varepsilon)^{1/K}) \leq 1.$$

This is the case if and only if

$$\frac{-1+\sqrt{5}}{2}(1/\varepsilon)^{1/K}\beta_i^*((1/\varepsilon)^{1/K}) \leq 1, \quad 1 \leq i \leq N.$$

We have thus proved the following theorem.

Theorem 4: If

$$\frac{-1+\sqrt{5}}{2}(1/\varepsilon)^{1/K}\beta_i^*((1/\varepsilon)^{1/K}) \leq 1, \tag{6}$$

for $1 \leq i \leq N$, it is granted that $\pi_{\text{loss}}(K,x) \leq \varepsilon$ provided that

$$\sum_{i=1}^{N} x_i a_i(\varepsilon) \leq 1,$$

where the *effective bandwidths* $a_i(\varepsilon)$ are given by:

$$a_i(\varepsilon) = \rho_i (1/\varepsilon)^{1/K} \beta_i^*((1/\varepsilon)^{1/K}), \quad 1 \leq i \leq N. \tag{7}$$

The practical interest of the above result is in that the equivalent bandwidth of a class depends only upon its characteristics. More precisely, it is worth noting that the equivalent bandwidth is equal to the mean rate (expressed in packet/s) of the traffic sources times an over-allocation factor. In the case of deterministic batches, this over-allocation factor depends on the ratio between the maximum burst size and the buffer capacity. This is really in line with the definition of the equivalent bandwidth, which has to capture through a single parameter the behavior of a source when multiplexed with other sources in a queue.

A practically important corollary can be obtained in the case of deterministic batch sizes..

Corollary 1: Assume $\beta_i(z) = z^{k_i}$ for some $k_i \geq 1$, $1 \leq i \leq N$, $\varepsilon = 10^{-n}$ for some $n \geq 1$, $K \geq 100n$. Then condition (6) is satisfied if

$$\max_{1 \leq i \leq N} \frac{k_i}{K} \leq \frac{0.37}{n}.$$

Proof: For $1 \leq i \leq N$, we have

$$\frac{-1+\sqrt{5}}{2}(1/\varepsilon)^{1/K}\beta_i^*((1/\varepsilon)^{1/K}) \leq \frac{-1+\sqrt{5}}{2} 10^{1/100} \frac{10^{nk_i/K}-1}{\ln(10)nk_i/K}.$$

Let

$$f(x) = \frac{-1+\sqrt{5}}{2} 10^{1/100} \frac{10^x - 1}{\ln(10)x}.$$

Function f is continuous and increasing on $\mathbb{R}_+$, and $f(x) = 1$ for $x = 0.37$. This completes the proof. ■

It is worth noting that the equivalent bandwidth given by equation (7) can be used when VBR connections with an MBS of, say, 150 cells are multiplexed in a queue with a capacity of 3000 cells if the target loss probability is $\varepsilon = 10^{-7}$. Such MBS values are commonly encountered in VBR services currently offered by network operators. Moreover, 3000 cells are usual buffer capacities implemented for VBR traffic in ATM switches available today on the market.

It is also important to note that when the buffer capacity is such that $K/k_i \gg n \log 10$ with $\varepsilon = 10^{-n}$, $(1/\varepsilon)^{(k_i/K)} \sim 1$ and then $a_i(\varepsilon)$ is closed to the mean rate of the source. Coming back to a VBR connection with a sustainable cell rate SCR (expressed in cells per second) and a maximum burst size MBS, we have $\lambda = SCR/MBS$. Assuming that $\eta \times MBS \ll 1$ where $\eta = n \log 10/K$, we have

$$a(\varepsilon) = SCR\left(1 + (MBS-1)\frac{\eta}{2}\right).$$

We hence deduce that under the condition $\eta MBS \ll 1$, the equivalent bandwidth given by equation (7), although conservative, is not far from the ideal value equal to SCR. (This value is ideal in the sense that it corresponds to a resource allocation based on the mean rate only, without introducing any over-allocation factor.)

When the above condition is not satisfied (for instance when $K = 3000$, $MBS = 150$, $n = 7$), we have

$$a(\varepsilon) = 1.532 \times SCR.$$

This means that an over-allocation factor of 53% has to be added to the sustainable cell rate. This is reasonable when the sustainable cell rate is small (say, a few hundreds of Kbit/s). The over-allocation factor may however be a potential drawback for high bit rate connections. It is important to note that the over-allocation factor depends only on the ratio MBS/K and not on the SCR and PCR values.

In spite of the over-allocation factor, the key advantage of the above equivalent bandwidth definition is in that the equivalent bandwidth defines an admissibility region, which is delimited by a hyper-plane. This is a crucial advantage when comparing the above admission control algorithm with other algorithms defining concave admissibility regions, which can be delimited by hyper-planes only through intricate computations.

Finally, let us discuss the admission control algorithm developed by Gravey *et al*, which puts a uniform upper-bound on MBS values [14] in order to allow a resource allocation based on the SCR. Roughly speaking, with that admission control algorithm, when connections are multiplexed in a FIFO queue with capacity K, then their MBS value should be less than or equal to $K/\kappa_\varepsilon(\rho)$, where $\kappa_\varepsilon(\rho)$ is the capacity needed in a finite capacity $M/D/1$ queue in order to obtain a rejection probability less than ε when the load is equal to ρ. For instance, in the case when $K = 3000$ cells, $\rho = 80\%$, and $\varepsilon = 10^{-7}$, this admission control would limit the MBS value to $K/32 = 93$ cells since $\kappa_\varepsilon(\rho) = 32$. Note that since the load in 80%, this is equivalent to introduce an over-allocation of 25% with respect to the sustainable cell rate value. In reality, the MBS values admissible by using the equivalent bandwidth definition (7) are larger than the limit imposed by the algorithm of Gravey *et al*. The counterpart is however that the over-allocation factor due to the equivalent bandwidth is greater.

From the above discussion, it turns out that the admission control based on the equivalent bandwidth definition (7) should be carefully compared with other admission control algorithms. Nevertheless, the key advantage of the method used in this paper and relying on a stochastic majorization of the system is in that a priority level can be easily accounted for. This point is addressed in the next section.

V. Addition of a higher priority $M/G/1$ queue

A. *The equivalent single-source model*

We now complete our multiclass model by adding a higher priority $M/G/1$ queue. Independently from the N source classes considered in the previous section, high priority packets arrive at the single-server channel according to a Poisson process with rate λ_0. Their service times have a general probability distribution function $F_0(x)$ with Laplace transform $F_0^*(s) = \int_0^\infty e^{-sx} dF_0(x)$, $s \geq 0$, and finite mean σ_0. This new stream has preemptive resume priority over the original sources, and has a separate buffer of infinite capacity. Letting $\rho_0 = \lambda_0\sigma_0$, it is assumed that the total offered load $\rho_0 + \sum_{i=1}^N x_i\rho_i$ is less than one.

As far as the original sources are concerned, all they see is that service is interrupted during the busy periods of this $M/G/1$ queue. The time between two consecutive busy periods is exponentially distributed with parameter λ_0; the durations of the busy periods have a probability distribution function $G_0(x)$ with Laplace transform $G_0^*(s) = \int_0^\infty e^{-sx} dG_0(x)$, $s \geq 0$; it is well-known (see for example Feller [17], Chap. XIV.4) that G_0^* may be characterized as the smallest root of the equation

$$G_0^*(s) = F_0^*\left(s + \lambda_0(1 - G_0^*(s))\right), \tag{8}$$

which yields the mean: $\pi_0 = -(G_0^*)'(0) = \sigma_0/(1-\rho_0)$.

Furthermore, it is easily seen that the overflow probability in the multiclass $M^{[X]}/M/1/K$ queue will be unchanged if we contract each busy period in one point and gather all the packets arrived meanwhile from the original sources in one single batch. It is a matter of elementary calculation to obtain that the generating function of this batch size is

$$\beta_0(z,x) = G_0^*\left(\sum_{i=1}^N x_i\lambda_i(1-\beta_i(z))\right), \tag{9}$$

and that its mean is

$$\bar\beta_0(x) = \pi_0 \sum_{i=1}^N x_i\lambda_i\bar\beta_i.$$

This procedure is equivalent to replacing the higher priority $M/G/1$ queue with one additional source of class, say, 0, which generates batches of packets according to a Poisson process with rate λ_0; the batch size has a probability distribution with generating function $\beta_0(z,x)$ and finite mean $\bar\beta_0(x)$; the service times of the packets are independent random variables having the usual exponential distribution with finite mean σ.

Finally, our new model is equivalent to the single-source model of section III where batches arrive at rate

$$\lambda(x) = \lambda_0 + \sum_{i=1}^N x_i\lambda_i,$$

have as generating function

$$\beta(z,x) = \frac{\lambda_0}{\lambda(x)}\beta_0(z,x) + \sum_{i=1}^N \frac{x_i\lambda_i}{\lambda(x)}\beta_i(z),$$

and mean

$$\bar\beta(x) = \frac{\lambda_0}{\lambda(x)}\bar\beta_0(x) + \sum_{i=1}^N \frac{x_i\lambda_i}{\lambda(x)}\bar\beta_i = \frac{\sum_{i=1}^N x_i\lambda_i\beta_i}{\lambda(x)(1-\rho_0)}.$$

Letting

$$\rho(x) = \lambda(x)\bar\beta(x)\sigma = \frac{\sum_{i=1}^N x_i\rho_i}{1-\rho_0},$$

the offered load $\rho(x)$ is less than one as expected.

B. Effective bandwidths

Still denoting by $\pi_{\text{loss}}(K,x)$ the overflow probability, we want to use the upper bounds of Section III, but the question arises of whether we are in the conditions of Proposition 1 or Proposition 2. First note that

$$\begin{aligned}\rho(x)z\beta^*(z,x) &= \lambda(x)\bar\beta(x)\sigma z\frac{1-\beta(z,x)}{(1-z)\bar\beta(x)}\\ &= \rho_0(x)z\beta_0^*(z,x)+\sum_{i=1}^N x_i\rho_i z\beta_i^*(z),\end{aligned}$$

where $\rho_0(x)=\lambda_0\bar\beta_0(x)\sigma$.

According to equation (9):

$$\rho_0(x)z\beta_0^*(z,x)=\lambda_0\sigma z\frac{1-G_0^*\left(\sum_{i=1}^N x_i\lambda_i(1-\beta_i(z))\right)}{1-z}$$

The r.h.s. of the above equation is equal to

$$\lambda_0\pi_0\frac{1-G_0^*\left(\sum_{i=1}^N x_i\lambda_i(1-\beta_i(z))\right)}{\left[\sum_{i=1}^N x_i\lambda_i(1-\beta_i(z))\right]\pi_0}\times\sum_{i=1}^N x_i\lambda_i\sigma z\frac{1-\beta_i(z)}{1-z},$$

which yields

$$\begin{aligned}&\rho_0(x)z\beta_0^*(z,x)\\ &=\frac{\rho_0}{1-\rho_0}\overline{G}_0^*\left(\sum_{i=1}^N x_i\lambda_i(1-\beta_i(z))\right)\sum_{i=1}^N x_i\rho_i z\beta_i^*(z),\end{aligned}$$

where $\overline{G}_0^*(s)=(1-G_0^*(s))/(s\pi_0)$ is the Laplace transform of the probability density function $(1-G_0(x))/\pi_0$.

We thus obtain:

$$\rho(x)z\beta^*(z,x)=\frac{\sigma z}{z-1}f_0\left(\sum_{i=1}^N x_i\lambda_i(\beta_i(z)-1)\right), \tag{10}$$

with

$$f_0(s)=s\left(1+\frac{\rho_0}{1-\rho_0}\overline{G}_0^*(-s)\right).$$

Notice that f_0 only depends on the characteristics of the $M/G/1$ queue.

At this stage we must make an assumption on the tail behaviour of $F_0(x)$. We assume that F_0^* is defined for $s\ge s^*$, where $-\infty\le s^*<0$ is a given real number. By convention and in order to simplify the notation in the following, we assume that F_0^* is infinite outside the interval (s^*,∞). Note that when $s^*>-\infty$, s^* is a singularity for the Laplace transform F_0^*. The following results are of important practical use (see the proof in Appendix A).

Proposition 3: Let

$$\begin{aligned}S &= \sup\{s: F_0^*(-s)<\infty\},\\ T &= \sup\{t: G_0^*(-t)<\infty\},\\ \varphi_t(s) &= F_0^*(t+\lambda_0(1-s))-s,\end{aligned}$$

for $s,t\in\mathbb{R}$. Note that $S=-s^*$ and we have the properties:

- $\{\varphi_t,\ t\in\mathbb{R}\}$ is a non-increasing family of convex functions which are positive on $\mathbb{R}_-$ with $\varphi_t(-\infty)=+\infty$. When $s^*=-\infty$, $\varphi_t(+\infty)=\infty$
- $-T=\inf\{t\mid \inf_{\mathbb{R}}\varphi_t\le 0\}$, and for $t\ge -T$, $G_0^*(t)$ is the smallest root of φ_t.
- T and $G_0^*(-T)$ are both finite.
- $T>0$ if and only if $S>0$.

In the following, we assume that $S>0$ and, to avoid trivialities, that $x\ne 0$. Since the power series $\beta_i(z)$, $1\le i\le N$, have infinite convergence radii, equation (10) clearly shows that the convergence radius $R(x)$ of $\beta^*(z,x)$ (or equivalently of $\beta(z,x)$) is the unique number $z>1$ such that:

$$\sum_{i=1}^N x_i\lambda_i(\beta_i(z)-1)=T. \tag{11}$$

Using equation (10) again, we deduce that:

$$\rho(x)R(x)\beta^*(R(x),x)=\frac{\sigma R(x)}{R(x)-1}f_0(T). \tag{12}$$

Set

$$C=\frac{\sigma(1/\varepsilon)^{1/K}}{(1/\varepsilon)^{1/K}-1}f_0(T).$$

We have to consider two cases, namely $C>1$ and $C<1$.

In the case when $C<1$, the probability $\pi_{\text{loss}}(K,x)\le\varepsilon$ if the convergence radius $R(x)\ge(1/\varepsilon)^{1/K}$, since then $\rho(x)R(x)\beta^*(R(x),x)<1$ and we fall into the framework of Proposition 2. In view of equation (11), condition $R(x)\ge(1/\varepsilon)^{1/K}$ is equivalent to

$$\sum_{i=1}^N x_i\lambda_i(\beta_i((1/\varepsilon)^{1/K})-1)\le T,$$

or

$$\sum_{i=1}^N x_i\rho_i(1/\varepsilon)^{1/K}\beta_i^*((1/\varepsilon)^{1/K})\le\frac{\sigma(1/\varepsilon)^{1/K}}{(1/\varepsilon)^{1/K}-1}T.$$

Notice that the right-hand side is less than $1-\rho_0$ since $T<(1-\rho_0)f_0(T)$. We have thus proved the following theorem.

Theorem 5: If $C < 1$, the probability $\pi_{\text{loss}}(K,x) \le \varepsilon$ provided that

$$\sum_{i=1}^{N} x_i a_i(\varepsilon) \le 1 - \rho_0,$$

where the *effective bandwidths* $a_i(\varepsilon)$ are given by:

$$a_i(\varepsilon) = (1-\rho_0)\frac{\lambda_i}{T}\left(\beta_i((1/\varepsilon)^{1/K}) - 1\right), \tag{13}$$

for $1 \le i \le N$.

If $C > 1$ and the convergence radius

$$R(x) < 1/(1-\sigma f_0(T))^+$$

(where $(a)^+ = \max(a,0)$ and $1/0 = +\infty$), then $\rho(x)R(x)\beta^*(R(x),x) > 1$ and we fall into the framework of Proposition 1. In view of equation (11), condition $R(x) < 1/(1-\sigma f_0(T))^+$ is equivalent to

$$\sum_{i=1}^{N} x_i\lambda_i(\beta_i(1/(1-\sigma f_0(T))^+) - 1) > T. \tag{14}$$

It is guaranteed that the probability $\pi_{\text{loss}}(K,x) \le \varepsilon$ provided that $\rho(x) \ge (-1+\sqrt{5})/2$ and

$$\rho(x)(1/\varepsilon)^{1/K}\beta^*((1/\varepsilon)^{1/K},x) \le 1,$$

which in view of equation (10) is equivalent to

$$f_0\left(\sum_{i=1}^{N} x_i\lambda_i(\beta_i((1/\varepsilon)^{1/K}) - 1)\right) \le \frac{(1/\varepsilon)^{1/K}-1}{\sigma(1/\varepsilon)^{1/K}} = \frac{1-\varepsilon^{1/K}}{\sigma}.$$

Function f_0 is continuous and increasing on $[0,T]$, with $f(0) = 0$ and $f_0(T) > (1-\varepsilon^{1/K})/\sigma$ (by assumption), hence the above condition is equivalent to

$$\sum_{i=1}^{N} x_i\lambda_i(\beta_i((1/\varepsilon)^{1/K}) - 1) \le f_0^{-1}\left((1-\varepsilon^{1/K})/\sigma\right), \tag{15}$$

or

$$\sum_{i=1}^{N} x_i\rho_i(1/\varepsilon)^{1/K}\beta_i^*((1/\varepsilon)^{1/K}) \le \frac{f_0^{-1}\left((1-\varepsilon^{1/K})/\sigma\right)}{(1-\varepsilon^{1/K})/\sigma}.$$

Notice that $f_0^{-1}(u)/u < 1-\rho_0$ since $f_0(t) > t/(1-\rho_0)$. The following lemma shows that conditions (14) and (15) are always compatible.

Lemma 1: If $\sigma f_0(T) < 1$ and

$$\sum_{i=1}^{N} x_i\lambda_i\left[\beta_i\left(\frac{1}{1-\sigma f_0(T)}\right) - 1\right] \le T,$$

then

$$\sum_{i=1}^{N} x_i\lambda_i\left[\beta_i\left((1/\varepsilon)^{1/K}\right) - 1\right] < f_0^{-1}\left((1-\varepsilon^{1/K})/\sigma\right).$$

Proof: It is clearly equivalent to show that

$$\lambda_i\left[\beta_i\left((1/\varepsilon)^{1/K}\right) - 1\right]T < \lambda_i\left[\beta_i\left(\frac{1}{1-\sigma f_0(T)}\right) - 1\right] f_0^{-1}\left((1-\varepsilon^{1/K})/\sigma\right)$$

for $1 \le i \le N$, or

$$f_0\left(\frac{\beta_i\left((1/\varepsilon)^{1/K}\right) - 1}{\beta_i\left(1/(1-\sigma f_0(T))\right) - 1}T\right) < \frac{(1/\varepsilon)^{1/K}-1}{\sigma(1/\varepsilon)^{1/K}}$$

for $1 \le i \le N$.

Notice that $(1/\varepsilon)^{1/K} < 1/(1-\sigma f_0(T))$ by assumption, and that f_0 is convex on $[0,T]$, with $f_0(0) = 0$. As a consequence:

$$f_0\left(\frac{\beta_i((1/\varepsilon)^{1/K}) - 1}{\beta_i(1/(1-\sigma f_0(T))) - 1}T\right) \le \frac{\beta_i((1/\varepsilon)^{1/K}) - 1}{\beta_i(1/(1-\sigma f_0(T))) - 1} f_0(T)$$

for $1 \le i \le N$. We hence just need to check that

$$\frac{\beta_i\left((1/\varepsilon)^{1/K}\right) - 1}{\beta_i\left(1/(1-\sigma f_0(T))\right) - 1} f_0(T) < \frac{(1/\varepsilon)^{1/K}-1}{\sigma(1/\varepsilon)^{1/K}}$$

for $1 \le i \le N$, or equivalently

$$(1/\varepsilon)^{1/K}\beta_i^*((1/\varepsilon)^{1/K}) < \frac{1}{1-\sigma f_0(T)}\beta_i^*\left(\frac{1}{1-\sigma f_0(T)}\right),$$

for $1 \le i \le N$, which is obviously the case. This completes the proof. ■

Arguing as in the proof of Theorem 4, we finally obtain the following theorem.

Theorem 6: If $C > 1$ and

$$\frac{-1+\sqrt{5}}{2}(1/\varepsilon)^{1/K}\beta_i^*\left((1/\varepsilon)^{1/K}\right) \le \frac{f_0^{-1}\left((1-\varepsilon^{1/K})/\sigma\right)}{(1-\varepsilon^{1/K})/\sigma}, \tag{16}$$

for $1 \leq i \leq N$, it is guaranteed that $\pi_{\text{loss}}(K, x) \leq \varepsilon$ provided

$$\sum_{i=1}^{N} x_i a_i(\varepsilon) \leq 1 - \rho_0,$$

where the *effective bandwidths* $a_i(\varepsilon)$ are given by:

$$\begin{aligned} a_i(\varepsilon) &= \lambda_i(1-\rho_0)\frac{\beta_i((1/\varepsilon)^{1/K})-1}{f_0^{-1}(f_0(T)/C)} \\ &= \lambda_i(1-\rho_0)\frac{\beta_i((1/\varepsilon)^{1/K})-1}{f_0^{-1}\left((1-\varepsilon^{1/K})/\sigma\right)}, \quad (17) \end{aligned}$$

for $1 \leq i \leq N$.

Notice that the effective bandwidths are continuous functions of C even when C tends to 1.

To conclude this section, let us note that in the case of ATM VBR connections, $a_i(\varepsilon) \sim SCR$ when $K \gg n \log 10 MBS$. Thus, the equivalent bandwidth is close to the SCR value. Moreover, note that the admissibility regions are still delimited by hyper-planes. Finally, it is worth noting that the equivalent bandwidths defined by (13) and (17) capture the impact of the high priority queue on the low priority queue. This impact is much more complicated than simply reducing the service rate of the low priority queue and depends on the distribution of the busy periods in the high priority queue.

C. Exponential case

We are now interested in the specific case of exponential service times for the $M/G/1$ queue. Thanks to Proposition 3, it is a matter of simple analysis to obtain that

$$\overline{G}_0^*((1-\rho_0)t/\sigma_0) = \frac{2}{1+t+\sqrt{(1+t)^2+4\rho_0 t/(1-\rho_0)}},$$

and

$$\begin{aligned} T &= \frac{(1-\sqrt{\rho_0})^2}{\sigma_0}, && (18) \\ f_0(T) &= \frac{1-\sqrt{\rho_0}}{\sigma_0}, && (19) \\ C &= \frac{\sigma}{\sigma_0}\frac{1-\sqrt{\rho_0}}{1-\varepsilon^{1/K}}. && (20) \end{aligned}$$

Trite calculations eventually lead to the surprisingly simple expression:

$$\frac{f_0^{-1}(u)}{u} = 1 - \frac{\rho_0}{1-\sigma_0 u}, \quad u \leq \frac{1-\sqrt{\rho_0}}{\sigma_0}.$$

Theorems 5 and 6 may then be reformulated as follows.

Theorem 7: If $C < 1$, it is guaranteed that the loss probability $\pi_{\text{loss}}(K, x) \leq \varepsilon$ provided

$$\sum_{i=1}^{N} x_i a_i(\varepsilon) \leq 1 - \rho_0,$$

where the *effective bandwidths* $a_i(\varepsilon)$ are given by:

$$a_i(\varepsilon) = (1-\rho_0)\frac{\lambda_i \sigma_0}{(1-\sqrt{\rho_0})^2}(\beta_i((1/\varepsilon)^{1/K})-1), \quad (21)$$

for $1 \leq i \leq N$.

Theorem 8: If $C > 1$ and

$$\frac{-1+\sqrt{5}}{2}(1/\varepsilon)^{1/K}\beta_i^*\left((1/\varepsilon)^{1/K}\right) + \frac{\rho_0}{1-(1-\sqrt{\rho_0})/C} \leq 1$$

for $1 \leq i \leq N$, it is guaranteed that $\pi_{\text{loss}}(K, x) \leq \varepsilon$ provided

$$\sum_{i=1}^{N} x_i a_i(\varepsilon) \leq 1 - \rho_0,$$

where the *effective bandwidths* $a_i(\varepsilon)$ are given by:

$$a_i(\varepsilon) = \rho_i(1/\varepsilon)^{1/K}\beta_i^*((1/\varepsilon)^{1/K}) \times \frac{\sigma - (1-\varepsilon^{1/K})\sigma_0}{\sigma - \dfrac{1-\varepsilon^{1/K}}{1-\rho}\sigma_0}, \quad (22)$$

for $1 \leq i \leq N$.

VI. CONCLUSION

We have introduced in this paper the $M^{[X]}/M/1/K$ queue to get a stochastic upper bound on the rejection probability in a queue fed with the superposition of worst case traffic sources. This model is then used to develop an admission control algorithm for traffic sources described by their peak rate, mean rate and maximum burst size. In a pure FIFO case, it is possible to define an equivalent bandwidth for a source. This equivalent bandwidth is proportional to the mean rate and the coefficient of proportionality depends only on the ratio of the maximum burst size to the buffer capacity. It has rapidly been checked that the results given by the above equivalent bandwidth are comparable with those obtained when using other admission control algorithms published earlier in the literature; exhaustive comparison studies are currently in progress.

In a second step, we have introduced a head of line priority level in the original system. The remarkable property of this new system is that it is possible to define an equivalent bandwidth for traffic sources in the presence of the head of line priority. The key feature of the equivalent bandwidth is that it fully captures the impact of the high priority queue on the low priority queue. It turns out that the interaction between the two queue is much more complex as in the so-called reduced service rate approximation.

Since the $M^{[X]}/M/1/K$ queue majorizes in a stochastic sense the real system, the obtained admission control algorithm is indeed conservative but the bandwidth unused by guaranteed bandwidth services could be consumed by elastic or best effort services. The real advantage of the algorithm obtained in this paper is that it is robust when introducing head of line priorities, while admissibility regions are delimited by hyper-planes.

To develop an admission control in a complete network, it is necessary to understand how traffic parameters are altered when a flow traverses a queue, especially in the presence of head of line priorities. Indeed, in such a case, several bursts of a stream could be aggregated and a new maximum burst size should be computed. This point will be addressed in further studies.

APPENDIX

I. PROOF OF PROPOSITION 3

It is obvious that $\{\varphi_t, t \in \mathbb{R}\}$ is a non-increasing family of convex functions which are positive on $\mathbb{R}_-$. By Jensen's inequality:

$$\varphi_t(s) \geq f_t(s) := e^{-(t+\lambda_0(1-s))\sigma_0} - s,$$

Hence $\varphi_t(-\infty) = \varphi_t(+\infty) = +\infty$.

Set $-\tau = \inf\{t \mid \inf_{\mathbb{R}} \varphi_t \leq 0\}$. It is easy to check that $\inf_{\mathbb{R}} f_t \leq 0$ if and only if $t \geq (\ln(\rho_0) + 1 - \rho_0)/\sigma_0$, therefore $\tau \leq (\rho_0 - 1 - \ln(\rho_0))/\sigma_0$, and since $G_0^*(t)$ is a root of φ_t for all t such that $G_0^*(t) < \infty$ (Feller [17], XIV.4):

$$T \leq \tau < \infty.$$

For $t > -T$, let $\pi(t)$ be the smallest root of φ_t.

$\pi(t) > 0$ and $\varphi_t(s) > 0$ for all $s < \pi(t)$. Now, let us define the sequence $\pi^n(t)$ as follows: $\pi^0(t) = 0$ and for $n \geq 0$, $\pi^{n+1}(t) = \varphi_t(\pi^n(t)) + \pi^n(t)$.

By induction $\pi^n(t) < \pi^{n+1}(t) < \pi(t)$ (since $F_0^*(\cdot)$ is decreasing), and the limit of the sequence $\{\pi^n(t)\}$ is a root of φ_t which is bounded above by $\pi(t)$, hence $\pi(t) = \lim \pi^n(t)$.

Now $\pi^1(-\tau + u) = F_0^*(\lambda_0 - \tau + u)$ $(u > 0)$ is completely monotone (see [17], XIII.4), which recursively shows ([17], Criterion XIII.4.2) that $\pi^2(-\tau + u)$, $\pi^3(-\tau + u), \ldots$ are completely monotone. By the continuity theorem ([17], Theorem XIII.1.2) the same is true for the limit $\pi(-\tau + u)$, hence ([17], Theorem XIII.4.1a) it is the Laplace transform of a positive measure on $\mathbb{R}_+$. But $\pi(-\tau + u)$ and $G_0^*(-\tau + u)$ coincide for $u > \tau$ ([17], Example XIII.4(a)), hence ([17], Theorem XIII.1.1a) they correspond to the same measure on $\mathbb{R}_+$ (namely $e^{\tau x} G_0\{dx\}$), which eventually shows that $G_0^*(-\tau + u) = \pi(-\tau + u)$ for all $u > 0$. In particular, $G_0^*(-t) < \infty$ for all $t < \tau$, therefore $T \geq \tau$, and finally $T = \tau$.

Now set $\pi_0^+(t) = \sup\{s \mid \varphi_t(s) \leq 0\}$ for $t > -T$. $\pi_0^+(t)$ is finite because $\varphi_t(+\infty) = +\infty$. Moreover, since φ_t decreases when t increases, we have:

$$G_0^*(t) < G_0^*(u) < \pi_0^+(u) < \pi_0^+(t) \quad \text{for } t > u > -T,$$

hence $G_0^*(-T) = \lim_{t \to T^-} G_0^*(-t)$ is finite.

Finally, since $G_0^*(0) = 1$ and $\varphi_0'(1) = \rho_0 - 1 < 0$, obviously $T > 0$ if and only if $\pi_0^+(0) > 1$, hence if and only if $S > 0$, which completes the proof of Proposition 3.

REFERENCES

[1] J. Roberts, U. Mocci, and J. Virtamo, Eds., *Broadband network teletraffic*, Lecture Notes in Computer Science, Springer Verlag, 1996, Cost 242 Final Report.

[2] R. Guérin, H. Ahmadi, and M. Nagshineh, "Equivalent capacity and its application to bandwidth allocation in high speed networks," *IEEE J. Sel. Areas Commun.*, vol. 6, pp. 968–981, 1991.

[3] F. Kelly, "Effective bandwidths at multi-type queues," *Queueing Systems*, vol. 10, pp. 5–15, 1991.

[4] A.I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 329–343, 1993.

[5] G. de Veciana, G. Kesidis, and J. Walrand, "Resource management in wide area networks using effective bandwidth," *IEEE JSAC*, vol. 13, pp. 1081–1090, 1995.

[6] A. Elwalid, D. Mitra, and R.H. Wentworth, "A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node," *IEEE J. Sel. Area Commun.*, vol. 13(6), pp. 1115–1127, August 1995.

[7] S. Borst and D. Mitra, "Asymptotically achievable performance in ATM networks," *Adv. Appl. Prob.*, vol. 30(2), pp. 568–585, 1998.

[8] A.W. Berger and W. Whitt, "Effective bandwidths with priorities," *IEEE/ACM Trans. on Networking*, vol. 6, no. 4, pp. 447–460, 1998.

[9] A.W. Berger and W. Whitt, "Extending the effective bandwidth concept to networks with priority classes," *IEEE Communications Magazine*, vol. 36, no. 8, pp. 78–84, 1998.

[10] S. Delas, R. R. Mazumdar, and C. Rosenberg, "Cell loss asymptotics in buffers handling a large number of sources with HOL service priorities," in *Proc. of INFOCOM'99*, New York, 1999, Longer version submitted to Queueing Systems.

[11] G. Kesidis and T. Konstantopoulos, "Worst case performance of a buffer with independent shaped arrival processes," *IEEE Comm. Lett.*, vol. 4, no. 1, pp. 26–28, 2000.

[12] J. Garcia, J.M. Barceló, and O. Casals, "An exact model for the multiplexing of worst case traffic sources," in *Proc. Sixth Int. Conf. on Data Comm. Syst. and their Performance*, Istanbul, October 1995, IFIP.

[13] I. Cidon, R. Guérin, I. Kessler, and A. Khamisy, "Analysis of a statistical multiplexer with generalized periodic sources," *Queueing Systems*, vol. 20, pp. 139–169, 1995.
[14] A. Gravey, J. Boyer, K. Sevilla, and J. Mignault, "Resource allocation for worst case traffic in ATM networks," *Performance Evaluation*, vol. 1-2, pp. 19–43, 1997.
[15] F. Baccelli and P. Brémaud, *Elements of queueing theory*, vol. 26 of *Applications of Mathematics*, Springer Verlag, New York, 1976.
[16] H. Timjs and J.V. Ommeren, "Asymptotic analysis for buffer behavior in communication systems," *Probability in the Engineering and Informational Sciences*, vol. 3, pp. 1–12, 1989.
[17] W. Feller, *An introduction to probability theory and its applications*, vol. II, John Wiley & Sons Ltd, 1971.