Telecommunication Systems 16:3,4, 271–296, 2001

A Nonstationary Offered-Load Model for Packet Networks

N.G. DUFFIELD duffield@research.att.com
AT&T Labs, Room B139, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA

W.A. MASSEY will@research.bell-labs.com
Bell Laboratories, Lucent Technologies, Room 2C-320, 700 Mountain Avenue, Murray Hill, NJ 07974-0636, USA

W. WHITT wow@research.att.com
AT&T Labs, Room A117, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA

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.

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

AMS subject classification: primary: 60K25; secondary: 60H05, 60K30

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 [37], Cáceres et al. [8], Leland et al. [27], Paxson and Floyd [38] and Willinger et al. [41]. 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 [38], and Feldman et al. [19]. 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 [12]. 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 [15–17] by considering *nonstationary* customer arrival processes. In [15–17] 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 [15–17] 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 [15–17] for additional motivation.

Second, the present paper extends the nonstationary Poisson-arrival-location model (PALM) for wireless networks investigated by Massey and Whitt [30,31] and Leung et al. [28] 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.

The nonstationary model for wired networks proposed here is also a generalization of the queueing network with M_t arrival processes considered in Massey and Whitt [30]. Instead of just customers, here we have customers with general evolving bandwidth requirements. Nevertheless, the theory here is similar to the theory for more elementary $M_t/G/\infty$ models, for which there is a long history. For background, including a review of previous literature, see [18,23,30].

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 [32] 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 construct and characterize the basic bandwidth stochastic processes, using the theory developed in section 6. We give expressions for the mean, variance, higher cumulants and covariances of the total-required-bandwidth process in the offered load model for a single link. We also show how we can model the cumulative packet arrival process for a single link and extend the model to cover the case of multiple links in a communication network. We characterize the finite-dimensional distributions of all these processes. In section 3 we establish 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 et al. [4], Choe and Shroff [10] and Norros [34,35].

In section 4 we give illustrative examples to highlight the insights that can be gained from the nonstationary offered-load model. We show how the peak offered load lags behind the peak connection arrival rate in two scenarios: first, with a quadratic arrival rate (which approximates the behavior near the peak of general arrival-rate functions) and, second, with a traffic surge (where the arrival rate has a constant higher value above a stationary background rate in a subinterval). Moreover, we can see the influence of the connection holding-time distribution (e.g., if it is heavy-tailed) upon the offered load. The traffic surge example is similar to the traffic accident example in Leung et al. [28].

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, as in Whitt [40].

We conclude in section 6 by giving proofs. We show that the various bandwidth processes can all be characterized by considering stochastic integration over Poisson processes with a special integrand.

2. Constructing and characterizing the bandwidth processes

In this section we define and characterize the stochastic processes 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* that is defined by a *Poisson random measure* A, where we let $A(s,t]$ count the *number of customer connections arriving* in the interval $(s,t]$ for all $s < t$. To say that A is a Poisson random measure is equivalent to assuming that the random number of arriving customer connections during disjoint time intervals are mutually independent. We also assume that A has an *intensity function* $\alpha(t)$. This is a locally integrable function such that

$$\mathsf{E}\big[A(s,t]\big] = \int_s^t \alpha(r)\,\mathrm{d}r < \infty \tag{2.1}$$

for all $s < t$. From this it follows that

$$\mathsf{P}\big(A(s,t] = n\big) = \frac{(\int_s^t \alpha(r)\,\mathrm{d}r)^n}{n!} \exp\biggl(-\int_s^t \alpha(r)\,\mathrm{d}r\biggr) \tag{2.2}$$

for all positive integers n.

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 $\{B(s,t) \mid t \geqslant s\}$ as being a collection of mutually independent random processes indexed by real s with probability laws depending on s. Just as in [30, p. 196], 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. See section 6 for more discussion.

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 [39], Klüppelberg and Mikosch [26] 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.,

$$\bigl\{B(s,t) \mid t \geqslant s\bigr\} \stackrel{\mathrm{d}}{=} \bigl\{B(0,t-s) \mid t \geqslant s\bigr\} \tag{2.3}$$

for all (s,t) with $s < t$. Note however that the stochastic process $\{B(t) \mid t \geqslant 0\}$ 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 [30,31]; see section 6 for more discussion. The total bandwidth (rate) required at time t is then

$$R(t) = \int_{-\infty}^{t} B(s,t)\,\mathrm{d}A(s) \equiv \lim_{\tau \downarrow -\infty} \int_{\tau}^{t} B(s,t)\,\mathrm{d}A(s), \tag{2.4}$$

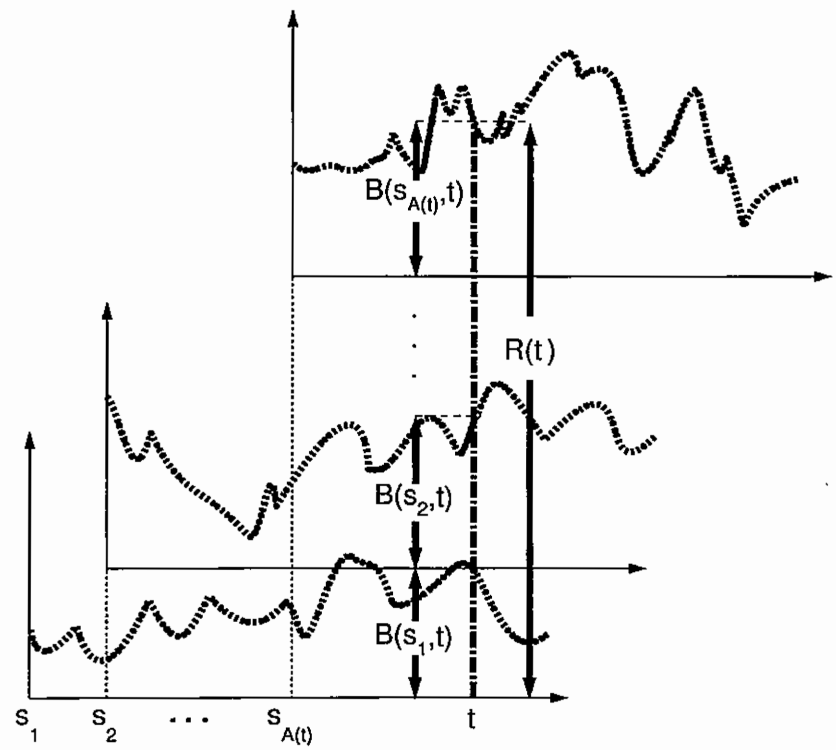

Figure 1. Typical realizations of individual bandwidth processes associated with $A(t)$ active connections at time t. Also displayed is $R(t)$, the total required bandwidth at time t.

where $\mathrm{d}A(s) \equiv \lim_{r\uparrow s} A(r, s]$. 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 [30,31]. 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.

We now give formulas for the characteristic function, mean, variance and higher cumulants of the total bandwidth (for this one customer class). Recall that the *characteristic function* of a random variable X (or its probability distribution) is $\mathsf{E}[\mathrm{e}^{\mathrm{i}\theta X}]$ for real θ where $\mathrm{i} = \sqrt{-1}$. The characteristic function uniquely characterizes the distribution of X. For a non-negative random variable, the characteristic function is complex analytic (has a convergent power series expansion) at zero if and only if $\mathsf{E}[\mathrm{e}^{\theta X}] < \infty$ for some strictly positive θ. Moreover, $\mathsf{E}[\mathrm{e}^{\theta X}]$ is the *moment generating function* (mgf) for

X as well as the analytic continuation of the characteristic function for X when θ is pure imaginary instead of real. Since the mgf is analytic in an open neighborhood of zero where its value is 1, it follows that $\log \mathsf{E}[e^{\theta X}]$ is also analytic at zero and has a unique convergent power series too. We can then define the *cumulants* or cumulant moments of X to be $\mathsf{C}^{(n)}[X]$ for all positive integers n, where

$$\log \mathsf{E}\big[\exp(\theta X)\big] = \sum_{n=1}^{\infty} \frac{\theta^n}{n!} \mathsf{C}^{(n)}[X], \tag{2.5}$$

for some sufficiently small but positive θ.

From (2.5) 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

$$\mathsf{C}^{(n)}[X+Y] = \mathsf{C}^{(n)}[X] + \mathsf{C}^{(n)}[Y] \quad \text{and} \quad \mathsf{C}^{(n)}[\lambda X] = \lambda^n \mathsf{C}^{(n)}[X]. \tag{2.6}$$

The first two cumulants are the mean and variance or

$$\mathsf{C}^{(1)}[X] = \mathsf{E}[X] \quad \text{and} \quad \mathsf{C}^{(2)}[X] = \mathsf{Var}[X]. \tag{2.7}$$

The covariance of X and Y can be written in terms of second order cumulants as

$$\mathsf{Cov}[X, Y] = \frac{1}{2}\big(\mathsf{C}^{(2)}[X+Y] - \mathsf{C}^{(2)}[X] - \mathsf{C}^{(2)}[Y]\big). \tag{2.8}$$

Finally, a Poisson distribution is uniquely characterized as a distribution whose cumulants are all equal to its mean, while a Gaussian distribution is uniquely characterized as a distribution whose cumulants of order 3 and greater are all equal to zero.

In section 6, we prove a fundamental theorem for a special class of stochastic integrals that provide the key result for our packet-network offered-load model. We use this result to characterize the finite-dimensional distributions of the stochastic process $\{R(t) \mid -\infty < t < \infty\}$.

Theorem 2.1. Let $t_1 < t_2 < \cdots < t_k$ be k increasing time points and $\theta_1, \ldots, \theta_k$ be k arbitrary real numbers. If we have

$$\int_{-\infty}^{t_j} \mathsf{E}\big[B(s, t_j)\big]\alpha(s)\,\mathrm{d}s < \infty \tag{2.9}$$

for all $j = 1, \ldots, k$, then the characteristic function for the joint distribution of $R(t_1), \ldots, R(t_k)$ is

$$\mathsf{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^{k}\theta_j R(t_j)\right)\right] = \exp\left(\int_{-\infty}^{t_k} \mathsf{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^{k}\theta_j B(s, t_j)\right) - 1\right]\alpha(s)\,\mathrm{d}s\right). \tag{2.10}$$

Now, assuming that the mgf exists as well, we can also obtain the associated cumulants.

Theorem 2.2. If, in addition to the hypothesis of theorem 2.1, we have

$$\int_{-\infty}^{t_k} \mathsf{E}\left[\exp\left(\sum_{j=1}^{k}\widehat{\theta}_j B(s,t_j)\right) - 1\right]\alpha(s)\,\mathrm{d}s < \infty \tag{2.11}$$

for some positive $\widehat{\theta}_1, \ldots, \widehat{\theta}_k$, then

$$\mathsf{C}^{(n)}\left[\sum_{j=1}^{k}\theta_j R(t_j)\right] = \int_{-\infty}^{t_k} \mathsf{E}\left[\left(\sum_{j=1}^{k}\theta_j B(s,t_j)\right)^n\right]\alpha(s)\,\mathrm{d}s \tag{2.12}$$

for all positive integers n and all real $\theta_1, \ldots, \theta_k$.

As simple consequences, we obtain 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 $\alpha(t)$.

Corollary 2.3. The nth cumulant of $R(t)$ in (2.4) is

$$\mathsf{C}^{(n)}\big[R(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[B(s,t)^n\big]\alpha(s)\,\mathrm{d}s, \tag{2.13}$$

from which we get

$$\mathsf{E}\big[R(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[B(s,t)\big]\alpha(s)\,\mathrm{d}s, \qquad \mathsf{Var}\big[R(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[B(s,t)^2\big]\alpha(s)\,\mathrm{d}s, \tag{2.14}$$

and, for all $t_1 < t_2$,

$$\mathsf{Cov}\big[R(t_1), R(t_2)\big] = \int_{-\infty}^{t_1} \mathsf{E}\big[B(s,t_1)B(s,t_2)\big]\alpha(s)\,\mathrm{d}s. \tag{2.15}$$

Given the values of the expectations in the integrands of (2.13)–(2.15), we can compute the displayed quantities by performing numerical integration, e.g., see Davis and Rabinowitz [13]. 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

$$I(t,t') \equiv \int_{t}^{t'} R(s)\,\mathrm{d}s = \int_{-\infty}^{t'} C_s(t,t')\,\mathrm{d}A(s), \tag{2.16}$$

where

$$C_s(t,t') \equiv \int_{t}^{t'} B(s,\tau)\,\mathrm{d}\tau. \tag{2.17}$$

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 [24], we define an *effective-bandwidth function* by

$$\beta_\theta(t, t') \equiv \frac{1}{\theta \cdot (t' - t)} \log \mathsf{E}\big[\exp\big(\theta \cdot I(t, t')\big)\big]. \tag{2.18}$$

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 [9] for further discussion of the nonstationary case.

Now we characterize the finite-dimensional distributions of the total cumulative input 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. If we have

$$\int_{-\infty}^{t_j} \mathsf{E}\big[C_s\big(t_j, t_j'\big)\big]\alpha(s)\,\mathrm{d}s < \infty \tag{2.19}$$

for all $j = 1, \ldots, k$, then the characteristic function for the joint distribution of $I(t_1, t_1'), \ldots, I(t_k, t_k')$ is

$$\mathsf{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^{k} \theta_j I\big(t_j, t_j'\big)\right)\right] = \exp\left(\int_{-\infty}^{t^*} \mathsf{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^{k} i\theta_j C_s\big(t_j, t_j'\big)\right) - 1\right]\alpha(s)\,\mathrm{d}s\right), \tag{2.20}$$

where $t^* = \max(t_1', \ldots, t_k')$.

Theorem 2.5. If, in addition to the hypothesis of theorem 2.4,

$$\int_{-\infty}^{t^*} \mathsf{E}\left[\exp\left(\sum_{j=1}^{k} \widehat{\theta}_j C_s\big(t_j, t_j'\big)\right) - 1\right]\alpha(s)\,\mathrm{d}s < \infty \tag{2.21}$$

for some positive $\widehat{\theta}_1, \ldots, \widehat{\theta}_k$, then

$$\mathsf{C}^{(n)}\left[\sum_{j=1}^{k} \theta_j I\big(t_j, t_j'\big)\right] = \int_{-\infty}^{t^*} \mathsf{E}\left[\left(\sum_{j=1}^{k} \theta_j C_s\big(t_j, t_j'\big)\right)^n\right]\alpha(s)\,\mathrm{d}s \tag{2.22}$$

for all real $\theta_1, \ldots, \theta_k$.

A simple consequence of this theorem is

Corollary 2.6. For all intervals (t_i, t_i') and (t_j, t_j'),

$$\mathsf{Cov}\big[I\big(t_i, t_i'\big), I\big(t_j, t_j'\big)\big] = \int_{-\infty}^{\min(t_i', t_j')} \mathsf{E}\big[C_s\big(t_i, t_i'\big)C_s\big(t_j, t_j'\big)\big]\alpha(s)\,\mathrm{d}s. \tag{2.23}$$

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

$$\mathbf{R}(t) = \left(R^{(1)}(t), \ldots, R^{(L)}(t)\right) \tag{2.24}$$

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

$$\mathbf{R}(t) = \int_{-\infty}^{t} \mathbf{B}(s, t)\, \mathrm{d}A(s), \tag{2.25}$$

where

$$\mathbf{B}(s, t) = \left(B^{(1)}(s, t), \ldots, B^{(L)}(s, t)\right) \tag{2.26}$$

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 $t \geqslant s$. The total bandwidth required at time t and link ℓ is then the same as for the single link case:

$$R^{(\ell)}(t) = \int_{-\infty}^{t} B^{(\ell)}(s, t)\, \mathrm{d}A(s). \tag{2.27}$$

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. Applying theorem 6.1, we obtain

Theorem 2.7. Let $t_1 < t_2 < \cdots < t_k$ be k time points and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ be k arbitrary L-dimensional vectors. If we have

$$\int_{-\infty}^{t_j} \mathsf{E}\left[B^{(\ell)}(s, t_j)\right]\alpha(s)\, \mathrm{d}s < \infty \tag{2.28}$$

for all $j = 1, \ldots, k$ and $\ell = 1, \ldots, L$, then the characteristic function for the joint distribution of $\mathbf{R}(t_1), \ldots, \mathbf{R}(t_k)$ is

$$\mathsf{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^{k} \boldsymbol{\theta}_j \cdot \mathbf{R}(t_j)\right)\right] = \exp\left(\int_{-\infty}^{t_k} \mathsf{E}\left[\exp\left(\mathrm{i}\sum_{j=1}^{k} \boldsymbol{\theta}_j \cdot \mathbf{B}(s, t_j)\right) - 1\right]\alpha(s)\, \mathrm{d}s\right). \tag{2.29}$$

Theorem 2.8. If, in addition to the hypothesis of theorem 2.7,

$$\int_{-\infty}^{t_k} \mathsf{E}\left[\exp\left(\sum_{j=1}^{k} \widehat{\boldsymbol{\theta}}_j \cdot \mathbf{B}(s, t_j)\right) - 1\right]\alpha(s)\,\mathrm{d}s < \infty \tag{2.30}$$

for some strictly positive L-dimensional vectors $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_k$, then we have

$$\mathsf{C}^{(n)}\left[\sum_{j=1}^{k} \boldsymbol{\theta}_j \cdot \mathbf{R}(t_j)\right] = \int_{-\infty}^{t_k} \mathsf{E}\left[\left(\sum_{j=1}^{k} \boldsymbol{\theta}_j \cdot \mathbf{B}(s, t_j)\right)^n\right]\alpha(s)\,\mathrm{d}s \tag{2.31}$$

for all L-dimensional vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$.

Corollary 2.9. For all links ℓ_1 and ℓ_2 and all times $t_1 \leqslant t_2$, we have

$$\mathsf{Cov}\big[R^{(\ell_1)}(t_1), R^{(\ell_2)}(t_2)\big] = \int_{-\infty}^{t_1} \mathsf{E}\big[B^{(\ell_1)}(s, t_1)B^{(\ell_2)}(s, t_2)\big]\alpha(s)\,\mathrm{d}s. \tag{2.32}$$

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 $\{R(t) \mid -\infty < t < \infty\}$. We prove this in section 6.

Theorem 3.1. Consider a family of models indexed by $\eta > 0$, with common bandwidth processes, where the arrival rate in model has the form

$$\alpha^{\eta}(t) = \eta\alpha^{(0)}(t) + \sqrt{\eta}\alpha^{(1)}(t) + \mathrm{o}\big(\sqrt{\eta}\big) \tag{3.1}$$

as $\eta \to \infty$ for all t, where

$$\lim_{\eta\to\infty} \int_{-\infty}^{t} \mathsf{E}\big[B(s, t)^k\big]\big(\alpha^{(0)}(s) + \big|\alpha^{(1)}(s)\big|\big)\,\mathrm{d}s < \infty \tag{3.2}$$

and

$$\lim_{\eta\to\infty} \int_{-\infty}^{t} \mathsf{E}\big[B(s, t)^k\big]\left|\frac{\alpha^{\eta}(s) - \eta\alpha^{(0)}(s)}{\sqrt{\eta}} - \alpha^{(1)}(s)\right|\mathrm{d}s = 0 \tag{3.3}$$

for all t and $k = 1, 2, 3$. If $\{R^{(0)}(t) \mid -\infty < t < \infty\}$ is a total required bandwidth process with a Poisson intensity function $\alpha^{(0)}$, then the finite-dimensional distributions of the process $\{X^{\eta}(t) \mid -\infty < t < \infty\}$ given by

$$X^{\eta}(t) \equiv \frac{R^{\eta}(t) - \eta\mathsf{E}[R^{(0)}(t)]}{\sqrt{\eta}} \tag{3.4}$$

converge in distribution as $\eta \to \infty$ to a Gaussian process with the drift given by $\int_{-\infty}^{t} \mathsf{E}[\phi(Z,s)]\alpha^{(1)}(s)\,\mathrm{d}s$ and the same covariances as formulas for $\{R^{(0)}(t) \mid -\infty < t < \infty\}$ in (2.15).

The form of α^{η} in (3.1) follows Mandelbaum et al. [29]. If instead we simply have $\alpha^{\eta} = \eta\alpha^{(0)}$ for η a positive integer, then $\{R^{\eta}(t) \mid -\infty < t < \infty\}$ is the sum of η i.i.d. processes, each distributed as $\{R^{1}(t) \mid -\infty < t < \infty\}$ and the classical CLT applies directly, yielding a zero-mean Gaussian limit after scaling.

Note that corresponding CLTs hold for normalized versions of the stochastic process $I(t,t')$ and $\mathbf{R}(t)$ by the same argument. Given the Gaussian approximation, we can compute the probability that demand exceeds the critical level, namely

$$\mathsf{P}\big(R(t) \geqslant L\big) \approx \Phi^{c}\left(\frac{L-m}{\sigma}\right) \equiv \frac{1}{\sqrt{2\pi}\,\sigma}\int_{L}^{\infty} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)\mathrm{d}x, \tag{3.5}$$

where $m = \mathsf{E}[R(t)]$, $\sigma^2 = \mathsf{Var}[R(t)]$, Φ^{c} is defined to be the *associated complementary distribution function* for the standard (mean 0, variance 1) Gaussian distribution. If the level L in (3.5) 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 et al. [22] in a refinement to their "equivalent capacity" scheme (in a stationary setting). With highly bursty traffic, we think that it may be appropriate to use only (3.5).

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

$$\mathsf{E}\big[\big(R(t)-L\big)^{+}\big] = \mathsf{E}\big[R(t)-L \mid R(t) > L\big]\mathsf{P}\big(R(t) > L\big). \tag{3.6}$$

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

$$\begin{aligned}\mathsf{E}\big[R(t)-L \mid R(t) > L\big] &\approx \mathsf{E}\big[N\big(m-L,\sigma^2\big) \mid N\big(m-L,\sigma^2\big) > 0\big]\\ &= m-L+\sigma\mathsf{E}\big[N(0,1) \mid N(0,1) > (L-m)/\sigma\big]\\ &= m-L+\sigma\frac{\phi((L-m)/\sigma)}{\Phi^{c}((L-m)/\sigma)}.\end{aligned} \tag{3.7}$$

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

$$\mathsf{P}\big(R_1(t)+\cdots+R_j(t) > L_j\big) < \varepsilon_j, \quad 1 \leqslant j \leqslant k, \tag{3.8}$$

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 Ahmadi et al. [5, pp. 608–609]. We contribute by providing a model leading to formulas for the required means and variances in a nonstationary setting.

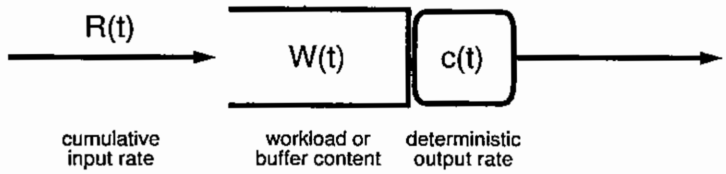

Figure 2. The fluid queue with infinite capacity buffer and a time-dependent deterministic output rate.

We now apply the Gaussian process approximation to the cumulative input process $\{I(t,t') \mid -\infty < t < t' < \infty\}$ 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 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 et al. [21]. We develop an approximation for the buffer-content distribution, which is necessarily time-dependent because the input and output are nonstationary. In figure 2 we depict the fluid queue model being considered.

Let $W(t)$ denote the *total workload or buffer content* at time t. Since (2.9) holds or $\int_{-\infty}^{t} \mathsf{E}[B(s,t)]\alpha(s)\,\mathrm{d}s < \infty$, $R(t)$ is almost surely finite. Assuming that $\int_{-\infty}^{t} c(s)\,\mathrm{d}s = \infty$ but $\int_{\tau}^{t} c(s)\,\mathrm{d}s < \infty$ for all $\tau < t$, we can deduce that $W(t)$ is also almost surely finite for all t and

$$W(t) = \sup_{s \leqslant t}\left(I(s,t) - \int_{s}^{t} c(\tau)\,\mathrm{d}\tau \right). \tag{3.9}$$

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

$$\mathsf{P}\big(W(t) > x\big) \geqslant \sup_{s \leqslant t} \mathsf{P}\left(I(s,t) - \int_{s}^{t} c(\tau)\,\mathrm{d}\tau > x \right) \quad \text{for all } x > 0, \tag{3.10}$$

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 [14] and Botvich and Duffield [7]. However, without resorting to simplifying asymptotics, even the lower bound in (3.10) is very complicated. As noted by, e.g., Addie and Zukerman [3], Addie et al. [4], Choe and Shroff [10] and Norros [34,35], the lower bound in (3.10) greatly simplifies if we approximate the total cumulative input process $I(s,t)$ by a Gaussian process, which is often reasonable, because we can apply the CLT. Then the *net cumulative input process* $\{I(s,t) - \int_{s}^{t} c(\tau)\,\mathrm{d}\tau \mid s \leqslant t\}$ also becomes a time-reversed Gaussian process in s. We can then find the maximizing s in the right side of (3.10) in terms of the time-dependent means and variances of $I(s,t)$. Recall that

$$\mathsf{E}\big[I(s,t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[C_\tau(s,t)\big]\alpha(\tau)\,\mathrm{d}\tau \quad \text{and} \quad \mathsf{Var}\big[I(s,t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[C_\tau(s,t)^2\big]\alpha(\tau)\,\mathrm{d}\tau. \tag{3.11}$$

Now let

$$Z(s,t) = \frac{I(s,t) - \mathsf{E}[I(s,t)]}{x - \mathsf{E}[I(s,t)] + \int_s^t c(\tau)\,\mathrm{d}\tau}, \quad s \leqslant t, \tag{3.12}$$

where x come from the level specified in (3.10) and note that

$$\left\{I(s,t) - \int_s^t c(\tau)\,\mathrm{d}\tau > x\right\} = \big\{Z(s,t) \geqslant 1\big\}. \tag{3.13}$$

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

$$\sup_{s\leqslant t} \mathsf{P}\left(I(s,t) - \int_s^t c(\tau)\,\mathrm{d}\tau > x\right) = \sup_{s\leqslant t} \mathsf{P}\big(Z(s,t) \geqslant 1\big) = \Phi^c\left(\frac{1}{\sqrt{\mathsf{Var} Z(s^*,t)}}\right), \tag{3.14}$$

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

$$\mathsf{Var}\big[Z(s,t)\big] = \frac{\mathsf{Var}[I(s,t)]}{(x - \mathsf{E}[I(s,t)] + \int_s^t c(\tau)\,\mathrm{d}\tau)^2} \tag{3.15}$$

over all $s \leqslant t$. (We assume that the supermum over s in (3.15) is attained.)

In the stationary context, $\mathsf{E}[I(s,t)] - \int_s^t c(\tau)\,\mathrm{d}\tau$ 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 $c(t) \geqslant \gamma$ for all t, than we can conclude that $E[I(s,t)] - \int_s^t c(\tau)\,\mathrm{d}\tau \geqslant -\gamma \cdot (t-s)$ and that $E[I(s,t)] - \int_s^t c(\tau)\,\mathrm{d}\tau$ 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 $\{R(t) \mid t \geqslant 0\}$ is a stationary stochastic process and $\{I(s,t) \mid s \leqslant t\}$ has stationary increments), the stochastic process $\{Z(s,t) \mid s \leqslant t\}$ is not itself a stationary process. In the stationary case, Choe and Shroff [10] have shown that the approximation (3.14) performs remarkably well, even outside the large-deviations asymptotic regime. Hence, (3.14) is a promising approximation in the nonstationary case as well.

By calculating (3.14) 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

$$\mathsf{P}\big(W(t^*) > x\big) \approx \sup_t \mathsf{P}\big(W(t) > x\big). \tag{3.16}$$

4. Physics: 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 et al. [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 nth arriving connection, and $\{T_n \mid n \geqslant 1\}$ is a sequence of i.i.d. non-negative random variables with finite mean. 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 $t \geqslant 0$,

$$\mathsf{P}(T_e \leqslant t) \equiv \frac{\int_0^t \mathsf{P}(T > s)\,\mathrm{d}s}{\mathsf{E}[T]}. \tag{4.2}$$

The next result shows that the cumulants $\mathsf{C}^{(n)}[R(t)]$ depend on the distribution T through its mean $\mathsf{E}[T]$ and the distribution of T_e.

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

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

for all positive integers n and real t.

Proof. Substituting (4.1) into (2.13), we have

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

Now we use the identity established in Massey and Whitt [33] for all functions f continuously differentiable on $[0,\infty)$ and $x \geqslant 0$,

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

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

For the next theorem, we introduce some notation. Let X and Y be any two non-negative random variables. If f is any non-negative function such that both $\mathsf{E}[f(X)]$

and $\mathsf{E}[Yf(X)]$ are finite and positive, then f induces a new probability measure or expectation E_f on Y such that

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

We also define $\mathsf{Var}_f[Y] \equiv \mathsf{E}_f[Y^2] - \mathsf{E}_f[Y]^2$, provided $\mathsf{E}[Y^2 f(X)] < \infty$.

We now consider the case of a quadratic arrival-rate function. We are especially interested in the case in which the quadratic function has a finite maximum, because then the quadratic function can serve as an approximation (e.g., Taylor series expansion) for a more general arrival-rate function near its peak. As in Eick et al. [18], we apply formula (4.3) directly even though $\alpha(t)$ will typically be negative for times t in the past. In applications, we need to check that the simple formula so produced is indeed a good approximation for the actual (non-negative) arrival-rate function. As discussed in [18], experience indicates that it often is. The insight gained justifies the simplifying assumption.

The next theorem shows how the cumulants of $R(t)$ depend on the *stochastic* distributions of T and the *deterministic* fluctuations of $b(t)$ in the case of a quadratic arrival rate.

Theorem 4.2. If, in addition to (4.1), $\alpha(t) = a_0 + a_1 t - a_2 t^2$, then for all integers $n \geqslant 1$

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

We now apply theorem 4.2 to describe the time lags and space shifts in the cumulants of $R(t)$ compared to the peak connection arrival rate.

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

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

and

$$\mathsf{C}^{(n)}\big[R(t_n^*)\big] = \big(\alpha^* - a_2 \mathsf{Var}_{b^n}[T_e]\big) \cdot \mathsf{E}\big[b(T_e)^n\big] \cdot \mathsf{E}[T]. \tag{4.9}$$

When $a_2 > 0$, the extremal value $\mathsf{C}^{(n)}[R(t_n^*)]$ is a maximum. This corollary tells us that there is always a non-negative *time lag* $\ell_n \equiv \mathsf{E}_{b^n}[T_e]$ between the time t_α^* of the extremal value α^* for α and t_n^*, the time for the extremal value of the nth 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

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

for all positive integers n and real t. When $\mathsf{C}^{(n)}[R(t_n^*)]$ is a maximum, it is strictly less than the *pointwise stationary value* (4.10); observe that T_e can never be a constant. Hence, just as in [18], we have space shifts as well as time lags.

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

$$\mathsf{C}^{(n)}\big[R(t)\big] = \big(\alpha\big(t - \mathsf{E}[T_e \mid T_e \in A]\big) - a_2 \mathsf{Var}[T_e \mid T_e \in A]\big) \cdot b^n \mathsf{P}(T_e \in A) \cdot \mathsf{E}[T], \tag{4.11}$$

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

Observe that, for any constant c,

$$\mathsf{E}[T_e \mid T_e \geqslant c] \geqslant c \quad \text{and} \quad \mathsf{E}[T_e \mid T_e \leqslant c] \leqslant c, \tag{4.12}$$

assuming respectively that $\mathsf{P}(T_e \geqslant c) > 0$ and $\mathsf{P}(T_e \leqslant c) > 0$. 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.

In order to establish some further consequences of theorem 4.2, we need the following result.

Lemma 4.5. Let X be any non-negative random variable with a finite mean. If f and g are two non-negative, 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

$$\mathsf{E}_f[X] \geqslant \mathsf{E}_g[X]. \tag{4.13}$$

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

$$\mathsf{E}_f[X] \leqslant \mathsf{E}_g[X]. \tag{4.14}$$

Proof. The identity function $I(x) = x$ is an increasing function. If ϕ is nondecreasing, then the result follows immediately from the FKG inequality (see Fortuin et al. [20]) since

$$\mathsf{E}_f[X] = \mathsf{E}_{\phi g}[X] = \frac{\mathsf{E}_g[X\phi(X)]}{\mathsf{E}_g[\phi(X)]} \geqslant \frac{\mathsf{E}_g[X]\mathsf{E}_g[\phi(X)]}{\mathsf{E}_g[\phi(X)]} = \mathsf{E}_g[X]. \tag{4.15}$$

Since $-I$ is a decreasing function, the reverse inequality will hold when ϕ is a non-increasing function. □

Corollary 4.6. Under the hypothesis of theorem 4.2, Let b_1 and b_2 be two bandwidth functions. If, in addition to the hypothesis of theorem 4.2, $b_1 = \phi b_2$, where ϕ is a non-negative, nondecreasing function on the support of T_e, then the time lags ℓ_n due to b_1, using the same α and T, are always larger than the ones due to b_2.

This corollary yields yet another result of interest.

Corollary 4.7. If b is a nondecreasing function on the support of T_e, then the time lags ℓ_n are ordered by

$$0 \leqslant \mathsf{E}[T_e] \leqslant \ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_n \leqslant \cdots. \tag{4.16}$$

Similarly, if b in non-increasing, then

$$0 \leqslant \cdots \leqslant \ell_n \leqslant \cdots \leqslant \ell_2 \leqslant \ell_1 \leqslant \mathsf{E}[T_e]. \tag{4.17}$$

These results say that if b is nondecreasing, α is quadratic, and $a_2 > 0$, then $t_1^* \leqslant t_2^*$ 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 $t_2^* \leqslant t_1^*$.

Now, to model a traffic surge, instead of a quadratic arrival-rate function, we let

$$\alpha(t) = \begin{cases} a, & \text{for } 0 \leqslant t \leqslant t_0, \\ 0, & t < 0 \text{ and } t > t_0. \end{cases} \tag{4.18}$$

The arrival-rate function in (4.18) represents the increment beyond a constant stationary arrival rate. We neglect the stationary part because we can treat the two components separately and add the resulting independent random variables.

In this setting we can apply proposition 4.1 to obtain the following result.

Corollary 4.8. Under (4.1) and (4.18),

$$\mathsf{C}^{(n)}\big[R(t)\big] = a\mathsf{E}\big[b(T_e)^n;\, t - t_0 < T_e \leqslant t\big] \cdot \mathsf{E}[T]. \tag{4.19}$$

If, in addition, $b(t) = b$, then

$$\mathsf{C}^{(n)}\big[R(t)\big] = ab^n\mathsf{P}(t - t_0 < T_e \leqslant t) \cdot \mathsf{E}[T]. \tag{4.20}$$

Formula (4.20) is a minor variation of the $M_t/G/\infty$ result in section 5 of Eick et al. [18].

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 [15–17], 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 [15–17].

We focus on a single customer class at a single link and assume that we can add the results for different classes. As in Whitt [40], 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 (2.4), (2.14) and (2.15) 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 $(B_i(t) \mid I_i(0))$ 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 Parthasarathy [36, chapter V]. In particular, assuming that the information $I_i(0)$ is observed, $(B_i(t) \mid I_i(0))$ 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 $(R(t) \mid I(0))$ 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

$$\mathsf{E}\big[R(t) \mid I(0)\big] = \sum_{i=1}^{n} \mathsf{E}\big[B_i(t) \mid I_i(0)\big] + \int_0^t \mathsf{E}\big[B(s, t)\big]\alpha(s)\,\mathrm{d}s, \tag{5.1}$$

$$\mathrm{Var}\big[R(t) \mid I(0)\big] = \sum_{i=1}^{n} \mathrm{Var}\big[B_i(t) \mid I_i(0)\big] + \int_0^t \mathrm{E}\big[B(s,t)^2\big]\alpha(s)\,\mathrm{d}s. \tag{5.2}$$

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

$$\begin{aligned}&\mathrm{Cov}\big[R(t_1), R(t_2) \mid I(0)\big]\\ &\quad= \sum_{i=1}^{n} \mathrm{Cov}\big[B_i(t_1)B_i(t_2) \mid I_i(0)\big] + \int_0^t \mathrm{E}\big[B(s,t_1)B(s,t_2)\big]\alpha(s)\,\mathrm{d}s.\end{aligned} \tag{5.3}$$

Following Duffield and Whitt [15,16] we propose as a first-order approximation for the conditional total required bandwidth $(R(t) \mid I(0))$ 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.

6. Fundamental theorems and proofs

Our offered-load model can be analyzed by considering stochastic integrals of a nonhomogenous Poisson process with integrands that have a special structure. We now develop the theory for such integrals. As in section 2 of [30], let $\{Z_n \mid n = 1, 2, \ldots\}$ be an i.i.d. sequence of random elements belonging to some complete separable metric space Σ and distributed as Z. The idea is that Σ can be the function space D of right-continuous real-valued functions on $(-\infty, \infty)$ with left limits, endowed with an appropriate topology, so that Σ contains the sample paths of an individual bandwidth process $\{B(s,t) \mid -\infty < t < \infty\}$. We say that $\phi : \Sigma \times \mathbb{R} \to \mathbb{R}$ is an *integrand* with respect to A and Z if it is a bounded, measurable function. If the values of ϕ are non-negative, then we call ϕ a *non-negative integrand*. If the values are equal 0 or 1, then we call ϕ a *binary integrand*.

Given A, $\{Z_n\}$ and an integrand ϕ, we define $Z_\phi^A(t)$ to be

$$Z_\phi^A(t) \equiv \lim_{\tau\downarrow-\infty} \int_\tau^t \phi\bigl(Z_{A(\tau,s]}, s\bigr)\,\mathrm{d}A(s), \tag{6.1}$$

where the last integral can be written as the random sum

$$\int_\tau^t \phi\bigl(Z_{A(\tau,s]}, s\bigr)\,\mathrm{d}A(s) = \sum_{n=1}^{A(\tau,t]} \phi\bigl(Z_n, \hat{A}_n\bigr), \tag{6.2}$$

where $\hat{A}_n$ is the arrival epoch of the nth arrival, starting from the last arrival and counting backwards in time. The random variables $\phi(Z_n, \hat{A}_n)$ in (6.2) are conditionally mutually independent, given the arrival epochs $\hat{A}_n$, $n \geqslant 1$.

We now state our key result for such stochastic integrals.

Theorem 6.1. If for all non-negative integrands ϕ of A and Z, we have

$$\int_{-\infty}^t \mathsf{E}\bigl[\phi(Z, s)\bigr]\alpha(s)\,\mathrm{d}s < \infty \tag{6.3}$$

for all real t, then

$$\mathsf{E}\bigl[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(t)}\bigr] = \exp\biggl(\int_{-\infty}^t \mathsf{E}\bigl[\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)} - 1\bigr]\alpha(s)\,\mathrm{d}s\biggr) \tag{6.4}$$

for all real t and θ.

Proof. Let $\tau < t$. Since $\mathsf{E}[A(\tau, t]] = \int_\tau^t \alpha(r)\,\mathrm{d}r < \infty$, the Poisson process induced by the Poisson random measure A has a finite number of jumps on the interval $(\tau, t]$. If f is any real-valued bounded measurable function on $\mathbb{R}$, we can express $f(Z_\phi^A(t))$ as

$$\begin{aligned} f\bigl(Z_\phi^A(t)\bigr) &= f\bigl(Z_\phi^A(\tau)\bigr) + \int_\tau^t \bigl[f\bigl(Z_\phi^A(s)\bigr) - f\bigl(Z_\phi^A(s-)\bigr)\bigr]\,\mathrm{d}A(s) \\ &= f\bigl(Z_\phi^A(\tau)\bigr) + \int_\tau^t \bigl[f\bigl(Z_\phi^A(s-) + \phi\bigl(Z_{A(\tau,s]}, s\bigr)\bigr) - f\bigl(Z_\phi^A(s-)\bigr)\bigr]\,\mathrm{d}A(s). \end{aligned}$$

Observing that $Z_\phi^A(s-)$ and $\phi(s, Z_{A(\tau,s]})$ are independent random variables when s is the time of a jump and letting $f(x) = \mathrm{e}^{\mathrm{i}\theta x}$, we obtain

$$\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(t)} = \mathrm{e}^{\mathrm{i}\theta Z_\phi^A(\tau)} + \int_\tau^t \bigl[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(s-)}\bigl(\mathrm{e}^{\mathrm{i}\theta\phi(Z_{A(\tau,s]},s)} - 1\bigr)\bigr]\,\mathrm{d}A(s). \tag{6.5}$$

Taking expectations of both sides, we get

$$\mathsf{E}\bigl[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(t)}\bigr] = \mathsf{E}\bigl[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(\tau)}\bigr] + \int_\tau^t \mathsf{E}\bigl[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(s-)}\bigr]\mathsf{E}\bigl[\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)} - 1\bigr]\alpha(s)\,\mathrm{d}s, \tag{6.6}$$

invoking lemma 2.2 of [30], which shows that the characteristic function of $Z_\phi^A(t)$ is an absolutely continuous function of t. It follows that it is the unique solution to the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(t)}\big] = \mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(t)}\big]\mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta\phi(Z,t)} - 1\big]\alpha(t), \tag{6.7}$$

whose solution is

$$\mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(t)}\big] = \mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(\tau)}\big]\exp\bigg(\int_\tau^t \mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)} - 1\big]\alpha(s)\,\mathrm{d}s\bigg). \tag{6.8}$$

We are done once we show that

$$\lim_{\tau\downarrow-\infty} \mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(\tau)}\big] = 1 \tag{6.9}$$

and

$$\lim_{\tau\downarrow-\infty}\int_\tau^t \mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(s)} - 1\big]\alpha(s)\,\mathrm{d}s = \int_{-\infty}^t \mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(s)} - 1\big]\alpha(s)\,\mathrm{d}s. \tag{6.10}$$

The first limit (6.9) follows from the fact that (6.3) implies

$$\lim_{\tau\downarrow-\infty} \mathsf{E}\big[Z_\phi^A(\tau)\big] = 0 \tag{6.11}$$

and

$$\big|\mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(\tau)}\big] - 1\big| \leqslant \mathsf{E}\big[\big|\mathrm{e}^{\mathrm{i}\theta Z_\phi^A(\tau)} - 1\big|\big] \leqslant |\theta|\mathsf{E}\big[Z_\phi^A(\tau)\big]. \tag{6.12}$$

The second limit (6.10) follows from combining (6.3) with the bound

$$\big|\mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)}\big] - 1\big| \leqslant |\theta|\mathsf{E}\big[\phi(Z,s)\big] \tag{6.13}$$

and applying dominated convergence. The proof is now complete. □

It is easy to apply theorem 6.1 to establish the theorems in section 2.

Proof of theorem 2.1. The proof follows simply from the observation that $B(s,t) = 0$ for all $s > t$ implies that for all $t > t_j$ and so

$$R(t_j) = \int_{-\infty}^t B(s,t_j)\,\mathrm{d}A(s), \tag{6.14}$$

where $j = 1, \ldots, k$. As a result, for all $t > t_k$ we have

$$\sum_{j=1}^k \theta_j R(t_j) = \int_{-\infty}^t \bigg(\sum_{j=1}^k \theta_j B(s,t_j)\bigg)\,\mathrm{d}A(s). \tag{6.15}$$

The rest follows from theorem 6.1. □

While the distribution of $Z_{\phi}^{A}(t)$ is in general not Poisson, theorem 6.1 allows us to characterize the distribution uniquely. The next theorem allows us to compute all its cumulants in terms of α, ϕ, and Z.

Theorem 6.2. For all non-negative integrands ϕ of A and Z, where

$$\int_{-\infty}^{t} \mathsf{E}\big[e^{\theta\phi(Z,s)} - 1\big]\alpha(s)\,ds < \infty \tag{6.16}$$

for some positive θ and all real t, we have

$$\mathsf{C}^{(n)}\big[Z_{\phi}^{A}(t)\big] = \int_{-\infty}^{t} \mathsf{E}\big[\phi(Z,s)^n\big]\alpha(s)\,ds \tag{6.17}$$

for all positive integers n.

Proof. Combining theorem 6.1 with (6.16), we can show that $\mathsf{E}[e^{i\theta\phi(Z,s)}]$ is an analytic function of θ when $|\theta|$ is sufficiently small. We then have

$$\log \mathsf{E}\big[e^{\theta Z_{\phi}^{A}(t)}\big] = \int_{-\infty}^{t} \mathsf{E}\big[e^{\theta\phi(Z,s)} - 1\big]\alpha(s)\,ds \tag{6.18}$$

and the rest follows from the definition of cumulants and a power series expansion of the right hand term. □

We now establish a CLT implying theorem 3.1. The proof exploits the characteristic function representation, just like classical proofs of CLTs for partial sums; see section 6.4 of Chung [11]. We only state the CLT for one-dimensional distributions. A corresponding limit holds for all finite-dimensional distributions by the same argument, applying the Cramer–Wold device, i.e., by considering the one-dimensional limits for arbitrary linear combinations of the finite-dimensional distributions [6, p. 49].

Theorem 6.3. Let $\{A^{\eta} \mid \eta > 0\}$ be a family of Poisson random measures, where A^{η} has intensity function α^{η}. Let ϕ be a non-negative integrand and let $Z_{\phi}^{\eta}(t)$ be the stochastic integral of ϕ with respect to A^{η}, as defined by (6.1). If $\alpha^{(0)}$ and $\alpha^{(1)}$ are two locally integrable functions on $\mathbb{R}$ such that $\alpha^{(0)}$ is non-negative and, for all real t,

$$\int_{-\infty}^{t} \mathsf{E}\big[\phi(Z,s)^k\big]\big(\alpha^{(0)}(s) + \big|\alpha^{(1)}(s)\big|\big)\,ds < \infty \tag{6.19}$$

and

$$\lim_{\eta\to\infty} \int_{-\infty}^{t} \mathsf{E}\big[\phi(Z,s)^k\big]\left|\frac{\alpha^{\eta}(s) - \eta\alpha^{(0)}(s)}{\sqrt{\eta}} - \alpha^{(1)}(s)\right| ds = 0 \tag{6.20}$$

for $k = 1, 2, 3$, then

$$\lim_{\eta\to\infty} Y^{\eta}(t) \stackrel{d}{=} N\left(\int_{-\infty}^{t} \mathsf{E}\big[\phi(Z,s)\big]\alpha^{(1)}(s)\,ds,\ \int_{-\infty}^{t} \mathsf{E}\big[\phi(Z,s)^2\big]\alpha^{(0)}(s)\,ds\right), \tag{6.21}$$

where the convergence is in distribution, and

$$Y^\eta(t) = \frac{Z_\phi^\eta(t) - \eta \int_{-\infty}^t \mathsf{E}[\phi(Z,s)]\alpha^{(0)}(s)\,\mathrm{d}s}{\sqrt{\eta}}. \tag{6.22}$$

Proof. Using theorem 6.1, we need only show that

$$\begin{aligned}\lim_{\eta\to\infty} \int_{-\infty}^t \mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)/\sqrt{\eta}} - 1\big]\alpha^\eta(s)\,\mathrm{d}s - \mathrm{i}\theta\sqrt{\eta}\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)\big]\alpha^{(0)}(s)\,\mathrm{d}s \\ = \mathrm{i}\theta \int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)\big]\alpha^{(1)}(s)\,\mathrm{d}s - \frac{\theta^2}{2}\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)^2\big]\alpha^{(0)}(s)\,\mathrm{d}s.\end{aligned} \tag{6.23}$$

The limit (6.21) follows from

$$\begin{aligned}&\int_{-\infty}^t \mathsf{E}\big[\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)/\sqrt{\eta}} - 1\big]\alpha^\eta(s)\,\mathrm{d}s - \mathrm{i}\theta\sqrt{\eta}\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)\big]\alpha^{(0)}(s)\,\mathrm{d}s \\ &\quad= \int_{-\infty}^t \mathsf{E}\left[\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)/\sqrt{\eta}} - 1 - \frac{\mathrm{i}\theta}{\sqrt{\eta}}\phi(Z,s) + \frac{\theta^2}{2\eta}\phi(Z,s)^2\right]\alpha^\eta(s)\,\mathrm{d}s \\ &\qquad+ \mathrm{i}\theta\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)\big]\left(\frac{\alpha^\eta(s) - \eta\alpha^{(0)}(s)}{\sqrt{\eta}}\right)\mathrm{d}s - \frac{\theta^2}{2\eta}\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)^2\big]\alpha^\eta(s)\,\mathrm{d}s\end{aligned}$$

$$= \int_{-\infty}^t \mathsf{E}\left[\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)/\sqrt{\eta}} - 1 - \frac{\mathrm{i}\theta}{\sqrt{\eta}}\phi(Z,s) + \frac{\theta^2}{2\eta}\phi(Z,s)^2\right]\alpha^\eta(s)\,\mathrm{d}s \tag{6.24}$$

$$+ \mathrm{i}\theta\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)\big]\left(\frac{\alpha^\eta(s) - \eta\alpha^{(0)}(s)}{\sqrt{\eta}} - \alpha^{(1)}(s)\right)\mathrm{d}s \tag{6.25}$$

$$- \frac{\theta^2}{2}\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)^2\big]\left(\frac{\alpha^\eta(s)}{\eta} - \alpha^{(0)}(s)\right)\mathrm{d}s \tag{6.26}$$

$$+ \mathrm{i}\theta\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)\big]\alpha^{(1)}(s)\,\mathrm{d}s - \frac{\theta^2}{2}\int_{-\infty}^t \mathsf{E}\big[\phi(Z,s)^2\big]\alpha^{(0)}(s)\,\mathrm{d}s. \tag{6.27}$$

The terms (6.25) and (6.26) go to zero as $\eta \to \infty$ given the condition (6.20). The first term (6.24) also goes to zero since by the Taylor series remainder formula we have

$$\left|\mathrm{e}^{\mathrm{i}\theta\phi(Z,s)/\sqrt{\eta}} - 1 - \frac{\mathrm{i}\theta}{\sqrt{\eta}}\phi(Z,s) + \frac{\theta^2}{2\eta}\phi(Z,s)^2\right| \leqslant \frac{|\theta|^3}{6\eta\sqrt{\eta}}\phi(Z,s)^3. \tag{6.28}$$

This completes the proof. □

References

[1] J. Abate and W. Whitt, Numerical inversion of probability generating functions, Operations Research Letters 12 (1992) 245–251.
[2] J. Abate and W. Whitt, Numerical inversion of Laplace transforms of probability distributions, ORSA Journal on Computing 7 (1995) 36–43.

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

[4] R.G. Addie, M. Zukerman and T. Neame, Fractal traffic-measurements, modeling and performance evaluation, in: *Proc. of IEEE Infocom '95* (1995) pp. 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 Journal 34 (1995) 604–628.

[6] J.P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).

[7] 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.

[8] 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.

[9] C.S. Chang, Stability, queue length, and delay of deterministic and stochastic queueing networks, IEEE Transactions on Automatic Control 39 (1994) 913–931.

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

[11] K.L. Chung, *A Course in Probability Theory*, 2nd ed. (Academic Press, New York, 1974).

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

[13] P.J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed. (Academic Press, New York, 1984).

[14] N.G. Duffield and N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications, Mathematical Proceedings of the Cambridge Philosophical Society 118 (1995) 363–374.

[15] 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.

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

[17] N.G. Duffield and W. Whitt, Network design and control using on–off and multi-level source traffic models with heavy-tailed distributions, in: *Self-Similar Network Traffic and Performance Evaluation*, eds. K. Park and W. Willinger (Wiley, New York, 2000) chapter 17, pp. 421–445.

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

[19] A. Feldmann, A.C. Gilbert, W. Willinger and T.G. Kurtz, The changing nature of network traffic: scaling phenomena, Computer Communications Review 28 (1998) 5–29.

[20] C.M. Fortuin, P.W. Kastelyn and J. Ginibre, Correlation inequalities on some partially ordered sets, Communications in Mathematical Physics 22 (1971) 89–103.

[21] A.G. Greenberg, R. Srikant and W. Whitt, Resource sharing for book-ahead and instantaneous-request calls, in: *Teletraffic Contributions for the Information Age, Proc. of ITC 15*, eds. V. Ramaswami and P.E. Wirth (Elsevier, Amsterdam, 1997) pp. 539–548. Longer version to appear in IEEE/ACM Transaction Networking.

[22] R. Guerin, H. Ahmadi and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, IEEE Journal on Selected Areas in Communications 9 (1991) 968–991.

[23] J. Keilson and L. Servi, Networks of non-homogeneous $M/G/\infty$ systems, Journal of Applied Probability 31A (1994) 157–168.

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

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

[26] C. Klüppelberg and T. Mikosch, Explosive Poisson shot noise processes with application to risk reserves, Bernoulli 1 (1995) 125–147.
[27] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Transactions on Networking 2 (1994) 1–15.
[28] K.K. Leung, W.A. Massey and W. Whitt, Traffic models for wireless communication networks, IEEE Journal on Selected Areas in Communications 12 (1994) 1353–1364.
[29] A. Mandelbaum, W.A. Massey and M.I. Reiman, Strong approximations for markovian service networks, Queueing Systems 30 (1998) 149–201.
[30] W.A. Massey and W. Whitt, Networks of infinite-server queues with nonstationary Poisson input, Queueing Systems 13 (1993) 183–250.
[31] W.A. Massey and W. Whitt, A stochastic model to capture space and time dynamics in wireless communication systems, Probability in the Engineering and Informational Sciences 8 (1994) 541–569.
[32] W.A. Massey and W. Whitt, Peak congestion in multi-server service systems with slowly varying arrival rates, Queueing Systems 25 (1997) 157–172.
[33] W.A. Massey and W. Whitt, A probabilistic generalization of Taylor's theorem, Statistics and Probability Letters 16 (1993) 51–54.
[34] I. Norros, A storage model with self-similar input, Queueing Systems 16 (1994) 387–396.
[35] I. Norros, On the use of fractal Brownian motion in the theory of connectionless networks, IEEE Journal on Selected Areas in Communications 13 (1995) 953–962.
[36] K.R. Parthasarathy, *Probability Measures on Metric Spaces* (Academic Press, New York, 1967).
[37] P.F. Pawlita, Traffic measurements in data networks, recent measurement results, and some implications, IEEE Transactions on Communications 29 (1981) 525–535.
[38] V. Paxson and S. Floyd, Wide-area traffic: the failure of Poisson modeling, IEEE/ACM Transactions on Networking 3 (1995) 226–244.
[39] J. Rice, On generalized shot noise, Advances in Applied Probability 9 (1977) 553–565.
[40] W. Whitt, Dynamic staffing in a telephone call center aiming to immediately answer all calls, Operations Research Letters 24 (1999) 205–212.
[41] 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*, eds. F.P. Kelly, S. Zachary and I. Ziedins (Clarendon Press, Oxford, 1996) pp. 339–366.