1

Notes on Effective Bandwidths

Frank Kelly

University of Cambridge

Abstract

This paper presents a personal view of work to date on effective bandwidths, emphasising the unifying role of the concept: as a summary of the statistical characteristics of sources over different time and space scales; in bounds, limits and approximations for various models of multiplexing under quality of service constraints; and as the basis for simple and robust tariffing and connection acceptance control mechanisms for poorly characterized traffic. The framework assumes only stationarity of sources, and illustrative examples include periodic streams, fractional Brownian input, policed and shaped sources, and deterministic multiplexing.

1 Introduction

Within a broadband network, the usage of a network resource may not be well assessed by a simple count of the number of bits carried. For example, to provide an acceptable performance to bursty sources with tight delay and loss requirements it may be necessary to keep the average utilization of a link below 10%, while for constant rate sources or sources able to accommodate substantial delays it may be possible to push the average utilization well above 90%.

This paper attempts a unified perspective on *effective bandwidth*, a concept that has been developed by several authors over recent years to provide a measure of resource usage which adequately represents the trade-off between sources of different types, taking proper account of their varying statistical characteristics and quality of service requirements. The concept has attracted much attention and some criticism, but in the author's view there is emerging an elegant and powerful mathematical theory with important technological applications. It seems appropriate to describe the paper as a personal view, since there is not

This paper is based on talks given to the INFORMS Telecommunications Conference held in Boca Raton, Florida in March 1995, and to the Royal Statistical Society Research Workshop in Stochastic Networks held at Heriot-Watt University, Edinburgh in August 1995; I am grateful to participants at these meetings for many comments. The figures for the paper were prepared by Damon Wischik, while supported by the Commission of the European Communities under ACTS project AC039. The paper is published in *Stochastic Networks: Theory and Applications* (ed. F.P. Kelly, S. Zachary and I. Ziedins), Volume 4 of *Royal Statistical Society Lecture Notes Series*, Oxford University Press(1996), 141–168. Pointers to related information resources will be available for a period on `http://www.statslab.cam.ac.uk/~frank/` .

yet a generally accepted definition of an effective bandwidth, and since other frameworks for the interpretation of the material are certainly possible.

In Section 2 we present our definition of the effective bandwidth of a source, describe some of its simpler properties, and present a variety of examples. The effective bandwidth of a source depends upon two free parameters, representing a space and time scaling respectively, and, as Gibbens (1996) demonstrates, this dependence provides a convenient tool for the description and analysis of real sources. The appropriate choice of space and time scale will depend upon characteristics of the resource such as its capacity, buffer size, traffic mix and scheduling policy.

In Section 3 we compare and contrast several multiplexing models, and describe how the effective bandwidth provides a measure associated with a source such that a resource can deliver a performance guarantee expressed in terms of loss or delay by limiting the sources served so that their effective bandwidths sum to less than a threshold. Under different models this result may be expressed as a conservative global bound, or as an asymptotic local limit, or as an approximation capable of successive refinements; but the ubiquity of the single functional form, described in Section 2, is striking.

The effective bandwidth of a source depends sensitively upon the statistical properties of the source, yet these properties may not be known with certainty, either to the user responsible for the source or to the network. It is sometimes thought that this limits the applicability of the concept. On the contrary, the concept is central to any understanding of just how well described a source needs to be, and to the discussion, in Section 4, of tariffing and connection acceptance control mechanisms for sources that may be poorly characterized.

Whitt (1993) and de Veciana and Walrand (1995) provide valuable reviews of earlier work on effective bandwidths. The term itself was first used by Gibbens and Hunt (1991) and Kelly (1991) in their investigation of linear acceptance regions for certain buffered resources, although the essential concept for unbuffered resources had been described earlier in the seminal paper of Hui (1988), and a closely related notion was described by Guérin *et al.* (1991).

2 Effective bandwidths

In this Section we define the effective bandwidth associated with a stationary source, describe some of its simpler properties, and illustrate the definition with several contrasting examples.

2.1 Definition

Let $X[0,t]$ be the amount of work that arrives from a source in the interval $[0,t]$. Assume that $X[0,t]$ has stationary increments. Define the *effective bandwidth* of the source to be

$$\alpha(s,t) = \frac{1}{st} \log \mathbb{E}\left[e^{sX[0,t]}\right] \qquad 0 < s,t < \infty. \tag{2.1}$$

2.2 Properties

(i) If $X[0,t]$ has independent increments, then $\alpha(s,t)$ does not depend upon t.

(ii) If there exists a random variable X such that $X[0,t] = Xt$ for $t > 0$, then $\alpha(s,t) = \alpha(st,1)$, and so $\alpha(s,t)$ depends on s,t only through the product st. Otherwise $\alpha(s/t,t)$ is strictly decreasing in t.

(iii) If $X[0,t] = \sum_i X_i[0,t]$, where $\big(X_i[0,t]\big)_i$ are independent, then

$$\alpha(s,t) = \sum_i \alpha_i(s,t). \tag{2.2}$$

(iv) For any fixed value of t, $\alpha(s,t)$ is increasing in s, and lies between the mean and peak of the arrival rate measured over an interval of length t: that is

$$\frac{\mathbb{E}X[0,t]}{t} \le \alpha(s,t) \le \frac{\bar{X}[0,t]}{t} \tag{2.3}$$

where $\bar{X}[0,t]$ is the (possibly infinite) essential supremum

$$\bar{X}[0,t] = \sup\Big\{ x : P\big\{ X[0,t] > x \big\} > 0 \Big\}.$$

The form of $\alpha(s,t)$ near $s = 0$ is determined by the mean, variance and higher moments of $X[0,t]$, while the form of $\alpha(s,t)$ near $s = \infty$ is primarily influenced by the distribution of $X[0,t]$ near its maximum: if $\alpha(s,t)$ is finite for some $s > 0$ then for given t

$$\alpha(s,t) = \frac{1}{t}\mathbb{E}X[0,t] + \frac{s}{2t}\mathrm{Var}\,X[0,t] + \mathrm{o}(s) \quad \text{as } s \to 0 \tag{2.4}$$

while if $\alpha(s,t)$ is bounded above as $s \to \infty$ then for given t

$$\alpha(s,t) = \frac{\bar{X}[0,t]}{t} + \frac{1}{st}\log P\big\{ X[0,t] = \bar{X}[0,t] \big\} + \mathrm{o}\left(\frac{1}{s}\right) \quad \text{as } s \to \infty. \tag{2.5}$$

Write $\alpha(0,t)$ and $\alpha(\infty,t)$ for the lower and upper bounds respectively of the range (2.3); note that the mean rate $\alpha(0,t)$ does not depend on t, since $X[0,t]$ has stationary increments.

The definition (2.1) may be motivated in several ways. The logarithmic moment generating function is naturally associated with the additive property (iii), while the scalings with t and s beget properties (i) and (iv) respectively. Properties (i)–(iv) are straightforward consequences of convexity and of results on moment generating functions – see Chang (1994) for several relevant observations, as well as a discussion of the case, not considered here, when the increments of $X[0,t]$ may be non-stationary. Although $(\alpha(s,t), 0 < s,t < \infty)$ does not in general determine the distribution of $(X[0,t], 0 < t < \infty)$, it follows from the analyticity of the moment generating function (Billingsley 1986, Exercise 26.7) that if $\alpha(s,t)$ is finite for $s = \varepsilon > 0$ then $(\alpha(s,t), 0 < s < \varepsilon)$ determines the distribution of $X[0,t]$ and, further, $\alpha(s,t)$ is infinitely differentiable with respect

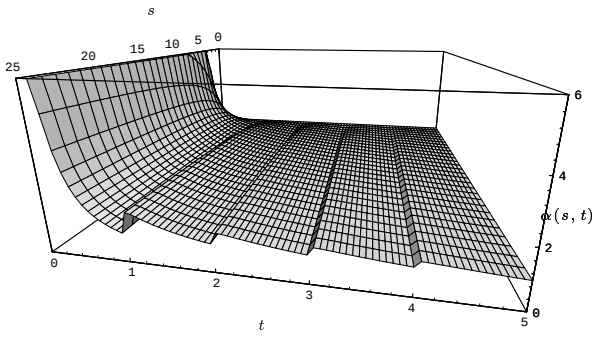

FIG. 1. Effective bandwidth of a periodic source. The source produces a single unit of workload at the end of every unit interval, but the phase of the source is random. Note the growth of the effective bandwidth over intervals shorter than the period of the source.

to s on the interior of the interval on which $\alpha(s,t)$ is finite.

Courcoubetis *et al.* (1995) and Duffield *et al.* (1995) have explored the estimation of $\alpha(s,t)$ for large values of t, and its relation to the tail behaviour of queues. In contrast, Sriram and Whitt (1986) have emphasised the importance of identifying the relevant time scale for queueing phenomena: from relation (2.4)

$$\alpha(s,t) = \alpha(0,t)\Big(1 + \frac{s}{2}I[0,t] + \mathrm{o}(s)\Big) \quad \text{as } s \to 0$$

where $I[0,t] = \mathrm{Var}\, X[0,t]/\mathbb{E}X[0,t]$ is their index of dispersion for counts.

2.3 Examples

2.3.1 *Periodic sources*

For a source which produces b units of workload at times $\{Ud+nd, n=0,1,\ldots\}$, where U is uniformly distributed on the interval $[0,1]$,

$$\alpha(s,t) = \frac{b}{t}\left\lfloor \frac{t}{d} \right\rfloor + \frac{1}{st}\log\left[1 + \left(\frac{t}{d} - \left\lfloor \frac{t}{d} \right\rfloor\right)\left(e^{bs}-1\right)\right]. \tag{2.6}$$

Observe that

$$\lim_{t\to 0}\alpha(s,t) = \frac{e^{bs}-1}{ds};$$

the growth of the effective bandwidth as t decreases is apparent in Fig. 1, which plots the function (2.6) with parameters $b=d=1$. The model has been used to describe the packet streams arising from constant rate information sources: for a review see Roberts (1992, Section 6). We shall consider the model further in Sections 3.5 and 3.6.1.

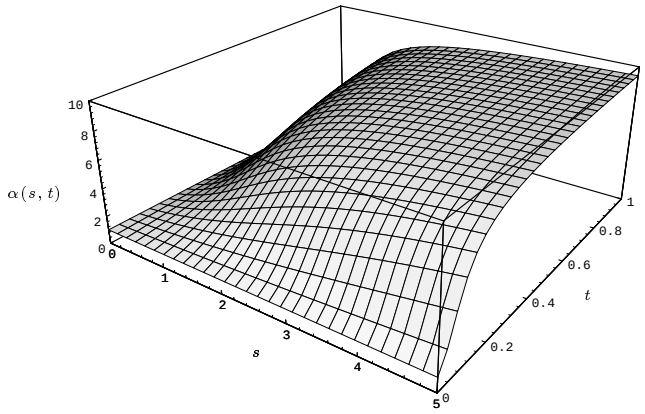

FIG. 2. Effective bandwidth of an on-off fluid source, with parameters $\lambda = 1, \mu = 9, h = 10$. The effective bandwidth $\alpha(s,t)$ approaches the mean rate $\lambda h/(\lambda+\mu)$ as either s or t approaches zero.

2.3.2 *Fluid sources*

Consider a stationary fluid source described by a two-state Markov chain. The transition rate from state 2 to state 1 is λ and the transition rate from state 1 to state 2 is μ. While the Markov chain is in state 1 workload is produced at a constant rate h; while it is in state 2 no workload is produced. Then

$$\alpha(s,t) = \frac{1}{st} \log\left\{ \left(\frac{\lambda}{\lambda+\mu}, \frac{\mu}{\lambda+\mu} \right) \exp\left[\begin{pmatrix} -\mu+hs & \mu \\ \lambda & -\lambda \end{pmatrix} t \right] \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\},$$

and

$$\lim_{t\to\infty} \alpha(s,t) = \frac{1}{2s}\left(hs - \mu - \lambda + \left((hs-\mu+\lambda)^2 + 4\lambda\mu \right)^{\frac{1}{2}} \right),$$

a central expression of Gibbens and Hunt (1991) and Guérin *et al.* (1991), and there obtained from the seminal work on stochastic fluid models of Anick *et al.* (1982). The function $\alpha(s,t)$ is illustrated in Fig. 2.

More generally, consider a stationary source described by a finite Markov chain with stationary distribution $\boldsymbol{\pi}$ and q-matrix Q, where workload is produced at rate h_i while the chain is in state i. Then from the backward equations for the Markov chain one can deduce (Kesidis *et al.* 1993, p.427) that

$$\alpha(s,t) = \frac{1}{st} \log\{ \boldsymbol{\pi} \exp[(Q + \mathbf{h}s)t] \mathbf{1} \} \tag{2.7}$$

where $\mathbf{h} = \mathrm{diag}(h_i)_i$, and

$$\lim_{t\to\infty} \alpha(s,t) = \frac{1}{s}\,\phi(s)$$

where $\phi(s)$ is the largest real eigenvalue of the matrix $Q + \mathbf{h}s$ (Elwalid and Mitra

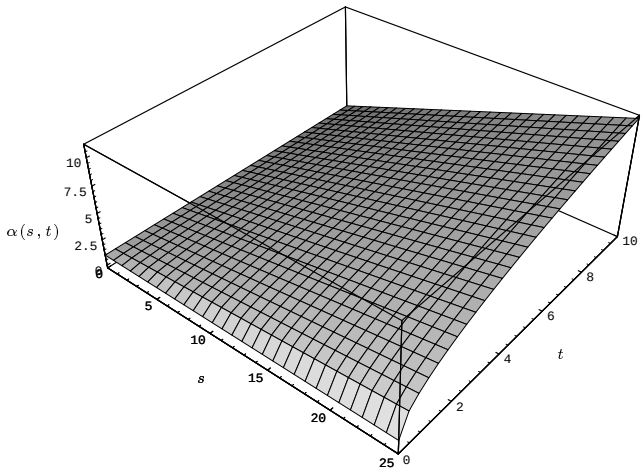

FIG. 3. Effective bandwidth of a Gaussian source. This example has Hurst parameter $H = 0.75$: long range order is indicated by the continued growth of the effective bandwidth with large t.

1993). If $h_1 > h_i, i \neq 1$, then relation (2.5) becomes

$$\alpha(s,t) = h_1 - \frac{1}{s}\left(\mu_1 - \frac{1}{t}\log \pi_1\right) + \mathrm{o}\left(\frac{1}{s}\right) \qquad \text{as } s \to \infty$$

where μ_1 is the transition rate out of the state with peak rate. Chang and Thomas (1995, p.1097) discuss this expansion in the case $t = \infty$: for a fluid source, the relevant limits in s and t commute.

2.3.3 *Gaussian sources*

Suppose that

$$X[0,t] = \lambda t + Z(t)$$

where $Z(t)$ is normally distributed with zero mean; as usual, the facility of calculation under Gaussian assumptions outweighs any problem of interpretation for negative increments. Then

$$\alpha(s,t) = \lambda + \frac{s}{2t}\,\mathrm{Var}\, Z(t),$$

and so $\alpha(s,t)$ is determined, for all s and t, by the first two terms of the expansion (2.4).

The case $\mathrm{Var}\, Z(t) = \sigma^2 t$ commonly arises from heavy traffic models (Harrison 1985). The more general case $\mathrm{Var}\, Z(t) = \sigma^2 t^{2H}$ arises when the process Z is fractional Brownian motion, with Hurst parameter $H \in (0,1)$. Then

$$\alpha(s,t) = \lambda + \frac{\sigma^2 s}{2}\, t^{2H-1}, \tag{2.8}$$

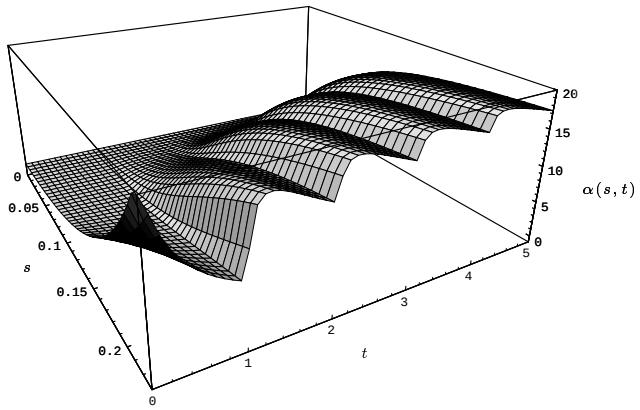

FIG. 4. Effective bandwidth of an on-off periodic source. Note the increase of the effective bandwidth as t either decreases below the period of the source, or increases towards the interval over which the source remains 'on' or 'off'.

and the behaviour of $\alpha(s,t)$ as $t \to \infty$ depends upon whether $H < \frac{1}{2}$, $H = \frac{1}{2}$ or $H > \frac{1}{2}$. Respectively $\lim_{t\to\infty} \alpha(s,t)$ is finite and does not depend upon s, as in example (i); or the limit depends upon s, as in example (ii); or $\alpha(s,t)$ grows as a fractional power of t. The third case exhibits long range order (Norros 1994; Willinger 1995), and has been proposed as a model for Ethernet traffic data (Willinger *et al.* 1995). Fig. 3 illustrates the form (2.8) with parameters $H = 0.75$, $\lambda = 1$ and $\sigma^2 = 0.25$.

Courcoubetis and Weber (1995), Weber (1994) discuss the approximation of the effective bandwidth of an arbitrary stationary source by a Gaussian source with general autocovariance structure, using the first two terms of the expansion (2.4).

We shall consider Gaussian models further in Sections 3.1.3, 3.2.2, 3.3.1, 3.5, 3.6.1 and 3.6.2.

2.3.4 *General on-off sources*

Suppose next that a source alternates between long periods in an 'on' state, where it behaves as a source with effective bandwidth $\alpha_1(s,t)$, and long periods in an 'off' state, where it produces no workload. Let p be the proportion of time spent in the 'on' state. Then for values of t small compared with periods spent in an 'on' or 'off' state,

$$\alpha(s,t) = \frac{1}{st} \log\Big[1 + p\Big(\exp\big(st\alpha_1(s,t)\big) - 1\Big)\Big]. \tag{2.9}$$

Fig. 4 illustrates this function when $p = 0.05$ and $\alpha_1(s,t)$ is given by expression (2.6) with $b = 20, d = 1$.

The above example shares similarities with examples 2.3.1, 2.3.3 or 2.3.2, over short, intermediate or long time scales respectively. Observe that the con-

struction can be endlessly repeated: the 'on' period of a fluid source may at a finer time scale appear as a periodic source, whose bursts may themselves have structure on a still finer time scale, and so on. By variations of this hierarchical construction it is possible to define sources whose effective bandwidth $\alpha(s,t)$ may resemble any or all of Figures 1, 2, 3 and 4, depending on the range of s and t values plotted.

2.4 Lévy processes

A process $X[0,t]$ with stationary independent increments is called a *Lévy process*; as noted in property (i), for such a process $\alpha(s,t)$ does not depend upon t. We have seen one example: the Gaussian source of Section 2.3.3 with $\mathrm{Var}\, Z(t) = \sigma^2 t$. A compound Poisson source will provide another example, and these two cases essentially exhaust the forms that $\alpha(s,t) = \alpha(s)$, say, may take.

2.4.1 *Compound Poisson sources*

If

$$X[0,t] = \sum_{n=1}^{N(t)} Y_n$$

where $Y_1, Y_2, \ldots$ are independent identically distributed random variables with distribution function F, and $N(t)$ is an independent Poisson process of rate ν, then

$$\alpha(s) = \frac{1}{s}\int (e^{sx} - 1)\nu dF(x).$$

For example, if $Y_1, Y_2, \ldots$ are exponentially distributed with parameter μ, then

$$\alpha(s) = \frac{\nu}{\mu - s} \quad \text{for } s < \mu. \tag{2.10}$$

2.4.2 *Infinitely divisible sources*

If $X[0,t]$ has stationary independent increments, then $X[0,1]$ is infinitely divisible. Hence, by the Lévy–Khinchin representation of any infinitely divisible random variable as the limit of a mixture of compound Poisson random variables (Feller 1971, Chapter XVII),

$$\alpha(s) = \lambda + \frac{\sigma^2 s}{2} + \frac{1}{s}\int_{-\infty}^{+\infty} (e^{sx} - 1)d\nu(x) \tag{2.11}$$

is the most general form possible for $\alpha(s)$, where $\nu(\cdot)$ is a measure on $(-\infty, \infty)$.

If, in addition, the increments of $X[0,t]$ are non-negative, then the most general form possible for $\alpha(s)$ is

$$\alpha(s) = \lambda + \frac{1}{s}\int_{0}^{\infty} (e^{sx} - 1)d\nu(x)$$

where $\nu(\cdot)$ is a measure on $(0, \infty)$: it follows that $\alpha(s)$ is convex, and indeed that

all derivatives are positive. For example, if $\lambda = 0$ and $d\nu(x) = x^{-1}e^{-x}dx$, then

$$\alpha(s) = -\frac{1}{s}\log(1-s)$$

and $X[0,t]$ is a gamma process, with increments distributed as gamma random variables (Kingman 1993, Chapter 8,9). If $\int_0^\infty d\nu(x) = \infty$, as for the gamma process, then jumps of $X[0,t]$ are everywhere dense. Jumps of height greater than δ form a Poisson process of rate $\int_\delta^\infty d\nu(x)$.

2.5 Policing and shaping

Say that a stationary source is *policed by parameters* (ρ, β) if

$$\bar{X}[0,t] \le \rho t + \beta \qquad 0 < t < \infty, \tag{2.12}$$

or, equivalently,

$$\alpha(\infty, t) \le \rho + \frac{\beta}{t} \qquad 0 < t < \infty.$$

The parameters ρ and $(\beta - 1)/\rho$ are the *peak cell rate* and *cell delay variation tolerance* of ITU Recommendation I371 (1994). For example, the periodic source described in Section 2.3.1 is policed by parameters (ρ, β) provided $\rho \ge \frac{b}{d}$ and $\beta \ge b$. A fluid source, as described in Section 2.3.2, is policed by parameters (ρ, β) provided $\rho \ge \max_i\{h_i\}$.

Sources which do not satisfy constraint (2.12) may be *shaped* to do so, by either delaying or discarding some of the arriving workload. In general, shaping will alter the effective bandwidth of a source for larger values of s, and on short, intermediate and long time scales, as we next illustrate.

Suppose that a source is shaped to conform with parameters (ρ, β) by passage through a device that delays or discards some of the workload. Let $X_{sh}[0,t], t > 0$, describe the stationary departure stream from the device, the *shaped* process, and let $\alpha_{sh}(s,t)$ be its effective bandwidth. Since $0 \le X_{sh}[0,t] \le \rho t + \beta$, a simple upper bound is

$$\alpha_{sh}(s,t) \le \frac{1}{st}\log\left[1 + \frac{t\alpha_{sh}(0,t)}{\rho t + \beta}\left(e^{s(\rho t+\beta)} - 1\right)\right]$$

where $\alpha_{sh}(0,t)$ is the mean rate of the shaped process. Example 2.3.1 illustrates that this bound may become tight as t approaches zero.

To explore the impact of shaping at intermediate time scales we describe a simple example. Consider a Lévy process, shaped by passage through a queue with service rate C. The queue size is a Markov process: assume further that it may be represented by a stationary Markov chain with q-matrix Q, for example the q-matrix of an M/M/1 queue for the arrival process leading to expression (2.10). Then the shaped process can be described as a fluid source, possibly with infinite state space. Indeed the shaped process is just an alternating renewal process, taking the level C for a busy period and the level 0 for an exponentially distributed idle period. If $\alpha_{sh}(s,t)$ is the effective bandwidth of the departure

stream from the queue, then $\alpha_{sh}(s,t)$ may be calculated from expression (2.7), where h_i takes values 0 or C.

De Veciana *et al.* (1994) have extensively explored the impact of shaping on the limiting form of the effective bandwidth as $t \to \infty$. An example of their results is that a Gaussian source, with effective bandwidth $\alpha(s,t) = \lambda + \sigma^2 s/2$, shaped by passage through a queue with service rate C, has

$$\alpha_{sh}(s,\infty) = \lambda + \frac{\sigma^2 s}{2} \qquad s < \frac{C-\lambda}{\sigma^2} \tag{2.13}$$

$$= C - \frac{(C-\lambda)^2}{2\sigma^2 s} \quad \text{otherwise.} \tag{2.14}$$

More generally, de Veciana *et al.* (1994) consider a wide class of arrival processes for which $\alpha(s,\infty) < \infty$ for some s, and show that $\alpha_{sh}(s,\infty) = \alpha(s,\infty)$ for values of s less than a critical level, while above this level the impact of the peak rate C is felt and $\alpha_{sh}(s,\infty) = C - \kappa/s$, where the constant κ may be calculated. See de Veciana and Walrand (1995) for further discussion of shaping.

3 Multiplexing models

In this Section we suppose the arrival process is

$$X[0,t] = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji}[0,t] \tag{3.1}$$

where $(X_{ji}[0,t])_{ji}$ are independent processes with stationary increments whose distributions may depend upon j but not upon i, and that there is a resource that has to cope with the aggregate arriving stream of work. We interpret n_j as the number of sources of type j, and shall write $\alpha_j(s,t)$ for the effective bandwidth of a source of type j. Thus

$$\alpha(s,t) = \sum_{j=1}^{J} n_j \alpha_j(s,t). \tag{3.2}$$

We shall explore several multiplexing models, and shall be interested in the relationship between constraints of the form

$$\sum_{j=1}^{J} n_j \alpha_j(s^*,t^*) \le C^* \tag{3.3}$$

for one or several choices of (s^*,t^*,C^*) and the acceptance region, defined as the set of vectors $(n_1,n_2,\ldots,n_J)$ for which a given performance, described in terms of queueing delay or buffer overflow, can be guaranteed.

In Section 3.1 we describe the result of Hui (1988, 1990) which establishes inequality (3.3) as a conservative bound on the non-linear acceptance region for a bufferless model. In Section 3.2, based on Kelly (1991), we see relation (3.3) emerge as the linear limiting form of, and as a conservative bound on, the accep-

tance region for a buffered model with Lévy input. A linear limiting form was established for more general input processes, including the fluid sources studied in detail by Gibbens and Hunt (1991) and Elwalid and Mitra (1993), by Kesidis *et al.* (1993): we review this result in Section 3.3, together with its recent generalization by Duffield and O'Connell (1996). In the models of Sections 3.1 and 3.3 time scales essentially degenerate: the time scale t^* appearing in (3.3) approaches zero or infinity. In Sections 3.4 and 3.5 we describe two tractable models illustrating phenomena when time scales do not degenerate. In Section 3.6 we discuss the important recent results of Botvitch and Duffield (1995), Simonian and Guibert (1995) and Courcoubetis and Weber (1996) on an asymptotic regime where the form (3.3) emerges, for finite values of t^*, as a tangent to the limiting acceptance region. In Section 3.7 we briefly discuss priority models, which provide further important examples where several constraints of the form (3.3) may be needed to approximate the acceptance region.

3.1 Bufferless models

We look first at a simple model where

$$X = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji}$$

and X_{ji} are independent random variables with scaled logarithmic moment generating functions

$$\alpha_j(s) = \frac{1}{s} \log \mathbb{E}[e^{sX_{ji}}]. \tag{3.4}$$

We might suppose that X_{ji} is the instantaneous arrival rate of work from a source of type j at a bufferless resource of capacity C, corresponding to the choice $\alpha_j(s) = \lim_{t \to 0} \alpha_j(s/t, t)$. Alternatively we might suppose that $X_{ji}[0,t] = X_{ji}t$, so that $\alpha_j(s/t, t) = \alpha_j(s)$ for all values of t.

Chernoff's bound gives

$$\log P\{X \geq C\} \leq \log \mathbb{E}\left[e^{s(X-C)}\right] = s\big(\alpha(s) - C\big) \tag{3.5}$$

where $\alpha(s) = \sum_j n_j \alpha_j(s)$. Thus the constraint $\log P\{X \geq C\} \leq -\gamma$ will certainly be satisfied if the vector $n = (n_1, n_2, \ldots, n_J)$ lies within the set

$$A = \left\{ n : \inf_s \left[s\big(\sum_{j=1}^{J} n_j \alpha_j(s) - C\big) \right] \leq -\gamma \right\} \tag{3.6}$$

where throughout $n \geq 0$. The region A has a convex complement in $\mathbb{R}_+^J$, since this complement is defined as the intersection of $\mathbb{R}_+^J$ with a family of half-spaces. The half-space touching at a point n^* on the boundary of the region A is

$$\sum_{j=1}^{J} n_j \alpha_j(s^*) \leq C - \frac{\gamma}{s^*}, \tag{3.7}$$

where s^* attains the infimum appearing in relation (3.6) with n replaced by n^*. Thus condition (3.7) is a conservative global bound, of the form (3.3), on the acceptance region: if n satisfies this condition then the performance guarantee $\log P\{X \geq C\} \leq -\gamma$ is assured.

Let $A(\gamma, C)$ be the subset of $\mathbb{R}_+^J$ such that $n \in A(\gamma, C)$ implies $\log P\{X \geq C\} \leq -\gamma$. Chernoff's theorem (Billingsley 1986) gives that

$$\lim_{N\to\infty} \frac{1}{N} \log P\{\sum_{j=1}^{J} \sum_{i=1}^{n_j N} X_{ji} \geq CN\} = \inf_s \Big[s\big(\sum_{j=1}^{J} n_j \alpha_j(s) - C\big)\Big]. \tag{3.8}$$

Except in the trivial circumstances where the infimum is zero or minus infinity or where a source type has zero mean rate, the infimum (3.8) is strictly increasing in each component of n. It follows that

$$\lim_{N\to\infty} \frac{A(\gamma N, CN)}{N} = A, \tag{3.9}$$

and in this sense the approximation leading to the region A becomes more accurate as the number of sources increases, and the tail probability decreases.

The convergence statement (3.9) requires a comment on topology. Throughout we shall use the Hausdorff distance (Csáskár 1978) to define a pseudo-metric over the set of subsets of $\mathbb{R}_+^J$. This induces a topology over the quotient space formed by identifying subsets of $\mathbb{R}_+^J$ which share the same closure. Our notation will use a subset of $\mathbb{R}_+^J$ to represent its equivalence class; the intersection operator, as will appear in for example relation (3.25), is defined on equivalence classes in the natural way using *closed* subsets of $\mathbb{R}_+^J$ as representatives. Thus limits such as (3.9) are uninformative about the limiting behaviour of performance measures at points on the boundary of the set A.

3.1.1 *Stream based measures*

The discussion leading to relation (3.9) concerned a resource-based congestion measure, the probability of resource overload, rather than a stream-based congestion measure, such as the proportion of work from an arriving stream that is lost. To convert from resource-based measures to stream-based measures requires two steps: relating the expected size of overloads to tail probabilities of overloads; and dividing by stream rates.

From Chernoff's bound the expected rate of load loss

$$\begin{aligned} \mathbb{E}(X-C)^+ &= \int_0^\infty P\{X \geq C + x\} dx \\ &\leq \int_0^\infty \exp\Big[s\big(\alpha(s) - (C+x)\big)\Big] dx \\ &= \frac{1}{s} \exp\Big[s\big(\alpha(s) - C\big)\Big]. \end{aligned}$$

We deduce that

$$\mathbb{E}(X-C)^+ \leq \frac{1}{s^*}\exp\Big[s^*\big(\alpha(s^*)-C\big)\Big] \tag{3.10}$$

where s^* attains the infimum in (3.8). Thus if condition (3.7) is satisfied, and hence $P\{X > C\} \leq e^{-\gamma}$ is assured, then also assured is that $\mathbb{E}(X-C)^+ \leq e^{-\gamma}/s^*$. Note that the proportion of load lost is just $\mathbb{E}(X-C)^+/\mathbb{E}X$. Let $A_{st}(\gamma, C)$ be the subset of $\mathbb{R}_+^J$, such that $n \in A_{st}(\gamma, C)$ implies that the proportion of work lost is not greater than $e^{-\gamma}$. Then a further consequence of Chernoff's theorem is that

$$\lim_{N\to\infty} \frac{A_{st}(\gamma N, CN)}{N} = A.$$

3.1.2 *Improved approximations*

The inequalities (3.5) and (3.10) provide bounds on probability of resource overload or the proportion of work lost: closely related *tilted approximations* may be developed by various techniques reviewed by Reid (1988, Section 6.3), Bucklew (1990, Chapter VII), and Jensen (1995). For example the estimates

$$P\{X \geq C) \sim \frac{1}{s^*(2\pi\sigma^2(s^*))^{\frac{1}{2}}}\, e^{s^*(\alpha(s^*)-C)} \tag{3.11}$$

and

$$\mathbb{E}(X-C)^+ \sim \frac{1}{s^{*2}(2\pi\sigma^2(s^*))^{\frac{1}{2}}}\, e^{s^*(\alpha(s^*)-C)}, \tag{3.12}$$

where $\sigma^2(s) = \frac{\partial^2}{\partial s^2}(s\alpha(s))$, have been discussed by Hui (1988), Roberts (1992, p. 154), and Hsu and Walrand (1995): the prefactor of the exponential term considerably improves accuracy. Note that the prefactor depends upon whether the measure of interest is resource-based or stream-based.

3.1.3 *Approximate linearity*

How well approximated is the region (3.6) by the linearly constrained region (3.7)? Some insight may be obtained from the Gaussian case, where explicit calculations are easy to perform. Suppose that

$$\alpha_j(s) = \lambda_j + \frac{s\sigma_j^2}{2},$$

corresponding to a normally distributed load with mean λ_j and variance σ_j^2. Then the region (3.6) becomes

$$\sum_j n_j\lambda_j + (2\gamma\sum_j n_j\sigma_j^2)^{\frac{1}{2}} \leq C. \tag{3.13}$$

The tangent plane at a point n^* on the boundary of the region (3.13) is of the form (3.7) with

$$s^* = \frac{C - \sum_j n_j^*\lambda_j}{\sum_j n_j^*\sigma_j^2}$$

and hence

$$\alpha_j(s^*) = \lambda_j + \frac{\gamma\sigma_j^2}{C(1-\delta^*)} \tag{3.14}$$

where $\delta^* = \sum_j n_j^* \lambda_j / C$, the traffic intensity. Thus the coefficients (3.14) will be relatively insensitive to the traffic mix n^*, provided $(1-\delta^*)^{-1}$ does not vary too greatly with n^*, or, equivalently, provided the traffic intensity is not too close to 1 on the boundary of the acceptance region.

Let $C^* = C - \gamma/s^*$, the *effective capacity* appearing on the right hand side of inequality (3.7). Then

$$\begin{aligned} C^* &= C - \gamma \frac{\sum_j n_j^* \sigma_j^2}{C(1-\delta^*)} \\ &= C - \gamma \frac{\text{variance of load}}{\text{mean free capacity}}. \end{aligned}$$

We shall refer again to this simple model in Sections 3.3.1 and 3.6.2.

3.2 M/G/1 models

Next suppose that each of the processes $X_{ji}[0,t]$ has independent increments, as discussed in Section 2.5, and write $\alpha_j(s) = \alpha_j(s,t)$, $\alpha(s) = \alpha(s,t)$. Let Q be distributed as the stationary workload in a queue with a server of capacity C and an infinite buffer, fed by the arrival stream $X[0,t]$. (More formally we could define the queue size at time τ as

$$Q(\tau) = (X[0,\tau] - C\tau) - \inf_{0<t<\tau} \{X[0,t] - Ct\},$$

and let $\tau \to \infty$—see Harrison 1985, p. 19, or Asmussen 1987, Chapter III, 7–8.) Then the Pollaczek–Khinchin formula (see, for example, Asmussen 1987, p. 206, or Kella and Whitt 1992) is simply

$$\mathbb{E}[e^{sQ}] = \frac{C - \alpha(0)}{C - \alpha(s)}. \tag{3.15}$$

Cramér's estimate (Feller 1971) describes the tail behaviour of the distribution for Q. Suppose there exists a finite constant κ such that $\alpha(\kappa) = C$, and suppose that κ is in the interior of the interval on which $\alpha(s)$ is finite, so that $\alpha'(\kappa)$ is necessarily finite. Then Cramér's estimate is

$$P\{Q \geq b\} \sim \frac{C - \alpha(0)}{\kappa\alpha'(\kappa)} e^{-\kappa b} \quad \text{as } b \to \infty. \tag{3.16}$$

Let $A(\gamma, b)$ be the subset of $\mathbb{R}_+^J$ such that $n \in A(\gamma, b)$ implies $\log P\{Q \geq b\} \leq -\gamma$. Then a consequence of Cramér's estimate is that

$$\lim_{N\to\infty} A(\gamma N, bN) = A, \tag{3.17}$$

where

$$A = \Big\{ n : \sum_j n_j \alpha_j \left(\frac{\gamma}{b} \right) \leq C \Big\}, \tag{3.18}$$

again a region defined by a constraint of the form (3.3). Kelly (1991) notes that $A \subset A(\gamma, b)$, and so the linearly constrained region A is a conservative global bound, as well as an asymptotic limit.

3.2.1 *Finite buffers*

The above discussion concerned the proportion of time the buffer occupancy exceeded a level b, in a queue with an infinite buffer. Next we consider what happens if there is a finite buffer of size b, and any excess workload over this level is lost. Note that we can construct a sample path of this process from the sample path of an M/G/1 queue with infinite buffer: just remove the time intervals when the workload is above b. That this construction works is a consequence of the simple rule for overflow, and the assumption that the arrival process has independent increments. The stationary distribution for the workload in an M/G/1 queue with finite buffer b is thus obtained from that for the infinite buffer case by conditioning on the event that the workload does not exceed b. From Cramér's estimate (3.16) it can be deduced that the proportion of workload lost with a finite buffer of size b, $L(b)$, satisfies

$$L(b) \sim \frac{C(C - \alpha(0))^2}{\kappa \alpha'(\kappa) \alpha(0)} \, e^{-\kappa b} \quad \text{as } b \to \infty.$$

It follows that if $A_{\rm prop}(\gamma, b)$ is the subset of $\mathbb{R}_+^J$ such that $n \in A_{\rm prop}(\gamma, b)$ implies $\log L(b) \leq -\gamma$ then

$$\lim_{N \to \infty} A_{\rm prop}(\gamma N, bN) = A.$$

3.2.2 *Brownian input*

Suppose that

$$X_{ji}[0, t] = \lambda_j t + \sigma_j Z(t)$$

where $Z(t)$ is a standard Brownian motion. Then superpositions can also be expressed in terms of a Brownian motion, Z^1, as

$$X[0, t] = \big(\sum_j n_j \lambda_j \big) t + \big(\sum_j n_j \sigma_j^2 \big)^{\frac{1}{2}} Z^1(t),$$

and hence, from basic results on reflected Brownian motion (Harrison 1985),

$$P\{Q \geq b\} = \exp \left\{ \frac{-2b\big(C - \sum_j n_j \lambda_j\big)}{\sum_j n_j \sigma_j^2} \right\}.$$

Thus the constraint $\log P\{Q \geq b\} \leq -\gamma$ becomes *precisely* the condition

$$\sum_j n_j \left(\lambda_j + \sigma_j^2 \frac{\gamma}{2b} \right) \leq C, \tag{3.19}$$

which is just the canonical constraint (3.3) with $s^* = \gamma/b$.

3.2.3 *Mean delays*

From the Pollaczek–Khinchin formula (3.15) it follows that $\mathbb{E}Q = \alpha'(0)/(C - \alpha(0))$, and hence that a constraint of the form $\mathbb{E}Q \leq L$ is satisfied if and only if

$$\sum_{j=1}^{J} n_j \left[\alpha_j(0) + \frac{\alpha_j'(0)}{L}\right] \leq C.$$

This provides a linear acceptance region which accords with the previous example in the case of Brownian input, but which is not, in general, of the canonical form (3.3). In Kelly (1991) this and other possible definitions of an effective bandwidth were considered, with emphasis on the linearity of the acceptance region under a variety of performance criteria. In this paper we explore a different perspective, one which emphasises the unifying role of the definition (2.1) under a variety of multiplexing models.

3.3 Buffer asymptotic models

Tail probabilities decay exponentially in models more general than the M/G/1 queue. Suppose that Q is distributed as the stationary workload in a queue with a server of capacity C and an infinite buffer, fed by an arrival stream $X[0,t]$ with stationary and ergodic increments. Thus we weaken the M/G/1 assumption of independent increments to an assumption of ergodic increments. Suppose that

$$\lim_{t\to\infty} \alpha(s,t) = \alpha(s) \tag{3.20}$$

and that there exists a finite constant κ such that $\alpha(\kappa) = C$, and $\alpha'(\kappa)$ is finite. Then

$$\lim_{b\to\infty} \frac{1}{b} \log P\{Q \geq b\} = -\kappa \tag{3.21}$$

(Kesidis *et al.* 1993; Chang 1994; Glynn and Whitt 1994). Thus the relations (3.17), (3.18) hold in this more general context.

The examples of Section 2.3 show that even if the limit (3.20) exists, convergence to the limit may be arbitrarily slow; further, for finite values of t, $\alpha(s,t)$ may be much smaller or larger than the limit $\alpha(s)$. The examples of Choudhury *et al.* (1994) can be interpreted as further illustrations of this phenomenon. The usefulness of the limit (3.21) thus depends on the rate of convergence to this limit, and whether convergence has essentially occurred on the time scales of interest. Interestingly, the GI/G/1 or M/G/1 models provide a natural choice of time scale. For example, suppose the limit (3.20) is approximately of the form (2.11) appropriate for an M/G/1 model. Under this model the time taken to empty a full buffer is of order $t_1 = b/(C - \alpha(0))$, while the time taken to fill an empty buffer is of order $t_2 = b/(\kappa\alpha'(\kappa))$, as b increases (see Tse *et al.* 1995 for a valuable discussion of regenerative structure in this model). For the asymptotic (3.21) to be appropriate, $\alpha(\kappa,t)$ should have essentially converged to its limit $\alpha(\kappa)$ by time scales t in the region of t_1, t_2; and $\alpha(s,t)$ evaluated in this region

should be used in estimates such as (3.16).

The limit (3.20) may not, however, be capable of representation in the form (2.11) appropriate for an M/G/1 model, or even in the form (3.4) for *any* random variable. The form (3.4) is differentiable, while, for example, the function (2.13),(2.14) obtained from a shaped process has a discontinuous derivative at the critical point $s = (C - \lambda)/\sigma^2$, where there is a transition away from a regenerative regime.

Duffield and O'Connell (1996) have extended buffer asymptotics to examples where the limit (3.20) does not exist, but where, with a suitable rescaling, a large deviation principle may still be applied. We illustrate their result with a simple example.

3.3.1 *Fractional Brownian input*

Suppose that $\alpha(s,t)$ is given by expression (2.8), corresponding to fractional Brownian motion with Hurst parameter H. Then Duffield and O'Connell (1996) show that

$$\lim_{b\to\infty} \frac{\log P\{Q \geq b\}}{b^{2(1-H)}} = -\frac{1}{2\sigma^2}\left(\frac{C-\lambda}{H}\right)^{2H}(1-H)^{-2(1-H)} \tag{3.22}$$

agreeing with an earlier bound of Norros (1994).

Next suppose that

$$\alpha_{ji}(s,t) = \lambda_j + \frac{\sigma_j^2}{2}\, s t^{2H-1} \tag{3.23}$$

so that $\alpha(s,t)$, given by relation (3.2), corresponds to a superposition of fractional Brownian sources, all sharing the same Hurst parameter H. Then from the result (3.22) it is possible to deduce (Duffield *et al.* 1994) that the condition $P\{Q \geq b\} \leq e^{-\gamma}$ becomes (asymptotically, as $\gamma, b \to \infty$ with $\gamma/b^{2(1-H)}$ held constant) the condition

$$2\gamma \sum_{j=1}^{J} n_j \sigma_j^2 \leq b^{2(1-H)} \left(\frac{C - \sum_{j=1}^{J} n_j \lambda_j}{H}\right)^{2H} (1-H)^{-2(1-H)}. \tag{3.24}$$

Observe that for $H = 1/2$ this is just the (exact) condition (3.19), while as $H \to 1$ it approaches the condition (3.13). The major effect of long range order is thus on the scaling relationship between γ, b and C, as discussed by Norros (1994), rather than on the geometrical form of the acceptance region A. We shall return to this point later, in Section 3.6.2, where we shall also discuss the connection between inequality (3.24) and the form (3.3).

3.4 Deterministic multiplexing

Suppose the arriving work is dealt with by a server of capacity C with a finite buffer of capacity b, initially empty. Under what condition is the capacity of the

buffer *never* exceeded? The condition is (Cruz 1991) that $n \in A$ where

$$A = \bigcap_{0<t<\infty} A_t, \tag{3.25}$$

an intersection of linearly constrained regions

$$A_t = \left\{ n : \sum_j n_j \alpha_j(\infty, t) \le C + \frac{b}{t} \right\}. \tag{3.26}$$

3.4.1 *Policed sources*

Recall that in Section 2.5 we discussed policed sources. If

$$\alpha_j(\infty, t) = \rho_j + \frac{\beta_j}{t}$$

corresponding to a source policed by parameters (ρ_j, β_j), then

$$A = A_0 \cap A_\infty$$

where

$$A_0 = \{n : \sum_j n_j \beta_j \le b\}, \quad A_\infty = \{n : \sum_j n_j \rho_j \le C\}.$$

Note that if β_j/ρ_j does not vary with the source type j, then the boundaries of the regions A_0 and A_∞ are parallel.

3.4.2 *Multiple policers*

If

$$\alpha_j(\infty, t) = \min_{k \in Kj} \left\{ \rho_{jk} + \frac{\beta_{jk}}{t} \right\}$$

corresponding to a source policed by a finite set K_j of parameter choices, then A can be written as an intersection of a finite collection of sets A_t. For example, if $K_j = \{1, 2, \ldots, K\}$, if $(\beta_{jk}, k = 1, 2, \ldots, K)$ is an increasing sequence for each $j = 1, 2, \ldots, J$, and if the ratios

$$t_k = \frac{\beta_{jk+1} - \beta_{jk}}{\rho_{jk} - \rho_{jk+1}} \tag{3.27}$$

do not vary with the source type j and are increasing in k, then

$$A = \bigcap_{k=0}^{K} A_{t_k},$$

where $t_0 = 0$ and $t_K = \infty$.

3.5 Brownian bridge models

When several independent periodic sources, of type (2.6), are superimposed, the resulting process can be approximated by a Brownian bridge (for a recent review

see Hajek 1994). This motivates study of the source

$$X_{ji}[0,t] = \lambda_j t + \sigma_j Z_0(t - \lfloor t \rfloor)$$

where $Z_0(t), 0 \leq t \leq 1$, is a standard Brownian bridge. Then

$$\alpha_j(s,t) = \lambda_j + \frac{\sigma_j^2 s}{2t}(t - \lfloor t \rfloor)(1 + \lfloor t \rfloor - t). \tag{3.28}$$

For example, a periodic source, with period 1 and burst size β_j, might be approximated by $\lambda_j = \beta_j, \sigma_j = \beta_j$: this example is of some interest as a conservative description of sources policed by parameters (ρ_j, β_j), where the ratio β_j/ρ_j does not vary with source type, and where setting the ratio to 1 simply fixes the time unit. Superpositions can also be expressed in terms of a Brownian bridge, Z_0^1, as

$$X[0,t] = \sum_j n_j \lambda_j t + \left(\sum_j n_j \sigma_j^2\right)^{\frac{1}{2}} Z_0^1(t - \lfloor t \rfloor).$$

The condition for the queue to be stable is just

$$\sum_j n_j \lambda_j < C. \tag{3.29}$$

Given this, the stationary probability $P\{Q \geq b\}$ is

$$P\left\{\max_{0 \leq t \leq 1} \{X[0,t] - Ct\} \geq b\right\} = \exp\left\{\frac{-2b}{\sum_j n_j \sigma_j^2}(b + C - \sum n_j \lambda_j)\right\}$$

(Hajek 1994, p. 150), and this probability is less than $e^{-\gamma}$ if and only if

$$\sum_j n_j \left(\lambda_j + \sigma_j^2 \frac{\gamma}{2b}\right) < b + C. \tag{3.30}$$

This constraint does not, in general, imply the condition (3.29).

Thus *two* linear constraints (3.29) and (3.30) are equivalent to the condition that $\log P\{Q \geq b\} < -\gamma$. Constraint (3.29) is of the canonical form (3.3) with $t^* = \infty$. Constraint (3.30) may be thrown into the form (3.3), with for example the choice $(s^*, t^*) = (2\gamma/b, 1/2)$. This example will be explored further in Section 3.6.1.

3.6 Buffer and source asymptotics

In Sections 3.1 and 3.3 we described asymptotic results when the number of sources or the buffer size, respectively, increased. Recently Botvich and Duffield (1995), Simonian and Guibert (1995) and Courcoubetis and Weber (1996) have obtained important results when the number of sources *and* the buffer size increase together, the regime considered in a key early paper of Weiss (1986).

Again suppose that the arrival process is given by the superposition (3.1), where the increments of $X_{ji}[0,t]$ are stationary. Let $L(C,b,n)$ be the proportion of workload lost, through overflow of a buffer of size $b > 0$, when the server has

rate C and as usual $n = (n_1, n_2, \ldots, n_J)$. Then the above authors establish that

$$\lim_{N\to\infty} \frac{1}{N} \log L(CN, bN, nN) = \sup_t \inf_s \big[st \sum_j n_j \alpha_j(s,t) - s(b + Ct)\big]. \quad (3.31)$$

Let $A(\gamma, C, b)$ be the subset of $\mathbb{R}_+^J$ such that $n \in A(\gamma, C, b)$ implies $\log L(C, b, n) \leq -\gamma$. As in Section 3.2, the limit (3.31) is strictly increasing in each component of n, and hence

$$\lim_{N\to\infty} \frac{A(\gamma N, CN, bN)}{N} = A$$

where

$$A = \bigcap_{0<t<\infty} A_t \quad (3.32)$$

with

$$A_t = \Big\{ n : \inf_s \big[st \sum_j n_j \alpha_j(s,t) - s(b + Ct)\big] \leq -\gamma \Big\}, \quad (3.33)$$

a region with convex complement in $\mathbb{R}_+^J$. Moreover, if the boundary of the region A is differentiable at a point n^*, then the tangent plane is

$$\sum_j n_j \alpha_j(s^*, t^*) = C + \frac{b}{t^*} - \frac{\gamma}{s^* t^*} \quad (3.34)$$

where (s^*, t^*) is an extremizing pair in relation (3.31) with n replaced by n^*. Thus a constraint of the canonical form (3.3) emerges as an asymptotic local limit, local in variations of the traffic mix n.

It is interesting to compare the regions (3.32) and (3.33) with corresponding regions obtained in earlier Sections. Consider the model of Section 3.1, where several formal comparisons are possible. If $b = 0$, then A_t is increasing in t, by the final remark of property (ii), and so $A = A_0$. We recover the region (3.6), with the interpretation $\alpha_j(s) = \lim_{t\to 0} \alpha_j(s/t, t)$. Or, if $b > 0$ and $\alpha_j(s,t)$ depends on s, t only through the product st, then A_t is decreasing in t and so $A = A_\infty$. Again we recover the region (3.6) with the interpretation $\alpha_j(s) = \alpha_j(s/t, t), t > 0$. If $\alpha_j(s,t)$ is independent of t, as discussed in Section 3.2, then the envelope of the regions A_t is the linear boundary of the region (3.18).

The results of Section 3.1 concern a regime where the time taken to fill a buffer is much *shorter* than the time periods over which sources fluctuate, while the results of Section 3.3 concern a regime where it is much *longer*. In both cases the limit results concern the behaviour of $\alpha(s,t)$ for t near zero or infinity. The great advantage of the limiting regime described in this Section is that allows the shape of the effective bandwidth $\alpha(s,t)$ to identify the relevant time scale implicitly, and, in general, the region (3.32) will depend upon $\alpha(s,t)$ evaluated at finite values of t. A simple illustration of this is provided by the limit as $\gamma \to \infty$, when the region (3.33) shrinks to the region (3.26) of Section 3.4; thus in example 3.4.2 there is a single linear constraint for each of the time constants

(3.27). A more subtle illustration is provided by examples 3.6.1 and 3.6.2 below.

Several further examples are discussed in detail by Botvitch and Duffield (1995), Simonian and Guibert (1995) and Courcoubetis and Weber (1996). Simonian and Guibert (1995) also describe bounds and estimates that parallel the tilted approximations (3.11), (3.12).

3.6.1 *A Brownian bridge model*

Suppose that $\alpha_{ji}(s,t)$ is given by expression (3.28). Then the set (3.32) becomes

$$A = \bigcap_{0<t<1} A_t \cap A_\infty = A_{(0,1)} \cap A_\infty$$

where $A_{(0,1)}$ and A_∞ are simply the regions (3.30) and (3.29) respectively; recall that for the Brownian bridge model of Section 3.5 the acceptance region A is exact for finite values of γ, C and b.

3.6.2 *Fractional Brownian input*

If $\alpha_{ji}(s,t)$ is given by expression (3.23), then the tangent plane (3.34) uses the space and time scales[1]

$$s^* = 2(1-H)\frac{\gamma}{b}, \quad t^* = \left(\frac{H}{1-H}\right)\frac{b}{C - \sum_j n_j^* \lambda_j}$$

and the acceptance region (3.32) becomes

$$H\left(\frac{1-H}{b}\right)^{\frac{1}{H}-1}\left(2\gamma\sum_j n_j\sigma_j^2\right)^{\frac{1}{2H}} + \sum_j n_j\lambda_j \le C. \tag{3.35}$$

This is just condition (3.24), although the limiting regime is different. Note that the region (3.35) is convex or concave (has convex complement) according as $H \le \frac{1}{2}$ or $H \ge \frac{1}{2}$. Regions that are neither convex nor concave can be constructed by allowing the Hurst parameter H to vary with source type.

If $H = \frac{1}{2}$ the condition (3.35) is just the linear constraint (3.19). Even the most extreme values of H produce rather well behaved acceptance regions: as $H \to 1$ the inequality (3.35) approaches the condition (3.13), and as $H \to 0$ it approaches the conditions

$$\sum_j n_j\lambda_j \le C, \quad 2\gamma\sum_j n_j\sigma_j^2 \le b^2,$$

[1]The published version of this paper has a mistake in its formula for t^* (the exponent of the first term in the published version should be $-1/2H$, not $1/2H$): I'm grateful to Yih-Choung Teh for pointing this out. But in addition the published expressions for s^* and t^* are not as simple as those above. I'm grateful to Bong Ryu for providing the expression above for t^*, an expression that appears in his paper with Anwar Elwalid entitled "The Importance of the Long-Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities," Proc. ACM SIGCOMM '96 Stanford University, CA, available at http://www.wins.hrl.com/people/ryu. The expressions above make clear that the space scale s^* is inversely proportional to b, and the time scale t^* is linear in b, for given values of γ, C, and n.

a limiting acceptance region with a similar geometrical form to that found in example 3.4.1.

3.7 Priorities

Multiple time and space scales may also arise for certain priority mechanisms. Suppose a single resource gives strict priority to sources $j \in J_1$, which have a strict delay requirement, but also serves sources $j \in J_2$, which have a much less stringent delay requirement. Then two constraints of the form

$$\sum_{j \in J_1} \alpha_j(s_1, t_1) \leq C_1, \qquad \sum_{j \in J_1 \cup J_2} \alpha_j(s_2, t_2) \leq C_2 \tag{3.36}$$

may be needed to ensure that both sets of requirements are met (for several examples, see Bean 1994, Elwalid and Mitra 1995, de Veciana and Walrand 1995). If the less stringent delay requirement becomes very weak, corresponding to a *very* large buffer and almost *no* sensitivity to delay, then s_2 will approach zero, and $\alpha_j(s_2, t_2)$ will approach $\mathbb{E}X[0,t]/t$, the mean load produced by source j. The second constraint of (3.36) then becomes the simple constraint that the mean loads of all sources should not exceed the capacity of the resource.

With several priority classes the key point remains that each priority class may have its own characteristic space and time scale: under strict priority a source is unaffected by lower priority sources, but will be affected by the behaviour of higher priority sources on its characteristic space and time scale. Kulkarni *et al.* (1995) study an alternative priority mechanism, where first-in-first-out scheduling is used and arriving work of low priority is rejected if the workload is above a threshold.

In Section 3 we have reviewed a variety of results, emphasising their interpretation in terms of effective bandwidths. Of course other perspectives are possible. In particular, Shwartz and Weiss (1995) explore several more detailed aspects of buffer behaviour for on-off fluid sources of the type defined in Section 2.3.2, using this model to illustrate the considerable power of large deviation theory.

4 Tariffs and connection acceptance

The effective bandwidth of a source depends sensitively upon its statistical characteristics. The source, however, may have difficulty providing such information. Uncertain characterization of sources raises challenging practical and theoretical issues for the design of tariffing and connection acceptance control mechanisms. Suppose, for example, that mechanisms are based on attempts to measure the effective bandwidth of a connection, perhaps by estimating expression (2.1) using an empirical averaging to replace the expectation operator. Is this satisfactory? Suppose a user requests a connection policed by a high peak rate, but then happens to transmit very little traffic over the connection. Then an *a posteriori* estimate of quantity (2.1) will be near zero, even though an *a priori* expectation may be much larger, as assessed by either the user or the network. If tariffing and connection acceptance control are primarily concerned with expectations of

future quality of service, and if sources may be non-ergodic over the relevant time scales, then the distinction matters.

In this Section we describe an approach to tariffing and connection acceptance control mechanisms that can make effective and robust use of both prior declarations and empirical averages. The key idea is the use of prior declarations to choose a linear function that bounds the effective bandwidth (as illustrated in Fig. 5); tariffs and connection acceptance can then be based upon the relatively simple measurements needed to evaluate this function.

Although an *individual* source may be poorly characterized, certain features of the *aggregate* load on a resource may be known. In this Section we assume that the key constraints (3.3), and the critical space and time scales appearing in these constraints, have been identified.

4.1 Charging mechanisms

Let

$$Z = \mathbb{E}e^{sX[\tau,\tau+t]}, \tag{4.1}$$

and rewrite expression (2.1) as

$$\alpha(Z) = \frac{1}{st}\log Z, \tag{4.2}$$

where the notation now emphasizes the dependence of the effective bandwidth on the summary Z of the statistical characteristics of the source.

Suppose that, before the call's admission, the network requires the user to announce a value z, and then charges for the call an amount $f(z;Z)$ per unit time, where Z is estimated by an empirical averaging. We suppose that the user is risk-neutral and attempts to select z so as to minimize the expected cost per unit time: call a minimizing choice of z, $\hat{z}$ say, an *optimal* declaration for the user. What properties would the network like the optimal declaration $\hat{z}$ to have? Well, first of all the network would like to be able to deduce from $\hat{z}$ the user's *a priori* expectation (4.1). A second desirable property would be that the expected cost per unit time under the optimal declaration $\hat{z}$ be proportional to the effective bandwidth (4.2) of the call. In Kelly (1994a) it is shown that these two requirements essentially characterize the tariff $f(z;Z)$ as

$$f(z;Z) = a(z) + b(z)Z, \tag{4.3}$$

defined as the tangent to the curve $\alpha(Z)$ at the point $Z = z$.

4.1.1 *On-off sources*

Consider the very simple case of an on-off source which produces workload at a constant rate h while in an 'on' state, and produces no workload while in an 'off' state. Suppose the periods spent in 'on' and 'off' states are large, so that the effective bandwidth is given by expression (2.9) with $\alpha_1(s,t) = h$ and $p = M/h$. Here M and h are respectively the mean and peak of the source. If h is fixed

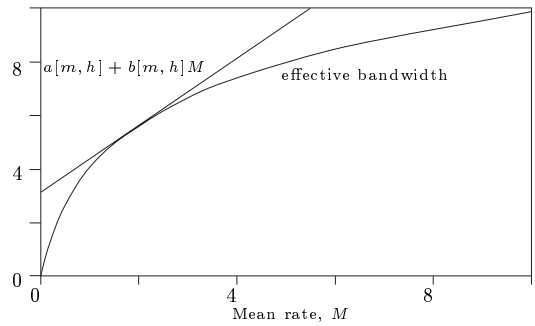

FIG. 5. Implicit pricing of an effective bandwidth. The effective bandwidth is shown as a function of the mean rate, M. The user is free to choose the declaration m, and is then charged an amount $a[m,h]$ per unit time, and an amount $b[m,h]$ per unit volume. The values of $a[m,h]$ and $b[m,h]$ are determined from the tangent at the point $M=m$.

and known (it may, for example, be policed) then

$$Z = 1 + \frac{M}{h}\left(e^{sth} - 1\right). \tag{4.4}$$

(Kelly 1994b provides a numerical illustration of the choice, discussed theoretically in Section 3.1, of the parameter st.) If we let z be defined by expression (4.4) with M replaced by m then the tariff (4.3) may be rewritten as

$$a[m,h] + b[m,h]M, \tag{4.5}$$

the tangent to the function

$$\alpha[M,h] = \frac{1}{st}\log\left[1 + \frac{M}{h}\left(e^{sth} - 1\right)\right]$$

at the point $M=m$ (see Fig. 5). Note the very simple interpretation possible for the tariff (4.5): the user is free to choose a value m, and then incurs a charge $a[m,h]$ per unit time, and a charge $b[m,h]$ per unit of volume carried.

4.1.2 *Priorities*

Next we consider an example where it may be important to tariff several constraints of the form (3.3) simultaneously. Consider the model of Section 3.7, where there are several priority classes. Let $Z_k, \alpha_k(Z_k), z_k, f_k, a_k, b_k$ be defined as in relations (4.1)–(4.3), but with (s,t) replaced by (s_k,t_k). Then a tariff for priority class j of

$$f^{(j)}\left((z_k)_k;(Z_k)_k\right) = \sum_{k\geq j} c_k\left[a_k(z_k) + b_k(z_k)Z_k\right]$$

has the required incentive properties, where c_k is a weight, or shadow price, attached to the k^{th} constraint from the collection (3.36).

4.2 Connection acceptance control

We now describe how the coefficients defined in Section 4.1 can be used as the basis of a simple and effective connection acceptance control.

Suppose that a resource has accepted connections $1, 2, \ldots, I$, and write (a_i, b_i) for the coefficients $(a(z_i), b(z_i))$ chosen by the user responsible for connection i at the time that the connection was accepted. Suppose also that the resource measures the load $X_i[\tau, \tau + t]$ produced by connection i over a period of length t, and let $Y_i = \exp(sX_i[\tau, \tau + t])$. Define the *effective load* on the resource to be

$$\sum_{i=1}^{I}(a_i + b_i Y_i).$$

Then a connection acceptance control may be defined as follows. A new request for a connection should be accepted or rejected according as the most recently calculated effective load is below or above a threshold value, with the proviso that if a request is rejected then later requests are also rejected until an existing connection terminates.

4.2.1 *On-off sources*

Consider again the simple case of on-off sources described in Section 4.1.1. Let h_i be the fixed and known peak of connection i, write (a_i, b_i) for the coefficients $(a[m_i, h_i], b[m_i, h_i])$ chosen by the user, and let the measured load from connection i be $M_i = X_i[\tau, \tau + t]/t$. Then the effective load on the resource becomes

$$\sum_{i=1}^{I}(a_i + b_i M_i),$$

to be compared with a threshold value.

An advantage of the on-off model, both for tariffing and connection acceptance control, is that it bounds other more complex source models. The reader surprised that schemes using only simple load measurements can guarantee strict quality of service requirements should see Gibbens *et al.* (1995), where issues of robustness and performance are investigated in some detail.

Of course on-off sources may, on a finer time scale, have more detailed structure, as in example 2.3.4. This may give rise to additional constraints of the form (3.3). There are a range of responses possible, ranging in complexity and conservatism. The models of Sections 2.3.1 or 3.5 might be appropriate as a conservative bound when source i is policed by parameters (ρ_i, β_i), where $\beta_i/\rho_i = 1$; this approach is described for a single source type in Gibbens *et al.* (1995, Section 6). A less conservative approach would use the same space and time scales, around $(s, t) = (2\gamma/b, 1/2)$, to assess the aggregate fine time scale load (4.1). Work in progress concerns how such connection acceptance controls might be implemented.

Bibliography

1. Anick, D., Mitra, D. and Sondhi, M.M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell Syst. Tech. J.*, **61**, 1871–1894.
2. Asmussen, S. (1987). *Applied Probability and Queues*. Wiley, Chichester.
3. Bean, N. (1994). Effective bandwidths with different quality of service requirements. In *IFIP Transactions, Integrated Broadband Communication Networks and Services* (ed. V.B. Iverson). Elsevier, Amsterdam, 241–252.
4. Billingsley, P. (1986). *Probability and Measure* (2nd edn). Wiley, New York.
5. Botvich, D.D. and Duffield, N. (1995). Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, **20**, 293–320.
6. Bucklew, J.A. (1990). *Large Deviation Techniques in Decision, Simulation and Estimation*. Wiley, New York.
7. Chang, C.-S. (1994). Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Trans. Automatic Control*, **39**, 913–931.
8. Chang, C.-S. and Thomas, J.A. (1995). Effective bandwidth in high-speed digital networks. *IEEE J. Selected Areas Commun.*, **13**, 1091–1100.
9. Choudhury, G., Lucantoni, D. and Whitt, W. (1994). On the effectiveness of effective bandwidths for admission control in ATM networks. In Labetoulle and Roberts (1994), 411–420.
10. Courcoubetis, C., Kesidis, G., Ridder, A., Walrand, J. and Weber, R. (1995). Admission control and routing in ATM networks using inferences from measured buffer occupancy. *IEEE Trans. Commun.*, **43**, 1778-1784.
11. Courcoubetis, C. and Weber, R. (1996). Buffer overflow asymptotics for a switch handling many traffic sources. *J. Appl. Prob.*, **33**.
12. Courcoubetis, C. and Weber, R. (1995). Effective bandwidths for stationary sources. *Prob. Eng. Inf. Sci.*, **9**, 285–296.
13. Cruz, R.L. (1991). A calculus for network delay. *IEEE Trans. Information Theory*, **37**, 114–141.
14. Császár, A. (1978) *General Topology*. Adam Hilger, Bristol.
15. de Veciana, G., Courcoubetis, C. and Walrand, J. (1994). Decoupling bandwidths for networks: a decomposition approach to resource management for networks. In *Proc. IEEE INFOCOM*, Vol. 2, 466–474.
16. de Veciana, G. and Walrand, J. (1995). Effective bandwidths: call admission, traffic policing and filtering for ATM networks. *Queueing Systems*, **20**, 37–59.
17. Duffield, N.G., Lewis, J.T., O'Connell, N., Russell, R. and Toomey, F. (1994). Predicting quality of service for traffic with long-range fluctuations.
18. Duffield, N.G., Lewis, J.T., O'Connell, N., Russell, R. and Toomey, F. (1995). Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE J. Selected Areas Commun.*, **13**, 981–990.

19. Duffield, N.G. and O'Connell, N. (1996). Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Camb. Phil. Soc.*
20. Elwalid, A.I. and Mitra, D. (1993). Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking*, **1**, 329–343.
21. Elwalid, A. and Mitra, D. (1995). Analysis, approximations and admission control of a multi-service multiplexing system with priorities. In *Proc. IEEE INFOCOM*, 463–472.
22. Feller, W. (1971) *An Introduction to Probability Theory and Its Applications, Volume II* (2nd edn). Wiley, New York.
23. Gibbens, R.J. (1996). Traffic characterisation and effective bandwidths for broadband network traces. In *Stochastic Networks: Theory and Applications* (ed. F. P. Kelly, S. Zachary and I. Ziedins). Volume 4 of *Royal Statistical Society Lecture Notes Series*, pp. 169–179. Oxford University Press, Oxford.
24. Gibbens, R.J. and Hunt, P.J. (1991). Effective bandwidths for the multi-type UAS channel. *Queueing Systems*, **9**, 17–28.
25. Gibbens, R.J., Kelly, F.P. and Key, P.B. (1995). A decision-theoretic approach to call admission control in ATM networks. *IEEE J. Selected Areas Commun.*, **13**, 1101–1114.
26. Glynn, P.W. and Whitt, W. (1994) Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*
27. Guérin, R., Ahmadi, H. and Naghshineh (1991). Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Selected Areas Commun.*, **9**, 968–981.
28. Hajek, B. (1994). A queue with periodic arrivals and constant service rate. In Kelly (1994c), 147–157.
29. Harrison, J.M. (1985). *Brownian Motion and Stochastic Flow Systems.* Krieger.
30. Hsu, I. and Walrand, J. (1995). Admission control for ATM networks. In Kelly and Williams (1995), 413–429.
31. Hui, J.Y. (1988). Resource allocation for broadband networks. *IEEE J. Selected Areas in Commun.*, **6**, 1598–1608.
32. Hui, J.Y. (1990). *Switching and Traffic Theory for Integrated Broadband Networks*. Kluwer, Boston.
33. ITU Recommendation I371 (1994). Traffic control and congestion control in B–ISDN. Geneva.
34. Jensen, J.L. (1995). *Saddlepoint Approximations*. Oxford University Press.
35. Kella, O. and Whitt, W. (1992). A tandem fluid network with Lévy input. In *Queueing and Related Models* (ed. U.N. Bhat and I.V. Basawa). Oxford University Press, 112–128.
36. Kelly, F.P. (1991). Effective bandwidths at multi-class queues. *Queueing*

Systems, **9**, 5–16.

37. Kelly, F.P. (1994a). On tariffs, policing and admission control of multiservice networks. *Operations Research Letters*, **15**, 1–9.
38. Kelly, F.P. (1994b). Tariffs and effective bandwidths in multiservice networks. In Labetoulle and Roberts (1994), 401–410.
39. Kelly, F.P. (ed.) (1994c). *Probability, Statistics and Optimisation: a Tribute to Peter Whittle*. Wiley, Chichester.
40. Kelly, F.P. and Williams R.J. (ed.) (1995). *Stochastic Networks*. Springer Verlag, New York.
41. Kesidis, G., Walrand, J. and Chang, C.-S. (1993). Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. Networking*, **1**, 424-428.
42. Kingman, J.F.C. (1993). *Poisson Processes*. Clarendon Press, Oxford.
43. Kulkarni, V.G., Gün, L. and Chimento, P.F. (1995) Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer. *IEEE J. Selected Areas Commun.*, **13**, 1039–1047.
44. Labetoulle, J. and Roberts, J.W. (ed.) (1994). *The Fundamental Role of Teletraffic in the Evolution of Telecommunication Networks*. Elsevier, Amsterdam.
45. Norros, I. (1994). A storage model with self-similar input. *Queueing Systems*, **16**, 387–396.
46. Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, **3**, 213–238.
47. Roberts, J.W. (ed.) (1992). *Performance Evaluation and Design of Multiservice Networks*. Office for Official Publications of the European Communities, Luxembourg.
48. Shwartz, A. and Weiss, A. (1995). *Large Deviations for Performance Analysis: Queues, Communication and Computing*. Chapman and Hall, London.
49. Simonian, A. and Guibert, J. (1995). Large deviations approximation for fluid sources fed by a large number of on/off sources. *IEEE J. Selected Areas Commun.*, **13**, 1017–1027.
50. Sriram, K. and Whitt, W. (1986). Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Selected Areas Commun.*, **4**, 833–846.
51. Tse, D.N.C., Gallager, R.E. and Tsitsiklis, J.N. (1995). Statistical multiplexing of multiple time-scale Markov streams. *IEEE J. Selected Areas in Commun.*, **13**, 1028–1038.
52. Weber, R. (1994). Large deviation and fluid approximations in control of stochastic systems. In Kelly (1994c), 159–171.
53. Weiss, A. (1986). A new technique for analyzing large traffic systems. *Adv. Appl. Prob.*, **18**, 506–532.
54. Whitt, W. (1993). Tail probabilities with statistical multiplexing and ef-

fective bandwidths in multi-class queues. *Telecommunication Systems*, **2**, 71–107.

55. Willinger, W. (1995). Traffic modelling for high-speed networks: theory versus practice. In Kelly and Williams (1995), 395–409.
56. Willinger, W., Taqqu, M.S., Leland, W.E. and Wilson D.V. (1995). Self-similarity in high-speed packet traffic: analysis and modelling of ethernet traffic measurements. *Statistical Science*, **10**, 67–85.