A NONSTATIONARY OFFERED-LOAD MODEL FOR PACKET NETWORKS

N. G. Duffield[1] William A. Massey[2]
AT&T Labs Bell Laboratories
Ward Whitt[3]
AT&T Labs

March 16, 1998

Abstract

Motivated by the desire to appropriately account for complex features of network traffic revealed in traffic measurements, such as heavy-tail probability distributions, long-range dependence, self similarity and nonstationarity, we propose a nonstationary offered-load model. Connections of multiple types arrive according to independent nonhomogeneous Poisson processes, and general bandwidth stochastic processes (not necessarily Markovian) describe the individual user bandwidth requirements at multiple links of a communication network during their connections. We obtain expressions for the moment generating function, mean and variance of the total required bandwidth of all customers on each link at any designated time. We justify Gaussian approximations by establishing a central limit theorem for the offered-load process. We also obtain a Gaussian approximation for the time-dependent buffer-content distribution in an infinite-capacity buffer with constant processing rate. The offered-load model can be used for predicting future bandwidth requirements; we then advocate exploiting information about the history of connections in progress.

Key Words: ATM, IP, source traffic models, communication networks, nonstationary offered-load models, congestion control, overload control, statistical multiplexing, Gaussian approximation, fluid queues, time-dependent behavior

[1] Room A175, AT&T Labs, 180 Park Avenue, Building 103, Florham Park, NJ 07932-0971, duffield@research.att.com

[2] Room 2C-320, Bell Laboratories, Murray Hill, NJ 07974-0636; will@research.bell-labs.com

[3] Room A117, AT&T Labs, 180 Park Avenue, Building 103, Florham Park, NJ 07932-0971, wow@research.att.com

1. Introduction

In the design and control of packet networks, it is important to appropriately account for complex features of network traffic revealed by traffic measurements. Traffic measurements have revealed heavy-tailed probability distributions, long-range dependence and self similarity; e.g., see Pawlita [34], Cáceres, Danzig, Jamin and Mitzel [7], Leland, Taqqu, Willinger and Wilson [25], Paxson and Floyd [35] and Willinger, Taqqu and Erramilli [37]. Also very important in the longer time scale of connection times is nonstationarity. As in the past, current network traffic measurements reveal a strong time-of-day effect.

In this paper, we propose a framework to capture all these features. In particular, we propose a nonstationary offered-load model. It is intended to describe the total bandwidth needed by all customers as a function of time, given a specification of the individual customer behavior. The offered load is the total required bandwidth that customers would use if there were no constraints, i.e., if there were always enough available bandwidth. We focus on the offered load unaltered by constraints because it is considerably more tractable than the required bandwidth after it has been modified by congestion, e.g., by loss, delay and congestion control algorithms such as TCP. We also focus on the offered load because we believe it can be very useful for network design and control.

We address traffic complexity in two ways. First, we allow the arrival rate of connection requests to be time dependent, in order to be able to capture potentially important time-of-day effects. Second, we allow very general "bandwidth" stochastic processes to represent user bandwidth requirements during their connections. Our model allows a rich class of bandwidth processes for active connections, including on-off models with general (possibly heavy-tail) on-time and off-time distributions, and hierarchical models with multiple sessions, each containing multiple flows, each containing multiple packets. Moreover, our analysis shows that much can be done with a limited partial characterization of these bandwidth processes. In particular, much can be done with only means, variances, and covariances.

Both in the long run (for design) and in the short run (for control), the offered-load models can help take measures to ensure that supply is adequate to meet demand. The general idea is to make decisions based on the probability that the instantaneous demand will exceed supply, or will exceed some other target level, at the time of interest. (Using the probability that demand exceeds supply is tantamount to focusing on the time-dependent loss probability in a bufferless queue.) We also develop an approximation for the time-dependent buffer-content distribution when this input comes to a queue with an infinite-capacity buffer and a constant output rate,

A key assumption here is that *customers* (users or connections) arrive according to a nonhomogeneous Poisson process. The Poisson process structure greatly helps achieve tractability and, at the same time, is realistic. It is important to note here that the Poisson property is *not* being assumed for packets, flows (collections of packets) or even sessions (collections of flows), but *only* for connections (which may include multiple sessions and flows). Traffic measurements show that it is reasonable to model the arrival process of connections as a Poisson process, e.g. see Paxson and Floyd [35]. Indeed, as is well known in telephony, it is natural to regard user connection requests as a Poisson process, because the connection-request process is the superposition of independent processes associated with individual users, where each user process tends to contribute only one point or only a few widely spaced points. For an overview of the theoretical basis for Poisson process models, see Çinlar [10]. The theorems that suggest why modelling connection level traffic by a nonhomogeneous Poisson process is reasonable, also suggest why using these models for packet level traffic is not reasonable. Extensive traffic measurements have demonstrated that a Poisson process is not appropriate for individual or aggregate packet arrival processes.

The present paper is an extension of two previous lines of research. First, the present paper extends the offered-load models for communication networks proposed by Duffield and Whitt [14]–[17] by considering *nonstationary* customer arrival processes. In [14]–[17] it was shown how the offered-load models can be used for network design and control. It was also shown how the conditional expected future bandwidth can be used to approximately describe buffer content when a buffer is imposed with a specified bandwidth. The nonstationarity introduced here can be an important addition to capture the time-of-day variations in arrival rates. Many of the ideas in [14]–[17] apply directly to the more general nonstationary setting considered here, once we show that the corresponding descriptive quantities can be computed, which we do here. Thus we refer to [14]–[17] for additional motivation.

Second, the present paper extends the nonstationary Poisson-arrival-location model (PALM) for wireless networks investigated by Massey and Whitt [27], [28] and Leung, Massey and Whitt [26] to produce versions for high-speed wired networks. For wireless networks, the PALM model captures customer mobility by allowing movement through space after arrival, but assumes common unit bandwidth requirements for all customers. In contrast, for emerging high-speed wired networks, it seems important to capture the variable bandwidth requirements, with variability applying to different customers and to what any one customer needs over time. It is possible to consider both mobility and variable bandwidth requirements for customers, but for wired networks it seems appropriate, at least initially, to leave out the mobility. Hence we do not consider mobility here.

A major reason for considering time-dependent arrival rates is to determine the value and time of the peak offered load. It is significant that the time of the peak offered load tends to lag behind (occur later than) the peak connection arrival rate. Moreover, the value of the peak offered load often is significantly less than the stationary offered load assuming a constant arrival rate equal to the peak value. For the $M_t/G/\infty$ model, these phenomena have been studied in Massey and Whitt (1997) and references therein. We provide a framework here for investigating these same questions for packet networks.

Here is how the rest of this paper is organized. In Section 2 we use stochastic integration over Poisson processes (with a special type of integrand) to specify the offered-load model for a single link. What follows immediately from the associated stochastic calculus are expressions for the mean, variance, higher cumulants and covariances of the total-required-bandwidth process. We also show how we can model the cumulative packet arrival process for a single link and extend the model to cover the case multiple links in a communication network. We give moment generating function solutions for all the finite-dimensional distributions of these processes. In Section 3 we prove a central limit theorem justifying a Gaussian process approximation, and discuss its application. We also develop a useful Gaussian approximation for the buffer-content distribution when the offered-load process is the input to an infinite-capacity buffer with a deterministic, possibly time-dependent processing rate. Here we extend the lower bound developed for stationary models with Gaussian input by Addie and Zukerman [3], Addie, Zukerman and Neame [4], Choe and Shroff [9] and Norros [31], [32].

In Section 4 we give an illustrative example to highlight the insights that can be gained from the offered-load model. We show that the peak offered load lags behind the peak connection arrival rate. Moreover, we can see the influence of the connection holding-time distribution (e.g. if it is heavy-tailed) upon the offered load after a sudden traffic surge. This example is similar to the traffic accident example in Leung et al. [26].

In Section 5 we show how the offered-load model can be used for predicting the future bandwidth requirements given current state information. In order to exploit information contained in the history of connections in progress, we separately analyze the bandwidth requirements for new arrivals and for previous arrivals still in the system.

Additional material is contained in our companion paper [12], including a review of Poisson integration and proofs omitted here.

2. Constructing and Characterizing the Bandwidth Processes

In this section we define and characterize the stochastic process under study. We start by defining a stochastic process $\{ R(t) \mid -\infty < t < \infty \}$ describing the *total required bandwidth or rate* on a single link of a communication network over time.

We have in mind multiple classes of customers each with their own time-dependent arrival rates and stochastic characteristics. Assuming that these classes are mutually independent, the total required bandwidth will simply be the sum of the required bandwidths over all classes, and the means, variances and covariances will add. Hence in the following discussion we restrict attention to a single customer class.

For the single class under consideration, we assume that customers arrive according to a nonhomogeneous Poisson process. Let $A(t)$ count the *number of customer connections arriving* in the interval $(-\infty, t]$. We assume that $\{ A(t) \,|\, t \geq 0 \}$ is a nonhomogeneous Poisson process with intensity function $\alpha(t)$. We assume that $\int_{-\infty}^{t} \alpha(s) ds < \infty$ for all t, so that $A(t)$ has a proper Poisson distribution for each t. (This does not cover a stationary model; the assumption can be changed to treat the stationary case.)

Let $B(s,t)$ be the *individual required bandwidth or rate* at time t for a customer that arrives at time s, with the convention that $B(s,t) = 0$ whenever $s > t$. We think of the collection $\{ B(s,t) \,|\, t \geq s \}$ as being a collection of mutually independent random processes indexed by real s with probability laws depending on s. Just as on page 196 of Massey and Whitt [27], we can formally define these quantities in terms of an underlying countably infinite sequence of independent random variables. That construction avoids measurability problems associated with assuming an uncountably infinite collection of independent random variables.

Clearly, there are many possibilities for $B(s,t)$. We could let the customer bandwidth $B(s,t)$ become 0 after a random connection time T_s, with a distribution possibly depending on s. Then we call T_s the connection holding time. We could have $B(s,t)$ be a fixed deterministic function of t, which could depend or not depend on s. Then the total-required-bandwidth process is a classical (nonstationary) shot-noise process; e.g., see Rice [36], Klüppelberg and Mikosch [24] and references therein. We could have $B(s,t)$ be deterministic with a form depending upon the customer class, which could be randomly selected. The bandwidth $B(s,t)$ could be constant over time, but randomly distributed.

A principal case for applications is the *homogeneous* case in which $B(s,t)$ depends on the pair (s,t) only through the difference $t-s$, i.e.,

$$\{ B(s,t) \mid t \geq s \} \stackrel{\mathrm{d}}{=} \{ B(0,t-s) \mid t \geq s \} \tag{2.1}$$

for all (s,t) with $s < t$. Note however that the stochastic process $\{ B(t) \mid t \geq 0 \}$ where $B(t) \equiv B(0,t)$ need *not* be a stationary process. For example, $B(t)$ might be an on-off process that always starts at the beginning of an on time.

It is significant that our framework also allows for *nonhomogeneous* individual bandwidth processes. Traffic measurements indicate that, not only is the connection arrival rate strongly time-dependent, but so also is the expected individual bandwidth usage. This phenomenon is consistent with previous measurements of telephone calls. Both the average holding times and the arrival rate have been observed to be time-dependent, with average holding times tending to be longer in the evenings.

With the framework above, we can define the *total-required-rate (or bandwidth) process* by stochastic integration with respect to the Poisson arrival process, just as in [27] and [28]. We elaborate in a companion paper [12]. The total bandwidth (rate) required at time t is then

$$R(t) = \int_{-\infty}^{t} B(s,t) dA(s) \ . \tag{2.2}$$

Figure 1 depicts a possible realization, with $A(t)$ connections active at time t. Sample paths of three of the $A(t)$ individual-bandwidth processes, with their appropriate start times, are displayed, displaced vertically, along with $R(t)$, the total required bandwidth at time t. Since we focus on the individual bandwidth processes, we only indicate the total required bandwidth at the single time t. The process $\{ R(t) \mid -\infty < t < \infty \}$ has a close connection to the network of infinite-server queues and the more general Poisson Arrival Location Model (PALM) in [27] [28]. With the PALM, however, the customers moved through space after arrival according to a location stochastic process. In contrast, here the customers do not move. Instead, the bandwidth required by each customer at each location evolves over time as a stochastic process. Otherwise, the supporting mathematics is essentially the same.

Formula (2.2) states that $R(t)$ is a Poisson sum of independent random variables, i.e.,

$$R(t) = \sum_{n=1}^{A(t)} X_n(\hat{A}_n, t) \ , \tag{2.3}$$

where $\hat{A}_n$ is the arrival epoch of the n^{th} arrival. The random variables $X_n(\hat{A}_n, t)$ in (2.3) are conditionally mutually independent, given the arrival epochs $\hat{A}_n$, $n \geq 1$.

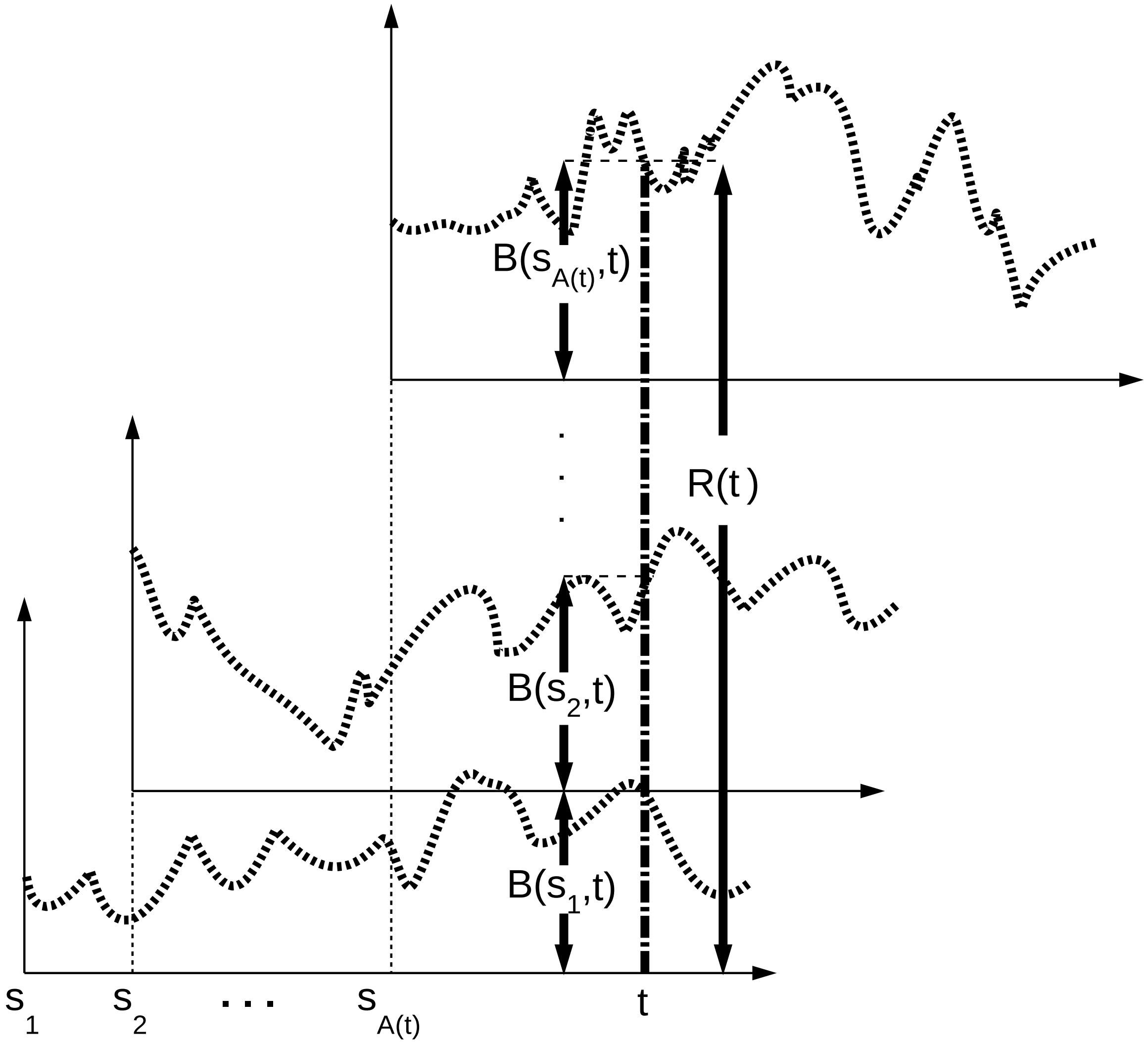

Figure 1: Typical realizations of individual bandwidth processes associated with $A(t)$ active connections at time t. Also displayed is $R(t)$, the total required bandwidth at time t.

We now give formulas for the logarithmic moment generating function, the mean, variance and higher cumulants of the total bandwidth (for this one customer class). Recall that *cumulants* of a random variable, here denoted by $\mathsf{C}^{(n)}[X]$, are the coefficients in the power-series expansion of the logarithmic moment generating function, i.e.,

$$\log \mathsf{E}\left[\exp(\theta X)\right] = \sum_{n=1}^{\infty} \frac{\theta^n}{n!} \mathsf{C}^{(n)}[X]. \tag{2.4}$$

As a technical regularity condition, we assume that the moment generating function $\mathsf{E}[\exp \theta X]$ is finite for some strictly positive θ; then the power series in (2.4) is valid (has a positive radius of convergence).

From (2.4) it follows that cumulants have the property that if X and Y are any two independent random variables and λ is any real number, then

$$\mathsf{C}^{(n)}[X+Y] = \mathsf{C}^{(n)}[X] + \mathsf{C}^{(n)}[Y] \quad \text{and} \quad \mathsf{C}^{(n)}[\lambda X] = \lambda^n \mathsf{C}^{(n)}[X]. \tag{2.5}$$

The first two cumulants are the mean and variance. For a Poisson random variable, all cumulants coincide with the mean. This property characterizes the Poisson distribution.

Our key result, proved in [12], is an expression for the logarithmic moment generating function of $R(t)$. We can also characterize the finite-dimensional distributions of the stochastic process $\{\, R(t) \mid -\infty < t < \infty \,\}$ by the same argument.

Theorem 2.1 *Let $t_1 < t_2 < \ldots < t_k$ be k increasing time points and $\theta_1, \ldots, \theta_k$ be k arbitrary real numbers. The logarithmic moment generating function for the joint distribution of $R(t_1), \ldots, R(t_k)$ is*

$$\log \mathsf{E}\left[\exp\left(\sum_{j=1}^{k} \theta_j R(t_j)\right)\right] = \int_{-\infty}^{t_k} \mathsf{E}\left[\exp\left(\sum_{j=1}^{k} \theta_j B(s,t_j)\right) - 1\right] \alpha(s) ds \ . \tag{2.6}$$

As simple consequences, we obtain fairly simple formulas for the cumulants of $R(t)$ and the covariance between $R(t_1)$ and $R(t_2)$, in terms of the moments of $B(s,t)$ and the arrival-rate function $\alpha(t)$. They follow by differentiation with respect to θ in (2.6).

Corollary 2.2 *The n-th cumulant of $R(t)$ in (2.2) is*

$$\mathsf{C}^{(n)}[R(t)] = \int_{-\infty}^{t} \mathsf{E}[B(s,t)^n] \alpha(s) ds \ , \tag{2.7}$$

from which we get

$$\mathsf{E}[R(t)] = \int_{-\infty}^{t} \mathsf{E}[B(s,t)] \alpha(s) ds \quad \text{and} \quad \mathsf{Var}[R(t)] = \int_{-\infty}^{t} \mathsf{E}[B(s,t)^2] \alpha(s) ds \ . \tag{2.8}$$

Corollary 2.3 *Let $t_1 < t_2 < \ldots < t_k$ be k increasing time points and $\theta_1, \ldots, \theta_k$ be k arbitrary real numbers. We then have*

$$\mathsf{C}^{(n)}\left[\sum_{j=1}^{k}\theta_j R(t_j)\right] = \int_{-\infty}^{t_k}\mathsf{E}\left[\left(\sum_{j=1}^{k}\theta_j B(s,t_j)\right)^n\right]\alpha(s)ds\ . \tag{2.9}$$

A consequence is that, for all $t_1 < t_2$,

$$\mathsf{Cov}\left[R(t_1), R(t_2)\right] = \int_{-\infty}^{t_1}\mathsf{E}\left[B(s,t_1)B(s,t_2)\right]\alpha(s)ds\ . \tag{2.10}$$

Given the values of the expectations in the integrands of (2.7)–(2.10), we can compute the displayed quantities by performing numerical integration, e.g., see Davis and Rabinowitz [11]. For that purpose, it is natural to simplify matters by requiring that $\alpha(s) = 0$ for $s < t_0$ for some t_0, so that all integrals are over the finite interval $[t_0, t]$.

We can also use the stochastic calculus to construct the *total cumulative input process* $I(t,t')$ for the interval (t,t') with $t < t'$, which equals

$$I(t,t') = \int_t^{t'} R(s)ds = \int_{-\infty}^{t'} C_s(t,t')dA(s)\ , \tag{2.11}$$

where

$$C_s(t,t') = \int_t^{t'} B(s,\tau)d\tau. \tag{2.12}$$

This makes $C_s(t,t')$ equal to the *individual cumulative input process* for a connection arriving at time s during the interval (t,t'). Closely paralleling the stationary setting, see Kelly [22], we define an *effective-bandwidth function* by

$$\beta_\theta(t,t') \equiv \frac{1}{\theta\cdot(t'-t)}\log\mathsf{E}\left[\exp\Big(\theta\cdot I(t,t')\Big)\right] \tag{2.13}$$

The effective-bandwidth function is additive for superpositions of independent sources and, in the stationary setting, gives a value between the peak and average rate. See Chang [8] for further discussion of the nonstationary case.

Now, closely paralleling Theorem 2.1, we characterize the finite-dimensional distributions of the stochastic process $\{\, I(t,t') \mid -\infty < t < \infty \,\}$ and the effective bandwidth function.

Theorem 2.4 *Let $(t_1,t_1'), \ldots, (t_k,t_k')$ and $\theta_1, \ldots, \theta_k$ be k time intervals and k arbitrary real numbers respectively. The logarithmic moment generating function for the joint distribution of $I(t_1,t_1'), \ldots, I(t_k,t_k')$ is*

$$\log\mathsf{E}\left[\exp\left(\sum_{j=1}^{k}\theta_j I(t_j,t_j')\right)\right] = \int_{-\infty}^{t^*}\mathsf{E}\left[\exp\left(\sum_{j=1}^{k}\theta_j C_s(t_j,t_j')\right) - 1\right]\alpha(s)ds\ . \tag{2.14}$$

where $t^* = \max(t'_1, \ldots, t'_k)$, *so that*

$$\mathsf{C}^{(n)} \left[\sum_{j=1}^{k} \theta_j I(t_j, t'_j) \right] = \int_{-\infty}^{t^*} \mathsf{E} \left[\left(\sum_{j=1}^{k} \theta_j C_s(t_j, t'_j) \right)^n \right] \alpha(s) ds \ . \tag{2.15}$$

A consequence of (2.15) is that for all intervals (t_i, t'_i) *and* (t_j, t'_j)*,*

$$\mathsf{Cov} \left[I(t_i, t'_i), I(t_j, t'_j) \right] = \int_{-\infty}^{\min(t'_i, t'_j)} \mathsf{E} \left[C_s(t_i, t'_i) C_s(t_j, t'_j) \right] \alpha(s) ds \ . \tag{2.16}$$

We can also model the total rate at each link of a multi-link packet network by defining a vector-valued stochastic process $\{ \mathbf{R}(t) \mid -\infty < t < \infty \}$, where

$$\mathbf{R}(t) = \left(R^{(1)}(t), \ldots, R^{(L)}(t) \right) \tag{2.17}$$

and $R^{(\ell)}(t)$ describes the *total required bandwidth* at *link* ℓ at time t. Each link ℓ is intended to represent a resource in the communication network. Since communication may involve multiple resources, bandwidth may be required at more than one link. The links might be part of a communication path; then the required bandwidth would usually be the same on all links. Our general framework allowing arbitrary subsets of all links encompasses multicast communication.

Using the stochastic calculus, we can define the overall bandwidth process to be

$$\mathbf{R}(t) = \int_{-\infty}^{t} \mathbf{B}(s, t) dA(s) \ , \tag{2.18}$$

where

$$\mathbf{B}(s, t) = \left(B^{(1)}(s, t), \ldots, B^{(L)}(s, t) \right) \tag{2.19}$$

and $B^{(\ell)}(s, t)$ is the *random bandwidth* required at time t and link ℓ for a customer that arrives at time s with $t \geq s$. The total bandwidth required at time t and link ℓ is then the same as for the single link case.

$$R^{(\ell)}(t) = \int_{-\infty}^{t} B^{(\ell)}(s, t) dA(s) \ . \tag{2.20}$$

In general, for every different set of time-link pairs (t, ℓ), the $R^{(\ell)}(t)$'s are dependent; i.e., $R^{(\ell_1)}(t)$ and $R^{(\ell_2)}(t)$ are in general dependent, as are $R^{(\ell)}(t_1)$ and $R^{(\ell)}(t_2)$. Now we characterize all the finite dimensional distributions for the vector bandwidth process $\{ \mathbf{R}(t) \mid -\infty < t < \infty \}$, which show the interactions of the bandwidth process across different links and points in time.

Theorem 2.5 *Let* $t_1 < t_2 < \ldots < t_k$ *be* k *time points and* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ *be* k *arbitrary* L*-dimensional vectors. The logarithmic moment generating function for the joint distribution of*

$\mathbf{R}(t_1), \ldots, \mathbf{R}(t_k)$ *is*

$$\log \mathsf{E}\left[\exp\left(\sum_{j=1}^{k} \boldsymbol{\theta}_j \cdot \mathbf{R}(t_j)\right)\right] = \int_{-\infty}^{t_k} \mathsf{E}\left[\exp\left(\sum_{j=1}^{k} \boldsymbol{\theta}_j \cdot \mathbf{B}(s, t_j)\right) - 1\right] \alpha(s) ds \ , \qquad (2.21)$$

so that

$$\mathsf{C}^{(n)}\left[\sum_{j=1}^{k} \boldsymbol{\theta}_j \cdot \mathbf{R}(t_j)\right] = \int_{-\infty}^{t_k} \mathsf{E}\left[\left(\sum_{j=1}^{k} \boldsymbol{\theta}_j \cdot \mathbf{B}(s, t_j)\right)^n\right] \alpha(s) ds \ . \qquad (2.22)$$

A consequence of 2.22 is that for all links ℓ_1 and ℓ_2 and all times $t_1 \leq t_2$,

$$\mathsf{Cov}[R^{(\ell_1)}(t_1), R^{(\ell_2)}(t_2)] = \int_{-\infty}^{t_1} \mathsf{E}\left[B^{(\ell_1)}(s, t_1) B^{(\ell_2)}(s, t_2)\right] \alpha(s) ds \ . \qquad (2.23)$$

3. Gaussian Approximations

In many applications there will be a large number of customers, with the bandwidth requirement of each being small compared to the total. Then it is natural to approximate the stochastic processes $\{ R(t) \mid -\infty < t < \infty \}$, $\{ I(t, t') \mid -\infty < t < t' < \infty \}$ and $\{ \mathbf{R}(t) \mid -\infty < t < \infty \}$ by Gaussian stochastic processes, by virtue of the central limit theorem (CLT). To illustrate, we give a version of the CLT for $\{ R(t) \mid -\infty < t < \infty \}$. It is not the most general, but it is adequate to justify normal approximation in most applications, and has an easy proof. We simply scale the arrival-rate function.

Theorem 3.1 *Consider a sequence of models indexed by $n \geq 1$, with common bandwidth processes, where the arrival rate in model is $\alpha^n(t) = n\alpha(t)$. Assume that the bandwidth processes are uniformly bounded; i.e., there is an M such that $B(s, t) \leq M$ a.s. for all pairs (s, t). Then the finite-dimensional distributions of the process $\{ X_n(t) \mid -\infty < t < \infty \}$ given by*

$$X_n(t) \equiv \frac{R_n(t) - n\mathsf{E}[R(t)]}{\sqrt{n}} \qquad (3.1)$$

converge in distribution as $n \to \infty$ to a Gaussian process with the same covariances as formulas for $\{ R(t) \mid -\infty < t < \infty \}$ in (2.10).

Proof. Note that the n^{th} model can be represented as the superposition of n i.i.d. versions of the model $R(t)$ corresponding to $n = 1$. First, the Poisson arrival process with arrival rate function $n\alpha(t)$ is the superposition of n i.i.d. Poisson processes with arrival rate functions $\alpha(t)$. Given that the arrival rate functions are identical, the random vector $(R_n(t_1), \ldots, R_n(t_k))$ is the sum of n i.i.d. random vectors, each distributed as $(R(t_1), \ldots, R(t_k))$. By the assumptions, the variable $R(t)$ has finite means and second moments. Hence, we can apply the classical CLT for i.i.d. random vectors with finite variances. ■

Given the normal approximation, we can compute the probability that demand exceeds the critical level. For that purpose, let Φ be the cumulative distribution function (cdf) of $N(0,1)$, i.e., $\Phi(t) \equiv \mathsf{P}(N(0,1) \leq t)$, and let $\Phi^c(t) \equiv 1 - \Phi(t)$ be the associated complementary cdf. Then

$$\mathsf{P}(R(t) \geq L) \approx \Phi^c \left(\frac{L - \mathsf{E}[R(t)]}{\sqrt{\mathsf{Var}[R(t)]}} \right) . \tag{3.2}$$

If the level L in (3.2) is the instantaneous output rate, then $\mathsf{P}(R(t) > L)$ is the time-dependent loss probability in a bufferless model. Such a normal approximation for connection admission control was proposed by Guerin, Ahmadi and Naghshineh [20] in a refinement to their "equivalent capacity" scheme. With highly bursty traffic, we think that it may be appropriate to use only (3.2).

In the setting of a bufferless model, we might also be interested in the expected quantity lost

$$\mathsf{E}\left[(R(t) - L)^+\right] = \mathsf{E}\left[\, R(t) - L \mid R(t) > L \,\right] \mathsf{P}(R(t) > L) . \tag{3.3}$$

To give a normal approximation for the conditional expectation, let ϕ be the density of Φ and let $m = \mathsf{E}[R(t)]$ and $\sigma^2 = \mathsf{Var}[R(t)]$. Then

$$\begin{aligned} \mathsf{E}\left[\, R(t) - L \mid R(t) > L \,\right] &\approx \mathsf{E}\left[N(m-L,\sigma^2) \,\middle|\, N(m-L,\sigma^2) > 0 \right] \\ &= m - L + \sigma \mathsf{E}\left[\, N(0,1) \mid N(0,1) > (L-m)/\sigma \,\right] \\ &= m - L + \sigma \frac{\phi((L-m)/\sigma)}{\Phi^c((L-m)/\sigma)} . \end{aligned} \tag{3.4}$$

We can also apply the method to treat multiple priority classes. Suppose that there are k classes with the lower indices having higher priority. When we consider class j, we can consider the total input rate for the first j classes; i.e., if $R_j(t)$ is the total rate for priority class j, then we obtain k constraints

$$\mathsf{P}(R_1(t) + \cdots + R_j(t) > L_j) < \epsilon_j, \quad 1 \leq j \leq k , \tag{3.5}$$

where the level L_j and target probability ϵ_j will depend on the class j. Indeed, a variant of this procedure is used in IBM's Network Broadband Services (NBBS) admission control algorithm; see pp. 608–609 of Ahmadi et al. [5]. We contribute by providing a model leading to formulas for the required means and variances in a nonstationary setting.

We now apply the Gaussian process approximation to $\{\, I(t,t') \mid -\infty < t < t' < \infty \,\}$, the cumulative input process, to obtain a Gaussian approximation for the buffer-content distribution in a fluid queue. Consider a single resource and assume that the offered load represents

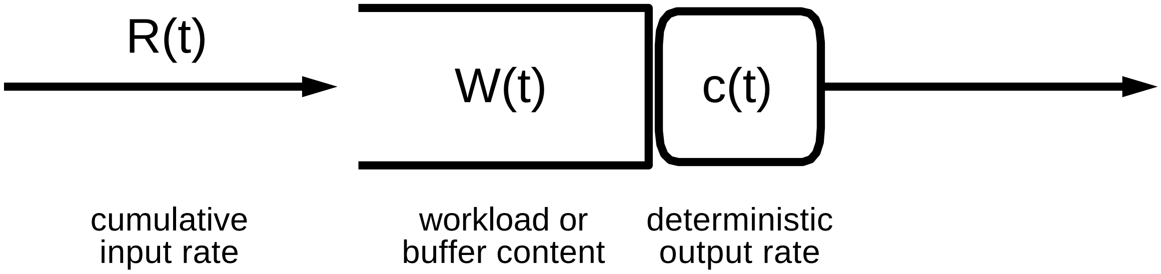

Figure 2: The fluid queue with infinite capacity buffer and a time-dependent deterministic output rate.

the rate of fluid coming to a single-server fluid queue with infinite buffer and time-dependent deterministic output rate (or channel capacity) $c(t)$. In the typical application $c(t)$ will be constant, but the result extends to the time-dependent case, which is of interest for example if part of the bandwidth is unavailable because of other uses, perhaps due to advance reservation; see Greenberg, Srikant, and Whitt [19]. We develop an approximation for the buffer-content distribution, which is necessarily time-dependent because the input and output are nonstationary.

Let $W(t)$ denote the *total workload or buffer content* at time t. Since $\int_{-\infty}^{t} \alpha(s)ds < \infty$, $A(t)$ and $R(t)$ are finite a.s. Assuming that $\int_{-\infty}^{t} c(s)ds = \infty$, we can deduce that $W(t)$ is finite a.s. for all t and

$$W(t) = \sup_{s \leq t} \left(I(s,t) - \int_{s}^{t} c(\tau)d\tau \right) \ . \tag{3.6}$$

In Figure 2 we depict the fluid queue model being considered.

As in many previous studies of stationary models, we exploit the lower bound for the tail probability,

$$\mathsf{P}(W(t) > x) \geq \sup_{s \leq t} \mathsf{P} \left(I(s,t) - \int_{s}^{t} c(\tau)d\tau > x \right) \tag{3.7}$$

as an approximation. For stationary models, this lower bound has been shown to be an asymptotically accurate approximation, and is used as a key step in establishing large deviation results; see Duffield and O'Connell [13] and Botvich and Duffield [6]. However, without resorting to simplifying asymptotics, even the lower bound in (3.7) is very complicated. As noted by, e.g., Addie and Zukerman [3], Addie, Zukerman and Neame [4], Choe and Shroff [9] and Norros [31], [32], the lower bound in (3.7) greatly simplifies if we approximate the total cumulative input process $I(s,t)$ by a Gaussian process, which we have indicated is often reasonable, because we can apply the CLT. Then the *net cumulative input process* $\left\{ I(s,t) - \int_{s}^{t} c(\tau)d\tau \mid s \leq t \right\}$

also becomes a time-reversed Gaussian process in s. We can then find the maximizing s in the right side of (3.7) in terms of the time-dependent means and variances of $I(s,t)$. Recall that

$$\mathsf{E}[I(s,t)] = \int_{-\infty}^{t} \mathsf{E}[C_\tau(s,t)]\alpha(\tau)d\tau \quad \text{and} \quad \mathsf{Var}[I(s,t)] = \int_{-\infty}^{t} \mathsf{E}[C_\tau(s,t)^2]\alpha(\tau)d\tau. \tag{3.8}$$

Now let

$$Z(s,t) = \frac{I(s,t) - \mathsf{E}[I(s,t)]}{x - \mathsf{E}[I(s,t)] + \int_s^t c(\tau)d\tau}, \quad s \le t\ , \tag{3.9}$$

and note that

$$\left\{ I(s,t) - \int_s^t c(\tau)d\tau > x \right\} = \{Z(s,t) \ge 1\}\ . \tag{3.10}$$

However, since $Z(s,t)$ has mean 0 for all $s \le t$, it suffices to consider only that s^* maximizing the variance of $Z(s,t)$. Consequently,

$$\sup_{s \le t} \mathsf{P}(I(s,t) - \int_s^t c(\tau)d\tau > x) = \sup_{s \le t} \mathsf{P}(Z(s,t) \ge 1) = \Phi^c\left(1/\sqrt{Var\, Z(s^*,t)}\right)\ , \tag{3.11}$$

where $\Phi^c(x) \equiv \mathsf{P}(N(0,1) > x)$ is again the standard (mean 0, variance 1) Gaussian complementary cdf and s^* maximizes

$$\mathsf{Var}[Z(s,t)] = \frac{\mathsf{Var}[I(s,t)]}{\left(x - \mathsf{E}[I(s,t)] + \int_s^t c(\tau)d\tau\right)^2} \tag{3.12}$$

over all $s \le t$. (We assume that the supermum over s in (3.12) is attained.)

In the stationary context, $\mathsf{E}[I(s,t)] - \int_s^t c(\tau)d\tau$ is always $-\beta(t-s)$ for some constant β, where $-\beta$ must be negative, in order to have model stability. Here, in the nonstationary case, if we assume that $c(t) \ge \gamma$ for all t, than we can conclude that $E[I(s,t)] - \int_s^t c(\tau)d\tau \ge -\gamma \cdot (t-s)$ and that $E[I(s,t)] - \int_s^t c(\tau)d\tau$ is eventually negative for all s sufficiently small because the total input over $(-\infty, t]$ is finite.

It is worth noting that even in the stationary context (when $\{ R(t) \mid t \ge 0 \}$ is a stationary stochastic process and $\{ I(s,t) \mid s \le t \}$ has stationary increments), the stochastic process $\{ Z(s,t) \mid s \le t \}$ is not itself a stationary process. In the stationary case, Choe and Shroff [9] have shown that the approximation (3.11) performs remarkably well, even outside the large-deviations asymptotic regime. Hence, (3.11) is a promising approximation in the nonstationary case as well.

By calculating (3.11) for a range of t, we can approximately determine how the tail probability $\mathsf{P}(W(t) > x)$ depends on t; e.g., we can identify the t^* for which

$$\mathsf{P}(W(t^*) > x) \approx \sup_t \mathsf{P}(W(t) > x)\ . \tag{3.13}$$

4. Example: The Physics of Time Lags and Space Shifts

In this section we consider an example to illustrate the insights that can be gained from the nonstationary offered-load model. The assumptions made here are general but idealized, making it possible to do the analysis analytically. We intend to do more specific modelling based on traffic data in the future. The discussion here parallels the analysis of the $M_t/G/\infty$ queue in Eick, Massey and Whitt [18].

Let b be a deterministic, non-negative real-valued function with support on $[0,\infty)$. If $s = \hat{A}_n$, then we define

$$B(s,t) \equiv b(t-s)1_{\{T_n > t-s\}} \tag{4.1}$$

where T_n is the *holding time* for the n-th arriving connection, and $\{\, T_n \mid n \geq 1 \,\}$ is a sequence of i.i.d. non-negative random variables. We call the deterministic function b the *bandwidth profile*. In practice b can be an effective bandwidth or an upper envelope for the stochastic (unpredictable) behavior of a connection's request for bandwidth. In this example we show how the model can be used to predict how the peak expected total required bandwidth will lag behind the peak connection arrival rate.

For the next theorem, let T denote a non-negative random variable with the same distribution as one of the T_n, and let T_e denote a non-negative random variable with the *stationary excess distribution* of T, i.e. for all $t \geq 0$,

$$\mathsf{P}\left(T_e \leq t\right) \equiv \frac{\int_0^t \mathsf{P}\left(T > s\right) ds}{\mathsf{E}[T]}. \tag{4.2}$$

Proposition 4.1 *If $B(s,t)$ is defined as above, then*

$$\mathsf{C}^{(n)}\left[R(t)\right] = \mathsf{E}\left[\alpha(t-T_e)b(T_e)^n\right] \cdot \mathsf{E}[T] \tag{4.3}$$

for all positive integers n and real t.

Proof: Substituting (4.1) into (2.7), we have

$$\mathsf{C}^{(n)}\left[R(t)\right] = \mathsf{E}\left[\int_{t-T}^{t} \alpha(s)b(t-s)^n ds\right] = \mathsf{E}\left[\int_0^T \alpha(t-s)b(s)^n ds\right] \tag{4.4}$$

Now we use the identity established in Massey and Whitt [30] for all functions f continuously differentiable on $[0,\infty)$ and $x \geq 0$,

$$\frac{\mathsf{E}[f(x+T)] - f(x)}{\mathsf{E}[T]} = \mathsf{E}[f'(x+T_e)]. \tag{4.5}$$

Applying (4.5) to (4.4) completes the proof. ▮

For the next theorem, we introduce some notation. Let X be any nonnegative random variable X and f be any nonnegative function such that both $\mathsf{E}[f(X)]$ and $\mathsf{E}[Xf(X)]$ are finite. The function f induces a new probability measure or expectation E_f on X such that

$$\mathsf{E}_f[X] \equiv \frac{\mathsf{E}[Xf(X)]}{\mathsf{E}[f(X)]}. \tag{4.6}$$

We also define $\mathsf{Var}_f[X] \equiv \mathsf{E}_f[X^2] - \mathsf{E}_f[X]^2$.

Theorem 4.2 *If $\alpha(t) = a_0 + a_1 t - a_2 t^2$, then for all integers $n \geq 1$*

$$\mathsf{C}^{(n)}\left[R(t)\right] = \Big(\alpha\left(t - \mathsf{E}_{b^n}[T_e]\right) - a_2 \mathsf{Var}_{b^n}[T_e]\Big) \cdot \mathsf{E}\left[b(T_e)^n\right] \cdot \mathsf{E}[T]. \tag{4.7}$$

Corollary 4.3 *Under the hypothesis of Theorem 4.2, if $a_2 \neq 0$, then α attains its unique extremal value α^* at $t^*_\alpha = a_1/(2a_2)$ and there is a unique extremal value for $\mathsf{C}^{(n)}\left[R(t)\right]$ at*

$$t^*_n = t^*_\alpha + \mathsf{E}_{b^n}[T_e] \tag{4.8}$$

and

$$\mathsf{C}^{(n)}\left[R(t^*_n)\right] = \Big(\alpha^* - a_2 \mathsf{Var}_{b^n}[T_e]\Big) \cdot \mathsf{E}\left[b(T_e)^n\right] \cdot \mathsf{E}[T], \tag{4.9}$$

When $a_2 > 0$, the extremal value $\mathsf{C}^{(n)}\left[R(t^*_n)\right]$ is a maximum. This corollary tells us that there is always a nonnegative *time lag* $\ell_n \equiv \mathsf{E}_{b^n}\left[T_e\right]$ between the time t^*_α of the extremal value α^* for α and t^*_n, the time for the extremal value of the n-th cumulant of R. Moreover, this time lag is the weighted average of T_e and *not* T. So even if $\mathsf{E}[T]$ is held fixed, changing the second moment $\mathsf{E}[T^2]$ will influence the time lags. Thus, in general, holding times with heavy-tail distributions have the potential of inducing large time lags between the time of the peak connection arrival rate and both the average peak load of total bandwidth in use and its variance. Also note that the values of the parameters a_0, a_1, and a_2 for the connection arrival rate α have no effect on these time lags.

If $\alpha(t)$ equals the constant rate rate α^*, then we have

$$\mathsf{C}^{(n)}\left[R(t)\right] = \alpha^* \cdot \mathsf{E}\left[b(T_e)^n\right] \cdot \mathsf{E}[T] \tag{4.10}$$

for all positive integers n and real t. When $\mathsf{C}^{(n)}\left[R(t^*_n)\right]$ is a maximum, it is strictly less (observe that T_e can never be a constant) than the *pointwise stationary value* (4.10), so just as in Eick, Massey and Whitt [18], along with our time lags we have space shifts.

Corollary 4.4 *Under the hypothesis of Theorem 4.2, if $b(x) = b \cdot 1_A(x)$ where A is some measurable subset of the positive reals such that $\mathsf{P}(T_e \in A) > 0$, then*

$$\mathsf{C}^{(n)}\left[R(t)\right] = \Big(\alpha\left(t - \mathsf{E}[\,T_e\,|\,T_e \in A\,]\right) - a_2 \mathsf{Var}[\,T_e\,|\,T_e \in A\,]\Big) \cdot b^n \mathsf{P}\left(T_e \in A\right) \cdot \mathsf{E}[T], \qquad (4.11)$$

and so $R(t)/b$ has a Poisson distribution.

Observe that for any constant c we have

$$\mathsf{E}[\,T_e\,|\,T_e \geq c\,] \geq c \quad \text{and} \quad \mathsf{E}[\,T_e\,|\,T_e \leq c\,] \leq c. \qquad (4.12)$$

assuming respectively that $\mathsf{P}(T_e \geq c) > 0$ and $\mathsf{P}(T_e \leq c) > 0$. Thus the times when a connection uses a non-zero amount of bandwidth during its "on" period can have an enormous effect on the time lag.

Corollary 4.5 *Under the hypothesis of Theorem 4.2, let b_1 and b_2 be two bandwidth functions. If $b_1 = \phi b_2$, where ϕ is a nonnegative, nondecreasing function on the support of T_e, then the using the same α and T, the time lags ℓ_n due to b_1 will always be larger than the ones due to b_2.*

This corollary yields yet another result of interest.

Corollary 4.6 *If b is a non-decreasing function on the support of T_e, then*

$$0 \leq \mathsf{E}\left[T_e\right] \leq \ell_1 \leq \ell_2 \leq \ldots \leq \ell_n \leq \ldots. \qquad (4.13)$$

Similarly, if b in non-increasing, then

$$0 \leq \ldots \leq \ell_n \leq \ldots \leq \ell_2 \leq \ell_1 \leq \mathsf{E}\left[T_e\right]. \qquad (4.14)$$

These results say that if b is non-decreasing, α is quadratic, and $a_2 > 0$, then $t_1^* \leq t_2^*$ or the time of the peak average total required bandwidth will precede the time of the peak variance for the total required bandwidth. Similarly, if b in non-increasing, then the opposite holds or $t_2^* \leq t_1^*$. The proofs for these last two results follows easily from the lemma below.

Lemma 4.7 *Let X be any nonnegative random variable with a finite mean. If f and g are two nonnegative, bounded, measurable functions defined on the support of X, and $f = g\phi$, where ϕ is a nondecreasing function on the support of X, then*

$$\mathsf{E}_f[X] \geq \mathsf{E}_g[X]. \qquad (4.15)$$

Similarly, if ϕ is a non-increasing function, then

$$\mathsf{E}_f[X] \leq \mathsf{E}_g[X]. \qquad (4.16)$$

Proof: To show that $\mathsf{E}_f[X] \geq \mathsf{E}_g[X]$ is equivalent to showing that

$$\mathsf{E}[Xf(X)]\mathsf{E}[g(X)] \geq \mathsf{E}[f(X)]\mathsf{E}[Xg(X)]. \tag{4.17}$$

This follows from

$$\begin{aligned}
&\mathsf{E}[Xf(X)]\mathsf{E}[g(X)] - \mathsf{E}[f(X)]\mathsf{E}[Xg(X)] \\
&\quad = \int_0^\infty \int_0^\infty \Big(xf(x)g(y) - f(x)yg(y)\Big)\mathsf{P}(X \in dx)\mathsf{P}(X \in dy) \\
&\quad = \int_0^\infty \int_0^\infty (x-y)f(x)g(y)\mathsf{P}(X \in dx)\mathsf{P}(X \in dy) \\
&\quad = 2\int_0^\infty \int_y^\infty (x-y)\Big(f(x)g(y) - f(y)g(x)\Big)\mathsf{P}(X \in dx)\mathsf{P}(X \in dy) \\
&\quad = 2\int_0^\infty \int_y^\infty (x-y)g(x)g(y)\Big(\phi(x) - \phi(y)\Big)\mathsf{P}(X \in dx)\mathsf{P}(X \in dy),
\end{aligned}$$

which completes the proof. ∎

5. Prediction Given Information

In this section we consider the problem of predicting the total required bandwidth at some future time, given information available at the present time. We believe that it will often be advantageous to appropriately exploit available information. As discussed in Duffield and Whitt [14]-[17], long-range dependence and heavy-tail distributions offer opportunities to do prediction, because past events can have a longer impact. The present information can take several forms. We initially assume that we know the full history, i.e., the numbers of customers of each class, the elapsed connection holding time for each customer and the history of each customer's bandwidth stochastic process. However, it remains to determine the critical information in each context. Fortunately, the model makes it possible to study the value of different kinds of information, as is illustrated by [14]–[17].

We focus on a single customer class at a single link and assume that we can add the results for different classes. We divide the future requirements for the designated customer class into two parts: (1) the requirements generated from new arrivals and (2) the requirements generated from previous arrivals already in the system.

Let the present time be 0 and let the future time of interest be $t > 0$. Depending on how close t is to 0, the new or previous arrivals can dominate in the prediction. It is significant that the calculations can reveal the contribution of each component to the total future bandwidth requirements.

The prediction of the total bandwidth requirements of new arrivals is just as presented in Section 2, except that now only arrivals in the interval $[0, t]$ are considered. Formulas (2.2), (2.8), and (2.10) all carry over once the integrals have been changed to be over $[0, t]$ instead of $(-\infty, t]$. Equivalently, we can apply Section 2 directly under the extra assumption that $\alpha(s) = 0$ for $s < 0$.

Now we turn to the future requirements due to previous arrivals, i.e., due to customers already in the system. The information available at time 0 should include the number of customers present, so that it is known. We assume that there are n customers in the system at time 0. Conditional on the n previous arrival times, also assumed known, the bandwidth processes and remaining holding times for different customers are mutually independent. We let $(B_i(t)|I_i(0))$ denote a random variable with the conditional distribution of the required bandwidth for customer i at time t given the information (history) for customer i at time 0. In view of the conditional independence, the variances as well as the means add. The important point is that there is great potential for the conditioning upon $I_i(0)$ to significantly improve our estimate of the future required bandwidth for connection i.

Assuming that both the bandwidth process and the information can be represented as random elements of complete separable metric spaces, the conditional probability distribution can be expressed via a regular conditional probability measure, i.e., by a kernel $\mathsf{P}(x, A)$ such that for each possible information state x, $\mathsf{P}(x, \cdot)$ is a probability measure, and, for any measurable set A, $\mathsf{P}(x, A)$ is a measurable function of x; see Chapter V of Parthasarathy [33]. In particular, assuming that the information $I_i(0)$ is observed, $(B_i(t)|I_i(0))$ can be regarded as a bonafide random variable.

Now we can combine the new and old customers to obtain expressions for the mean and variance of the total required bandwidth at time t. Let $(R(t)|I(0))$ be a random variable representing the conditional total required bandwidth at time t given all available information at time 0. Then the mean and variance are

$$\begin{aligned} \mathsf{E}[R(t)|I(0)] &= \sum_{i=1}^{n} \mathsf{E}[B_i(t)|I_i(0)] + \int_0^t \mathsf{E}[B(s,t)]\alpha(s)ds \ , \qquad (5.1) \\ \mathsf{Var}[R(t)|I(0)] &= \sum_{i=1}^{n} \mathsf{Var}[B_i(t)|I_i(0)] + \int_0^t \mathsf{E}[B(s,t)^2]\alpha(s)ds \qquad (5.2) \end{aligned}$$

Similarly, for the covariances at times t_1 and t_2, we obtain the formula

$$\mathsf{Cov}[R(t_1), R(t_2)|I(0)] = \sum_{i=1}^{n} \mathsf{Cov}[B_i(t_1)B_i(t_2)|I_i(0)] + \int_0^t \mathsf{E}[B(s,t_1)B(s,t_2)]\alpha(s)ds \ . \qquad (5.3)$$

Following Duffield and Whitt [14], [15] we propose as a first-order approximation for the conditional total required bandwidth $(R(t)|I(0))$ its expected value in (5.1) and as a second-order approximation the normal distribution with mean and variance in (5.1)–(5.2). Instead of using the normal approximation, we can also calculate the distribution of the total required bandwidth by performing numerical transform inversion. Calculations of the full distribution in several representative cases can reveal how well the normal approximation and the more elementary deterministic mean-value approximation actually perform. If these approximations are adequate, then they can be used. Otherwise, it is possible to use the inversion.

Numerical inversion is effective when the probability distribution is either discrete or has a smooth probability density function. In the discrete case we can apply numerical inversion of generating functions, as in Abate and Whitt [1]. In the case of a continuous probability density function, we can apply numerical inversion of Laplace transforms, as in Abate and Whitt [2]. Both approaches could be used, but it seems more natural to work with generating functions. To work with generating functions, we assume that all bandwidth values are integer multiples of some basic unit, which we take to be 1. Then the random variables $B(s,t)$, $B_i(t)$ and $R(t)$ are all integer valued.

References

[1] J. Abate and W. Whitt, Numerical inversion of probability generating functions, *Operations Res. Letters* **12** (1992) 245–251.

[2] J. Abate and W. Whitt, Numerical inversion of Laplace transforms of probability distributions, *ORSA J. Computing* **7** (1995), 36–43.

[3] R. G. Addie and M. Zukerman, An approximation for performance evaluation of stationary single server queues, *IEEE Trans. Commun.* **42** (1994), 3150–3160.

[4] R. G. Addie, M. Zukerman, and T. Neame, Fractal traffic-measurements, modeling and performance evaluation, *IEEE Infocom '95*, 1995, 977–984.

[5] H. Ahmadi, P. F. Chimento, R. A. Guerin, L. Gün, B. Lin, R. O. Onvural and T. E. Tedijanto, NBBS traffic management overview, *IBM Systems J.* **34** (1995), 604–628.

[6] D. D. Botvich and N. G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexes, *Queueing Systems* **20** (1995), 293–320.

[7] R. Cáceres, P. G. Danzig, S. Jamin and D. J. Mitzel, Characteristics of wide-area TCP/IP conversations, *Computer Communications Review* **21** (1991), 101–112.

[8] C. S. Chang, Stability, queue length, and delay of deterministic and stochastic queueing networks, *IEEE Trans. Aut. Control*, **39** (1994) 913–931.

[9] J. Choe and N. B. Shroff, A central limit theorem based approach for analyzing queue behavior in high-speed networks, *Teletraffic Contributions for the Information Age, Proceedings of ITC 15*, V. Ramaswami and P. E. Wirth (eds.), Elsevier, Amsterdam, 1997, 1129–1138.

[10] E. Çinlar, Superposition of point processes, in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P. A. W. Lewis (ed.), Wiley, New York, 1972, 549–606.

[11] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, second ed., Academic, New York, 1984.

[12] N. G. Duffield, W. A. Massey, and W. Whitt, An offered-load model for packet networks with a nonhomogeneous Poisson connection arrival process, in preparation.

[13] N. G. Duffield and N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications, *Math. Proc. Camb. Phil. Soc.* **118** (1995), 363–374.

[14] N. G. Duffield and W. Whitt, Control and recovery from rare congestion events in a large multi-server system, *Queueing Systems* **26** (1997), 69–104.

[15] N. G. Duffield and W. Whitt, A source traffic model and its transient analysis for network control, *Stochastic Models* **14** (1998), 51–78.

[16] N. G. Duffield and W. Whitt, A source traffic model and its steady-state analysis for network design, AT&T Labs, 1997.

[17] N. G. Duffield and W. Whitt, Network design and control using on-off and multi-level source traffic models with long-tailed distributions, AT&T Labs, 1997.

[18] S. G. Eick, W. A. Massey and W. Whitt, The physics of the $M_t/G/\infty$ queue, *Operations Res.* **41** (1993), 731–742.

[19] A. G. Greenberg, R. Srikant, and W. Whitt, Resource sharing for book-ahead and instantaneous-request calls. *Teletraffic Contributions for the Information Age, Proceedings of ITC 15*, V. Ramaswami and P. E. Wirth (editors), Elsevier, Amsterdam, 1997, 539–548.

[20] R. Guerin, H. Ahmadi and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE J. Sel. Areas Commun.* **SAC9** (1991), 968–991.

[21] O. Kella and W. Whitt, Linear stochastic fluid networks, *J. Appl. Prob.* (1998), to appear.

[22] F. P. Kelly, Notes on effective bandwidths, in *Stochastic Networks, Theory and Applications*, F. P. Kelly, S. Zachary and I. Ziedins (eds.), Clarendon Press, Oxford, 1996, 141–168.

[23] T. G. Kurtz, Limit theorems for workload input models, in *Stochastic Networks, Theory and Applications*, F. P. Kelly, S. Zachary and I. Ziedins, Clarendon Press, Oxford, 1996, 119–139.

[24] C. Klüppelberg and T. Mikosch, Explosive Poisson shot noise processes with application to risk reserves, *Bernoulli* **1** (1995), 125–147.

[25] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, On the self-similar nature of Ethernet traffic, *IEEE/ACM Trans. Networking* **2** (1994), 1–15.

[26] K. K. Leung, W. A. Massey and W. Whitt, Traffic models for wireless communication networks, *IEEE J. Sel. Areas Commun.* **12** (1994), 1353–1364.

[27] W. A. Massey and W. Whitt, Networks of infinite-server queues with nonstationary Poisson input, *Queueing Systems* **13** (1993), 183–250.

[28] W. A. Massey and W. Whitt, A stochastic model to capture space and time dynamics in wireless communication systems, *Prob. Eng. Inf. Sci.* **8** (1994), 541–569.

[29] W. A. Massey and W. Whitt, Peak congestion in multi-server service systems with slowly varying arrival rates, *Queueing Systems* **25** (1997), 157–172.

[30] W. A. Massey and W. Whitt, A probabilistic generalization of Taylor's theorem, *Statistics and Probability Letters*, **16** (1993), 51–54.

[31] I. Norros, A storage model with self-similar input, *Queueing Systems* **16** (1994), 387–396.

[32] I. Norros, On the use of fractal Brownian motion in the theory of connectionless networks, *IEEE J. Sel. Areas Commun.* **13** (1995), 953–962.

[33] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

[34] P. F. Pawlita, Traffic measurements in data networks, recent measurement results, and some implications, *IEEE Trans. Commun.* **29** (1981), 525–535.

[35] V. Paxson and S. Floyd, Wide-area traffic: the failure of Poisson modeling, *IEEE/ACM Trans. Networking* **3** (1995), 226–244.

[36] J. Rice, On generalized shot noise, *Adv. Appl. Prob.* **9** (1977), 553–565.

[37] W. Willinger, M. S. Taqqu and A. Erramilli, A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks, in *Stochastic Networks*, F. P. Kelly, S. Zachary and I. Ziedins (eds.), Oxford University Press, Oxford, 1996, 339–366.