On the Relevance of Time Scales in Performance Oriented Traffic Characterizations *

M. Montgomery and G. de Veciana

Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas 78712-1084

Abstract

A key problem for modern network designers is to characterize/model the "bursty" traffic arising in broadband networks with a view on predicting and guaranteeing performance. In this paper we attempt to unify several approaches ranging from histogram/interval based methods to "frequency domain" approaches by further investigating the asymptotic behavior of a multiplexer carrying a large number of streams. This analysis reveals the salient traffic/performance relationships which should guide us in selecting successful methods for traffic management and network dimensioning.

Keywords

Traffic modeling, performance evaluation, management, congestion and admission control.

1 Introduction

Efficient methods for congestion control in high-speed communication networks will be based on reasonable characterizations for traffic flows and time scale decompositions of the network dynamics. With a view on resolving the question of admission control, including bandwidth allocation and routing, as well as other traffic management activities, researchers have developed several approaches to modeling traffic and predicting performance. The first step towards resolving problems in traffic management on high-speed networks is that of obtaining a reasonable description for traffic statistics that directly relates to the performance characteristics of network elements subject to such loads. Such a characterization should

- provide a concise summary of the traffic statistics (data compression),
- translate to *accurate* performance predictions for network elements, and
- be easy to measure.

The most complete, albeit unwieldy, characterization of a (continuous or discrete time) stationary traffic flow are its statistics. We will let $A(0,t]$ (with $t \in \mathbb{R}^+$ or $t \in \mathbb{N}$) denote the distribution for the cumulative arrivals

*This work was supported by the National Science Foundation under Grant NCR-9409722 and appeared in part at Infocom 96. M. Montgomery is supported by an NSF Graduate Fellowship. Tel: (512) 471-1573 Fax: (512) 471-5532 E-mail: mcm@mail.utexas.edu

(packets or work) over the time interval $(0,t]$. The mean arrivals per unit time are given by $\mu = \mathbb{E}A(0,1]$, where in discrete time this is referenced to the time-slot interval.

Recent analytical and empirical studies have suggested a variety of approaches to capturing traffic characteristics and queuing dynamics. In this paper we attempt to unify these approaches, starting from an understanding of buffer asymptotics. In brief, §2.1, 2.2 review and extend recent work suggesting that the cumulative log-moment generating function

$$\Lambda_t(\theta) := \log \mathbb{E} \exp[\theta A(0,t]], \quad \theta \in \mathbb{R}, t \geq 0,$$

of the arrivals over time intervals is a useful representation of the traffic which translates directly to buffer overflow characteristics [15, 1]. We will see that two parameters θ_{t^*}, t^* define the "space" and time scales of interest to determine overflow probabilities in a buffered link.

Next we consider the case where only the second order characteristics are important in determining performance, and, in particular, the key traffic properties are captured by a marginal distribution and the covariance of the *rate process*. This approach recently developed by Li *et. al.* [13, 17, 14] has shown to be a remarkably good modeling tool. To further study these ideas, in §2.3 we consider Gaussian processes, where the second order characteristics completely determine $\Lambda_t(\theta)$ above. For such processes we show that t^* can be related to a so-called *cutoff frequency* ω_c [17] through a relationship between the variance and the traffic rate processes' power spectral density (PSD) developed in Appendix B.

Finally in §3 we consider various approximations. We argue that for multiplexers with a known time scale t^* of interest, only the distribution of $A(0,t^*]$ is relevant. Predicated on knowing t^*, interval-based methods as suggested by [6, 16] and histogram-based methods [18] may very well achieve high utilizations. Moreover when θ_{t^*} assumes "moderate" values, a Gaussian approximation may in fact suffice to appropriately predict performance. In practice determination of the appropriate time scale t^* depends on traffic characteristics and desired operational constraints, and may in turn be a difficult task.

2 Characterizing packet streams and performance asymptotics

In this section we discuss large deviations and Bahadur-Rao approximations for overflow probabilities in multiplexers supporting large numbers of streams. We briefly consider the accuracy and practical significance of such results. Finally we consider the case of Gaussian arrivals processes where results take a particularly simple form.

2.1 Large deviations

We begin our study by briefly discussing an asymptotic study of a multiplexer supporting a large number N of i.i.d. streams [1, 3]. Let $A_t^N = \sum_{i=1}^N A_i(-t,0]$ denote the aggregate arrivals over an interval of length t. As shown in Fig. 1 resources are also scaled in N, thus we identify a buffer and capacity *per stream*, b, c.

Using large deviations, for large N we can estimate the probability that over a time interval of length t an amount of work A_t^N sufficient to overcome the potential service Nct and further exceed a buffer level Nb enters the system:

$$\mathbb{P}(A_t^N > N(ct+b)) \approx \exp[-N\Lambda_t^*(ct+b)],$$

where $\Lambda_t^*(\alpha) = \sup_\theta[\theta\alpha - \Lambda_t(\theta)]$ [8]. For stationary and ergodic traffic the queue's steady state distribution is given by $Q^N = \sup_{t>0}[A_t^N - Nct]$. Thus the steady state probability of exceeding Nb can be approximated by

$$\mathbb{P}(Q^N > Nb) \approx \sup_{t>0} \mathbb{P}(A_t^N > N(ct+b)) \approx \exp[-N \inf_{t>0} \Lambda_t^*(ct+b)]. \tag{1}$$

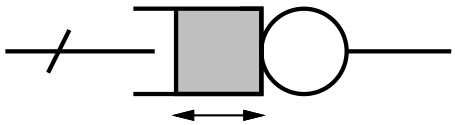

Figure 1: Scaling for large numbers of streams.

This argument is made rigorous in [1].

Intuitively a minimizer $t^* \in \text{arginf}_{t>0}\Lambda_t^*(ct+b)$ would correspond to a likely time scale on which overflows occur in this system. A performance constraint on buffer overflows of the form

$$\mathbb{P}(Q^N > Nb) \approx \exp[-N\Lambda_{t^*}^*(ct^*+b)] \leq \exp[-N\delta] \ \ (\text{e.g.}, \ \sim 10^{-6}),$$

translates to

$$\Lambda_{t^*}^*(ct^*+b) = \sup_\theta[\theta(ct^*+b) - \Lambda_{t^*}(\theta)] = \theta_{t^*}(ct^*+b) - \Lambda_{t^*}(\theta_{t^*}) > \delta,$$

for the associated maximizer θ_{t^*}. In order to satisfy the QoS requirement we need that:

$$c \ > \ \frac{\Lambda_{t^*}(\theta_{t^*})+\delta}{t^*\theta_{t^*}} - \frac{b}{t^*} \ =: \ \hat{\alpha}(c,b,\delta). \tag{2}$$

Notice that this is a normalized constraint depending on the resources allocated per stream and the traffic characteristics on the appropriate time scale t^*. The "space" scale parameter θ_{t^*} determines the impact (of higher order moments, see §3) of the tail distribution of $A(0,t^*]$ on overflows.

Example: Brownian motion model. Suppose the net input per stream is modeled by a Brownian motion with parameters σ^2 and drift $\mu - c < 0$, i.e. $A(0,t] - ct \sim N((\mu-c)t, t\sigma^2)$. Using the formulae in §2.3 one finds that $t^* = b/(c-\mu)$ and $\theta_{t^*} = 2(c-\mu)/\sigma^2$. In this special case the space scale is independent of the buffer size. Moreover (2) gives the following requirement

$$c > \mu + \frac{\delta\sigma^2}{2b}.$$

Recall that this is a normalized constraint, but it exhibits the role of shared buffering. We interpret this relationship as follows: for small buffers we essentially require an infinite capacity; this is due to the fast local variations of Brownian motion. As the buffer b increases, we can essentially get away with serving at the mean rate, but only because we are scaling the number of streams, service rate, and buffer size together, thus maintaining the space scale of interest θ_{t^*} constant while increasing the buffer to infinity! Based on a more refined analysis, one can find the exact invariant distribution and show how this approximation fares versus the finite buffer queue, see e.g. [7].

2.2 Bahadur-Rao improvements

A further improvement upon the large deviations result can be obtained via Bahadur-Rao asymptotics. While the large deviations result estimates the magnitude of the exponent, the improved asymptotics account for possible pre-factor contributions which in this case are found to be order $\sqrt{N}$.

The Bahadur-Rao result for sums of i.i.d. random variables [8, 12] gives the following improvement upon the large deviations bound considered previously:

$$\mathbb{P}(A_t^N > N(ct+b)) \approx \frac{1}{\sigma_t\theta_t\sqrt{2\pi N}}\exp[-N\Lambda_t^*(ct+b)],$$

where $\theta_t = \text{argsup}_\theta[\theta(ct+b) - \Lambda_t(\theta)]$ and $\sigma_t^2 = \frac{\partial^2}{\partial^2\theta}\Lambda_t(\theta_t)$. The large deviations heuristic of the previous section suggests that there is a dominant time scale of interest t^* whence,

$$\begin{aligned}\mathbb{P}(Q^N > Nb) &\approx \mathbb{P}(A^N_{t^*} > N[ct^* + b]) \\ &\approx \exp[-N\Lambda^*_{t^*}(ct^*+b) - \log[\sqrt{2\pi N\sigma^2_{t^*}\theta^2_{t^*}}]\ ]. \end{aligned} \tag{3}$$

The precise statement and proof of this result has been relegated to Appendix A; in general t^* may not be unique and the pre-factor may be off by a constant independent of N. However, we expect this is typically not the case, see Remark A.1.

A further simplification gives an ad hoc approximation, which is exact for the case of Gaussian arrivals processes. Assuming the log-moment has the required derivatives (true if arrivals are almost surely bounded), and noting that $\frac{\partial\Lambda_{t^*}(\theta_{t^*})}{\partial\theta} = ct^* + b$ we have by applying Taylor's Theorem at θ_{t^*} that

$$\begin{aligned} 0 = \Lambda_{t^*}(0) &= \Lambda_{t^*}(\theta_{t^*}) - \theta_{t^*}\ \frac{\partial\Lambda_{t^*}(\theta_{t^*})}{\partial\theta} + \frac{\theta^2_{t^*}}{2}\ \frac{\partial^2\Lambda_{t^*}(\theta_{t^*})}{\partial\theta^2} + R \\ &= -\Lambda^*_{t^*}(ct^*+b) + \frac{\theta^2_{t^*}\sigma^2_{t^*}}{2} + R, \end{aligned}$$

where $R = -\frac{\theta^3_{t^*}}{3!}\ \frac{\partial^3\Lambda_{t^*}(r)}{\partial\theta^3}$ for some $r \in (0, \theta_{t^*})$. Thus if the remainder R is small ($R = 0$ for Gaussian distributions) we then have that $2\Lambda^*_{t^*}(ct^*+b) \approx \theta^2_{t^*}\sigma^2_{t^*}$ giving the following approximation:

$$\mathbb{P}(Q^N > Nb) \approx \exp[-N\Lambda^*_{t^*}(ct^*+b) - \frac{1}{2}\log[4\pi N\Lambda^*_{t^*}(ct^*+b)]\ ] \tag{4}$$

Note that in (4) no further computation is needed to obtain the pre-factor since it depends on N and the large deviations exponent. The plots in Fig. 2 show that the simple ad-hoc approximation (4) is essentially the same as (3) and gives a marginal improvement upon the large deviations bounds (1) in [1]. In particular we base these directly on their results for superpositions of On/Off Markov fluids. For each source, discrete-time transitions from Off to On occur with probability a, from On to Off with probability d. When the source is On a unit of work is generated, and no work is generated in the Off state. Note that the sources in the three superpositions range from bursty ($a + d < 1$) to sub-bursty ($a + d > 1$), see [1] for details.

A QoS requirement on buffer overflows of the type $\mathbb{P}(Q^N > Nb) \leq \exp[-N\delta]$ translates to an adjusted operational constraint:

$$c > \hat{\alpha}(\delta', c, b) \quad \text{where} \quad \delta' = \delta - \frac{1}{2N}\log[2\pi N\theta^2_{t^*}\sigma^2_{t^*}] \approx \delta - \frac{1}{2N}\log[4\pi N\Lambda^*_{t^*}(ct^*+b)].$$

Notice that the constraint now depends on N and can be expressed as a simple adjustment in the target QoS δ which accounts for further multiplexing effects.

This result suggests when large deviations asymptotics might give accurate estimates. In particular suppose the target QoS is $\exp[-N\delta]$, and we are willing to put up with less than one (0.99) order of magnitude error. In this case, assuming the ad-hoc approximation holds, we can show that, *very roughly*, QoS requirements in the range $.98 > \exp[-N\delta] > 3.4\times10^{-4}$ can be handled via large deviations asymptotics. Perhaps a more telling remark is the apparent tradeoff between N and δ in determining the accuracy, that is based on this discussion we might argue that for sufficiently "large" N and "small" δ or vice-versa, good approximations are obtained. These types of effects are born out by simulation [11] and indicate in part why some performance studies using effective bandwidths give excellent results while others give rather bad performance estimates [2, 10].

For simplicity we have only discussed the case of multiplexing N homogeneous streams. For heterogeneous mixes of traffic, we can consider i.i.d. sources which are each a mix of the appropriate number

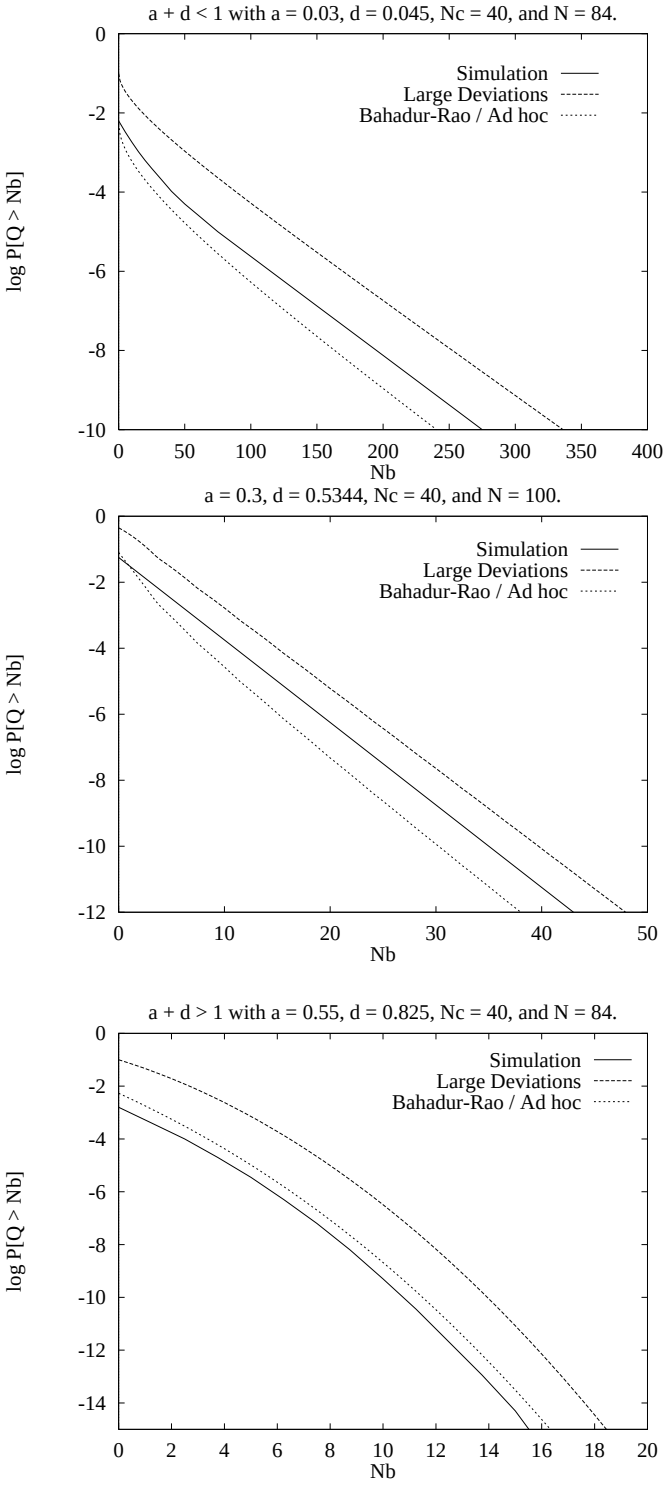

Figure 2: Bahadur-Rao pre-factor corrections.

of heterogeneous streams, see [12], and the asymptotics follow immediately. Thus, qualitatively, the picture for mixes of heterogeneous sources is quite similar to the i.i.d. case. For completeness we note a related work [9] where a slightly different approach to approximation was taken.

2.3 Gaussian processes and frequency domain characteristics

Gaussian traffic models for the net input into a queuing system often arise as heavy traffic approximations or aggregation limits. Herein we take the point of view that they adequately model more general overflow characteristics, and later look at the possibility of using Gaussian processes, which are more amenable to analysis, to approximate non-Gaussian traffic.

Suppose the underlying arrivals rate process is Gaussian. In this case $A(0,t]$ is Gaussian and the cumulative log-moment generating function and rate function can be computed explicitly as a function of $\mu = \mathbb{E}A(0,1]$ and $\sigma_t^2 = \mathrm{Var}(A(0,t])$:

$$\Lambda_t(\theta) = \mu t\theta + \frac{\theta^2\sigma_t^2}{2} \quad \text{and} \quad \Lambda_t^*(ct+b) = \frac{[(c-\mu)t+b]^2}{2\sigma_t^2}.$$

In the sequel we will also use the fact that the variance can in turn be expressed as the filtered power spectral density of the rate process; see (13),(14).

Fortunately for Gaussian traffic we can find revealing expressions for θ_{t^*}, t^*, the space-time scales of interest in (2). In particular from (1) and assuming the minimizer t^* is unique we write

$$t^* = \mathrm{argmax}_{t>0}\frac{\sigma_t^2}{[(c-\mu)t+b]^2} = \mathrm{argmax}_{t>0}\frac{\sigma_t^2}{[t+\alpha]^2} \quad \text{and} \quad \theta_{t^*} = \frac{(c-\mu)t^*+b}{\sigma_{t^*}^2} \tag{5}$$

where $\alpha = b/(c-\mu)$. For a Gaussian process with independent increments, i.e. $\sigma_t^2 = t\sigma^2$, the maximum is attained at $t^* = \alpha$ (the discrete time case needs to be appropriately quantized) and is independent of the input variance. We can also write $\alpha = d/(1-\rho)$ where $d = b/c$ is to be understood as the maximum delay permitted for packets that are successfully transmitted while ρ is the utilization of the buffer. The key point then is that α is an invariant, in the sense that for fixed α we get the same overflow t^* behavior, so we write $t^*(\alpha)$.

Introducing these quantities into the QoS constraint (2) gives the following simplified requirement:

$$c + \frac{b}{t^*} \;>\; \mu + \sqrt{2\delta\frac{\sigma_{t^*}^2}{t^{*2}}} \tag{6}$$

to be interpreted as a bandwidth c constraint to meet the QoS δ-constraint given a buffer of size b. We can interpret b/t^* as an effective increase in the capacity due to buffering traffic. Now using the alternative representation for variance (14) for continuous time processes (or (13) in discrete time) we have that

$$\frac{\sigma_{t^*}^2}{t^{*2}} \;=\; \int_{-\infty}^{+\infty}\frac{2\sin^2(\omega t^*/2)}{\pi(\omega t^*)^2}S_A(\omega)d\omega \;=\; \frac{1}{2\pi}\int_{-\omega_c}^{+\omega_c}S_A(\omega)d\omega, \tag{7}$$

where $\omega_c = 2\pi/t^*$ corresponds to a rough *cutoff frequency*. This approximation assumes that due to the $1/\omega^2$ decay of the filtering function the PSD does not contribute significantly beyond the cutoff frequency, and the gain is roughly that at DC, i.e. $1/2\pi$. In practice the effectiveness of this approximation will of course depend on the characteristics of the PSD, i.e. it should have a well-founded low frequency component. This approximation corresponds to allocating bandwidth based on the power in the output process from a low pass filter with cutoff at ω_c [17].

Example: The Ornstein-Uhlenbeck (OU) model reveals the basic behavior of a positively correlated arrival process with exponentially decaying correlations. Its characteristics are similar to bursty superpositions of On/Off Markov sources or first order auto-regressive models. Indeed the "rate" (velocity) process is Gaussian with mean μ and exponentially decaying covariance $k_A(\tau) = v \exp[-a|\tau|]$ for some constants $v, a > 0$. From (14) one can show that

$$\sigma_t^2 = \frac{2v}{a^2}(at - 1 + \exp[-at]). \tag{8}$$

Using (5) one can compute t^* numerically and obtain the following simple bounds:

$$\alpha \leq t^* \leq \alpha + \frac{2}{a}. \tag{9}$$

Thus the overflow time scale for positively correlated processes will exceed that of i.i.d. arrivals, α, by no more than twice the correlation time scale $1/a$ — positive correlations lead to "slower" overflows which exploit dependencies.

Using (5) and (6) in the case of the OU model we can investigate the maximum utilization ρ_{max} that can be achieved as a function of the loss constraint δ and the buffer size b in the multiplexer. Fig. 3 below exhibits the QoS vs. efficiency tradeoffs for an OU input process with parameters $\mu = 0.5$, $a = 1$, and $v = 1$.

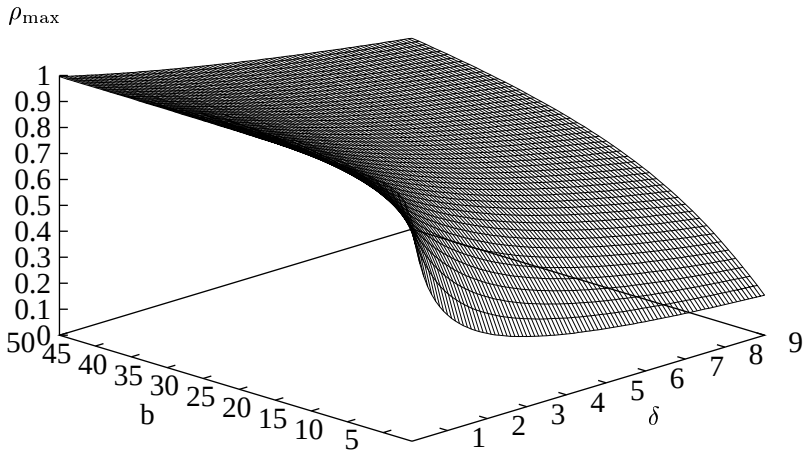

Figure 3: Efficiency versus buffer size and loss constraints.

Next we consider the impact of the traffic's parameters v, a. Motivated by the simulations in [13] for aggregations of On/Off sources, we consider two regimes: one in which the buffer is sufficiently large to effectively smooth correlations and one in which it is not. Consider (8); when $at^*(\alpha) < 1$, we can approximate $\sigma^2_{t^*(\alpha)} \approx vt^{*2}(\alpha)$. Inserting these approximations into (6) we find that the maximum utilization is approximately given by

$$\rho_{max} \approx \frac{1}{1 - ba/\mu + \sqrt{2\delta v}/\mu}.$$

The right hand side is found to vary slowly once a is sufficiently small, meaning that the buffer will no longer be effective. Numerically we can show that $at^*(\alpha) < 1$ when $\alpha = b/(c-\mu) < 0.1/a$ in which case we say that we are in the buffer non-effective regime, see [13]. Alternatively, we find that the buffer is effective once

$$d > 0.1 * 1/a,$$

i.e. the "maximum" delay exceeds a fraction of the correlation time scale.

Finally, what do these asymptotics tell us about the case with a *single* Gaussian input stream? Suppose for example that a single (aggregate) Gaussian arrival process enters a buffer link of size b with capacity c. We can represent it as a sum of N independent Gaussian processes with means and variances which are scaled by N^{-1}. The asymptotic results presented previously, for appropriately rescaled buffers and bandwidth per sub-stream, shows that the overflow asymptotic is independent of N. This probability of overflow is given by

$$\mathbb{P}(Q > b) \approx \frac{1}{\sqrt{4\pi\Lambda^*_{t^*}(ct^* + b)}} \exp[-\Lambda^*_{t^*}(ct^* + b)],$$

where all quantities refer to the single stream entering the link.

3 Approximations for non-Gaussian processes

In general the focus of traffic modeling should lie, not in precise modeling of the traffic statistics, but rather in capturing the "relevant" characteristics, so as to allow prediction or efficient management.

3.1 Approximating t^*

The previous discussion brings to the fore the importance of estimating t^* (or cutoff frequency) associated with a given buffered traffic load. One approach is to simulate or monitor a buffer subject to the desired traffic load and attempt to estimate t^* based on observation of the queuing dynamics. A second approach, which we believe will be effective, is to estimate t^* based on the observation of second order traffic characteristics. That is, based on estimates for the mean and variance of $A(0,t]$, say $\hat{\mu}, \hat{\sigma}^2_t$, we can numerically estimate $t^*(\alpha)$ from (5) for a variety of possible α. Alternatively, as suggested previously, the magnitude of t^* might be estimated via bounds such as (9) or heuristics such as inspection of the PSD to determine an appropriate cutoff frequency. Typically the PSD of traffic, such as VBR video, will have the majority of its power concentrated in a well founded low frequency component [17], permitting a rough evaluation of the essential time scales, e.g. scene changes, frame correlations, or picture blocks. Such time scales however depend on the type of compression and nature of the media, thus teleconferencing applications have different scales than say MPEG coded video [9].

3.2 Interval-based approximations for bandwidth requirements

Once an estimate for t^* is available, we may consider using interval-based traffic descriptors and bandwidth allocation schemes such as [7, 18, 16, 17]. There are at least two "simple" options: 1) to approximate $A(0,t^*]$ by a Gaussian distribution, and 2) to allocate the peak rate on the appropriate time scale.

We argue that $A(0,t^*]$ can be for practical purposes assumed to be Gaussian if

$$\Lambda_{t^*}(\theta) \approx \mu t^* \theta + \frac{\sigma^2_{t^*}\theta^2}{2} + R(\theta), \quad \forall \theta \in (0, \theta_{t^*})$$

where $\sigma^2_{t^*} = \mathrm{Var}(A(0,t^*])$. This approximation follows from a Taylor expansion of $\Lambda_{t^*}(\theta)$ at $\theta = 0$ and is accurate if the remainder R depending on θ_{t^*} and higher order moments of $A(0,t^*]$ is relatively small. A rough way of establishing if such an approximation is reasonable is to consider whether θ_{t^*}, as given by (5), is small ($\theta_{t^*} < 1$). If such an approximation is well founded we can use the estimates based on Gaussian distributions, e.g. (3) or (6) to evaluate the performance of the multiplexer.

When second order approximations break down, we can resort to peak rate allocation. Starting from the large deviations QoS requirement (2), we note that $\Lambda_{t^*}(\theta)$ is convex and eventually increasing so the right hand side is eventually non-decreasing in θ. This gives the following conservative requirement:

$$c + b/t^* > \lim_{\theta\to\infty} \frac{\log \mathbb{E}\exp\left[\theta(A(0,t^*]\right]}{\theta t^*} = \alpha(\infty, t^*). \tag{10}$$

The quantity $\alpha(\infty, t^*)$ is the *sustainable peak rate* over an interval of length t^*, i.e. the supremum over rates that can be sustained with non-zero probability; see [5] for a discussion in the case where $t^* \to \infty$. In general $\alpha(\infty, t^*)$ may be infinite, e.g. Gaussian distributions, however in practice traffic will have a bounded sustainable peak rate for any time scale since there are physical bounds on the ability of a source to sustain a maximal rate during a given period of time. Thus (10) corresponds to peak rate allocation on the appropriate time scale with b/t^* reflecting an increase in capacity due to buffering. Alternatively we can interpret (10) as allocating the peak rate of a low-pass filtered (cutoff ω_c) rate process [17].

3.3 Relevance of time scale

This work substantiates the premise that traffic needs to be modeled on the appropriate time scale. By developing methods for identifying the "relevant" time scales, we can in turn parametize traffic management and network dimensioning so as to make them effective over a wide range of traffic statistics. We expect that as traffic characteristics stabilize, buffer sizes and link capacities become standard, and the required QoS constraints are determined, estimating the relevant times scales will become an easier task.

Acknowledgment: We thank N. Duffield and S.Q. Li for many interesting discussions related to this topic.

A Bahadur-Rao asymptotics

In the theorem below we will assume the following hypotheses taken from [1].

Hypothesis A.1 [1]

1. *For each $\theta \in \mathbb{R}$, $\lambda_t(\theta) = t^{-1}\Lambda_t(\theta)$ and $\lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta)$ exist as extended real numbers.*
2. *λ_t and λ are essentially smooth.*
3. *There exists $\theta > 0$ for which $\lambda_t(\theta) < 0$ for all $t \in \mathbb{Z}^+$.*
4. *λ_t is second order differentiable in its domain D_{λ_t}.*

Roughly speaking the first two permit us to use the Gärtner-Ellis large deviations result. The third corresponds to a stability condition, i.e. more departures than arrivals on any time interval t. Finally, an additional hypothesis was introduced so as to use the Bahadur-Rao result. In the theorem below we distinguish between the lattice case, which for example corresponds to the case where the arrivals process $A(0,t]$ counts fixed sized packets, and the non-lattice case.

Theorem A.1 *Assuming Hypothesis A.1 and assuming $A(0,t]$, $\forall t \in \mathbb{Z}^+$, are non-lattice random variables then*

$$1 \;\leq\; \liminf_{N\to\infty}\; c_N^*\mathbb{P}(Q^N > Nb) \;\leq\; \limsup_{N\to\infty}\; c_N^*\mathbb{P}(Q^N > Nb) \;\leq\; K.$$

Let $\mathcal{T}^* = \text{arginf}_{t\in\mathbb{Z}^+}\Lambda_t^*(ct+b) = \text{arginf}_{t\in\mathbb{Z}^+} t\lambda_t^*(c+(b/t))$. *In the above, K is a constant independent of N,* $t^* = \text{argmin}_{t\in\mathcal{T}^*} c_N(t)$, $c_N(t) = \sigma_t\theta_t\sqrt{2\pi N}\exp[N\Lambda_t(ct+b)]$, $c_N^* = c_N(t^*)$, $\theta_t = \text{argsup}_\theta[\theta(ct+b)-\Lambda_t(\theta)]$, *and* $\sigma_t^2 = \frac{\partial^2}{\partial^2\theta}\Lambda_t(\theta_t)$ *(derivative exists by Hyp. A.1.4).*

Suppose $A(0,t]$ (e.g. packet arrivals), $\forall t \in \mathbb{Z}^+$ has a lattice law, i.e. for some a_0, d, the random variable $(A(0,t]-a_0)/d$ is an integer number, and d is the largest number with this property. Then we have

$$\frac{\theta_{t^*}d}{1-\exp[-\theta_{t^*}d]} \;\leq\; \liminf_{N\to\infty}\; c_N^*\mathbb{P}(Q^N > Nb) \;\leq\; \limsup_{N\to\infty}\; c_N^*\mathbb{P}(Q^N > Nb) \;\leq\; \frac{K\theta_{t^*}d}{1-\exp[-\theta_{t^*}d]}.$$

Proof: We show the proof for the non-lattice case. The proof for the lattice-case is similar. By noting that $\Lambda_t^*(ct+b) \geq 0$ and $\Lambda_t^*(ct+b) \overset{t\to\infty}{\to} \infty$, we see that $\inf_{t>0}\Lambda_t^*(ct+b), t\in\mathbb{Z}^+$ is attained. Furthermore, $\mathcal{T}^*$ has a finite number of elements K, independent of N.

Lower bound: $\mathbb{P}(Q^N > Nb) \geq \mathbb{P}(A_{t^*}^N > N[ct^*+b])$ and using the Bahadur-Rao result

$$\lim_{N\to\infty} c_N^*\mathbb{P}(A_{t^*}^N > N[ct^*+b]) \;=\; 1.$$

Upper bound: $\mathbb{P}(Q^N > Nb) \leq \sum_{t>0}\mathbb{P}(A_t^N > N[ct+b])$ by the union bound. Using the Bahadur-Rao result,

$$\lim_{N\to\infty} c_N^*\sum_{t>0}\mathbb{P}(A_t^N > N[ct+b]) \;\leq\; K + \lim_{N\to\infty} c_N^*\sum_{t>0,t\notin\mathcal{T}^*}\mathbb{P}(A_t^N > N[ct+b]).$$

Since $\Lambda_{t^*}^*(ct^*+b) < \Lambda_t^*(ct+b)$ for $t\notin\mathcal{T}^*$, it can be shown that $\lim_{N\to\infty} c_N^*\mathbb{P}(A_t^N > N[ct+b]) = 0$ for $t\notin\mathcal{T}^*$.

Consider $t' > \max\{t|t\in\mathcal{T}^*\}$. We aim to show

$$\lim_{N\to\infty} c_N^*\sum_{t>t'}\mathbb{P}(A_t^N > N[ct+b]) \;=\; 0.$$

By Chernoff's bound,

$$c_N^*\sum_{t>t'}\mathbb{P}(A_t^N > N[ct+b]) \;\leq\; c_N^*\exp[-\theta Nb]\sum_{t>t'}\exp[-\theta Nct+t\lambda_t(\theta)].$$

Since $\lambda_t(\theta) \overset{t\to\infty}{\to} \lambda(\theta)$ (Hyp. A.1.1), and $\lambda(\theta) < 0$ on $(0,\delta)$ (Hyp. A.1.3), we can find $\theta > 0$ and $\epsilon < 0$ such that $\lambda_t(\theta) < \epsilon$ for t' sufficiently large. This means that for such $t > t'$, the geometric series is summable, so we have

$$\begin{aligned}
c_N^*\exp[-\theta Nb]\sum_{t>t'}\exp[-\theta Nct+t\lambda_t(\theta)] &\leq c_N^*\exp[-\theta Nb]\sum_{t>t'}\exp[-\theta Nct+t\epsilon] \\
&= c_N^*\exp[-\theta Nb]\frac{\exp[t'(-\theta Nc+\epsilon)]}{1-\exp[-\theta Nc+\epsilon]} \\
&= \frac{\sigma_{t^*}\theta_{t^*}\sqrt{2\pi N}\exp[N(\Lambda_{t^*}^*(ct^*+b)-\theta[ct'+b])+t'\epsilon]}{1-\exp[-\theta Nc+\epsilon]}.
\end{aligned}$$

For t' sufficiently large, $\Lambda_{t^*}^*(ct^*+b) < \theta[ct'+b]$. Now taking limits as $N\to\infty$, the result follows. □

Remark A.1 *Note that we would typically expect the set $\mathcal{T}^*$ corresponding to the most likely overflow time scales to consist of a unique point. It is relatively easy to show this is the case for traffic processes with independent increments, or special cases such as Markov Fluids or the OU processes. We have been unable however to obtain a reasonable set of assumptions on the traffic processes such that this is indeed the case.*

The case of continuous-time processes is somewhat more involved. Note that the upper bound from the discrete time case has been multiplied by two.

For brevity, we will use the notation $W_t = A_t - ct$ corresponding to the net input process and as before we denote the sum of N independent copies by W_t^N.

Hypothesis A.2 (See [1]) *For all $t \geq r \geq 0$ define $\tilde{W}_{t,r} = \sup_{0<|r'|<r} |W_{t-r'} - W_t|$. Then for all $\theta \in \mathbb{R}$*

$$\limsup_{r \to 0} \sup_{t \geq 0} \log \mathbb{E}[\exp(\theta \tilde{W}_{t,r})] = 0.$$

Hypothesis A.3 *For $t \in \mathbb{R}$ we assume that $|\mathcal{T}^*|$ is finite, that is, there are only a finite number of t^* in the set $\mathcal{T}^* = \operatorname{arginf}_{t>0} \Lambda_t(ct+b)$.*

Proof of Upper Bound in Theorem A.1 for $t \in \mathbb{R}$:

This argument is based on that in [1]. For any $\epsilon > 0$ and $n \in \mathbb{N}$ define

$$\hat{W}_n^N = \sup_{(n-1)\epsilon < t \leq n\epsilon} W_t^N \quad \text{and} \quad \hat{\lambda}_n^N = \frac{1}{nN} \log \mathbb{E}[\exp(\theta \hat{W}_n^N)]$$

and for later convenience we define

$$\check{W}_n^N = \sup_{(n-1)\epsilon < t \leq (n+1)\epsilon} W_t^N.$$

Noting that

$$\hat{W}_n^N \leq W_{n\epsilon}^N + \sup_{0<|r'|<\epsilon} |W_{n\epsilon - r'}^N - W_{n\epsilon}^N| \leq W_{n\epsilon}^N + \tilde{W}_{n\epsilon,\epsilon}^N$$

by Hölder's inequality we have that for any $p \in (0,1)$,

$$n\hat{\lambda}_n^N(\theta) \leq n\epsilon p \lambda_{n\epsilon}(\theta/p) + (1-p) \log \mathbb{E}[\exp\left(\frac{\theta \tilde{W}_{n\epsilon,\epsilon}}{1-p}\right)]. \tag{11}$$

Using Hypothesis A.2, for any $p \in (0,1)$, we can make the second term on the right hand side of (11) as small as we like by choosing ϵ sufficiently small.

The goal is to show

$$\limsup_{N \to \infty} c_N^* \mathbb{P}(Q^N > Nb) \leq 2|\mathcal{T}^*|.$$

For a fixed small enough ϵ and large $t' > \max\{t | t \in \mathcal{T}^*\}$ we can write

$$\mathbb{P}(\sup_{t>0} W_t^N > Nb) \leq \mathbb{P}(\sup_{n>0} \hat{W}_n^N > Nb) \leq$$
$$\underbrace{\sum_{n \in \mathcal{T}_\epsilon^*} \mathbb{P}(\check{W}_n^N > Nb)}_{A} + \underbrace{\sum_{n > \lceil t'/\epsilon \rceil} \mathbb{P}(\hat{W}_n^N > Nb)}_{B} + \underbrace{\sum_{n \notin \mathcal{T}^*, n \leq \lceil t'/\epsilon \rceil} \mathbb{P}(\hat{W}_n^N > Nb)}_{C}$$

where $\mathcal{T}_\epsilon^* = \{n \in \mathbb{N} \mid (n-1)\epsilon < t < (n+1)\epsilon \text{ for some } t \in \mathcal{T}^*\}$. By the definition of $\mathcal{T}^*$ for all t not within ϵ of $\mathcal{T}^*$ we have that

$$\Lambda_t^*(ct+b) > \Lambda_{t^*}^*(ct^*+b). \tag{12}$$

Using an argument similar to the discrete time case one shows that for t' large enough (guaranteeing the summability of the geometric series) and the fact that exponents decay faster than the optimal (12) that $c_N^* B$ and $c_N^* C \to 0$ as $N \to \infty$.

The terms in $c_N^* A$ correspond to intervals containing times in $\mathcal{T}^*$. For each $t^* \in \mathcal{T}^*$ will have a neighborhood $(n^*-1)\epsilon < t^* < (n^*+1)\epsilon$ with $n^* \in \mathcal{T}_\epsilon^*$. To conclude the proof we can upper bound $\check{W}_{n^*}^N$ by sums of i.i.d. random variables,

$$\check{W}_{n^*}^N \leq W_{t^*}^N + \tilde{W}_{t^*,2\epsilon}^N.$$

and apply the Bahadhur-Rao result to the latter,

$$\check{c}_{N,\epsilon}(t^*) P(W_{t^*}^N + \tilde{W}_{t^*,2\epsilon}^N > Nb) = 1 \quad \text{as} \quad N \to \infty$$

where $\check{c}_{N,\epsilon}(t^*)$ are appropriately defined. Next we show that $\check{c}_{N,\epsilon}(t^*) \to c_N(t^*)$ as $\epsilon \to 0$. This follows by bounding the cumulative log moment generating function as in (11) and letting $\epsilon \to 0$ and then $p \to 1$. All the terms in A have similar properties, and there are at most $2|\mathcal{T}^*|$ such terms, thus we have shown that $\limsup_{N\to\infty} c_N^* A \leq 2|\mathcal{T}^*|$. □

B Representing variance via power spectral density

This appendix discusses a representation of the variance of an arrivals process $A(0,t]$ in terms of the power spectral density of its *rate process*. This permits interpreting performance and traffic characteristics in terms of standard frequency domain concepts.

B.1 Discrete-time processes

Let $\{R_A(t), t \in \mathbb{Z}\}$ be a stationary process denoting the packets (or work) arriving per time slot. Thus in this case, $A(0,t] = \sum_{i=1}^t R_A(i)$. The covariance is given by $k_A(\tau) = \mathbb{E}[(R_A(t)-\mu)(R_A(t+\tau)-\mu)]$, from which we define the transform pairs and corresponding power spectral density

$$S_A(e^{j\Omega}) = \sum_{\tau=-\infty}^{+\infty} k_A(\tau) e^{-j\Omega\tau} \quad \text{and} \quad k_A(\tau) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} S_A(e^{j\Omega}) e^{j\Omega\tau} d\Omega.$$

The spectral density is periodic with period 2π, for the frequency has been renormalized with respect to the slot time interval. Also we can show that

$$\mathrm{Var}(A(0,t]) = \frac{1}{2}tk_A(0) + \sum_{\tau=-t}^{+t} \frac{1}{2}(t-|\tau|)k_A(\tau) = \frac{1}{2}tk_A(0) + \frac{1}{2\pi}\int_{-\pi}^{+\pi} \frac{\sin^2(\Omega t/2)}{2\sin^2(\Omega/2)} S_A(e^{j\Omega}) d\Omega \tag{13}$$

using the definition of the covariance and power spectral density.

B.2 Continuous-time processes

Consider $A()$ as a random measure on $\mathbb{R}$; for random measures which are absolutely continuous with respect to Lebesgue measure, the cumulative arrivals over intervals can be represented as $A(0,t] = \int_0^t R_A(\tau) d\tau$ where

$\{R_A(t), t \in \mathbb{R}\}$ is a stationary random process corresponding to the *rate* of arrival of packets (or work). We can define $k_A(\tau) = \mathbb{E}[(R_A(t) - \mu)(R_A(t+\tau) - \mu)]$ to be the covariance of the arrivals *rate process*. In this case we define Fourier transform pairs and corresponding power spectral density

$$S_A(\omega) = \int_{-\infty}^{+\infty} k_A(\tau) e^{-j\omega\tau} d\tau \quad \text{and} \quad k_A(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S_A(\omega) e^{j\omega\tau} d\omega.$$

The variance of the cumulative arrivals on a time interval can then be expressed as

$$\mathrm{Var}(A(0,t]) = \int_{-t/2}^{+t/2} \int_{-t/2}^{+t/2} k_A(\tau - \gamma) d\tau d\gamma = \int_{-\infty}^{+\infty} \frac{2\sin^2(\omega t/2)}{\pi\omega^2} S_A(\omega) d\omega \tag{14}$$

by using stationarity and Parseval's relation.

The above mentioned continuity condition is somewhat restrictive, e.g. it includes Markov fluids and Gaussian processes with continuous sample paths but excludes point processes and white Gaussian noise. The natural identities corresponding to such processes can be developed in a more abstract setting, see e.g. [4, page 411].

B.3 Comments

While for stationary Gaussian processes the covariance and mean suffice to specify the traffic statistics, in general one should use caution in using second order properties for other processes. As seen in (13), (14), the variance of the processes is obtained by low pass filtering the PSD. As a final comment, we note that processes with long-range correlations have a covariance function which is not absolutely summable, leading to PSD's having most of their power concentrated about a singularity at zero frequency.

References

[1] D.D. Botvich and N.G. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. Technical Report DIAS-APG-94-12, Dublin Institute for Advanced Studies, 1994.

[2] G.L. Choudhury, D.M. Lucantoni, and W. Whitt. Squeezing the most out of ATM. Preprint, 1993.

[3] Costas Courcoubetis and Richard Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Prob.*, September 1996. To appear.

[4] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer-Verlag, 1988.

[5] G. de Veciana, C. Courcoubetis, and J. Walrand. Decoupling bandwidths for networks: A decomposition approach to resource management for networks. In *Proc. IEEE INFOCOM*, volume 2, pages 466–474, 1994. Also submitted to IEEE/ACM Trans. Networking.

[6] G. de Veciana and J. Walrand. Traffic shaping for ATM networks: Asymptotic analysis and simulations. Submitted to *IEEE/ACM Trans. Networking*, 1992.

[7] G. de Veciana and J. Walrand. Effective bandwidths: Call admission, traffic policing and filtering for ATM networks. *Queueing Syst.*, 1995. To appear.

[8] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Jones & Bartlett, Boston, 1992.

[9] A. Elwalid, D. Heyman, T.V. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE J. Selec. Areas Commun.*, 13(6):1004–16, Aug. 1995.

[10] A.I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking*, 1:329–43, 1993.

[11] C-F. Fong, 1994. Personal communication.

[12] I. Hsu and J. Walrand. Admission control for ATM networks. In *Proc. IMA Workshop on Stochastic Networks*, 1994.

[13] C-L. Hwang and S-Q. Li. On input state space reduction and buffer noneffective region. In *Proc. IEEE INFOCOM*, pages 1018–28, June 1994.

[14] Chia-Lin Hwang and S-Q. Li. On the convergence of traffic measurement and queueing analysis: A statistical-match queueing (SMAQ) tool. In *Proc. IEEE INFOCOM*, 1995.

[15] F.P Kelly. Modelling ATM networks: Effective bandwidths, pricing and admission control. Plenary talk INFORMS, Mar. 1995.

[16] E.W. Knightly and H. Zhang. Traffic characterization and switch utilization using a deterministic bounding interval dependent traffic model. In *Proc. IEEE INFOCOM*, 1995.

[17] S.-Q. Li, S. Chong, and C.-L. Hwang. Link capacity allocation and network control by filtered input rate in high-speed networks. *IEEE/ACM Trans. Networking*, 3(1), 1995.

[18] P. Skelly, M. Schwartz, and S. Dixit. A histogram based model for video traffic behavior in an ATM multiplexer. *IEEE/ACM Trans. Networking*, 1:446–59, Aug. 1993.