Available online at www.sciencedirect.com
SCIENCE DIRECT®
Performance Evaluation 59 (2005) 103–120
PERFORMANCE EVALUATION
An International Journal
www.elsevier.com/locate/peva

Quality of service parameters and link operating point estimation based on effective bandwidths

Laura Aspirot*, Pablo Belzarena, Paola Bermolen, Andrés Ferragut, Gonzalo Perera, María Simon

Grupo ARTES (Análisis de Redes, Tráfico y Estadística de Servicios), Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay

Available online 11 September 2004

Abstract

This work addresses the estimation and calculation of the *operating point* of a network's link in a digital traffic network. The notion of operating point comes from effective bandwidth (EB) theory. The results shown are valid for a wide range of traffic types. We show that, given a good EB estimator, the operating point, i.e. the values of time and space parameters in which the EB is related with the asymptotic overflow probability, can also be accurately estimated. This means that the operating point (and other parameters) inherits the statistical properties of the EB estimation. This affirmation is not an obvious one, because operating point parameters are related with the EB through an implicit function involving extremal conditions computations.

Imposing some regularity conditions, a consistent estimator and confidence regions for the operating point and Quality of Service parameters are developed. These conditions are very general, and they are met by commonly used estimators as the averaging estimator presented in [C. Courcoubetis, R. Weber, Buffer overflow asymptotics for a switch handling many traffic sources, J. Appl. Probability 33 (1996)] or the Markov Fluid model estimator presented in [J. Pechiar, G. Perera, M. Simon, Effective bandwidth estimation and testing for Markov sources, Perform. Eval. 48 (2002) 157–175].

Using a software package developed by our group that estimates the EB and other relevant parameters from traffic traces, simulation results are compared with the analytical results, showing very good fitting.

Keywords: Traffic modelling; Effective bandwidth; QoS parameter estimation; MPLS networks

* Corresponding author.
E-mail address: aspirot@fing.edu.uy (L. Aspirot).

0166-5316/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.peva.2004.07.006

1. Introduction

The usage of digital networks for carrying variable bit rate (VBR) and real time (RT) or time sensitive services is growing. New control mechanisms and protocols are added to existing data oriented networks to give an appropriate support to such services. The state of the art in traffic engineering is briefly described in Section 2.

Resource sharing in these networks is absolutely needed for an economic usage. This issue leads to the problem of estimating the resources needed for guaranteed VBR communications, which cannot be the peak rate nor the mean rate. Indeed, the mean rate would be a too optimistic estimation, that would cause frequent losses. On the other side, the peak rate would be too pessimistic and would lead to resource waste.

Effective bandwidth (EB) defined by F. Kelly in [7] is an useful and appropriate measure of channel occupancy. The EB is defined as follows:

$$\alpha(s,t) = \frac{1}{st}\log \boldsymbol{E}(\mathrm{e}^{sX_t}), \quad 0 < s, t < \infty.$$

where X_t is the total amount of work arriving from a source in the time interval $[0, t]$, which is supposed to be a stochastic process with stationary increments. $\alpha(s, t)$ lies between the mean rate (for $s \to 0$) and the peak rate (for $s \to \infty$) of the input process.

Parameters s and t are referred to as the space and time parameters, respectively. These parameters depend not only on the source itself, but on the context on which this source is involved. More specifically, s and t depend on the capacity, buffer size, scheduling policy of the multiplexer, the QoS parameter to be achieved and the actual traffic mix (i.e. characteristics and number of other sources). The EB concept can be applied to sources or to aggregated traffic, as we find in a network's core link.

Under the *many sources asymptotic regime* discussed in [3], where it is assumed that the link capacity and buffer size increase proportionally to the number of incoming sources, the EB is related with the stationary buffer overflow probability by the so called *inf sup* formula:

$$\Gamma = \inf_{t\geq 0} \sup_{s\geq 0}((B + Ct)s - Nst\alpha(s,t)),$$

where C is the link capacity, B is its buffer size and N the number of incoming multiplexed sources of effective bandwidth $\alpha(s, t)$. If Q_N represents the stationary amount of work in the queue, the buffer overflow probability or loss probability is approximately given by

$$\log \boldsymbol{P}(Q_N > B) \approx -\Gamma$$

as shown in [3] and [16].

We call s^* and t^* to the values of parameters s and t in which the *inf sup* is attained. These values s^* and t^* are called the link's *operating point*. A good estimation of s^* and t^* is useful for network design and routing procedures.

The technical relevance of the issue is pointed out in Section 2, where we briefly present the traffic engineering in Multiprotocol Label Switching (MPLS) networks. We point out the need of a good estimation of the bandwidth in order to optimise resource sharing.

In Section 3 we summarise the different equivalent bandwidth approaches to network performance estimation.

In Section 4 we show how the operating point of a link can be estimated, the consistency of this estimation and its confidence interval.

Our group has developed a software package that estimates the EB from general traffic traces, calculates it when an explicit model is given, estimates the values of s^* and t^* and deduces then the overflow or loss probability and other relevant design parameters like buffer size and link capacity. Using this tool EB and operating point of several traces were obtained, and their dispersion was estimated. Analytical results are compared with numerical data in Section 5.

In Section 6 we present several conclusions and indicate ways for further work.

2. Motivation

Convergence of the different telecommunications services on a unique network is an already old aspiration. Integrated Service Digital Network (ISDN) has been a goal for the Telecommunications community from long time ago. General functionalities for broadband networks were defined, as in [5] or [6]. Nevertheless, many candidates have failed to implement such integrated network. However, the research and proposals lead to some basic design principles and measurement methods that prevail and apply to the new proposals. One of the most important of them, which is a central point in this work, is the concept of Quality of Service (QoS), and how to define, establish and guarantee a QoS level.

Currently, the most promising approach to an integrated service network are the already old IP networks. And maybe the most attractive new service to be given is voice, the most traditional telecommunication service. But in this case, the service and the network come from different worlds. In fact, services as voice are often given on IP platforms, but without any guarantee of quality. This is because plain IP networks are based in a “best effort” policy, well suited for data but not for time sensitive applications. In order to support variable bit rate (VBR) and time sensitive services, the concept of QoS must be implemented over IP.

The first serious approach was the IntServ (Integrated Services) model. This model proposed access control for each individual flow, and resource reservation along the full path across the network. The drawback is the lack of scalability; the model cannot be applied in the Internet core, where the number of individual flows grows high.

During the last years, the most promising architecture proposal is Differentiated Services (DiffServ). This model goes around the scalability problems of IntServ by working with aggregate flows. Data packets are classified, and each node manages each class according to appropriate policies.

However, the DiffServ model does not guarantee by itself QoS on the Internet. Traffic Engineering is also needed. Some authors as [10], have proposed architectures integrating DiffServ and admission control at the network edges based on traffic classes.

More recently, the MPLS architecture appeared as an appropriate technology where the DiffServ model can be implemented. This architecture incorporates traffic engineering through explicit routing establishing tunnels named label switched paths (LSPs). MPLS also introduces the notion of forwarding equivalence class (FEC), giving the network operator the possibility of partitioning the traffic in aggregated flows according to the service model.

Using MPLS, the network operator can establish for each FEC one or more LSPs. But, given a LSPs configuration defined by the operator: is it possible to ensure the service level required? And if it is not possible: how can the operator use traffic engineering to give the required QoS for each aggregated flow?

The main goal of this work is to give some theoretical insight about traffic engineering in MPLS, and develop a practical tool for network design and performance optimisation topics. We are especially interested in traffic engineering methodologies based on statistical characterisation of different flows.

In this framework, overflow probability estimation is a key topic, which leads to the notion of an "equivalent" bandwidth required by the aggregated flows.

3. Equivalent Bandwidth applications in network analysis

The notion of equivalent bandwidth was formerly used to study the access control of some networks, for instance ATM networks. Many contributions following this approach were done during the 1990s to analyse the access control in some networks based on the IntServ model. In that situation, the access node receives a connection request and has to estimate the resources it requires, in order to allow or deny the new connection. Kelly's Effective Bandwidth (EB) may be used in such situations as the "equivalent capacity" needed by the new connection. In this context, the flow to be statistically characterised is an individual flow, and may be directly related with the data source (for instance voice or video codes).

The situation of access node-individual flows was studied using the so called *large buffer asymptotics*, in which we take an infinite buffer and study its filling above some large threshold. This approach cannot be used in backbone nodes, where buffers are devised to resolve simultaneous packet arrival, but not to store bursts, and they are consequently small.

The application of large deviations theory to the analysis of the MPLS backbone must be performed on the basis of the *many sources asymptotic*. In this regime we take buffer size $B = Nb$ (N being the number of sources) and output capacity $C = Nc$ and make N go to infinity. Results about loss probability and delay in this regime can be found in [3,16,15]. Recently, a different asymptotic with many sources and small buffer characteristics was proposed in [11]. In all these works, EB is related with the relevant QoS parameters through the notion of operating point of the link as mentioned in 1. Estimating the EB and QoS accordingly is the main goal of the next section.

Depending on technical possibilities, these estimations can be used for off-line design, admission control (at the LSP aggregated level) or even for routing. These applications are successively more demanding in computing capacity of the switches. At present time we focus primarily on design issues.

Even though this work was motivated by MPLS traffic engineering, its results are useful for a general network that handles VBR communications and has to provide a fixed or at least consistent QoS. It is valid more generally, for problems of limited resource sharing in which some guarantees of loss and delay are intended to be met.

4. Estimation

Estimating the operating point of a link, as defined in Section 1 is closely related with its defining equation which we rewrite here on a *per source* basis:

$$\gamma = \inf_{t\geq 0} \sup_{s\geq 0}((b + ct)s - st\alpha(s, t)) \tag{4.1}$$

where γ is the asymptotic decay rate of the overflow probability as the number of sources increases, c and b are the link's capacity and buffer size *per source* and $\alpha(s, t)$ the effective bandwidth function of the incoming traffic, also defined in Section 1:

$$\alpha(s, t) = \frac{1}{st} \log \boldsymbol{E}(\mathrm{e}^{sX_t}), \quad 0 < s, t < \infty. \tag{4.2}$$

With the present notation, stationary overflow probability in a switch multiplexing N sources, having capacity $C = Nc$ and buffer size $B = Nb$ verifies

$$\lim_{N\to\infty} \frac{1}{N} \log \boldsymbol{P}(Q_N > B) = -\gamma. \tag{4.3}$$

In general, the effective bandwidth function $\alpha(s, t)$ is unknown, and shall be estimated from measured traffic traces. The problem is how to estimate the moment generating function $\Lambda(s, t) = \boldsymbol{E}(\mathrm{e}^{sX_t})$ of the incoming traffic process X_t for each s and t.

Different approaches have been presented to solve this problem. One of them, presented in [2] and [13] is to estimate the expectation $\boldsymbol{E}(\mathrm{e}^{sX_t})$ as the time average given by

$$\Lambda_n(s, t) = \frac{1}{n} \sum_{k=1}^{n} \mathrm{e}^{s(X_{kt} - X_{(k-1)t})} \tag{4.4}$$

which is valid if the process increments are stationary and satisfy any weak dependence hypothesis that guarantees ergodicity. To estimate $\Lambda(s, t)$ a traffic trace of length $T = nt$ is needed. We can construct an appropriate estimator of the EB as $\alpha_n(s, t) = \frac{1}{st} \log(\Lambda_n(s, t))$.

When a model is available for incoming traffic, a parametric approach can be taken. In the case of a Markov Fluid model, i.e. when the incoming process is modulated by a continuous time Markov chain which dictates the rate of incoming work, explicit computation can be made as shown by Kesidis et al. in [8]. In this case, an explicit formula is given for $\Lambda(s, t)$ and $\alpha(s, t)$ in terms of the infinitesimal generator or Q-matrix of the Markov chain. In a previous work of our group [12], and based on the maximum likelihood estimators of the Q-matrix parameters presented in [9], an EB estimator and confidence intervals are developed.

Having an estimator of the function $\alpha(s, t)$, it seems natural to estimate γ, and the operating point s^*, t^* substituting the function $\alpha(s, t)$ by $\alpha_n(s, t)$ in Eq. (4.1) and solving the remaining optimisation problem. The output would be some values of γ_n, s_n^* and t_n^*, and the question is under what conditions these values are good estimators of the real γ, s^* and t^*.

Therefore, we may discuss two different problems concerning estimation. The first one is, given a "good" estimator $\alpha_n(s, t)$ of $\alpha(s, t)$, find sufficient conditions under which the estimators s_n^*, t_n^* and γ_n^* obtained by solving the optimisation problem:

$$\gamma_n = \inf_{t\geq 0} \sup_{s\geq 0}((b + ct)s - st\alpha_n(s, t)) \tag{4.5}$$

are "good" estimators of the operating point s^*, t^* and the overflow probability decay rate γ of a link. This affirmation is not an obvious result because s^* and t^* are found from a non linear and implicit function. We remark that the reasoning applied to s^* and t^* can be also applied to other parameters that are deduced from the EB. Further in the article the parameters B and C are also studied.

The second problem is finding this good estimator of the EB and determining whether the conditions are met, so that the operating point can be estimated using Eq. (4.5).

The remaining part of the section addresses the first problem, where a complete answer concerning consistency and Central Limit Theorem (CLT) properties of estimators is given by theorem 1, based on regularity conditions of the EB function. At the end of the section we discuss the validity of the theorem for some known estimators and in Section 5 we compare our analytical results with numerical ones.

Let us define

$$g(s,t) = s(b+ct) - st\alpha(s,t)$$

which can be rewritten in terms of $\Lambda(s,t) = \boldsymbol{E}(\mathrm{e}^{sX_t})$ as

$$g(s,t) = s(b+ct) - \log(\Lambda(s,t))$$

Then we have that $(\partial/\partial s)g(s,t) = 0$ if and only if

$$\frac{\partial}{\partial s}g(s,t) = b + ct - \frac{\partial/\partial s\Lambda(s,t)}{\Lambda(s,t)} = 0 \tag{4.6}$$

Assuming that for each t there exists $s(t)$ such that

$$\frac{\partial}{\partial s}g(s(t),t) = 0,$$

it is easy to show that $\sup_{s\geq 0} g(s,t) = g(s(t),t)$ because $g(s,t)$ is convex as a function of s. In that case, $\gamma = \inf_{t\geq 0} g(s(t),t)$, and

$$\frac{\partial}{\partial t}g(s(t),t) = \frac{\partial}{\partial s}g(s(t),t)\dot{s}(t) + \left.\frac{\partial}{\partial t}g(s,t)\right|_{s=s(t)}.$$

If there exists t^* such that

$$\frac{\partial}{\partial t}g(s(t^*),t^*) = 0$$

and the infimum is attained, it follows that

$$\gamma = g(s(t^*),t^*).$$

If we define $s^* = s(t^*)$, we have that $\gamma = g(s^*,t^*)$ where

$$\frac{\partial}{\partial s}g(s^*,t^*)\,\dot{s}(t^*) + \frac{\partial}{\partial t}g(s^*,t^*) = 0,$$

$$\frac{\partial}{\partial s}g(s^*,t^*) = 0$$

and then we have the relations

$$\frac{\partial}{\partial s} g(s^*, t^*) = \frac{\partial}{\partial t} g(s^*, t^*) = 0 \tag{4.7}$$

Since

$$\frac{\partial}{\partial t} g(s, t) = cs - \frac{(\partial/\partial t)\Lambda(s, t)}{\Lambda(s, t)}, \tag{4.8}$$

it follows from (4.6), (4.7) and (4.8) that the operating point must satisfy the equations:

$$b + ct^* - \frac{(\partial/\partial s)\Lambda(s^*, t^*)}{\Lambda(s^*, t^*)} = 0, \tag{4.9a}$$

$$cs^* - \frac{(\partial/\partial t)\Lambda(s^*, t^*)}{\Lambda(s^*, t^*)} = 0. \tag{4.9b}$$

If we make the additional assumptions that interchanging the order of the differential and expectation operators is valid, and that $\dot{X}_t$ exists for almost every t we can write

$$\frac{\partial}{\partial s}\Lambda(s, t) = \boldsymbol{E}(X_t \mathrm{e}^{sX_t}), \quad \frac{\partial}{\partial t}\Lambda(s, t) = \boldsymbol{E}(s\dot{X}_t \mathrm{e}^{sX_t}) \tag{4.10}$$

Replacing the expressions of (4.10) in Eqs. (4.9) we deduce an alternative expression for the solutions s^* and t^*:

$$b + ct^* - \frac{\boldsymbol{E}(X_{t^*} \mathrm{e}^{s^* X_{t^*}})}{\boldsymbol{E}(\mathrm{e}^{s^* X_{t^*}})} = 0, \tag{4.11a}$$

$$cs^* - \frac{\boldsymbol{E}(s^* \dot{X}_{t^*} \mathrm{e}^{s^* X_{t^*}})}{\boldsymbol{E}(\mathrm{e}^{s^* X_{t^*}})} = 0. \tag{4.11b}$$

Therefore, we can reformulate the optimisation problem presented in (4.1). The operating point of the link can be calculated solving the system of equations (4.9a), or (4.10) if the additional assumptions are valid. The first formulation, which is more general, is the one used in the main result of this work, which follows:

Theorem 1. *If $\Lambda_n(s, t)$ is an estimator of $\Lambda(s, t)$ such that both are C^1 functions and*

$$\Lambda_n(s, t) \underset{n}{\longrightarrow} \Lambda(s, t), \tag{4.12a}$$

$$\frac{\partial}{\partial s}\Lambda_n(s, t) \underset{n}{\longrightarrow} \frac{\partial}{\partial s}\Lambda(s, t), \tag{4.12b}$$

$$\frac{\partial}{\partial t}\Lambda_n(s, t) \underset{n}{\longrightarrow} \frac{\partial}{\partial t}\Lambda(s, t) \tag{4.12c}$$

almost surely and uniformly over bounded intervals, and if we denote s_n^* and t_n^* the solutions of

$$b + ct_n^* - \frac{(\partial/\partial s)\Lambda_n(s_n^*, t_n^*)}{\Lambda_n(s_n^*, t_n^*)} = 0, \tag{4.13a}$$

$$cs_n^* - \frac{(\partial/\partial t)\Lambda_n(s_n^*, t_n^*)}{\Lambda_n(s_n^*, t_n^*)} = 0, \tag{4.13b}$$

then (s_n^*, t_n^*) *are consistent estimators of* (s^*, t^*). *Moreover, if a functional Central Limit Theorem (CLT) applies to* $\Lambda_n - \Lambda$, *i.e.,*

$$\sqrt{n}(\Lambda_n(s,t) - \Lambda(s,t)) \underset{n}{\overset{w}{\Longrightarrow}} G(s,t),$$

where $G(s,t)$ *is a continuous Gaussian process, then*

$$\sqrt{n}((s_n^*, t_n^*) - (s,t)) \underset{n}{\overset{w}{\Longrightarrow}} N(\vec{0}, \Sigma), \tag{4.14}$$

where $N(\vec{0}, \Sigma)$ *is a centered bivariate normal distribution with covariance matrix* Σ.

Proof. From Eq. (4.9a), we know that (s^*, t^*) is the solution of the equation

$$K((s,t), \Lambda) = \vec{0},$$

where

$$K((s,t), \Lambda) = \begin{pmatrix} b + ct - \dfrac{(\partial/\partial s)\Lambda(s,t)}{\Lambda(s,t)} \\ cs - \dfrac{(\partial/\partial t)\Lambda(s,t)}{\Lambda(s,t)} \end{pmatrix} \tag{4.15}$$

and (s_n^*, t_n^*) is the solution of $K((s_n, t_n), \Lambda_n) = \vec{0}$.

Recall that if $E,\ F$ are normed spaces (or more in general semi-normed spaces), $f : E \to F$ is said to be differentiable at $e \in E$ if there exists a continuous linear map $df(e) : E \to F$ such that for any e' in a neighborhood at e we have that

$$f(e') = f(e) + df(e)(e' - e) + \mathrm{o}(\|e' - e\|) \tag{4.16}$$

Let also recall that if $E_1,\ E_2,\ F$ are normed spaces and $f : E_1 \times E_2 \to F$ is differentiable so are, for any $e_1 \in E_1$ and $e_2 \in E_2$, $f_{e_1} : E_2 \to F$ defined by $f_{e_1}(e_2) = f(e_1, e_2)$ end $f^{e_2} : E_1 \to F$ defined by $f^{e_2}(e_1) = f(e_1, e_2)$.

Finally, let us recall the general form of the implicit function derivative formula. If f is differentiable, and for each $e_2 \in E_2$ there exists an unique element $v(e_2)$ of E_1 such that $f(v(e_2), e_2) = \vec{0}$ and $df^{e_2}(v(e_2))$ is invertible, then

$$df^{e_2}(v(e_2))\, dv(e_2) + df_{v(e_2)}(e_2) = 0,$$

or equivalently

$$dv(e_2) = -\left(df^{e_2}(v(e_2))\right)^{-1} df_{v(e_2)}(e_2) \tag{4.17}$$

We will apply (4.17) to $E_1 = (\mathbb{R}^+)^2$ equipped with the Euclidean norm and

$$E_2 = \{f : (\mathbb{R}^+)^2 \to \mathbb{R} \text{ of class } C^1\}$$

equipped with the seminorm

$$\|f\| = \sum_{n=1}^{\infty} \frac{1}{2^n}\left(\frac{\|f\|_n}{1+\|f\|_n}\right),$$

where $\|f\|_n$ is the Sobolev-type norm

$$\sup\left\{|f(s,t)| + \left|\frac{\partial}{\partial s} f(s,t)\right| + \left|\frac{\partial}{\partial t} f(s,t)\right| : (s,t) \in [0,n]^2\right\}.$$

It is easy to check that, for a sequence $\{f_m\}$ of functions of E_2, $\|f_m\| \to 0$ when $m \to \infty$ if and only if $f_m \to 0$, $(\partial/\partial s) f_m \to 0$, $(\partial/\partial t) f_m \to 0$ uniformly over bounded intervals when $m \to \infty$.

Applying the preceding paragraph, we may think of (s^*, t^*) as $(s,t)(\Lambda) = v(\Lambda)$ and by (4.17)

$$dv(\Lambda) = -(dK^{\Lambda}(v(\Lambda)))^{-1} dK_{v(\Lambda)}(\Lambda).$$

Therefore, since $(s_n^*, t_n^*) = v(\Lambda_n)$,

$$(s_n^*, t_n^*) - (s,t) = v(\Lambda_n) - v(\Lambda) = dv(\Lambda)(\Lambda_n - \Lambda) + \mathrm{o}(\|\Lambda_n - \Lambda\|). \tag{4.18}$$

But the hypotheses of the theorem imply that $\|\Lambda_n - \Lambda\| \longrightarrow_n 0$ a.s. and $\mathrm{d}v(\Lambda)$ is continuous, therefore:

$$(s_n^*, t_n^*) \underset{n}{\longrightarrow} (s^*, t^*) \text{ a.s.}$$

showing consistency.

If in addition, a functional CLT applies to $\Lambda_n - \Lambda$, i.e.,

$$\sqrt{n}(\Lambda_n(s,t) - \Lambda(s,t)) \underset{n}{\overset{w}{\Longrightarrow}} G(s,t),$$

where $G(s,t)$ is a C^1 Gaussian process, we have that

$$\sqrt{n}((s_n^*, t_n^*) - (s,t)) = dv(\Lambda)(\sqrt{n}(\Lambda_n - \Lambda)) + \mathrm{o}(\sqrt{n}\|\Lambda_n - \Lambda\|).$$

Since $\sqrt{n}\|\Lambda_n - \Lambda\| \Longrightarrow_n^w \|G\|$, it is bounded in probability and, therefore,

$$o(\sqrt{n}\|\Lambda_n - \Lambda\|) \underset{n}{\overset{(p)}{\longrightarrow}} 0.$$

On the other hand, since $\mathrm{d}v(\Lambda)$ is continuous, we have that

$$\sqrt{n}\, \mathrm{d}v(\Lambda)(\Lambda_n - \Lambda) \underset{n}{\overset{w}{\Longrightarrow}} \mathrm{d}v(\Lambda)(G).$$

Therefore, we have finally shown that

$$\sqrt{n}((s_n^*, t_n^*) - (s,t)) \underset{n}{\overset{w}{\Longrightarrow}} \mathrm{d}v(\Lambda)(G)$$

which, being $\mathrm{d}v(\Lambda)$ a linear transformation into $\mathbb{R}^2$, implies that $\mathrm{d}v(\Lambda)(G)$ is a normal $N(\vec{0}, \Sigma)$ bivariate random variable, where Σ may be computed in terms of the covariances of G and the transformation $\mathrm{d}v(\Lambda)$. □

Remark 1. The method used in the precedent proof is commonly known as the δ-method. It is utilised to obtain CLT results for differentiable functionals of asymptotically Gaussian processes. An appropriate topology must be chosen in order to guarantee differentiability. Not every functional of an asymptotically gaussian process is asymptotically gaussian. Consider for instance a sequence of bounded *iid* random variables and F_n its empirical distribution sequence. Then, $\sqrt{n}(F_n - F)$ is asymptotically a gaussian process on the real line, but $\sqrt{n}(T(F_n) - T(F))$ is not gaussian when $T(F) = x_F = \inf\{t : F(t) = 1\}$. Indeed, the asymptotic distribution of x_F is a max-stable distribution.

Remark 2. As can be seen from the proof, computation of Σ may not be trivial. However, if replication is possible (for instance by taking large traces of weak-dependent signals), the previous result allows the estimation of Σ in terms of empirical covariances. Arguments of this type are used in Section 5.

Remark 3. The above theorem was stated for the operating point. However, if we define γ_n by

$$\gamma_n = s_n^*(b + ct_n^*) - \Lambda_n(s_n^*, t_n^*),$$

we have that $\gamma = F(s^*, t^*, \Lambda)$ where F is a differentiable function, and $\gamma_n = F(s_n^*, t_n^*, \Lambda_n) = F(v(\Lambda_n), \Lambda_n)$. Therefore, if the estimator Λ_n verifies a functional CLT we have for γ_n

$$\sqrt{n}(\gamma_n - \gamma) \overset{w}{\underset{n}{\Longrightarrow}} N(0, \sigma^2).$$

Moreover, in a many sources environment, expressions for the buffer size b and the link capacity c obtained by Courcoubetis [2] are similar to the *inf sup* equation. Therefore, the reasoning used in the previous theorem extends consistency and CLT results to b^* and c^*. Also, confidence intervals for these design parameters can be constructed in this way as studied in Section 5.

We address now the second question posed at the beginning of the section. As we can see, for the validity of theorem 1 it is necessary that the estimator $\Lambda_n(s, t)$ converge uniformly to the moment generating function over bounded intervals, as well as its partial derivatives. These conditions are reasonably general, and it can be verified that they are met by the estimator (4.4) presented in [2] and [13], and by the estimator for Markov Fluid sources presented in [12]. In both cases a CLT can be obtained so the CLT conclusion of the theorem is also valid. It should be noticed that a consistent but non-smooth estimator can be used with this procedure, if it is previously regularized by convolution with a suitable kernel.

5. Simulation and numerical results

5.1. Introduction

As written before, in the many sources regime the loss probability could be estimated from the solution of the *inf sup* formula (4.1). To solve this equation a double optimisation (in time and space parameters) is needed, in order to obtain the link operating point (s^*, t^*). The first problem is that in real cases, when is not assumed a model for the source, there is not an explicit formula for the EB. In the general case the information available is from traffic traces, and the equation (4.1) must be solved in terms of $\alpha_n(s, t)$ (an EB estimator) instead of $\alpha(s, t)$. From the previous section we know that the link operating point estimation obtained from a good estimation $\alpha_n(s, t)$ is consistent and has CLT properties. In this work we carry out the analysis with simulated traces from a known theoretical model in order to evaluate our results. In the what follows we explain the model and the EB estimation. After that an estimator of the link operating point will be obtained, and it will be used to calculate the QoS parameters and some link design parameters.

5.2. EB estimation

To validate the results obtained in the previous section, we simulated traffic using a two state (ON-OFF) Markov Fluid model. In that model, a continuous time Markov chain drives the process. When the chain is in the ON state, the workload is produced at constant rate h_0, and when it is in the OFF state no workload is produced ($h_1 = 0$). Denoting by Q the Markov chain infinitesimal generator, by $\vec{\pi}$, its invariant distribution, and by H, the diagonal matrix with the rates h_i in the diagonal. The effective bandwidth for a source of this type is [8,7]:

$$\alpha(s, t) = \frac{1}{st} \log\{\vec{\pi} \exp[(Q + Hs)t]\vec{\mathbf{1}}\}, \tag{5.19}$$

where $\vec{\mathbf{1}}$ is a column vector of ones.

In our simulations we generated 300 traffic traces of length T samples, with the following Q-matrix:

$$Q = \begin{pmatrix} -0.02 & 0.02 \\ 0.1 & -0.1 \end{pmatrix}$$

The effective bandwidth for this process calculated through Eq. (5.19) is shown in Fig. 1.

For each traffic trace we estimated EB using the following procedure. We divided the trace in blocks of length t and constructed the following sequence:

$$\tilde{X}_k = \sum_{i=(k-1)t}^{kt} x(i), \quad 1 \le k \le \lfloor T/t \rfloor,$$

where $x(i)$ is the amount of traffic arrived between samples and $\lfloor c \rfloor$ denotes the largest integer less than or equal to c.

EB can then be estimated by the time average proposed in [2,13] as

$$\alpha_n(s, t) = \frac{1}{st} \log\left[\frac{1}{\lfloor T/t \rfloor} \sum_{j=1}^{\lfloor T/t \rfloor} \mathrm{e}^{s\tilde{X}_j}\right], \tag{5.20}$$

where $n = \lfloor T/t \rfloor$. This is merely an implementation of the time average estimator in Eq. (4.4) based on a finite length traffic trace. When the values of t verify that $t \ll T$, the number of replications of the increment process within the trace is good enough to get a good estimation.

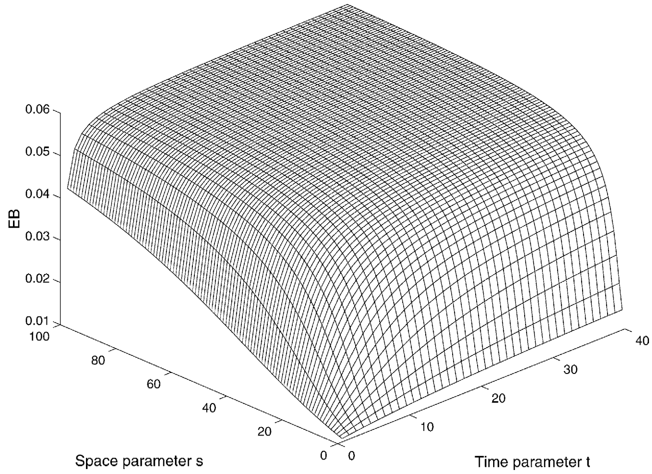

Fig. 1. Effective bandwidth of a Markov fluid source.

In order to find the operating point (s^*, t^*) of the theoretical Markov model, and its estimator (s_n^*, t_n^*) for each simulated trace, we solve the *inf sup* optimisation problem of Eq. (4.1). In our case $\alpha(s, t)$ will be the previous theoretical Eq. (5.19) for the Markovian source or the $\alpha_n(s, t)$ estimated for each trace. The numerical solution has two parts. First, for a fixed t we find the $s^*(t)$ that maximise $g(s, t)$ as a function of s. It can be shown that $st\alpha(s, t)$ is a convex function of s. This convexity property is used to solve the previous optimisation problem, that is reduced to find the maximum difference between a convex function and a linear function of s, and it can be done very efficiently. After the $s^*(t)$ is found for each t, it is necessary to minimise the function $g(s^*(t), t)$ and find t^*. For this second optimisation problem, there are no general properties that let us make the search algorithm efficient and a linear searching strategy is used.

One of the goals is to develop a confidence region for (s^*, t^*). We simulated 300 traces of length $100000(T)$ samples and constructed, for each simulated trace indexed by $i = 1, \dots, K$ the corresponding estimator $(s_n^*(i), t_n^*(i))$. By Theorem 1 the vector $\sqrt{n}((s_n^*, t_n^*) - (s^*, t^*))$ is asymptotically bivariate normal with $(0, 0)$ mean and covariance matrix Σ. We estimated the matrix Σ using the empirical covariances of the observations

$$\{\sqrt{n}((s_n^*(i), t_n^*(i)) - (s^*(i), t^*(i)))\}_{i=1,\dots,K}$$

given by

$$\Sigma_K = \frac{n}{K}\begin{pmatrix} \sum_{i=1}^{K}\left(s_n^*(i) - \overline{s_n^*}\right)^2 & \sum_{i=1}^{K}\left(s_n^*(i) - \overline{s_n^*}\right)\left(t_n^*(i) - \overline{t_n^*}\right) \\ \sum_{i=1}^{K}\left(s_n^*(i) - \overline{s_n^*}\right)\left(t_n^*(i) - \overline{t_n^*}\right) & \sum_{i=1}^{K}\left(t_n^*(i) - \overline{t_n^*}\right)^2 \end{pmatrix},$$

where $\overline{s_n^*} = (1/K)\sum_{i=1}^{K} s_n^*(i)$ and $\overline{t_n^*} = (1/K)\sum_{i=1}^{K} t_n^*(i)$.

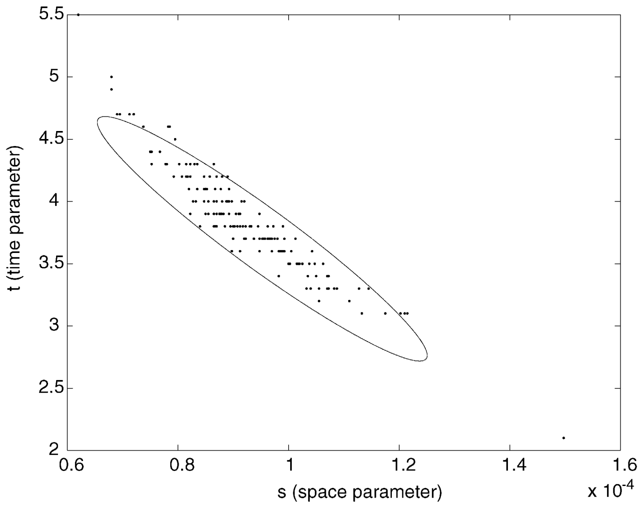

Fig. 2. Estimated operating points and confidence region.

Therefore, we can say that approximately

$$(s_n^*, t_n^*) \approx N\left((s^*, t^*), \frac{1}{n}\Sigma_K\right)$$

from where a level α confidence region can be obtained as

$$R_\alpha = (s_n^*, t_n^*) + \frac{A_K^t B(\vec{0}, \sqrt{\chi_\alpha^2(2)})}{\sqrt{n}}$$

being A_K the matrix that verifies $A_K^t A_K = \Sigma_K$, while $B(x, r)$ is the ball of center x and radius r.

To verify our results, we calculated the theoretical operating point (s^*, t^*) and simulated another 300 traces independent of those that were used to estimate Σ_K. We constructed then the 95% confidence region. If the results are right, approximately 95% of the times, (s^*, t^*) must fall inside that region, or equivalently and easier to check, approximately 95% of the simulated (s_n^*, t_n^*) must fall inside the region $R = (s^*, t^*) + 1/\sqrt{n} A_K^t B(\vec{0}, \sqrt{\chi_{0.05}^2(2)})$. Numerical results, plotted in Fig. 2, verify that the confidence level is attained, 95.33% of the estimated values fall inside the predicted region.

5.3. QoS parameters estimation

We estimate the link operating point in order to estimate loss probability and other QoS parameters, such as delay. In the many sources asymptotic regime, the real delay of packets that flows through a link coincides with its virtual delay [15]. The virtual delay is the delay value obtained through the queue size. If the link sends C packets per unit of time and the probability of having a queue size larger than B is q, then the probability of having a delay higher than B/C will be q. In this regime if we have an estimator of the probability of having a queue size larger than B, we have an estimator of the real delay. We will

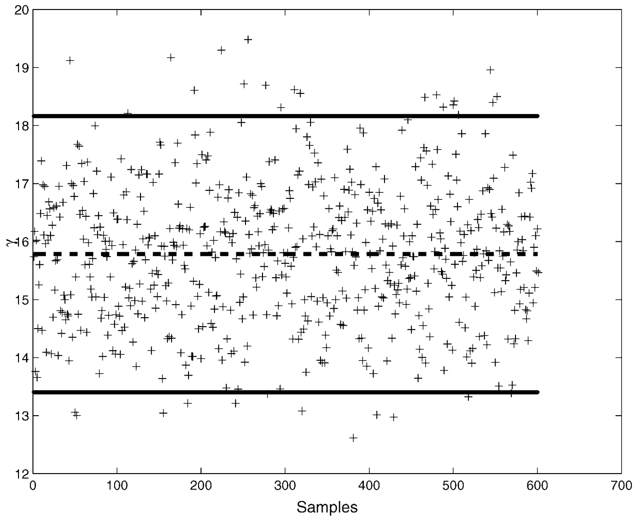

Fig. 3. Estimation of γ_n, theoretical γ and confidence interval.

focus on the estimation of loss probability, because the delay could be deduced from the same equation. As was said in Section 4, if we have an EB estimator that verifies the hypotheses of theorem 1, then

$$\gamma_n = \inf_t \sup_s ((b + ct)s - st\alpha_n(s, t)) \tag{5.21}$$

is a consistent estimator and has CLT properties. From this estimator loss probability could be approximated by

$$q_n = P_n(Q_N > B) \approx \mathrm{e}^{-N\gamma_n}, \tag{5.22}$$

where Q_N is the queue size and N is the number of sources. Fig. 3 shows the estimations of γ_n for 600 simulated traces, its theoretical value and its confidence interval. Numerical results show that in this case 94.8% of the values fall in the 95% confidence interval.

5.4. Link design based on EB estimation

Previous results could be extended to link design, when some QoS requirements are given. The goal is to know, for a certain link, the smallest buffer size when the capacity C, the input traffic traces and the maximum loss probability desired (or the maximum delay) are given. The same reasoning could be done in order to calculate the smallest necessary link capacity to guarantee the desired loss probability when the same information as before is available but the buffer size is fixed. The answers to these design problems are obtained from equations such as the *inf sup* formula. The smallest buffer size to guarantee loss probability γ is given [2] by the following equation:

$$B_n = \sup_t \inf_s (G_n(s, t)), \tag{5.23}$$

$$G_n(s, t) = \frac{(Nst\alpha_n(s, t) + N\gamma)}{s} - Ct, \tag{5.24}$$

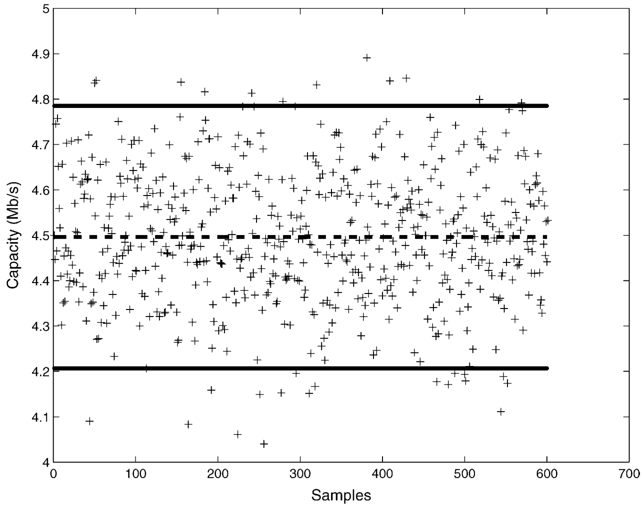

Fig. 4. Estimated capacity, theoretical capacity and confidence interval.

and the smallest capacity to guarantee loss probability γ is

$$C_n = \sup_t \inf_s (K_n(s,t)), \tag{5.25}$$

$$K_n(s,t) = \frac{(Nst\alpha_n(s,t) + N\gamma)}{st} - \frac{B}{t}. \tag{5.26}$$

In Figs. 4 and 5 smallest capacity and buffer size estimations are shown. For each one of the 600 simulated traces, B and C have been estimated using the previous equations. The theoretical values and the confidence intervals are also indicated.

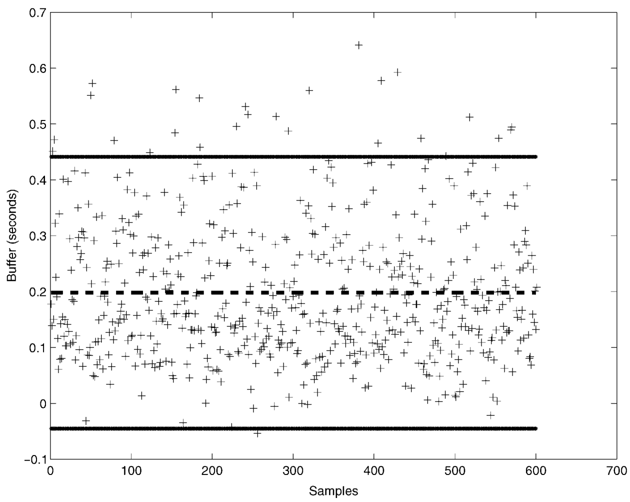

Fig. 5. Estimated buffer size, theoretical buffer size and confidence interval.

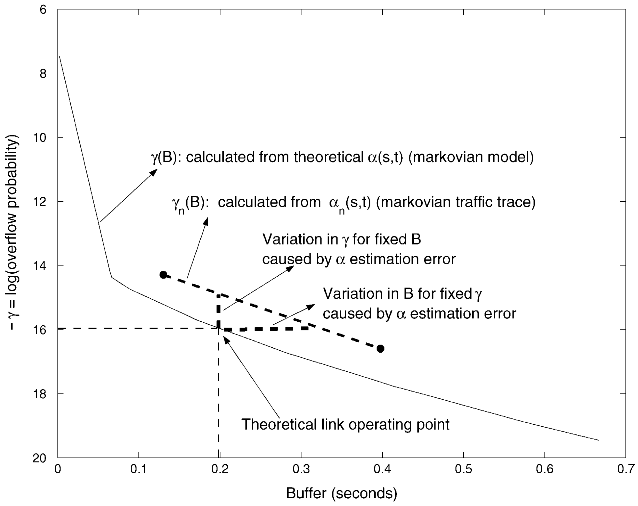

Fig. 6. $-\gamma$ variation vs. buffer size.

Numerical results for the capacity verify that the confidence level is attained, 95% of the estimated values fall inside the predicted interval. Negative values of buffer size for some traces show only that no buffer is needed to satisfy the desired QoS requirements.

Notice that buffer size has an important variation. This fact is related with the operating point of the link under design. In Fig. 6 we plot $-\gamma$ versus buffer size. In this curve there are two distinguishable zones. The first one is for low buffer sizes, where small changes in buffer size leads to important changes in loss probability. The second one shows that to have important changes in loss probability large changes in buffer size are needed. In our case the link is operating in the second zone. Solid and dotted curves

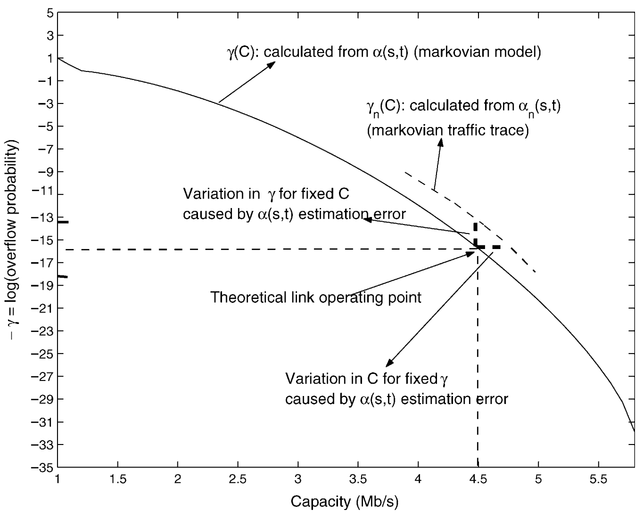

Fig. 7. $-\gamma$ variation vs. link capacity.

correspond to theoretical α and estimated α_n respectively. With fixed B, γ has little variation when we move from one of these curves to the other. However, if γ is fixed B has large variation.

The variation of γ with C can be studied as before. In Fig. 7 this variation is shown. The curve slope grows rapidly as C increases, and loss probability goes to zero (so $-\gamma \to -\infty$) when C goes to the peak rate. In this case the link capacity is 75% of the input traffic peak rate. γ has little variation when C is fixed. On the other side, if γ is fixed, C has little variation. This fact explains the small variation of C in Fig. 4.

Our group has developed a software package that estimates the effective bandwidth of a source from traffic traces by means of different estimators and finds the operating point, as well as different QoS parameters of a link using our previous framework. The Java source code of this software is available upon request to the authors.

6. Conclusions

We have shown that consistency and CLT properties of effective bandwidth estimators can be extended to the operating point estimation through a natural procedure under very general hypothesis. This kind of estimations are necessary in the analysis and design of networks that must guarantee some degree of QoS to the traffic they carry, in order to support time sensitive services, and to optimise the resource usage. This is an important problem in the development of modern networks, where convergence of different services is a main goal.

We have also checked that the numerical estimations constructed by simulation fit very well with the theoretical predictions.

Moreover, the same kind of asymptotic relations that are dealt here with, appear in the estimation of some other QoS parameters, besides the loss probability. In particular we analyze minimum buffer size and capacity estimation that guarantee link operating with a bounded loss probability.

We are also working in the extension of this results, as well as the software tools, to the case of a complete network.

Acknowledgements

The authors wish to thank our colleagues Darío Buschiazzo and Alejandro Vázquez for their careful review of this work.

References

[1] D. Awduche, J. Malcolm, Requirements for Traffic Engineering Over MPLS, Internet Engineering Task Force RFC 2702, 1999.

[2] C. Courcoubetis, V.A. Siris, Procedures and tools for analysis of network traffic measurements, Perform. Eval. 48 (2002) 5–23.

[3] C. Courcoubetis, R. Weber, Buffer overflow asymptotics for a switch handling many traffic sources, J. Appl. Probability 33 (1996).

[4] A. Dembo, O. Zeitouni, Large Deviations Techniques and its Applications, Jones and Bartlett, New York, 1993.
[5] ITU Recommendation I.211, General Service Aspects of B-ISDN, ITU-T, Geneva.
[6] ITU Recommendation I.311, B-ISDN General Network Aspects, ITU-T, Geneva.
[7] F. Kelly, Notes on effective bandwidth, in: F. Kelly, S. Zachary, I. Ziedins (Eds.), Stochastic Networks: Theory and Applications, Oxford University Press, Oxford, 1996.
[8] G. Kesidis, J. Walrand, C.S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, IEEE/ACM Trans. Network. 1 (1993) 424–428.
[9] E.A. Lebedev, L.I. Lukashuk, Maximum likelihood estimation of the infinitesimal matrix of a Markov chain with continuous time (in Russian), Dokl. Akad. Nauk Ukr. SSR Ser. A 1 (1986) 12–14.
[10] H.A. Mantar, J. Hwang, I.T. Okumus, S.J. Chapin, Edge-to-edge resource provisioning and admission control in diffserv networks, IEEE Softcom, 2001.
[11] O. Ozturk, R. Mazumdar, N. Likhanov, Many sources asymptotics in networks with small buffers, Queueing Syst. 46 (1/2) (2004) 129–147.
[12] J. Pechiar, G. Perera, M. Simon, Effective bandwidth estimation and testing for Markov sources, Perform. Eval. 48 (2002) 157–175.
[13] P. Rabinovitch, Statistical estimation of effective bandwidth, M.Sc. thesis, University of Cambridge, 2000.
[14] E. Rosen, A. Viswanathan, Multiprotocol Label Switching Architecture, Internet Engineering Task Force RFC 3031, 2001.
[15] S. Shakkottai, R. Srikant, Many-sources delay asymptotics with applications to priority queues, Queueing Syst. 39 (2001) 183–200.
[16] D. Wischik, Sample path large deviations for queues with many inputs, Ann. Appl. Probability, 11 (2000) 389–404.

ARTES Group is a joint research group of the Electrical Engineering and Mathematics and Statistics Departments of the School of Engineering of the universidad de la República, Montevideo, Uruguay. Some of its members wrote this work and they are briefly presented below:

Pablo Belzarena is Associate Professor and **María Simon** is Professor at the Electrical Engineering Department of the School of Engineering. Pablo Belzarena received his MS degree in 2003 and María Simon is an engineer since 1980 and she is also the Dean of the School of Engineering since 1998.

Gonzalo Perera is Professor at the Mathematics and Statistics Department at the School of Engineering and a Professor at the Mathematics Department of the School of Sciences of the Universidad de la República, Montevideo, Uruguay. He has a PhD in Mathematics and he is member of the Latin American Regional Committee of the Bernoulli Society of Probability and Mathematical Statistics.

Laura Aspirot, Paola Bermolen and **Andrés Ferragut** are Assistants in the Mathematics and Statistics Department of the School of Engineering. Laura Aspirot and Paola Bermolen received their degrees in Mathematics in 2003 and Andrés Ferragut is an engineer since 2004.