Journal of Computer Science 3 (12): 965 - 979, 2007
ISSN 1549-3636

Jitter Characterization in Admission Control and Pricing Issues in Integrated Multiservice Networks

Fabien Houéto and Samuel Pierre
Mobile Computing and Networking Research Laboratory (LARIM
Department of Computer Engineering, Ecole Polytechnique of Montreal, P.O. Box 6079
Station Centre-ville, Montréal, Que., Canada H3C 3A7

Abstract: This paper analyses the pricing framework of multiservice networks and proposes an improved pricing scheme based on the effective bandwidth concept for taking into account quality of service parameters. Based on the deficiencies noted in the classical effective bandwidth scheme (intolerance to user uncertainty and no guarantee on jitter), we propose an improved charging function which gives more flexibility to the user and we introduce an additional constraint to take into account an eventual guarantee on the jitter or delay variation. We also extended the effective bandwidth pricing scheme to the case with guaranteed jitter, in order to take into account and better deal with the various QoS parameters to be considered in 3G networks. Our proposed charging function improves the classical effective bandwidth scheme, while remaining simple in that it requires that the network only monitors the average rate and duration of each connection. It is also fairer than the classical effective bandwidth scheme as it is more flexible related to user uncertainty and the incentive to an efficient use of network resource is preserved. The constraint on the guaranteed jitter was also tested and proved to be viable.

Key words: Guaranteed jitter, connection admission control, effective bandwidth pricing, multiservice networks, third-generation networks

INTRODUCTION

In recent years, a variety of mobile computers equipped with wireless communication devices have become popular. Mobile users want to have diverse services and applications handy. These services include file transfer, videoconferencing, electronic mail, graphical user interfaces, remote file systems, etc. They require different levels of quality of service (QoS) measured in terms of delay, jitter, transmission error rate. Broadband technologies have emerged, making possible the integration of data, voice and multimedia applications.

The concept of multiservice cellular network is then used to denote the integrated infrastructure dedicated to support these applications and services. In fact, multiservice networks refer to networks carrying multimedia, voice, data and video traffic[1,2]. Their architecture integrates broadband and wireless mobile networks such that they are suitable for multimedia and mobile applications with bursty traffic[1].

The convergence of the Internet, the multiservice network and the traditional POTS (Plain Old Telephone Service), raised the problem of charging and pricing. Learning from the Internet, network operators are now aware that pricing is needed not only to recover costs, but also as a method of control. Therefore, in the last five years, a whole series of research[3-6] has been done in the pricing and charging area. However, these works are more or less adapted to the third-generation solutions with various QoS guarantees.

With the end of the public funding of the Internet, the problem of tariffs in computer networks became crucial. The transition towards a commercial Internet emphasized the need to cover the infrastructure costs through an access cost and possibly a usage-based cost. Therefore, numerous papers have been published on Internet engineering and economics; they give a good overview and introduction to the various models and frameworks[7]. We refer the reader to references[3-6] for a more detailed and complete review of pricing schemes.

Generally, users can be billed according to several factors going from the type of service to the usage via the allocated resources (or a measurement such as effective bandwidth), the duration of the call, the flow,

Corresponding Author: Samuel Pierre, Department of Computer Engineering, Ecole Polytechnique of Montreal, P.O. Box 6079, Station Centre-ville, Montréal, Que., Canada H3C 3A7

the call beginning, the distance, the number of calls, etc. Often, the price is given as a function of a combination of several of these factors.

Globally, pricing schemes can be gathered in two categories:

* Static pricing: this kind of pricing sets a flat fee per unit of resource that is independent of the resource usage or the network state. The fixed price can also be an average resulting from a dynamic scheme.
* Dynamic pricing: the price is in general per unit of resource and varies with the network state.

The effective bandwidth pricing is a dynamic pricing scheme based on the concept of effective bandwidth[8-14] . Basically, the idea is to apply a tariff including a price per unit of time, a price per unit of volume and a connection price. At connection time, the user chooses a tariff corresponding to its expected mean rate. If the user sends more or less than the declared rate, it pays a higher price. Even if in[12] the authors show the efficiency of this model from the competing point of view, some disadvantages are noted. For instance, the user must declare its expected flow, the connection admission algorithm uses the parameters declared by the user; if this latter send more than the declared rate, he/she is penalized at the same time by the loss of its packages and by a higher tariff. A more detailed presentation of the model will be done in later.

This paper analyses the pricing framework of multiservice networks and proposes an improved pricing scheme based on the effective bandwidth concept by taking into account more QoS parameters.

2. Basic concepts and background: In integrated networks, pricing schemes are strongly related to connection admission. Indeed, pricing requires having a good idea of the resources used (or expected to be used) by each connection, which enables a control of connection admission. Generally, connection admission is based on an estimation of the queue/buffer length probability. In this section, we present some concepts related to the estimation of the queue/buffer length probability and its application to connection admission control[15,16].

Cell loss asymptotic: With the development of ATM, there has been a lot of work to compute the buffer length probability or the asymptotic overflow probability. In this section, we summarize results obtained by Likhanov and Mazumdar[17], extending earlier works from Courcoubetis and Weber[18].

Consider N independent, identical, stationary and ergodic sources, each with input rate $\lambda_{n,j}$ where j refers to the j^{th} source and n to the n^{th} time slot. We assume that the time is discrete and that the input rates have some regularity properties[17]. The total amount of work offered by the j^{th} source during time interval $[0,t)$ is $X_{t,j} = \sum_{n=0}^{t-1} \lambda_{n,j}$ and the amount of work offered from all sources is $X_t^{(N)} = \sum_{k=0}^{N} X_{t,k}$.

Let $M_t(s)$ denote the moment generating function of $X_{t,1}$, i.e. $M_t(s) = E[e^{sX_{t,1}}]$. Suppose that the sources access a buffer of size NB with output rate NC. The stationary workload $W^{(N)}$ is given by: $W^{(N)} = \sup_{t \in \{1,\ldots\}} \left(X_{-t}^{(N)} - NCt\right)$ where $X_{-t}^{(N)}$ denotes the total numbers of cells arriving in the interval $(-t,0]$. Then as $N \to \infty$,

$$P\{W^{(N)} > NB\} = \frac{1}{\sqrt{2\pi N \sigma^2 s_0^2}} e^{-NI_{t_o,s_o}(C,B)} \left(1 + O\left(\frac{1}{N}\right)\right) \quad (1)$$

where:

$$\begin{aligned} -I_{t_o,s_o}(C,B) &= ln(M_{t_o}(s_0)) - (Ct_0 + B)s_0 \\ &= \sup_t \inf_s \left[ln(M_t(s)) - (Ct + B)s\right] \end{aligned} \quad (2)$$

and

$$\sigma^2 = \frac{\partial^2}{\partial s^2} \log\left(E[e^{sX_{t,1}}]\right) = \frac{M_t''(s_0)}{M(s_0)} - (Ct + B)^2 \quad (3)$$

In the previous definition, by abuse of notation, we suppose that $0 \le s_0, t_0 \le +\infty$.

In[19], the term $\sigma^2 s_0^2$ is approximated by $2I_{t_o,s_o}(C,B)$. Thus Eq. (1) can be rewritten:

$$P\{W^{(N)} > NB\} = \frac{1}{\sqrt{4\pi N I_{t_o,s_o}(C,B)}} e^{-NI_{t_o,s_o}(C,B)} \left(1 + O\left(\frac{1}{N}\right)\right) \quad (4)$$

Connection admission control and pricing: The effective bandwidth concept was introduced by Hui[20] and Guerin *et al.*[21]. Combined with the work from[17-19], it is used by Kelly *et al.*[8,10,22] in their investigation of linear acceptance region for certain buffered resources and in the design of pricing schemes. We recall some results obtained by Courcoubetis *et al.*[23].

Many source types without priority: Suppose that the arrival process at a broadband link is the superposition of independent identically distributed sources of J types. Let $N_i = Nn_i$, $i=1,\ldots, J$, be the number of sources of type i and let $n=(n_1, \ldots, n_i, \ldots, n_j)$ (the n_i are not necessarily integers). The link is serviced by a buffer

NB at rate *NC*. Parameter *N* is the scaling parameter (size of the system).

Adapting the notations from the previous section, we define $X_{t,j}$ as the total amount of work offered by a source of type *j* during time interval *[0,t)* and $M_{t,j}(s)$ as the moment generating function of $X_{t,j}$, i.e., $M_{t,j}(s) = E[e^{sX_{t,j}}]$, where *E* is the expected value operator. We recall that $X_{t,j}$ has stationary increments. Then the *effective bandwidth* of a source of type *j* is defined as follows:

$$\alpha_j(s,t) = \frac{1}{st} log\, E\left[e^{sX_{t,j}}\right] = \frac{1}{st} log(M_{t,j}(s))$$

where *s, t* are system parameters which are defined by the context of the source.

More precisely, *s* and *t* are defined by Eq. (2), which can be rewritten;

$$\begin{aligned} -I_{t_o,s_o}(C,B) &= s_0 t_0 \sum_{j=1}^{J} n_j \alpha_j(s_0,t_0) - (Ct_0 + B)s_0 \\ &= \sup_t \inf_s \left[st \sum_{j=1}^{J} n_j \alpha_j(s,t) - (Ct + B)s \right] \end{aligned} \quad (5)$$

Consider the QoS constraint on the overflow probability to be $P(overflow) \le e^{-\gamma}$ and assume $\gamma = N\gamma_0$. If a point $(N_1, \ldots, N_j) = (Nn_1, \ldots, Nn_j)$ satisfies:

$$\sum_{j=1}^{J} n_j \alpha_j(s,t) \le C + \frac{1}{t}(B - \frac{\gamma_0}{s}) = C^* \quad (6)$$

where s_0 and t_0 are defined by Eq. (5), then the QoS constraint on the overflow probability $P(overflow) \le e^{-\gamma}$ is satisfied. Thus Eq. (6) defines an acceptance region.

By taking into account the results presented in the previous section (Eq.(4)), the acceptance region can be improved as follows:

$$\sum_{j=1}^{J} n_j \alpha_j(s,t) \le C + \frac{1}{t}(B - \frac{\gamma_0'}{s}) = C_{B-R}^* \quad (7)$$

where $I_{t_o,s_o}(C,B) \approx \gamma_0 - \dfrac{\frac{1}{2}\log(4\pi N\gamma_0)}{N + \frac{1}{2\gamma_0}} = \gamma_0'$

Let's suppose that we have one type j_1 of traffic which is invoiced at a per unit time charge f_1. If the network operator has n_1 connections j_1, then it chooses its price f_1 so that it satisfies:

(O1): $\max f_1 n_1$ under $n_1\alpha_1(s_1,t_1) \le C + \frac{1}{t_1}(B - \frac{\gamma}{s_1}) = K_1$

The solution is $f_1 = \lambda_1 \alpha_1$ where λ_1 is the Lagrange coefficient associated with the constraint.

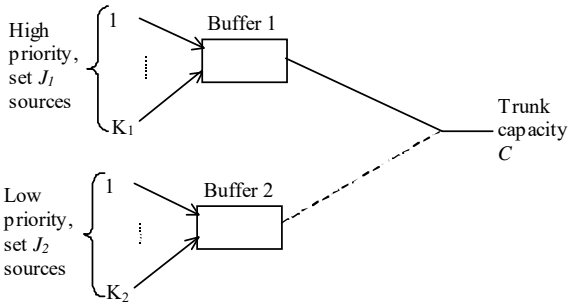

Fig. 1: Priority queuing system

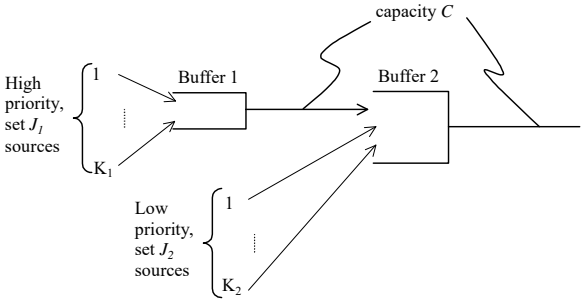

Fig. 2: Equivalent priority queuing system

More generally, suppose we have two types j_1 and j_2 of traffic. The connection admission constraint can be expressed at least locally as a linear constraint $n_1\alpha_1 + n_2\alpha_2 \le C$, where n_1, n_2 are the number of connections of types j_1 and j_2, whereas α_1, α_2 and *C* depend upon the quantities of network resources and the statistical characteristics of types j_1 and j_2 connections. The coefficients α_1, α_2 are the "effective bandwidths". In a competitive equilibrium, the overall social welfare $u(n_1, n_2)$ is maximized subject to this linear constraint. By social welfare, we mean the sum of all user benefits. By formulating the Lagrangian optimization problem, i.e., maximize $u(n_1, n_2) - \lambda\alpha_1 n_1 - \lambda\alpha_2 n_2$ with respect to n_1 and n_2, one can see that the optimum is reached when usage price $\lambda\alpha_1$ and $\lambda\alpha_2$ are posted and n_1 and n_2 are chosen in a decentralized way. Therefore, the proposed tariff should be proportional to the effective bandwidth and hence to the consumed resources.

Many source types with priority: Sometimes, different levels of quality of service are provided to different classes of traffic. Suppose for example, that traffic classes are partitioned into two sets, J_1 and J_2.

Service is FCFS (First Come, First Served), except that traffic sources with $i \in J_1$ are always given priority over traffic sources with $j \in J_2$. Figure 1 is an illustration.

In Fig. 1, Buffer 1 is serviced at rate *C*, whenever it is not empty or a J_1 source is transmitting. Only time-varying residual bandwidth, if any, is available for Buffer 2 service. In[24], the system presented in Fig. 1 is shown to be equivalent to the one in Fig. 2.

If we suppose, that for $i \in J_1$, there is a QoS guarantee on the delay of the form:

$$P(delay > B_1 / C) \le e^{-\gamma_1},$$

and that for all sources, there is a QoS guarantee on cell loss rate:

$$P(buffer\ overflow) \le e^{-\gamma_2},$$

and by noticing that $P(delay > B_1 / C) = P(buffer\ B_1\ overflows)$, we can repeatedly apply Eq. (6) or (7) and obtain two constraints of the form:

$$\sum_{j \in J_1} n_j \alpha_j(s_1, t_1) \le K_1 \quad (8)$$

$$\sum_{j \in J_1 \cup J_2} n_j \alpha_j(s_2, t_2) \le K_2 \quad (9)$$

where K_1 and K_2 are defined by Eq. (6) or (7). The s_i, t_i are the appropriate extremising values.

Effective bandwidth pricing implementation: The proposed tariff, even though simple and efficient, is not easy to implement in a decentralized way. In[8,9], the authors investigated a simple implementation in which the charge per unit time for a connection of type *j* can be expressed as a linear function of the form:

$$f(X) = a_0 + a_1 g_1(X) + \ldots + a_L g_L(X) = a_0 + a^{\perp} g(X)$$

where $g_1(X), \ldots, g_L(X)$ are measurements taken from the observation of $X = (X_1, \ldots, X_T)$, the subscript of *X* referring to the discrete time 1,…,T, or some functions of those measurements and $^{\perp}$ is the transpose symbol. Here *X* and $a_0, \ldots, a_L$ depend on *j* and hence perhaps on policing parameters like the mean rate for sources of type *j*, but the dependence on *j* is not written for greater simplicity.

Suppose that *X* has stationary increments, $X \in \chi(h)$, for a given set $\chi(h)$ parameterized by some vector *h* and the measurements satisfies $E[g(X)] = m$ for a given tuple *m*. Let $\overline{Q}_{g(X)}(m,h)$ be the following upper bound:

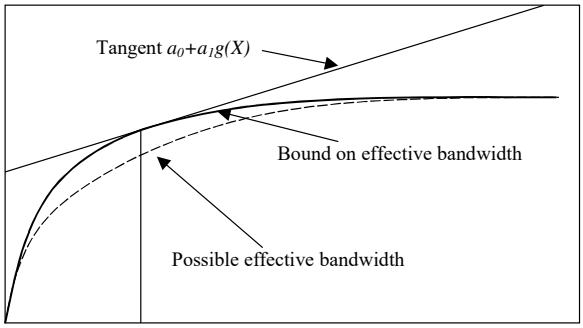

Fig. 3: Effective bandwidth and bounds as a function of *m*

$$\overline{Q}_{g(X)}(m,h) = \sup_{X: E(g(X))=m; X \in \chi(h)} \left\{ E[e^{sX[0,t]}] \right\} \quad (10)$$

with $g(X) = \frac{1}{T}\sum_{i=1}^{T} X_i$, *X[0,t]*, the total load generated within the *[0,t]* interval and

$$\overline{\alpha}(m,h) = \frac{1}{st} \log \overline{Q}_{g(X)}(m,h) \quad (11)$$

$\overline{\alpha}(m,h)$ is concave in *m*.

Figure 3 shows how the effective bandwidth might be bounded by a linear function of the measured parameter, $E(g(X))$. m and *h* can be thought of as respectively the mean and peak rate of the source.

The rest of the paper will be constrained to the case *L=1* and $g_1(X) = \frac{1}{T}\sum_{i=1}^{T} X_i$ [8,9]. In this case, the total charge is just a function of the total number of cells carried and through a_0, the duration of the connection. Thus the network operator publishes a set of possible tariff pairs ($a_0(m)$, $a_1(m)$) defining tangents to the bound on the effective bandwidth $\overline{\alpha}(m,h)$ and given his contract, the user chooses the pair ($a_0(m_e)$, $a_1(m_e)$) corresponding to the expected mean rate m_e of his connection in order to minimize his charge. This pair ($a_0(m_e)$, $a_1(m_e)$) corresponds to the tangent of $\overline{\alpha}$ at point $m_e = \frac{X_t}{t}$. The user price per unit of time is then $\overline{\alpha}$.

Effective bandwidth pricing analysis: The effective bandwidth scheme has been developed as a measure of a connection's resource usage at one switch. However, it can also be used as a measurement along a route through the network because, often in the network,

there is a bottleneck which provides the binding constraint. Moreover, the effective bandwidth at the bottleneck link is often not affected by the multiplexing which takes place in previous buffers. Finally, the *s* and *t* parameters can be used to fine tune the effective bandwidth measurement and capture the overall performance of the network[9].

The effective bandwidth pricing has some intrinsic properties of fairness and incentive compatibility. The fairness properties are implemented by the fact that the charge for a connection is proportional to the bound on the effective bandwidth $\overline{\alpha}$. However, the computation of $\overline{\alpha}$ is generally extremely hard. Therefore, in practice, an approximation $\overline{\alpha}'$ of $\overline{\alpha}$ is used. In[25], Siris presented three simple approximations: an on-off bound which depends solely on the connection's peak rate *h* and its mean rate *m*, a simple and tighter bound based on the leaky bucket algorithm and an inverted *T* approximation. Fortunately, the usage of $\overline{\alpha}'$ instead of $\overline{\alpha}$ preserves the fairness properties.

The incentive compatibility is reached through a cycle. The networks operator posts tariffs that have been computed for the current operating point on the basis of the parameters *s* and *t*. These tariffs provide incentives to the customers to change their contracts in order to minimize their anticipated costs. Under these new contracts, the operating point of the system moves since the network operator must guarantee the performance requirements of these new contracts. The network operator will calculate new tariffs for the new operating point and this interaction will continue until an equilibrium is reached. In[9], it is shown that this equilibrium exists even when the bound $\overline{\alpha}$ is replaced by $\overline{\alpha}'$ and this equilibrium is a point maximizing social welfare.

Finally, the pricing scheme is simple as it requires only the monitoring of the mean rate of a connection.

Nevertheless, in the case of a user whose measured mean rate does not correspond to his declared mean rate, the connection is penalized twice, once by being charged a higher rate ($E(f_{m,h}(X)) \geq \overline{\alpha}(E(g(X)),h)$) and secondly by some of his packets being dropped by the conformance algorithm. Moreover, we recall that the excess charge is done based on an already conservative bound $\overline{\alpha}(E(g(X)),h)$. We will thus try to propose a charging function whose excess charge is smaller while preserving the "good" properties of the linear charging function.

The present charging scheme is certainly fair since the tariff is proportional to the user's resource usage in the basic case of guaranteed cell loss probabilities. Section 2.2.2 shows how the constraints could be extended in the case of guaranteed delay and cell loss probability (Eq. (8) and (9)). However, there is no constraint to take into account an eventual guarantee on the jitter or delay variation. In the following, we will also propose a constraint for the case of guaranteed jitter and then will extended the effective bandwidth pricing to the case with guaranteed delay, jitter and cell loss probability.

3. Our proposed charging function: Suppose that *X* has stationary increments, $X \in \chi(h)$, for a given set $\chi(h)$ parameterized by some vector *h* and the measurement satisfy $E[g(X)] = m$. We recall that $\overline{Q}_{g(X)}(m,h)$ is the following upper bound:

$$\overline{Q}_{g(X)}(m,h) = \sup_{X:E(g(X))=m;X\in\chi(h)} \{E[e^{sX[0,t]}]\}$$ with

$$g(X) = \frac{1}{T}\sum_{i=1}^{T} X_i$$

and $\overline{\alpha}(m,h) = \frac{1}{st}\log \overline{Q}_{g(X)}(m,h)$ (*s, t* are considered fixed).

We also recall that we firstly consider a tariff function of the form:

$$f(X) = a_0 + a_1 g(X) \quad (12)$$

with $a_0 = \overline{\alpha}(m,h) - m\frac{\partial}{\partial m}\overline{\alpha}(m,h)$, $a_1 = \frac{\partial}{\partial m}\overline{\alpha}(m,h)$. *m* and *h* could respectively be interpreted as the mean and the peak rate of the source. We suppose that the peak rate *h* is fixed by the type of the connection as specified by the 3GPP group[26] and the user is allowed to choose *m*. For the sake of simplicity, we note $\overline{Q}_{g(X)}(m) = \overline{Q}_{g(X)}(m,h)$ and

$$\overline{\alpha}(\overline{Q}_{g(X)}(m)) = \overline{\alpha}(m,h) = \frac{1}{st}\log \overline{Q}_{g(X)}(m,h).$$

Our goal is then to propose a charging function whose excess charge is smaller while preserving the "good" properties of the linear charging function. Therefore, we will try to use the concavity of the logarithm function to develop a "better" charging function as it would be less sensitive to the user imprecision without requiring more measurements. Consider

$$h_{\overline{Q},m}(m_r) = \overline{\alpha}(\overline{Q}_{g(X)}(m)) + \lambda_{1,\overline{Q}}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m)) + \lambda_{2,\overline{Q}}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^2 + \lambda_{3,\overline{Q}}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^3 \quad (13)$$

with

$$\lambda_{1,\overline{Q}} = \frac{\partial \overline{\alpha}}{\partial \overline{Q}}\bigg|_{\overline{Q}_{g(X)}(m)} = \frac{1}{st\overline{Q}_{g(X)}(m)},$$

$$\lambda_{2,\overline{Q}} = \frac{1}{2!}\frac{\partial^2 \overline{\alpha}}{\partial \overline{Q}^2}\bigg|_{\overline{Q}_{g(X)}(m)} = -\frac{1}{2st(\overline{Q}_{g(X)}(m))^2},$$

$$\lambda_{3,\overline{Q}} = \frac{1}{3!}\frac{\partial^3 \overline{\alpha}}{\partial \overline{Q}^3}\bigg|_{\overline{Q}_{g(X)}(m)} = \frac{1}{3st(\overline{Q}_{g(X)}(m))^3}$$ and m is the mean rate declared by the user at the connection set-up, while m_r is the real mean rate measured by the network. Figure 4 shows how our charging function bounds the effective bandwidth and the linear tariff function within a given neighborhood.

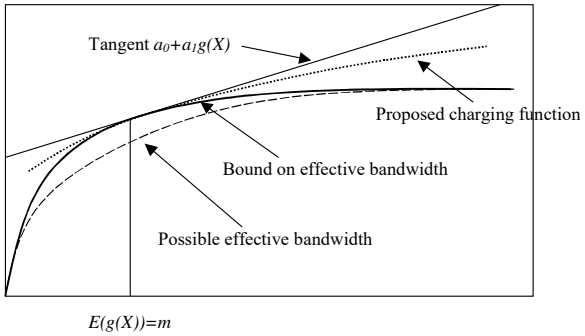

Fig. 4: Proposed charging function

$h_{\overline{Q},m}$ is the Taylor expansion of $\overline{\alpha}$ in the neighbourhood of $\overline{Q}_{g(X)}(m)$.

Notice that the proposed function is as simple as the initial linear function given that the network only needs to monitor the connection mean rate. Indeed, if we have an expression for $\overline{Q}_{g(X)}(m,h)$, we can calculate $h_{\overline{Q},m}$ directly from the measured mean rate *g(X)* by using a polynomial function. The polynomial function used is of higher degree than in the case of Eq. (12), but it is only a function of the measured rate *g(X).* In the cycle to reach the incentive compability, the coefficients of the function can be computed once at the beginning of each round. In fact, we used a series expansion which is a function of the effective bandwidth bound with the supremum $\overline{Q}_{g(X)}(m,h)$ of the moment generating being the free parameter rather than the mean rate *m*, because $\overline{Q}_{g(X)}(m,h)$ does not have a closed form. In practice, many approximations of $\overline{Q}_{g(X)}(m,h)$ are used (see Appendix). Therefore, the proposed function is more general since it can be adapted to the available approximation of $\overline{Q}_{g(X)}(m,h)$. Otherwise, for each approximation of $\overline{Q}_{g(X)}(m,h)$, we need to develop a new charging function. The other properties of the proposed functions are:

Lemma 1: $h_{\overline{Q},m}(m_r)$ is greater or equal to $\overline{\alpha}(\overline{Q}_{g(X)}(m_r))$ with equality when $m_r = m$.

The proof is given in the Appendix.

Lemma 2: $h_{\overline{Q},m}(m_r)$ is increasing in m_r.

The proof is given in the Appendix.

This lemma means that $h_{\overline{Q},m}$ preserves a certain fairness as a customer who makes more use of the network is charged more. The quantification of this fairness is left for further studies. However, as already said, in practice, an approximation $\overline{\alpha}'$ of $\overline{\alpha}$ (and hence an approximation $\overline{Q}'_{g(X)}(m)$ of $\overline{Q}_{g(X)}(m)$) is used. The approximations $\overline{Q}'_{g(X)}(m)$ of $\overline{Q}_{g(X)}(m)$ should be increasing so that $h_{\overline{Q}',m}(m_r)$ could still be guaranteed as an increasing function. In[25], Siris presented three simple approximations $\overline{\alpha}'$ of $\overline{\alpha}$. The corresponding approximations $\overline{Q}'_{g(X)}(m)$ of $\overline{Q}_{g(X)}(m)$ can be directly deduced. In practice, these approximations are "generally" increasing in *m*.

Lemma 3: $h_{\overline{Q},m}(m_r) \le f_{m,h}(X)\big|_{m_r}$ for m_r in a certain neighbourhood of *m*.

Proof: For this proof, we will use the fact that $\overline{Q}_{g(X)}(m)$ is concave in *m*. The proof of the concavity of $\overline{Q}_{g(X)}(m)$ is given in the Appendix. Consider the difference:

$$h_{\overline{Q},m}(m_r) - f_{m,h}(X)\big|_{m_r} = \frac{1}{st\overline{Q}_{g(X)}(m)}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m) - (m_r - m)\frac{\partial \overline{Q}}{\partial m}\bigg|_m)$$

$$-\frac{1}{2st(\overline{Q}_{g(X)}(m))^2}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^2$$

$$+\frac{1}{3st(\overline{Q}_{g(X)}(m))^3}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^3$$

As $\overline{Q}_{g(X)}(m)$ is concave in *m*, then $\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m) - (m_r - m)\frac{\partial \overline{Q}}{\partial m}\Big|_m \le 0$. If we note

$Y_r = \overline{Q}_{g(X)}(m_r)$ and $Y = \overline{Q}_{g(X)}(m)$, then the sum of the two last terms of the difference has the same sign as $R = \frac{2Y_r - 5Y}{6stY}$ and is negative in the neighborhood $[0, \frac{5}{2}Y] = [0, \frac{5}{2}\overline{Q}_{g(X)}(m)]$ of $\overline{Q}_{g(X)}(m)$. And as $\overline{Q}_{g(X)}(m)$ is increasing in *m,* this corresponds to an interval [0, *m'*] such that $m \in [0, m']$. Thus in this neighbourhood of *m*, $h_{\overline{Q},m}(m_r) \le f_{m,h}(X)\big|_{m_r}$, i.e. the user has a better charging function than $f_{m,h}(X)$ at no additional expense on the measurements while keeping a good incentive for the user. However, if the measured rate is not in the interval [0, *m'*], then the user pays a higher rate than $f_{m,h}(X)$. Consequently, even if the user is not obliged to give a precise expected rate *m* for his/her connection, he/she is strongly incited to give a reasonable expected rate so that his/her measured rate m_r verifies $m_r \in [0, m']$. In[25], the approximations $\overline{Q}'_{g(X)}(m)$ of $\overline{Q}_{g(X)}(m)$ presented are linear thus concave; in these cases, the tolerance interval *[0, m']* is equal to $[0, \frac{5}{2}m]$. Even with these simple approximations (on-off bound, simple bound), the proposed tariff function $h_{\overline{Q}',m}(m_r)$ verifies the three previous lemmas and therefore is better than the $f_{m,h}(X)$ tariff function. Even if at point *m,* the proposed tariff function $h_{\overline{Q}',m}(m_r)$ and $f_{m,h}(X)$ are equal, the improvement introduced by $h_{\overline{Q}',m}(m_r)$ is worth as it is done at the same measurement complexity, i.e., the network still only needs to monitor the mean rate of the connection. The slight calculation increase is not even worth mentioning unless the accounting manager has very scarce computation resources. Moreover, this improvement eases (but does not remove) the incentive on the user to strictly respect the declared mean rate (the network is thus more "tolerant" to user's incertitude).

4. Guaranteed Jitter QOS and extended effective bandwidth pricing: In Section 2.2.2, we recalled how the constraints could be extended in the case of guaranteed delay and cell loss probability (Eq. (8) and (9)). However, as previously noted, there is no constraint to take into account an eventual guarantee on the jitter or delay variation. In this section, we mainly propose a constraint for the case of guaranteed jitter and then extend the effective bandwidth pricing to the case where delay, jitter and cell loss probability are guaranteed.

4.1 Cell delay variation bound with a single switch: For real-time applications, the variation of inter-cell arrival times or jitter is important and must be a guaranteed QoS parameter. Therefore, it is important to have an accurate estimate of the cell delay variation of the traffic stream. In[23], the term jitter is used to capture the burstiness of traffic and is defined as the maximum number of packets in an averaging interval. In[27], it is used to capture the magnitude of the distortion to the traffic pattern caused by the network and is defined as the maximum difference between the delays experienced by any two packets on the same connection. In[28], the jitter is referred to as the *peak-to-peak cell delay variation* as specified by the ATM Forum. It is then defined as the (1-α) quantile of the cell transfer delay (CTD) minus the fixed CTD, that could be experienced by any delivered cell on a connection during the entire connection holding time. The term "peak-to-peak" refers to the difference between the best and the worst case of CTD, where the best case is equal to the fixed delay due to the propagation delays on links, switching delays, transmission delays and the worst case is equal to a value likely to be exceeded with probability no greater than α.

In this paper, we define the cell delay variation (CDV) as the (1-α) quantile of the difference between the delays experienced by any two consecutive packets on the same connection during the entire connection holding time. The local CDV is the CDV in one particular switch. End-to-end CDV is the CDV along a connection path from source to destination. For methods to evaluate the end-to-end CDV with local CDVs, the reader is referred to[28].

Let's note B_1 and B_2 the switch buffer size respectively at the arrival of two consecutive packets *P1* and *P2* of the same connection *j* in a switch multiplexing *N+1* connections. If the packets' sizes are small compared to the buffers' sizes, the variation on the delay incurred by these two packets at the switch could be defined as the absolute value $J = \left|\frac{B_2 - B_1}{(N+1)C}\right|$.

Using the notations of Section 2.1, we have:

$$|B_2 - B_1| \le \sup_t(0, \lambda_t^N - (n+1)C\Delta t) \le \sup_t(X_{-t}^{(N)} - (N+1)Ct)$$

$$= \sup_t(X_{-t}^{(N)} - (N)(\frac{N+1}{N}C)t) = W^{(N)}(\frac{N+1}{N}C),$$

where $X_t^{(N)} = \sum_{k=0}^{N} X_{t,k}$ is the aggregation of the N other connections different from the connection j and Δt is the time between two consecutive arrivals on the same connection. Thus, using Eq. (1), the jitter "overflow" probability could be defined by

$$P(J > \frac{NB}{(N+1)C}) = P(|B_2 - B_1| > NB) \le P(W^{(N)}(\frac{N+1}{N}C) > NB)$$

$$= \frac{1}{\sqrt{2\pi N\sigma^2 s_0^2}} e^{-NI_{t_0,s_0}(\frac{N+1}{N}C,B)} \left(1 + O\left(\frac{1}{N}\right)\right)$$

where:

$$-I_{t_o,s_o}(\frac{N+1}{N}C,B) = ln(M_{t_o}(s_0)) - (\frac{N+1}{N}Ct_0 + B)s_0$$

$$= \sup_t \inf_s \left[ln(M_t(s)) - (\frac{N+1}{N}Ct + B)s \right]$$

and

$$\sigma^2 = \frac{\partial^2}{\partial s^2} \log\left(E[e^{sX_{t,1}}]\right) = \frac{M_t''(s_0)}{M(s_0)} - (\frac{N+1}{N}Ct + B)^2 .$$

By replacing $I_{t_o,s_o}(\frac{N+1}{N}C,B)$, we can simply write:

$$P(J > \frac{NB}{C}) = P(|B_2 - B_1| > NB) \le P(W^{(N)}(\frac{N+1}{N}C) > NB)$$

$$\approx e^{-NI_{t_0,s_0}(\frac{N+1}{N}C,B) - \frac{1}{2}log(2\pi N\sigma^2 s_0^2)}$$

$$\le e^{N\ln(M_{t_o}(s_0)) - (N+1)Cs_0t_0 - NBs_0 - \frac{1}{2}\log(2\pi N\sigma^2 s_0^2)} = Le^{-s_0 NB} \quad (14)$$

where $L = e^{N\ln(M_{t_o}(s_0)) - (N+1)Cs_0t_0 - \frac{1}{2}\log(2\pi N\sigma^2 s_0^2)}$ is a constant which does depend on *NB*. If we use Eq. (4) instead of Eq. (1), we will have $L = e^{N\ln(M_{t_o}(s_0)) - (N+1)Cs_0t_0 - \frac{1}{2}\log(4\pi NI_{t_o,s_o}(\frac{N+1}{N}C,B))}$. This equation could be extended to the case of multiple connection type, in a way similar to that used in Section 2.3.1.

4.2 Connection admission control: Consider the following QoS constraint on the jitter *J*:

$$P(J > \frac{NB}{C}) \le e^{-N\alpha_0} = e^{-N\alpha} .$$

Suppose that the sources access a buffer of size *NB* with output rate *NC*. We assume that we have only one type of sources and that they are identically distributed. The previous section has shown that a sufficient condition for respecting the QoS constraint on the jitter is that $P(W^{(N)}(\frac{N+1}{N}C) > NB) \le e^{-N\alpha_0}$. Therefore, Eq. (6) and (7) give respectively the following constraints for the acceptance region:

$$\alpha(s_0,t_0) \le \frac{N+1}{N}C + \frac{1}{t}(B - \frac{\alpha_0}{s}) = C' \quad (15)$$

$$\alpha(s_0,t_0) \le \frac{N+1}{N}C + \frac{1}{t}(B - \frac{\alpha_0'}{s}) = C'_{B-R} \quad (16)$$

where $\alpha(s_0,t_0)$ is the effective bandwidth of one the $N+1$ connections, $I_{t_o,s_o}(\frac{N+1}{N}C,B) \approx \alpha_0 - \frac{\frac{1}{2}\log(4\pi N\alpha_0)}{N + \frac{1}{2\alpha_0}} = \alpha_0'$ and s_0, t_0 are the extremising values of $-I_{t_o,s_o}(\frac{N+1}{N}C,B) = ln(M_{t_o}(s_0)) - (\frac{N+1}{N}Ct_0 + B)s_0$ $= \sup_t \inf_s \left[ln(M_t(s)) - (\frac{N+1}{N}Ct + B)s \right]$.

This process could also be extended to the case of multiple connection type, in a way similar to that used in Section 2.3.1.

Assume that the radio link is the bottleneck in the UMTS network, which provides the binding constraint. Consider a user who wants to connect to such a network. It issues a connection request which is forwarded to an RNC (Radio Network Controller). The RNC applies a connection admission algorithm to verify the availability of resources. When enough resources are available, the connection is allowed, otherwise it is denied. In the case where DCH (Dedicated Channel) channels are used, the user request corresponds to a bandwidth reservation request. In[29], for example the amount of noise rise due to an user *i* for the uplink is $(1+i)L_j$, where L_j is the ratio of the received signal power form user *j* to the total received wideband power including the thermal noise power in the base station and *i* is the other cell to own cell interference ratio as seen by the base station receiver. Therefore there is a connection admission algorithm based on the power resources. In this case, the tariffs for user *i* should be directly proportional to a linear combination of his/her loading factor on uplinks and downlinks.

In the case of DSCH (Downlink Shared Channel) channels, there is also a connection admission scheme at the link level based on the power resources. However, this admission algorithm can not take into account users QoS requirements at the application level as on the DSCH, applications are aggregated on the same link. A solution could be to add a second admission algorithm applied to the radio link considered as the bottleneck to assess the resources needed by each connection request or the possible guarantees on QoS at the application level.

Consequently the user is priced based on his/her effective bandwidth.

4.3 Charging with guaranteed delay, jitter and cell loss probability: Suppose for example that traffic classes are partitioned into two sets, J_1 and J_2. Service is FCFS (First Come, First Served), except that traffic sources with $i \in J_1$ are always given priority over traffic sources with $j \in J_2$. We suppose that, for $i \in J_1$, there is a QoS guarantee on the delay of the form:

$$P(delay > B_1 / C) \le e^{-\gamma_1},$$

and a QoS guarantee on the jitter of the form:

$$P(jitter > B_1 / C) \le e^{-\gamma_2}$$

and that for all sources, there is a QoS guarantee on cell loss rate:

$$P(buffer\ overflow) \le e^{-\gamma_3}.$$

As in Section 2.3.2, we can repeatedly apply Eq. (6) or (7) and obtain constraints of the form:

$$\sum_{j \in J_1} \alpha_j(s_{1'}, t_{1'}) \le K_{1'} \quad (17)$$

$$\sum_{j \in J_1} \alpha_j(s_{1''}, t_{1''}) \le K_{1''} \quad (18)$$

$$\sum_{j \in J_1 \cup J_2} \alpha_j(s_2, t_2) \le K_2 \quad (19)$$

Depending on the values of $s_{1'}, t_{1'}, s_{1''}, t_{1''}, K_{1'}, K_{1''}$ the constraint (17) or (18) will be active and will determine the region in which the network provider can expect to operate[8].

Suppose a network operator charges f_i per unit of time for a connection of type i, $i=1,2$. The revenue $n_1 f_1 + n_2 f_2$ is maximized by operating, if possible, at some point on the boundary of the admission region. If Eq. (19) is active, then it will be appropriate to charge both types 1 and 2 connections at price proportional to $\alpha_1(s_2, t_2)$ and $\alpha_2(s_2, t_2)$ respectively. If one of the equations (17) or (18) is active, then it will be appropriate to charge type 1 connections at a price proportional to $\alpha_1(s_1, t_1)$. If one of the equations (17) or (18) is active simultaneously with equation (19), then type 1 connections should be invoiced at a price $\lambda_1 \alpha_1(s_1, t_1) + \lambda_2 \alpha_1(s_2, t_2)$, where λ_1, λ_2 are shadow prices respectively associated with constraints (17) or (18) and constraint (19), while type 2 connections still incur $\lambda_2 \alpha_2(s_2, t_2)$.

These pricing could be efficiently implemented using the tariff function proposed in Section 3. If we suppose for example that the operator uses one of the linear approximations (14 or 16), for type 2 connections, the network operator proposes a n-uple *(a, b, c, d)* and the user pays for the total duration T of his connection:

$$h(m_r) = aT + bV + cVm_r + dVm_r^2$$

where *a, b, c, d* depend on $m_1, h_1, m_2, h_2, s_2, t_2$.

This formula was obtained by multiplying the per unit of time tariff function $h_{\overline{Q}',m}(m_r)$ by the connection duration T and then by replacing the upper bound $\overline{Q}_{g(X)}(m)$ by its approximations $\overline{Q}'_{g(X)}(m)$ which are linear function of m. Therefore, this charge is proportional to both the duration T and the volume V (two first terms), but introduces some corrections to deal with the uncertainty of the user. These corrections give more flexibility to the user compared to the previously proposed time-volume charging.

In the case of type 1 connections which should be invoiced at a price $\lambda_1 \alpha_1(s_1, t_1) + \lambda_2 \alpha_1(s_2, t_2)$, the charging function for a connection which lasts T units of time and transfers a volume V of data is:

$$\begin{aligned} h(m_r) &= \lambda_1(a_1 T + b_1 V + c_1 V m_r + d_1 V m_r^2) \\ &+ \lambda_2(a_2 T + b_2 V + c_2 V m_r + d_2 V m_r^2), \\ &= aT + bV + cVm_r + dVm_r^2 \end{aligned}$$

where a_1, b_1, c_1, d_1 depend on $m_1, h_1, s_{1'}, t_{1'}, s_{1''}, t_{1''}$ and a_2, b_2, c_2, d_2 depend on $m_1, h_1, m_2, h_2, s_2, t_2$. The form of the charge remains the same.

5. Numerical results: Firstly, we test the results of the previous section concerning the jitter-guarantee constraint (Eq. 17), by using the Comnet ® simulator. The model, as displayed in Comnet, is shown in Fig. 6.

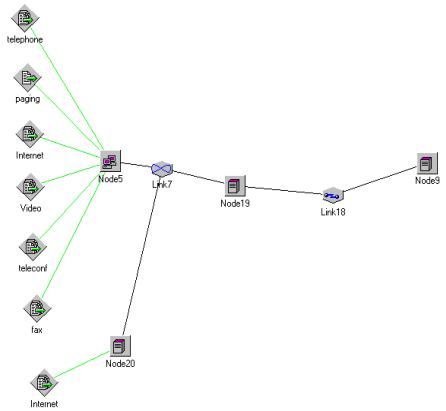

Fig. 5: Comnet test model

Table 1: Characterization of the source models

	Number of messages per session	Messages interarrival time (sec)	Message size (bytes)
Application 1 (Internet)	Pareto (location=33.0, shape=1.16)	Bursty: Interarrival of bursts Pareto (location=1.0, shape=1.5) Interarrival within bursts: Weibull (shape=0.382, scale=1.46) Number of messages within bursts : Pareto (location=1.0, shape=2.43)	Mixed : 93% of messages : Lognormal (mean=8897.0, std=37009.0) Remaining messages Pareto (location=3328.0, shape=1.383)
Application 2 (Voice)	Geometric (min=0.0, mean=3000.0)	Bursty: Interarrival of bursts Exponential (mean=1.002) Interarrival within bursts: 0.02 Number of messages within bursts : Geometric (min=0.0, mean=17.0)	70
Application 3 (Fax)	Exponential (mean=1000.0)	Exponential (mean=0.156)	Exponential (mean=281.0)
Application 4 (Video)	Triangular (min=648.0, mode=1282.0, max=1916.0)	0.04166666	Bursty: First message: Lognormal (mean=1897, std=800) Eleven next messages: Lognormal (mean=637, std=467)

In this model, we use a computer group (*Node5*) to model many similar sources. A processing node (*Node20*) is used to model the source whose jitter is to be measured. The traffic generated by all the sources are multiplexed at *Node19* and then are sent to a sink (*Node9*) through a link (*Link18*), which is the equivalent of the radio link. The capacity of this link is chosen to be 2 Mbps[30]. The network supports packets with a maximum size of 1518 bytes. We simulate a wide variety of traffic models with distributions ranging from uniform to Pareto, via burst distributions, Weibull distributions. Table 1 summarizes the various distributions used in our session models.

The cell delay variation is calculated at *Node20* as this node supports only one instance of the application. We study both the case where the delay variation is defined as the variation between any two consecutive packets on the same connection and the case where the delay variation is defined as the variation relatively to the minimum delay incurred by any packet on the connection. These data are used to deduce the probability $P(jitter > x)$ (which is the complementary of the cumulative distribution function (CDF)).

To obtain the theoretic bounds (Eq. (14)), we gather for each traffic type a trace from one single source and calculate the extremising values s_0, t_0, using the *msa* software developed by Courcoubetis *et al.*[22], available at http://www.ics.forth.gr/netgroup/msa/, with the appropriate buffer and capacity values (*(N+1)C and (N+1)B for N sources*). We suppose that the total mean rate of the *N* sources is less than the link capacity.

On the following figures, we plot the probability $P(jitter > x)$ obtained by the simulation and noted by *Psimul* in the case where jitter is defined as variations between consecutive packets and *Psimul_abs* in the case where the variations are taken relatively to the minimum delay incurred by any packet on the connection. We also plot three approximations of the bound $Le^{-s_0 NB}$ (Eq. (14)). The first approximation, *Psimple*, is obtained with $L = e^{N\ln(M_{t_o}(s_0)) - (N+1)Cs_0t_0}$, i.e., *Psimple* is equal to $e^{-NI_{t_o,s_o}(\frac{N+1}{N}C,B)}$; the second approximation *B-R-approx* is obtained with $L = e^{N\ln(M_{t_o}(s_0)) - (N+1)Cs_0t_0 - \frac{1}{2}\log(4\pi NI_{t_o,s_o}(\frac{N+1}{N}C,B))}$, i.e., using the Bahadur-Rao improvement and the approximation in[19] and the last one, *B-R*, is obtained with $L = e^{N\ln(M_{t_o}(s_0)) - (N+1)Cs_0t_0 - \frac{1}{2}\log(2\pi N\sigma^2 s_0^2)}$. Figures 6-10 show the probability $P(jitter > x)$ as a function of *x* and its various upper bounds for different traffic types with various number of connections.

Firstly, we can notice that both the definitions of delay variation as variation between consecutive packets or relatively to the minimum delay incurred on the same connection yield approximately the same real jitter overflow probabilities. In the case of the traffic type used for the two previous figures, the bounds are tighter when *N* is small compared to greater values of *N*. Also in the absolute, the bounds become more precise for greater values of the jitter. Finally, note that for small values of the jitter, the *B-R* and *B-R-approx*

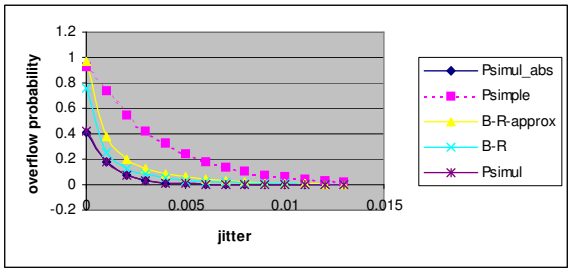

Fig. 6: Jitter overflow probability with "application 3" traffic, 25 connections and 2 Mbps link

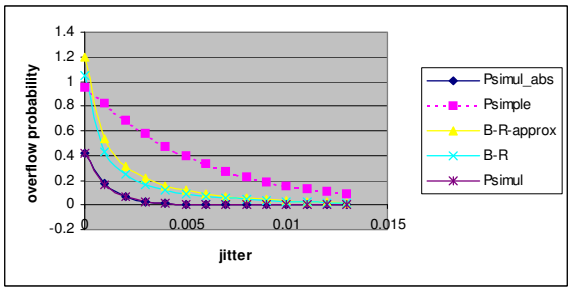

Fig. 7: Jitter probability with "application 3" traffic, 40 connections and 2 Mbps link

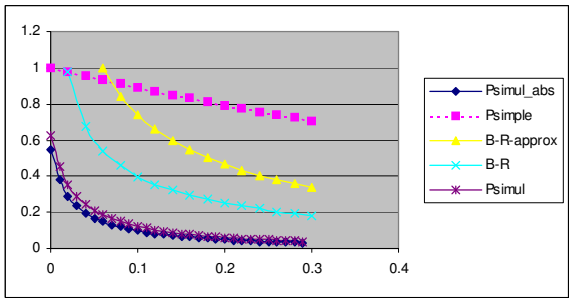

Fig. 8: Jitter probability with "application 1" traffic, 8 connections and 2 Mbps link

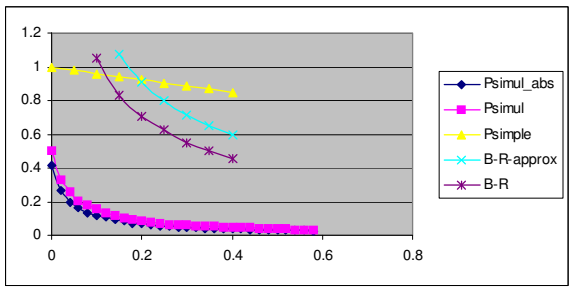

Fig. 9: Jitter probability with "application 1" traffic, 50 connections and 2 Mbps link

bounds are not very useful as they produce figures greater than one. In this case, it is better to use the simple bound.

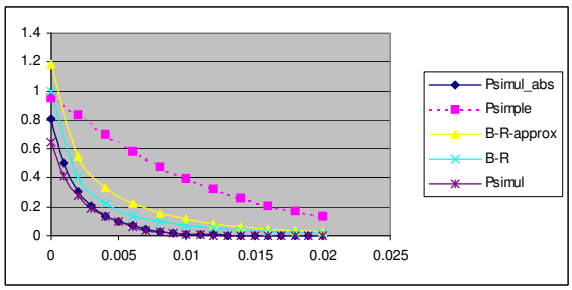

Fig. 10: Jitter probability with "application 4" traffic, 5 connections and 2 Mbps link

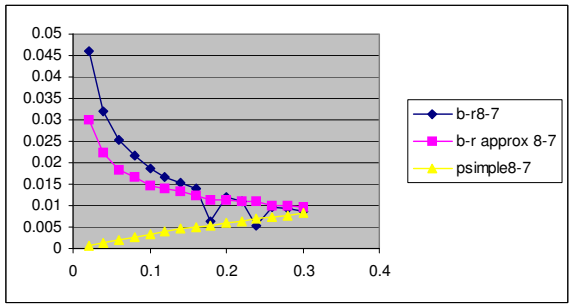

Fig. 11: Absolute improvement on the bound "delay-bound" for "application 1" traffic with 8 connections and 2 Mbps link

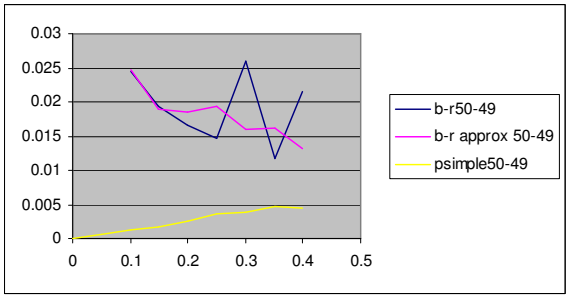

Fig. 12: Absolute improvement on the bound "delay-bound" for "application 1" traffic with 50 connections and 2 Mbps link

These last figures confirm the fact that our bounds are better when *N* is small.

Another comparison is to assess the improvement of our bounds compared to a simple bound *delay-bound* obtained by considering the jitter like a delay (*NC and NB for N sources*). Figures 11-13 show the absolute improvement obtained using ours bounds for the applications 1 and 3 traffics.

In terms of absolute improvement, our bounds seem not to be efficient enough. However, when converted to relative improvement, the usage of our bounds can bring improvement up to 10% as shown in Fig. 14. The best relative improvement is almost always

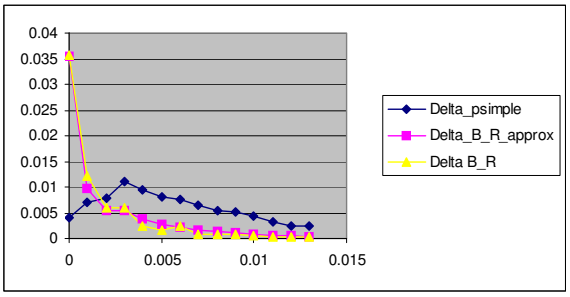

Fig. 13: Absolute improvement on the bound "delay-bound" for "application 3"traffic with 25 connections and 2 Mbps link

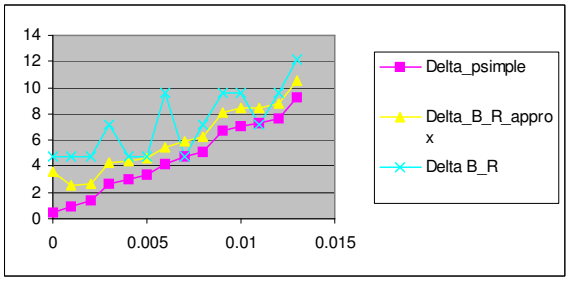

Fig. 14: Relative improvement on the bound "delay-bound" for "application 3" traffic with 25 connections and 2 Mbps link

obtained with the Bahadur-Rao bound and the improvement is more noticeable when *N* increases.

In the process of evaluating the theoretic bounds, we need to determine the epochs granularity used for the trace file, this influences the determination of the effective bandwidth by the *msa* software and also the determination of the extremum value of t[22]. A tradeoff should be obtained between the level of details of the source trace file (trace file should not be too big) and the accuracy of the results obtained (which directly depends on the level of details). When t is small, the traffic details are well captured and the algorithm for solving Equation (5) spans a large domain of t. The results obtained are thus more precise. Also, a priori, fast time scales are important to delay variation. Therefore, a good empiric value for the level of details of the trace is 1 or 2 ms. Figure 15 shows the bounds for the voice traffic for 2 time granularities.

The bounds obtained with finer time granularity are better than those with coarse time granularity for big jitter values. Note also that the bounds obtained are very loose. This is due to the fact that *N* is very high. Recall that we have already noticed that our bound is tighter for smaller *N*. This behavior is due to the fact that the theoretic bounds consider the worst case where

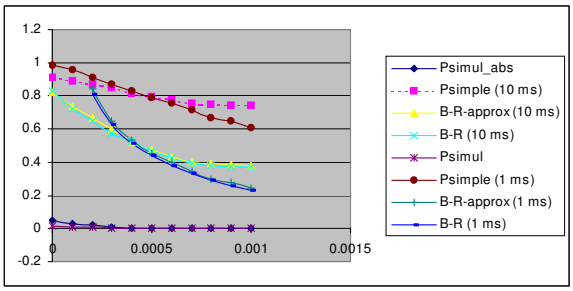

Fig. 15: Bounds for the "application 2"traffic with 120 connections and 2 Mbps link and two different time granularities

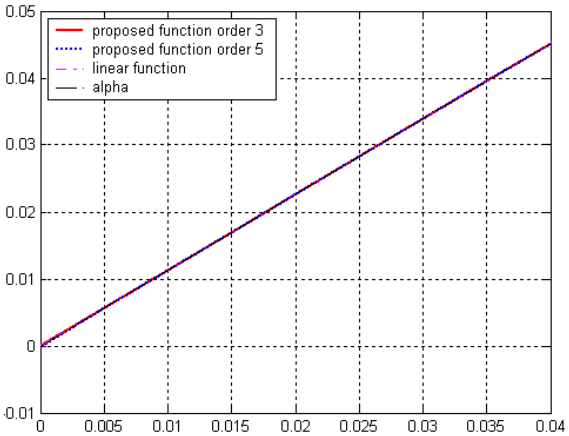

Fig. 16: Charging functions for "application 2" and 0 byte buffer

all *N* sources inject data in the buffer. However, in practice when *N* becomes larger, only a smaller proportion of the sources are active at the same time, yielding less important delay variations.

Another validity condition of the jitter bounds is that the observations used by the *msa* software should be made over a long enough period.

Finally, we test the proposed charging scheme. Using the Matlab ®, we plot on the same graph, the upper bound of the effective bandwidth denoted *alpha*, the linear pricing function and our proposed charging function. The upper bound used is the on-off upper bound introduced in[25]. As already said, the calculation of the charging function depends only on the measured mean rate. As the proposed charging function can be easily generalized, we plot it both at degree 3 and 5. Figures 16-18 show the results for applications 2 and 4 with various buffer sizes.

For small buffers, there is no difference between the charging functions. However, as the buffer size is increased, our charging function reflects more precisely

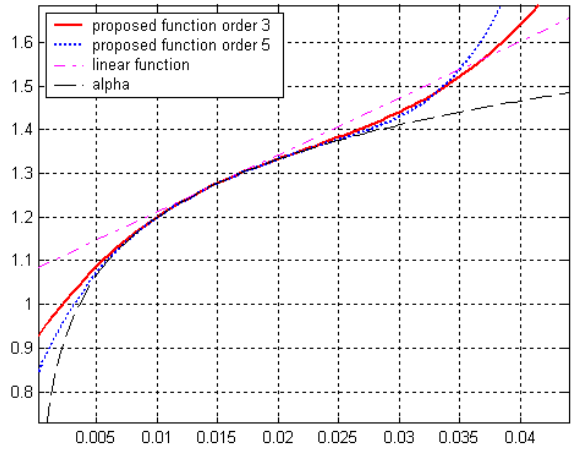

Fig. 17: Charging functions for "application 2" and 2000 bytes buffer

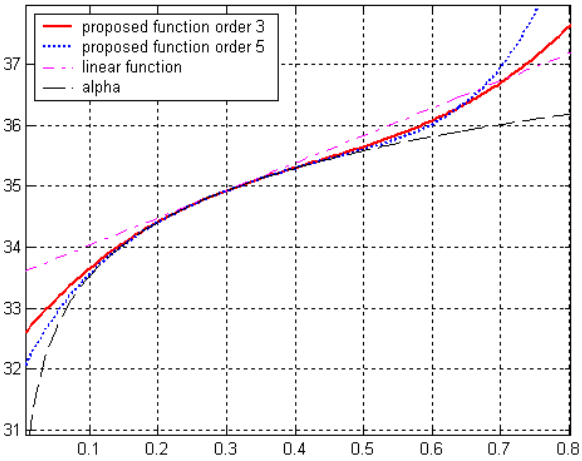

Fig. 18: Charging functions for "application 4" and 5000 bytes buffer

the bound on the effective bandwidth within a certain neighborhood as defined in Lemma 3. Finally, using a higher degree charging function does not yield a noticeable improvement. However, as the increase in complexity due to a higher degree charging function is minimal, the operator can determine its trade-offs.

CONCLUSION

In this paper, we analyzed the pricing framework of multiservice networks and proposed an improved pricing scheme based on the effective bandwidth concept and taking into account more quality of service parameters. After introducing some basic theory, we analysed the effective bandwidth pricing scheme. Based on the deficiencies of the classical scheme (intolerance to user uncertainty and no guarantee on jitter), we proposed an improved charging function which gives more flexibility to the user and we introduced an additional constraint for taking into account an eventual guarantee on the jitter or delay variation. We also extended the effective bandwidth pricing scheme to the case with guaranteed jitter.

Our proposed charging function is shown to improve on the classic effective bandwidth scheme. It remains simple as it requires that the network monitors only the average rate and duration of each connection. It is also fairer than the classic effective bandwidth scheme as "uncertain" but reasonable users are less penalized. Moreover, the incentive to an efficient use of network resource is preserved and the connection admission control is broadened to cases with guaranteed jitter. The constraint on the guaranteed jitter was tested and proved to be viable.

As further investigation, one can study the integration of this pricing scheme in the global design process of integrated networks[31] especially from the management point of view. For example, one can consider various thresholds on the income to recover costs and allow future extension of the network. It would also be interesting to test the jitter guarantee constraint with real network traffic.

Appendix: In this section, we present the relevant proofs of the lemmas 1 and 2 of Section 3 and of the concavity of $\overline{Q}_{g(X)}(m)$.

Proof of Lemma 1: Suppose the necessary regularity conditions, the Taylor expansion of $\overline{\alpha}(\overline{Q}_{g(X)}(m_r)) = \overline{\alpha}(Y_r)$ in the neighbourhood of $Y = \overline{Q}_{g(X)}(m)$ is

$$\overline{\alpha}(Y_r) = \overline{\alpha}(\overline{Q}_{g(X)}(m)) + \frac{1}{st\overline{Q}_{g(X)}(m)}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))$$

$$- \frac{1}{2st(\overline{Q}_{g(X)}(m))^2}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^2$$

$$+ \frac{1}{3st(\overline{Q}_{g(X)}(m))^3}(\overline{Q}_{g(X)}(m_r) - \overline{Q}_{g(X)}(m))^3$$

$$- \frac{1}{4st(\overline{Q}_{g(X)}(m))^4}(\overline{Q}_{g(X)}(m^*) - \overline{Q}_{g(X)}(m))^4$$

where $\overline{Q}_{g(X)}(m) \le \overline{Q}_{g(X)}(m^*) \le \overline{Q}_{g(X)}(m_r)$

$$s > 0,\ t > 0 \Rightarrow \overline{\alpha}(Y_r) - h_{\overline{Q},m}(m_r) =$$

$$- \frac{1}{4st(\overline{Q}_{g(X)}(m))^4}(\overline{Q}_{g(X)}(m^*) - \overline{Q}_{g(X)}(m))^4 \le 0, \forall m, \forall m_r$$

Thus $h_{\overline{Q},m}(m_r) \geq \overline{\alpha}(\overline{Q}_{g(X)}(m_r))$ and the equality is reached if $m_r = m$. Thus the user has an incentive to declare his/her real mean rate.

Proof of Lemma 2: Let's define $Y_r = \overline{Q}_{g(X)}(m_r)$ and $Y = \overline{Q}_{g(X)}(m)$.

$$\frac{\partial h_{\overline{Q},m}}{\partial Y_r} = \frac{1}{stY} - \frac{1}{stY^2}(Y_r - Y) + \frac{1}{stY^3}(Y_r - Y)^2$$

$$= \frac{1}{st}\left(\frac{{Y_r}^2 - 3YY_r + 3Y^2}{Y^3}\right) = \frac{1}{st}\left(\frac{(Y_r - \frac{3}{2}Y)^2 + \frac{3}{4}Y^2}{Y^3}\right)$$

As *s, t and Y* are positive, then $\frac{\partial h_{\overline{Q},m}}{\partial Y_r} \geq 0$ and $h_{\overline{Q},m}(m_r)$ is increasing in $\overline{Q}_{g(X)}(m_r)$.

To prove that $\overline{Q}_{g(X)}(m_r)$ is increasing in m_r, let's consider two real m_1 and m_2 such that $m_1 \leq m_2$. Let consider $X^S = \sup_{X \in \chi(h)} X$. We can suppose that X^S exists (i.e. is finite), otherwise it means that the policing $\chi(h)$ is inefficient but the demonstration remains similar. Suppose that this supremum X^S is attained (if not attained, we can find a function X^a arbitrary close to X^S which will play a role similar to X^S in this proof). By definition of X^S and $g(X)$, we have $m_S = E(g(X)) = g(X^S) = \sup_{X \in \chi(h)} g(X)$.

Therefore, for $m_1, m_2 \in [0, m_s]$, we have $\overline{Q}_{g(X)}(m_1) = \sup_{X:E(g(X))=m_1; X \in \chi(h)} \left\{E[e^{sX[0,t]}]\right\}$ and $\overline{Q}_{g(X)}(m_2) = \sup_{X:E(g(X))=m_2; X \in \chi(h)} \left\{E[e^{sX[0,t]}]\right\}$. For any function X_1 such that $E(g(X_1)) = m_1; X_1 \in \chi(h)$, we can find a function X_2 so that $E(g(X_2)) = m_2; X_2 \in \chi(h)$. This function X_2 is constructed by augmenting the function X_1 wherever $X_1 < X^S$ while respecting the constraint $X_2 < X^S$ until $E(g(X_2)) = m_2$. The resulting function X_2 is necessarily an element of $\chi(h)$ as it corresponds to a load less than the maximum allowed X^S. Moreover, we have by construction:

$X_2 \geq X_1 \Rightarrow E[e^{sX_2[0,t]}] \geq E[e^{sX_1[0,t]}]$ (as the expectation operator is essentially a time expectation operator).

Consequently by taking the supremum of both parts of the inequality, we find that $\overline{Q}_{g(X)}(m_1) \geq \overline{Q}_{g(X)}(m_2)$. And then $h_{\overline{Q},m}(m_r)$ is increasing in m_r.

Next, we give the proof of the concavity of $\overline{Q}_{g(X)}(m)$, which is used in the proof of Lemma 3.

Proof of the concavity of $\overline{Q}_{g(X)}(m)$: Suppose $X, Y \in \chi(h)$ and $E(g(X))=m_1$, $E(g(Y))=m_2$. Let Z be X or Y with probabilities θ and $1-\theta$ respectively, where $0 < \theta < 1$. This corresponds to the practical circumstance of being unsure of the type of a connection. Then

$E(g(Z)) = \theta E(g(X)) + (1-\theta)E(g(Y)) = \theta m_1 + (1-\theta)m_2$

So

$\overline{Q}_{g(X)}(\theta m_1 + (1-\theta)m_2) \geq E[e^{sZ[0,t]}]$ by definition of $\overline{Q}_{g(X)}(m)$

$= \theta E[e^{sX[0,t]}] + (1-\theta)E[e^{sY[0,t]}]$

Since this holds for all *X[0,t]* and *Y[0,t]* satisfying the constraints, we have after maximizing the right hand size:

$\overline{Q}_{g(X)}(\theta m_1 + (1-\theta)m_2) \geq \theta \overline{Q}_{g(X)}(m_1) + (1-\theta)\overline{Q}_{g(X)}(m_2)$

Thus, $\overline{Q}_{g(X)}(m)$ is concave in *m*.

REFERENCES

1. Lee, D.C., S.J. Park and J.S. Song, 2000. Performance analysis of queuing strategies for multiple priority calls in multiservice personal communications services. Comp. Commun., 23: 1069-1083.
2. Li, V.O.K., W. Liao, X. Qiu and E.W.M. Wong, 1996. Performance model of interactive video-on-demand systems. IEEE J. Selected Areas in Commun., 14: 6.
3. DaSilva, L.A., 2000. Pricing for QoS-enabled networks: A survey. IEEE Commun. Surveys and Tutorials, 3: 14-20.
4. Falkner, M., M. Devetsikiotis and I. Lambadaris, 2000. An overview of pricing concepts for broadband IP networks. IEEE Commun. Surveys and Tutorials, 3: 2.
5. Gupta, A. and S. Kalyanaraman, 2003. A Two-Component Spot Pricing Framework for Loss-Rate Guaranteed Internet Service Contracts. Proc. 35th Winter Simulation Conf., S. Chick, P.J. Sanchez, D. Ferrin and D. J. Morrice, (Eds.), New Orleans, Louisiana.

6. Li, T., Y. Iraqi and R. Boutaba, 2004. Pricing and admission control for QoS-enabled Internet. Computer Networks, Special issue on Internet economics: Pricing and Policies, 46: 87-110.
7. McKnight, L.W. and J.P. Bailey, 1997. Introduction to Internet Economics. In: Internet Economics, (Eds. McKnight, L.W. and J.P. Bailey), Cambridge, Mass: MIT Press.
8. Courcoubetis, C., F. Kelly and R. Weber, 2000. Measurement-based usage charges in communications networks. Oper. Res., 48: 535-548.
9. Courcoubetis, C., F.P. Kelly, V.A. Siris and R. Weber, 2000. A study of simple usage-based charging schemes for broadband networks. Telecommun. Sys., 15: 323-343.
10. Kelly, F.P., 1997. Charging and Accounting for Bursty Connections. In: Internet Economics, (L.W. McKnight and J. P. Bailey, Eds.), Cambridge, Mass: MIT Press.
11. Lindberger, K., 1997. Cost-based charging principles in ATM networks. Teletraffic Contributions for the Information Age. Proc. 15th Intl. Teletraffic Cong., ITC 15: 771-780.
12. Siris, V.A., D.J. Songhurst, G.D. Stamoulis and M. Stoer, 1999. Usage-based charging using effective bandwidths: Studies and reality. Proc. 16th Intl. Teletraffic Cong., (ITC-16), Edinburgh, UK, pp: 929-940.
13. Pechiar, J., G. Perera and M. Simon, 2002. Effective bandwidth estimation and testing for Markov sources. Performance Evaluation J., 48: 157–175.
14. Aspirot, L., P. Belzarena, P. Bermolen, A. Ferragut, G. Perera and M. Simon, 2005. Quality of service parameters and link operating point estimation based on effective bandwidths. Performance Evaluation J., 59: 103-120.
15. Ghaderi, M., R. Boutaba and G.W. Kenward, 2005. Stochastic admission control for quality of service in wireless packet networks. Networking 2005: Proc. 4th Intl. IFIP-TC6 Networking Conf., Waterloo, Canada, pp: 1309-1320.
16. Ghaderi, M. and R. Boutaba, 2006. Call Admission control for voice/data integration in broadband wireless networks. IEEE Trans. Mobile Comp., 5: 193-207.
17. Likhanov, N. and R.R. Mazumdar, 1998. Cell loss asymptotics in buffers fed with a large number of independent stationary sources. Proc., IEEE INFOCOM '98, pp: 339-346.
18. Courcoubetis, C. and R. Weber, 1996. Buffer overflow asymptotics for a buffer handling many traffic sources. J. Appl. Probab., 33: 886-903.
19. Montgomery, M. and G. De Veciana, 1996. On the relevance of time scales in performance oriented traffic characterizations. Proc. IEEE INFOCOM '96, pp: 513-520.
20. Hui, J.Y., 1988. Resource allocation for broadband networks. IEEE J. Selected Areas in Commun., 6: 1598-1608.
21. Guerin, R., H. Ahmadi and M. Naghshineh, 1991. Equivalent capacity and its application to bandwidth allocation in high-speed networks. IEEE J. Selected Areas in Commun., 9: 968-981.
22. Courcoubetis, C., V.A. Siris and G.D. Stamoulis, 1999. Application of the many sources asymptotic and effective bandwidths to traffic engineering. Telecommunication Systems- Modeling, Analysis, Design and Management, 12: 167-191.
23. Kalmanek, C.R., H. Kanakia and S. Keshav, 1990. Rate controlled servers for very high-speed networks", GLOBECOM '90: IEEE Global Telecommun. Conf. Exhib., pp: 12-20.
24. Elwalid, A.I. and D. Mitra, 1993. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. IEEE/ACM Trans. Networking, 1: 329-343.
25. Siris, V.A., 1997. Performance Analysis and Pricing in Broadband Networks, University of Crete, Heraklion, Crete.
26. 3GPP Technical Specification Group Services and System Aspects, 3GPP TS 23.107 v5.3.0 (2002-01) QoS Concept and Architecture, 2002.
27. Ferrari, D., 1990. Client requirements for real-time communication services. IEEE Commun. Mag., 28: 65-72.
28. Korpeoglu, I., S.K. Tripathi and X. Chen, 1998. Estimating end-to-end cell delay variation in ATM networks. Intl. Conf. Commun. Technol. Proc., pp: 472-483.
29. Holma, H. and A. Toskala, 2000. WCDMA for UMTS: Radio Access for Third Generation Mobile Communications. Chichester, England, New York, Wiley.
30. UMTS Forum. UMTS Forum Web Page. 2002.
31. Gavish, B. and S. Sridhar, 1995. Economic aspects of configuring cellular networks. Wireless Networks, 1: 115-128.