# Diffusion Approximation Modeling for Markov Modulated Bursty Traffic and Its Applications to Bandwidth Allocation in ATM Networks

Qiang Ren and Hisashi Kobayashi

***Abstract*—We consider a statistical multiplexer model, in which each of $K$ sources is a Markov modulated rate process (MMRP). This formulation allows a more general source model than the well studied "on–off" source model in characterizing variable bit rate (VBR) sources such as compressed video. In our model we allow an arbitrary distribution for the duration of each of $M$ states (or levels) that the source can take on. We formulate Markov modulated sources as a closed queueing network with $M$ infinite-server nodes. By extending our earlier results [17] we introduce an $M$-dimensional diffusion process to approximate the aggregate traffic of such Markov modulated sources. Under a set of reasonable assumptions we then show that this diffusion process can be expressed as an $M$-dimensional Ornstein–Uhlenbeck (O–U) process.**

**The queueing behavior of buffer content is analyzed by applying a diffusion process approximation to the aggregate arrival process. We show some numerical examples which illustrate typical sample paths, and autocorrelation functions of the aggregate traffic and its diffusion process representation. Simulation results validate our proposed approximation model, showing good fits for distributions and autocorrelation functions of the aggregate rate process and the asymptotic queueing behaviors. We also discuss how the analytical formulas derived from the diffusion approximation can be applied to compute equivalent bandwidth for real-time call admission controls, and how the model can be modified to characterize traffic sources with long-range dependence.**

## I. Introduction

In Broadband Integrated Services Digital Networks (B-ISDN), multiple types of information services are provided by means of fast packet switching with statistical multiplexing. The traffic into a statistical multiplexer is a superposition of cell streams from a large number of sources of different types.

There have been a number of noteworthy efforts to characterize multiple "on–off" sources that are statistically multiplexed in a packet-switching node. Anick, Mitra, and Sondhi [2] present a comprehensive analysis of a *fluid-flow* model with single type "on–off" sources. Kosten [19] extends [2] to multiple types of traffic. The "on–off" source model may be an appropriate model, when the source is a single voice or data source. As an alternative approach to modeling voice and data traffic in packet-switching environment, Heffes and Lucantoni [12] and others discuss applications of the Markov modulated Poisson process (MMPP) representation for superposed traffic. Norros *et al.* [23] consider a fluid input process, which is the sum of homogeneous "on–off" sources with general holding-time distributions. The Markov modulated source model by Elwalid *et al.* [9] generalizes the MMPP model by incorporating multiple Markovian states for each source.

In the B-ISDN environment, however, video traffic will make a significant part of the network traffic. Unlike voice and data sources, the well studied two-valued "on–off" representations will not be appropriate to characterize variable bit rate (VBR) traffic such as compressed video. It has been reported [21], [22] that video traffic can be modeled more appropriately as a multivalued rate process.

In this paper we introduce a diffusion process approximation for the superposition of multiple-state Markov modulated rate process (MMRP). This approach has an advantage that we can assume an arbitrary distribution for the duration of each state. The diffusion approximation model also allows us to obtain the transient solution as well as the steady state solution with much less computational complexity than is possible with the previous solution methods, where computational complexity grows exponentially as the number of states and the number of multiplexed sources increase.

The idea of approximating a discrete-state process by a diffusion process with continuous path was discussed by Cox and Miller [7] and others. It has been shown in Kobayashi [15] and Gelenbe [10] that diffusion approximations can be successfully applied to characterize and analyze general queueing networks. Using a diffusion process to study a queueing system enables us to deal with more general types of traffic than the conventional queueing theory can handle. It also allows us to incorporate serial dependency inherent in the superposed traffic, as will be shown in this study. The dependent features are represented by the second-order properties of arrival process, i.e., autocorrelations.

The diffusion process model we discuss in this paper is a multivariate version of the Ornstein–Uhlenbeck (O–U) process, which was originally introduced as a refinement of the Brownian motion. The use of the O–U process to a machine servicing model is discussed by Iglehart [13]. Kobayashi *et al.* [16] discuss the O–U representation of the multiple access scheme of the ALOHA channel. Knessl and Morrison [14]

Manuscript received December 1997; revised February 1998. The work was supported in part by the National Science Foundation, and the Ogasawara Foundation for the Promotion of Science and Engineering.

Q. Ren is with NEC USA, Inc., Princeton, NJ 08540 USA.

H. Kobayashi is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

Publisher Item Identifier S 0733-8716(98)04109-2.

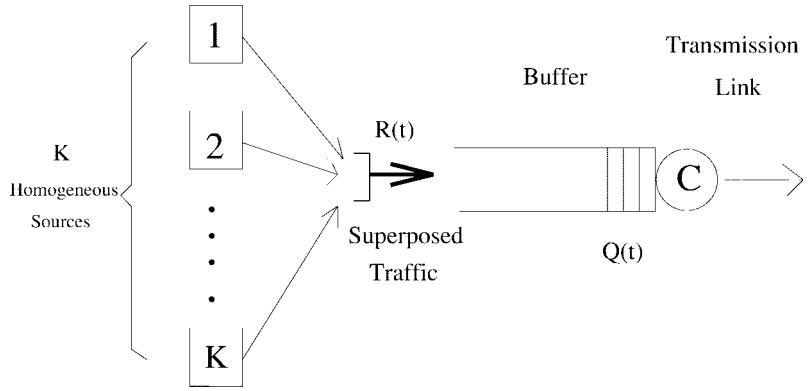


Fig. 1. Buffered statistical multiplexer and $K$ homogeneous sources.

use an O–U process to characterize the superposed traffic stream of multiple homogeneous "on–off" sources as in [2] and provide a heavy-traffic analysis for multiplexer buffer behavior. Similarly, Simonian [30] considers a fluid queue with an O–U input process and carries out the analysis in a different manner. In [17], we have shown that the O–U process provides a good approximation to characterize the superposition of traffic from a heterogeneous set of "on–off" sources.

In this paper we generalize the O–U process approximation to the multistate MMRP source models. In Section II we formulate the model by observing that the behavior of $K$ superposed MMRP's with $M$ states can be represented as a closed queueing network with $K$ customers and $M$ nodes, with each node being represented as ample ($K$ or more) servers. In Section III we develop a multidimensional O–U process to characterize the superposed MMRP's by capturing their first- and second-order statistics. In Section IV we calculate some statistical measures from our diffusion approximation analysis for the aggregate rate process and integrated arrival process. With such statistical measures, we then in Section V approximate the buffer content process by a diffusion process as discussed by Kobayashi [18] and analyze its asymptotic queueing behavior. The simple and closed form formulas derived for the performance measures are used in Section VI to compute equivalent bandwidth for a real-time call admission control problem. In Section VII we provide some numerical examples and simulation results, and discuss the accuracy and limitation of the diffusion approximation method. In Section VIII we show how our modeling technique can be modified to approximately characterize MMRP sources with long-range dependence. The main contributions of our work are summarized in Section IX.

## II. Formulation of the Diffusion Process Model

The system to be studied is composed of a statistical multiplexer and $K$ independent homogeneous sources (Fig. 1). Each source is governed by an $M$-state Markov chain with probability transition matrix $\boldsymbol{P} = \{p_{lm}\}, l, m = 0, 1, \cdots, M-1$. When a source is in state $m$, it generates cells at rate $R_m$ [cells/s]. The duration (or holding time) of state $m$ has a general distribution with mean $\alpha_m^{-1}$ and variance $\sigma_m^2$. When the source exits state $l$, it moves to state $m$ with probability $p_{lm}$. Fig. 2 depicts the state transition diagram of a single source, and a data stream from such a source.

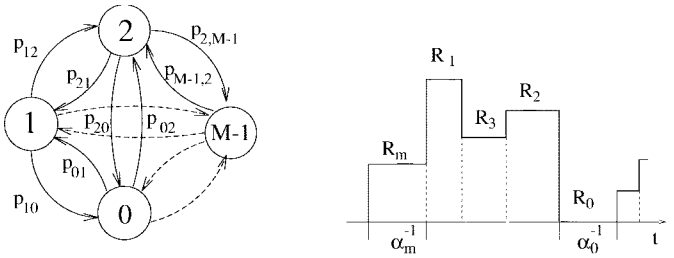


Fig. 2. State transition diagram and cell generation rate process from a single source. (a) State transition diagram for a single source. (b) A typical cell generation process from a single source.

Let us define an $M$-dimensional process $\boldsymbol{N}(t)$

$$\boldsymbol{N}(t) = [N_0(t), N_1(t), \cdots, N_{M-1}(t)]' \tag{1}$$

where $N_m(t)$ denotes the number of sources in state $m$ at time $t, m = 0, 1, \cdots, M-1$, and the superscript prime $(')$ denotes the transpose of a matrix. Clearly the superposed traffic to the multiplexer input at time $t$ can be defined as

$$R(t) = \sum_{m=0}^{M-1} R_m N_m(t). \tag{2}$$

The transmission link has a constant capacity $C$ [cells/s]. Hence, the change of the buffer content $Q(t)$ can be represented by the stochastic differential equation:

$$\frac{dQ(t)}{dt} = \begin{cases} R(t) - C, & \text{when } R(t) > C \text{ or } Q(t) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Now we represent the state transitions of the $K$ $M$-state MMRP sources by a closed queueing network with $M$ nodes and a total of $K$ customers. A source in state $m$ can be viewed as a customer attended by one of the $K$ parallel servers at node $m$ with mean service time $\alpha_m^{-1}$ (i.e., mean holding time in state $m$ for this MMRP source). In this closed queueing network representation, there will be no queue at any of $M$ nodes since $K$, the number of parallel servers, is the same as the numbers of the customers in the network. This is depicted in Fig. 3, in which $A_m(t)$ and $D_m(t)$ are the arrival and departure counting processes to node $m$ (i.e., the total numbers of arrivals at, and

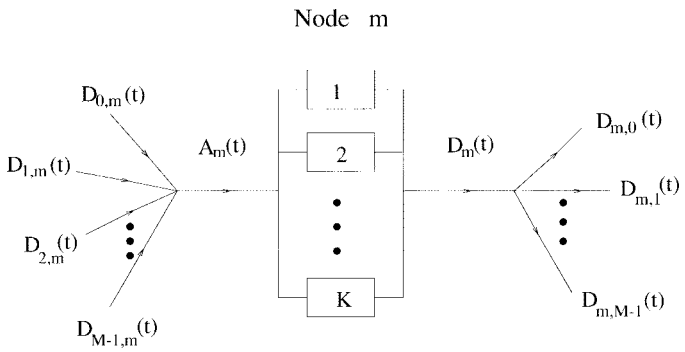


Fig. 3. Arrivals $A_m(t)$ and departures $D_m(t)$ to node $m$ in a queueing network with $K$ customers.

departures from, node $m$ up to time $t$) in the queueing network. The queueing network of Fig. 3 is a variant of the machine servicing model well discussed in the literature. When there are $N_m$ customers at node $m$, the mean departure rate is given by $\alpha_m N_m$, and $N_m(t)$ in (1) can be viewed as the number of customers in node $m$ at time $t$.

Upon the completion of service at node $l$, customers route to node $m$ with probability $p_{lm}, m = 0, 1, \cdots, M-1$, i.e., the MMRP source moves to state $m$ with probability $p_{lm}$ after leaving state $l$. Thus, the arrival process at node $m, A_m(t)$, is the aggregation of those departures from other nodes which route to node $m$.

Let $D_m(t)$ be the departure process from node $m$, and let $D_{l,m}(t)$ represent the counting process of customers that move from node $l$ to node $m$. Clearly,

$$A_m(t) = \sum_{l=0}^{M-1} D_{l,m}(t) \quad \text{and} \quad D_m(t) = \sum_{i=0}^{M-1} D_{m,i}(t). \quad (4)$$

Let an $M$-dimensional process $\boldsymbol{X}(t) = [X_0(t), X_1(t), \cdots, X_{M-1}(t)]'$ be a continuous-state Markov process approximation of the integer vector function $\boldsymbol{N}(t)$ in (1). The process $\boldsymbol{X}(t)$ must satisfy the constraint $\Sigma_{m=0}^{M-1} X_m(t) = K$. When $X_m(t) = x_m$, its mean departure rate is given by $1/\tau_m(x_m) = \alpha_m x_m$, where $\tau_m(x_m)$ denotes the mean interdeparture time when there are $x_m$ customers in state $m$.

We denote the variance of interdeparture time as $\sigma_m^2(x_m)$, and its squared coefficient of variation as $c_m(x_m) = \sigma_m^2(x_m)/\tau_m^2(x_m)$. The variance $\sigma_m^2(x_m)$ is the variance of the superposed stream of $x_m$ i.i.d. renewal processes and its expression depends on the holding-time distribution of the source model. Let $f(x)$ and $F(x)$ denote the probability density function and the distribution function for the holding-time of each individual source with mean equal to $1/u$. Then the probability density function, $g(x)$, of the interdeparture time for the superposed traffic stream of $x_m$ such sources has the following expression [8]:

$$g(x) = u^{x_m - 1}\left( f(x)\left[\int_x^\infty (1 - F(y))\, dy\right]^{x_m - 1} + (x_m - 1)(1 - F(x))^2 \cdot \left[\int_x^\infty (1 - F(y))\, dy\right]^{x_m - 2}\right). \quad (5)$$

Although it is derived for $x_m$ of integer values, the formula (5) can be applied directly to calculate $c_m(x_m)$ of real-valued $x_m$ because it is an analytical probability density function for any real-valued $x_m$. For example, if the holding time of burst period at state $m$ is exponentially distributed, we have $c_m(x_m) = 1$; if the holding time of burst period at state $m$ is deterministic with constant length, we have $c_m(x_m) = (x_m - 1/x_m + 1)$. For general cases, we may calculate $c_m(x_m)$ exactly or numerically using (5). Approximation methods are also helpful to evaluate $c_m(x_m)$. Two basic methods of such approximations are discussed in Whitt [32].

We represent the infinitesimal mean vector $\boldsymbol{b}(\boldsymbol{x})$ of the vector process $\boldsymbol{X}(t)$ by

$$\boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{x}$$

where $\boldsymbol{x} = [x_0, x_1, \cdots, x_{M-1}]'$ and

$$\boldsymbol{B} \stackrel{\text{def}}{=} \{\beta_{mn}\}_{M\times M} = \begin{bmatrix} -\alpha_0 & \alpha_1 p_{10} & \cdots & \alpha_{M-1} p_{M-1,0} \\ \alpha_0 p_{01} & -\alpha_1 & \cdots & \alpha_{M-1} p_{M-1,1} \\ \vdots & \vdots & & \vdots \\ \alpha_0 p_{0,M-1} & \alpha_1 p_{1,M-1} & \cdots & -\alpha_{M-1} \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-1} \end{bmatrix}. \quad (6)$$

Similarly, the infinitesimal covariance matrix $\boldsymbol{A}(\boldsymbol{x}) = \{a_{mn}(\boldsymbol{x})\}_{M\times M}$ can be approximated as [15]

$$a_{mn}(\boldsymbol{x}) = \sum_{l=0}^{M-1} \{(c_l(x_l) - 1)/\tau_l(x_l)\} p_{lm} p_{ln} + \left\{ c_m(x_m)/\tau_m(x_m) + \sum_{l=0}^{M-1} (p_{lm}/\tau_l(x_l)) \right\} \delta_{mn} - \left(\frac{c_m(x_m)}{\tau_m(x_m)}\right) p_{mn} - \left(\frac{c_n(x_n)}{\tau_n(x_n)}\right) p_{nm}. \quad (7)$$

Note that in the Markov modulated source, we have $p_{mm} = 0$ for $m = 0, 1 \cdots, M-1$. In other words, when the holding time in one state expires, the source always shifts to a different state. Then the expression for $\boldsymbol{A}(\boldsymbol{x})$ can be simplified as

$$\boldsymbol{A}(\boldsymbol{x}) = \sum_{l=0}^{M-1} \left(\frac{c_l(x_l)}{\tau_l(x_l)}\right) \boldsymbol{v}_l \cdot \boldsymbol{v}_l' + \mathcal{W}(\boldsymbol{x}) \quad (8)$$

where $\boldsymbol{v}_l$ is an $M$-dimensional column vector whose $l$-th element is unity and the $m$-th element $(m \neq l)$ is $-p_{lm}$, i.e., $\boldsymbol{v}_l = [-p_{l0}, \cdots, 1, \cdots, -p_{l,M-1}]'$. $\mathcal{W}(\boldsymbol{x})$ is an $M \times M$ matrix whose element is

$$w_{mn}(\boldsymbol{x}) = \sum_{l=0}^{M-1} [p_{lm}(\delta_{mn} - p_{ln})/\tau_l(x_l)], \quad 0 \leq m, n \leq M-1.$$

It is easy to show that $\mathcal{W}$ is nonnegative-definite. Note that $\mathcal{W}(\boldsymbol{x})$ depends only on the Markov chain $\boldsymbol{P} = \{p_{lm}\}$, the mean holding-times $\alpha_m^{-1}, 0 \le m \le M-1$, and the system state $\boldsymbol{x}$. For more detailed discussions of the above derivations, the reader is referred to [15] and references therein.

## III. Diffusion Approximation Analysis

With the formulation given in Section II, we now approximate the $M$-dimensional process $\boldsymbol{N}(t)$ using an $M$-dimensional diffusion process $\boldsymbol{X}(t)$ governed by the stochastic differential equation

$$d\boldsymbol{X}(t) = \boldsymbol{B}\cdot\boldsymbol{X}(t)\,dt + \sqrt{\boldsymbol{A}(\boldsymbol{x})}\cdot d\boldsymbol{W}(t), \quad \text{with} \quad \sum_{m=0}^{M-1} X_m(t) = K \tag{9}$$

where $\boldsymbol{W}(t)$ is an $M$-dimensional Brownian motion with zero mean and the covariance (matrix) function $\boldsymbol{I}\delta(t)$, with $\boldsymbol{I}$ being the $M \times M$ identity matrix, and $\delta(t)$ being Dirac's delta function.

Both $\boldsymbol{B}$ and $\boldsymbol{A}(\boldsymbol{x})$ are singular matrices due to the fact $\Sigma_{m=0}^{M-1}\ x_m = K$. By definition, $\boldsymbol{A}(\boldsymbol{x})$ is a symmetric and positive semidefinite matrix. Hence, $\sqrt{\boldsymbol{A}(\boldsymbol{x})}$ always exists and is uniquely defined.

From (9) it follows (see, e.g., [7]) that the conditional probability density function of $\boldsymbol{X}(t)$

$$f(\boldsymbol{x},t;\boldsymbol{x}_0,0)\,d\boldsymbol{x} = P[x_m \le X_m(t) < x_m + dx_m, 0 \le m \le M-1 | \boldsymbol{X}(0) = \boldsymbol{x}_0] \tag{10}$$

satisfies the following generalized multivariate Fokker–Planck equation:

$$\frac{\partial f}{\partial t} = -\sum_{m=0}^{M-1} \frac{\partial}{\partial x_m}\,[b_m(\boldsymbol{x})f] + \frac{1}{2}\sum_{m=0}^{M-1}\sum_{n=0}^{M-1} \frac{\partial^2}{\partial x_m \partial x_n}\,[a_{mn}(\boldsymbol{x})f]. \tag{11}$$

Let $\boldsymbol{x}^* = (x_0^*, x_2^*, \cdots, x_{M-1}^*)$ be the equilibrium state of the process $\boldsymbol{X}(t)$ such that

$$\boldsymbol{b}(\boldsymbol{x}^*) = \boldsymbol{B}\boldsymbol{x}^* = 0. \tag{12}$$

Then we can write

$$\boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{B}(\boldsymbol{x} - \boldsymbol{x}^*). \tag{13}$$

If we consider a narrow region around $\boldsymbol{x} = \boldsymbol{x}^*$, we can approximate $\boldsymbol{A}(\boldsymbol{x})$ by its value at $\boldsymbol{x} = \boldsymbol{x}^*$:

$$\boldsymbol{A}(\boldsymbol{x}) \approx \boldsymbol{A}(\boldsymbol{x}^*) \stackrel{\text{def}}{=} \boldsymbol{A}. \tag{14}$$

With the linear infinitesimal mean $\boldsymbol{b}(\boldsymbol{x})$ of (13) and the constant covariance $\boldsymbol{A}$ of (14), (9) becomes

$$d\boldsymbol{X}(t) = \boldsymbol{B}\cdot(\boldsymbol{X}(t) - \boldsymbol{x}^*)\,dt + \sqrt{\boldsymbol{A}}\cdot d\boldsymbol{W}(t) \tag{15}$$

which is a multivariate *Ornstein–Uhlenbeck* process, leading to the following differential equation:

$$\frac{\partial f}{\partial t} = -\sum_{m=0}^{M-1}\sum_{n=0}^{M-1} \beta_{mn}\,\frac{\partial}{\partial x_m}\,[(x_n - x_n^*)f] + \frac{1}{2}\sum_{m=0}^{M-1}\sum_{n=0}^{M-1} a_{mn}\,\frac{\partial^2 f}{\partial x_m \partial x_n} \tag{16}$$

where $\beta_{mn}$ and $a_{mn}$ are the $(m,n)$th entries of the $M \times M$ matrices $\boldsymbol{B}$ and $\boldsymbol{A}$, respectively.

The solution for (16) is the following multivariate Gaussian distribution:

$$f(\boldsymbol{x},t) = (2\pi)^{-(M/2)}(\det|\Xi(t)|)^{-(1/2)} \cdot \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}(t) - \overline{\boldsymbol{x}}(t))'\Xi^{-1}(t)(\boldsymbol{x}(t) - \overline{\boldsymbol{x}}(t))\right\} \tag{17}$$

where

$$\overline{\boldsymbol{x}}(t) = \boldsymbol{x}^* + e^{t\boldsymbol{B}}\boldsymbol{x}_0$$
$$\Xi(t) = \int_0^t e^{(t-\tau)\boldsymbol{B}}\boldsymbol{A}e^{(t-\tau)\boldsymbol{B}'}\,d\tau. \tag{18}$$

Note that $\Xi(t)$ is a symmetric positive semidefinite matrix, and therefore $\Xi^{-1}(t)$ denotes its generalized (or pseudo) inverse.

When we impose the reflecting barriers at hyperplanes $\Sigma_{m=0}^{M-1}\ x_m = K$ and $x_m = 0, m = 0, 1, \cdots, M-1$, then the corresponding solution (17) is a truncated Gaussian distribution.

If we limit ourselves to the steady-state traffic behavior, we let $t \to +\infty$, obtaining

$$\overline{\boldsymbol{x}} = \boldsymbol{x}^* \quad \text{and} \quad \Xi = \int_0^{+\infty} e^{\boldsymbol{B}t}\boldsymbol{A}e^{\boldsymbol{B}'t}\,dt. \tag{19}$$

As we noted earlier, $\boldsymbol{B}$ is a singular matrix due to the fact $\Sigma_{m=0}^{M-1}\ X_m(t) = K$. To overcome this problem, we *redefine* $\boldsymbol{X}(t) = [X_1(t), X_2(t), \cdots, X_{M-1}(t)]'$ as $(M-1)$-dimensional vector process, using the same notion. Then, we consider $(M-1)$ free variates $x_1, x_2, \cdots, x_{M-1}$ and find the following relationship between the density functions using the indicator function $\mathcal{X}_{\{\cdot\}}$:

$$f(x_0, x_1, \cdots, x_{M-1}) = f(x_1, \cdots, x_{M-1}) \cdot \mathcal{X}_{\{x_0 = (K - \Sigma_{m=1}^{M-1} x_m)\}}. \tag{20}$$

Thus, $\boldsymbol{B}$ and $\boldsymbol{A}$ in (6) and (10) are modified and *redefined* to $(M-1)\times(M-1)$ matrices, using the same notions $\boldsymbol{B} = \{\beta'_{mn}\}_{(M-1)\times(M-1)}$ and $\boldsymbol{A} = \{a'_{mn}\}_{(M-1)\times(M-1)}$, where

$$\beta'_{mn} = \beta_{mn} - \beta_{m0}, \quad \text{and} \quad a'_{mn} = a_{mn}, \quad m,n = 1, \cdots, M-1. \tag{21}$$

Then we have a unique solution of a symmetric positive-definite covariance matrix $\Xi$ for $f(x_1, \cdots, x_{M-1})$ by replacing new matrices $\boldsymbol{B}$ and $\boldsymbol{A}$ of (21) into (19). The autocovariance matrix $\boldsymbol{\rho}_{\boldsymbol{X}}(s,t), s \le t$, can be readily derived from the governing stochastic differential equation (15) as

$$\boldsymbol{\rho}_{\boldsymbol{X}}(s,t) = \Xi e^{(t-s)\boldsymbol{B}}. \tag{22}$$

Note that the original $M$-dimensional process $\boldsymbol{N}(t)$ is generally non-Markovian unless the holding time at each state has an exponential distribution. Here we use an $M$-dimensional diffusion process $\boldsymbol{X}(t)$ in (15), which has the exponential autocovariance function matrix of (22), to characterize $\boldsymbol{N}(t)$ in such a way that the process $\boldsymbol{X}(t)$ approximately captures the first- and second-order statistics of $\boldsymbol{N}(t)$ through the parameters derivable from matrices $\boldsymbol{B}$ and $\boldsymbol{A}$.

## IV. Performance Analysis for Aggregate Rate Process and Arrival Process

As we have assumed earlier, a source at state $m$ generates cells at a constant rate $R_m$. Without loss of generality, we can further assume $R_0 = 0$ (i.e., a source is always off at state 0) so that we only consider processes $[X_1(t), \cdots, X_{M-1}(t)]$ as we have just redefined in the previous section. So we can represent the multiplexer input process as

$$\tilde{R}(t) = \sum_{m=1}^{M-1} R_m X_m(t) \tag{23}$$

which is a diffusion process approximation of $R(t)$ in (2). Note that $\Xi$ is a symmetric positive-definite matrix, thus it can be diagonalized as $\Xi = \boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with element $\lambda_i, i = 1, \cdots, M-1$, being the eigenvalues of $\Xi$, and the row vectors of matrix $\boldsymbol{Q}$ are the associated orthonormal eigenvectors.

Define $\boldsymbol{Z}(t) = \boldsymbol{Q}\boldsymbol{X}(t)$, then $\boldsymbol{Z}(t)$ is a Gaussian process with mean $\boldsymbol{Q}\boldsymbol{x}^*$ and covariance matrix $\boldsymbol{\Lambda}$, which implies $Z_1(t), Z_2(t), \cdots, Z_{M-1}(t)$ are orthogonal (hence independent) Gaussian processes. Therefore, we have

$$\tilde{R}(t) = \sum_{m=1}^{M-1} R'_m Z_m(t) \tag{24}$$

where $R'_m = \Sigma_{i=1}^{M-1} R_i Q_{mi}$. Thus, the process $\tilde{R}(t)$ is also a Gaussian process, whose mean $\mu_{\tilde{R}}$, variance $\sigma^2_{\tilde{R}}$, and autocovariance function $\rho_{\tilde{R}}(\tau)$ in the steady state are

$$\begin{aligned}
\mu_{\tilde{R}} &= \lim_{t\to\infty} E\Big[\tilde{R}(t)\Big] = \sum_{m=1}^{M-1} R_m x^*_m \\
\sigma^2_{\tilde{R}} &= \lim_{t\to\infty} \mathrm{Var}\Big[\tilde{R}(t)\Big] = \sum_{m=1}^{M-1} \lambda_m R'^2_m \\
\rho_{\tilde{R}}(\tau) &= [R_1, R_2, \cdots, R_{M-1}]\Xi e^{\tau \boldsymbol{B}}[R_1, R_2, \cdots, R_{M-1}]'.
\end{aligned} \tag{25}$$

Note, however, that the $\tilde{R}(t)$ is generally not a Markov process.

We define $I(t)$ as the integrated arrival process for the aggregate traffic $\tilde{R}(t)$:

$$I(t) \stackrel{\mathrm{def}}{=} \int_0^t \tilde{R}(u)\, du = \sum_{m=1}^{M-1} R_m \int_0^t X_m(u)\, du. \tag{26}$$

Since $\boldsymbol{X}(t)$ satisfies (15), we can show that for a sufficiently large $t$, the mean and covariance matrices of $\int_0^t \boldsymbol{X}(u)\, du$ are

$$\begin{aligned}
E\left[\int_0^t \boldsymbol{X}(u)\, du\right] &\sim \boldsymbol{x}^* t + o(t), \\
\mathrm{Cov}\left[\int_0^t \boldsymbol{X}(u)\, du\right] &\sim \boldsymbol{B}^{-1}\boldsymbol{A}\big(\boldsymbol{B}^{-1}\big)' t + o(t)
\end{aligned} \tag{27}$$

which implies that the long-term infinitesimal mean and long-term infinitesimal variance of the arrival process $I(t)$ are given by

$$\begin{aligned}
\lim_{t\to\infty} \frac{E[I(t)]}{t} &= \sum_{m=1}^{M-1} R_m x^*_m = \mu_{\tilde{R}}, \\
\lim_{t\to\infty} \frac{\mathrm{Var}[I(t)]}{t} &= [R_1, \cdots, R_{M-1}]\boldsymbol{B}^{-1}\boldsymbol{A}\big(\boldsymbol{B}^{-1}\big)' \\
&\quad \cdot [R_1, \cdots, R_{M-1}]' \stackrel{\mathrm{def}}{=} a.
\end{aligned} \tag{28}$$

## V. A Diffusion Approximation Model and Performance Analysis for Queue Process $Q(t)$

In this section we form a diffusion process $\tilde{Q}(t)$ to approximate the buffer content process $Q(t)$ of (3) by using the approach discussed in [18].

With (28) we approximate the integrated arrival process $I(t)$ by a diffusion process $\tilde{I}(t)$. This diffusion process captures the original process through its first- and second-order statistics in equilibrium, i.e.,

$$d\tilde{I}(t) = \mu_{\tilde{R}} \cdot dt + \sqrt{a} \cdot dW(t) \tag{29}$$

where $\mu_{\tilde{R}}$ and $a$ are constants given in (28).

From (3) the output counting process $J(t)$ for transmission can be approximated by

$$dJ(t) = \begin{cases} C\, dt, & \text{when } \tilde{Q}(t) > 0 \text{ or } \tilde{R}(t) > C \\ \eta\, dt, & \text{otherwise,} \end{cases} \tag{30}$$

where $\eta$ is an unknown constant and may be approximated by $E[\tilde{R}|\tilde{R} < C]$ as suggested in [18].

The diffusion process $\tilde{Q}(t)$ therefore satisfies the following stochastic differential equation:[1]

$$\begin{aligned}
\tilde{Q}(t) &= dI(t) - dJ(t) \\
&= \begin{cases} (\mu_{\tilde{R}} - C)\, dt + \sqrt{a} \cdot dW(t), & \tilde{Q}(t) > 0 \\ (\mu_{\tilde{R}} - \eta)\, dt + \sqrt{a} \cdot dW(t), & \tilde{Q}(t) \le 0. \end{cases}
\end{aligned} \tag{31}$$

If we define the probability density function $f(\tilde{q}, t; \tilde{q}_0, 0)$ of the process, $\tilde{Q}(t)$ is $f(\tilde{q}, t; \tilde{q}_0, 0)\, d\tilde{q} = P[\tilde{q} \le \tilde{Q}(t) < \tilde{q} + d\tilde{q} | \tilde{Q}(0) = \tilde{q}_0]$, then we have

$$\frac{\partial f}{\partial t} = -(\mu_{\tilde{R}} - C)\frac{\partial f}{\partial \tilde{q}} + \frac{a}{2}\frac{\partial^2 f}{\partial \tilde{q}^2}$$

where the boundary condition for the reflecting barrier at $\tilde{q} = 0$ [15] is $(\mu_{\tilde{R}} - C)f = (a/2)(\partial f/\partial \tilde{q})$ at $\tilde{q} = 0$ for all $t > 0$.

Therefore, the stationary probability $P[\tilde{Q} > x]$ $(= \lim_{t\to\infty} P[\tilde{Q}(t) > x])$ is given by

$$P\Big[\tilde{Q} > x\Big] = \frac{\mu_{\tilde{R}} - \eta}{C - \eta} \exp\left(-\frac{2(C - \mu_{\tilde{R}})}{a}\, x\right). \tag{32}$$

[1] Note that $\tilde{Q}(t)$ is a diffusion approximation of $Q(t)$ and can take negative values.

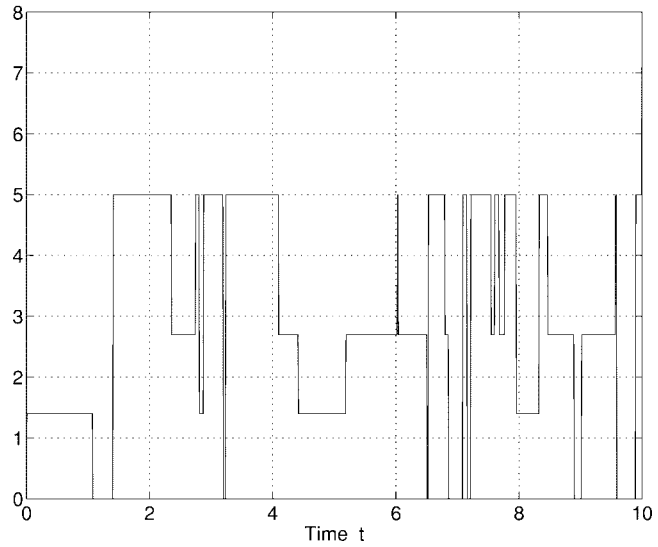


Fig. 4. Sample path of the traffic process from a single 4-state Markov modulated source with the parameters provided in Section VII.

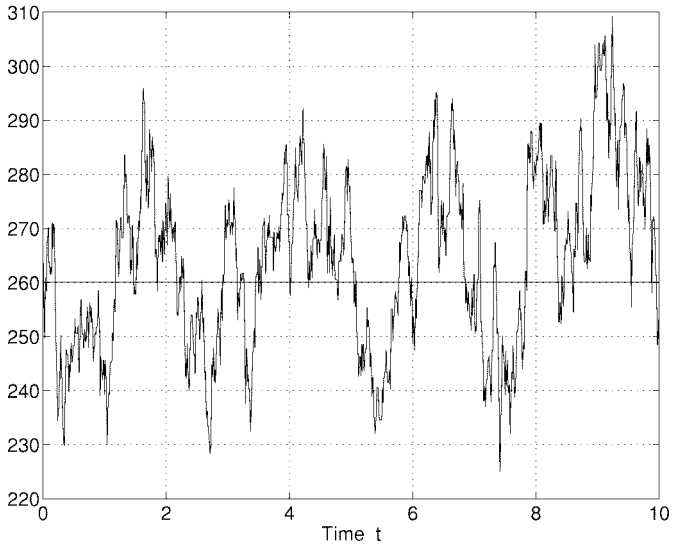


Fig. 5. Sample path of the aggregate traffic process $R(t)$ of $K = 100$ Markov modulated sources. Each source generates a pattern statistically similar to Fig. 4.

The exponent approximates the *tail-end* distribution (i.e., for a large value of $x$) of $P[\tilde{Q}(t) > x]$ and the coefficient $(\mu_{\tilde{R}} - \eta/C - \eta)$ is used to approximate $P[\tilde{Q}(t) > 0]$ as derived in [18]. We have discussed in [17] that $P[\tilde{R} > C]$ can serve as a lower-bound approximation of $P[\tilde{Q} > 0]$ and can be approximated by

$$P[\tilde{Q} > 0] \geq P[\tilde{R} > C] \approx \frac{e^{(-\theta^2/2)}}{\theta\sqrt{2\pi}} \tag{33}$$

where

$$\theta \stackrel{\text{def}}{=} \frac{C - \mu_{\tilde{R}}}{\sigma_{\tilde{R}}}. \tag{34}$$

Therefore, we can approximate the asymptotic complementary queue length distribution by

$$P[Q > x] \approx \left(\frac{e^{(-(\theta^2)/2)}}{\theta\sqrt{2\pi}}\right) \cdot \exp\left(-\frac{2\sigma_{\tilde{R}}\theta}{a}\, x\right). \tag{35}$$

It is interesting to note that the formula in (35) derived from our diffusion approximation has the same exponent $\exp(-2\sigma_{\tilde{R}}\theta/a)x$ as the asymptotic upper-bound for the complementary queue length distribution derived by Simonian [30] but has a different constant coefficient $\exp(-\theta^2)/\theta\sqrt{2\pi}$.

The parameters $\theta, \sigma_{\tilde{R}}$, and $a$ in (35) can either be obtained from the source declarations of $\mu_{\tilde{R}}$, $\{p_{lm}\}$, and $\{R_m\}$ [which are then computed by (25), (28), and (34)] or can be estimated from the direct measurements of the aggregate rate process $R(t)$. It is not difficult to measure the mean $\mu_{\tilde{R}}$ and the variance $\sigma^2_{\tilde{R}}$ of $R(t)$. As for the infinitesimal variance $a$ of the aggregate arrival process, it can be estimated by

$$a \approx 2\int_0^\infty \rho_R(\tau)\, d\tau \tag{36}$$

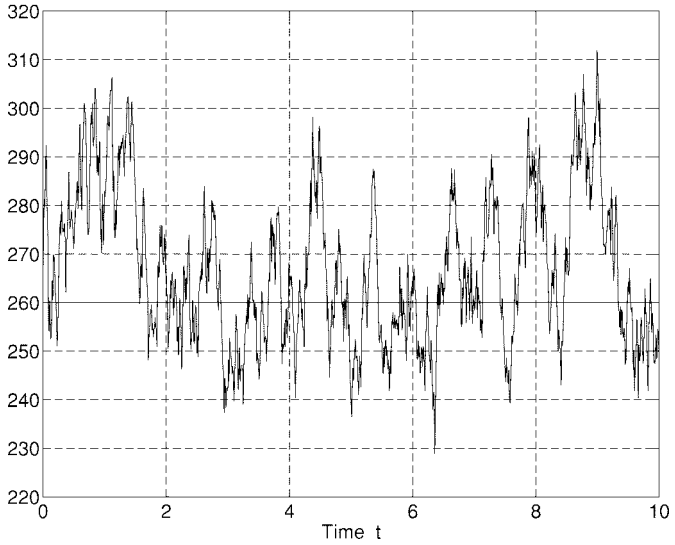


Fig. 6. Sample path of the diffusion process $\bar{R}(t)$, which approximates $R(t)$ in Fig. 5.

where $\rho_R(\tau)$ is the empirical autocovariance function of aggregate rate process $R(t)$. In [1] Addie and Zukerman have derived a formula similar to (36) for a discrete-time stationary Gaussian process.

## VI. Applications to Real-Time Call Admission Control

The analysis and simple formulas derived in Sections IV and V can be applied to compute an equivalent bandwidth for real-time call admission controls and bandwidth allocations. Assume that we have $K$ such $M$-state MMRP sources offered to a statistical multiplexer as shown in Fig. 1. The quality of service (QoS) requirement for the cell loss ratio (CLR) is

$$\text{CLR} < \epsilon. \tag{37}$$

We then wish to determine the required bandwidth $C_{\text{eqv}}$.

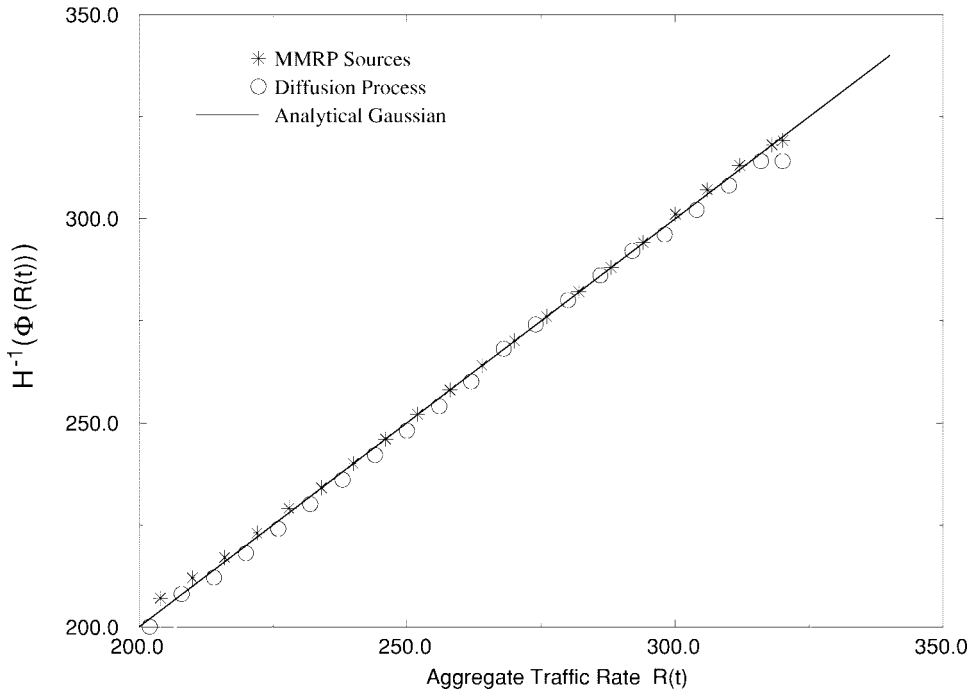


Fig. 7. Fractile diagrams (Q–Q plot) for the distribution functions of the aggregate traffic process $R(t)$, its diffusion process representation $\bar{R}(t)$, and the analytically derived Gaussian process. The state durations are exponentially distributed.

If the multiplexer has little or no buffer, we can model the multiplexer as a loss system, and can calculate CLR using the diffusion approximation of $R(t)$ discussed in Section IV

$$\mathrm{CLR} \stackrel{\mathrm{def}}{=} \frac{E[R(t)-C]^+}{E[R(t)]} = \int_C^\infty \frac{r-C}{\mu_{\tilde{R}}\sqrt{2\pi\sigma_{\tilde{R}}^2}} \exp\left[-\frac{(r-\mu_{\tilde{R}})^2}{2\sigma_{\tilde{R}}^2}\right] dr.$$

For a given $\mathrm{CLR} < \epsilon$, we can solve from the above equation for an equivalent bandwidth as shown in [26]

$$C_{\mathrm{eqv}} = \mu_{\tilde{R}} + \theta \cdot \sigma_{\tilde{R}} \tag{38}$$

where

$$\theta \approx 1.8 - 0.46 \cdot \log_{10}\left(\frac{\mu_{\tilde{R}}\sqrt{2\pi}}{\sigma_{\tilde{R}}}\,\epsilon\right).$$

If the multiplexer has large buffers of capacity $B$, then we can model the multiplexer as a queueing system with infinite buffer, and approximate CLR as the probability that $Q(t)$ exceeds B. Then, for a given $\mathrm{CLR} < \epsilon$, we solve (35) for the equivalent bandwidth as follows:

$$C_{\mathrm{eqv}} = \mu_{\tilde{R}} + \theta \cdot \sigma_{\tilde{R}} \tag{39}$$

where

$$\theta \approx \sqrt{\xi - 2\ln\left(\sqrt{\xi} - \frac{2\sigma_{\tilde{R}}}{a}\,B\right)} - \frac{2\sigma_{\tilde{R}}}{a}\,B$$

$$\xi = -2\ln(\sqrt{2\pi}\epsilon) + \frac{4\sigma_{\tilde{R}}^2}{a^2}\,B^2.$$

The applicability of (38) and (39) is not restricted to the MMRP sources. They can be applied to any traffic stream as long as its mean, variance, and autocovariance function are provided.

## VII. Numerical Examples, Simulation Results, and Discussion

We now give some numerical examples to illustrate and verify our modeling and analysis method. Simulations are conducted for both the superposed traffic of MMRP sources and its corresponding diffusion process approximation.

We consider a superposed traffic stream of 100 independent four-state MMRP sources, i.e., $K = 100$ and $M = 4$. The probability transition matrix for each MMRP source is given by

$$\boldsymbol{P} = \begin{bmatrix} 0, & 0, & \frac{1}{2}, & \frac{1}{2} \\ \frac{1}{6}, & 0, & \frac{1}{3}, & \frac{1}{2} \\ \frac{2}{5}, & \frac{1}{5}, & 0, & \frac{2}{5} \\ \frac{1}{5}, & \frac{3}{10}, & \frac{1}{2}, & 0 \end{bmatrix}.$$

Let

$$R_0 = 0, \quad R_1 = 1.4, \quad R_2 = 5.0, \quad \text{and} \quad R_3 = 2.7;$$
$$\alpha_0 = 7, \quad \alpha_1 = 1, \quad \alpha_2 = 3, \quad \text{and} \quad \alpha_3 = 2.0$$

and we assume at first that the holding times of burst periods are exponentially distributed (with means $\alpha_0^{-1}, \alpha_1^{-1}, \alpha_2^{-1}$, and $\alpha_3^{-1}$, respectively).

Fig. 4 shows a simulated sample path of a typical cell generation process from a single source as defined above. Fig. 5

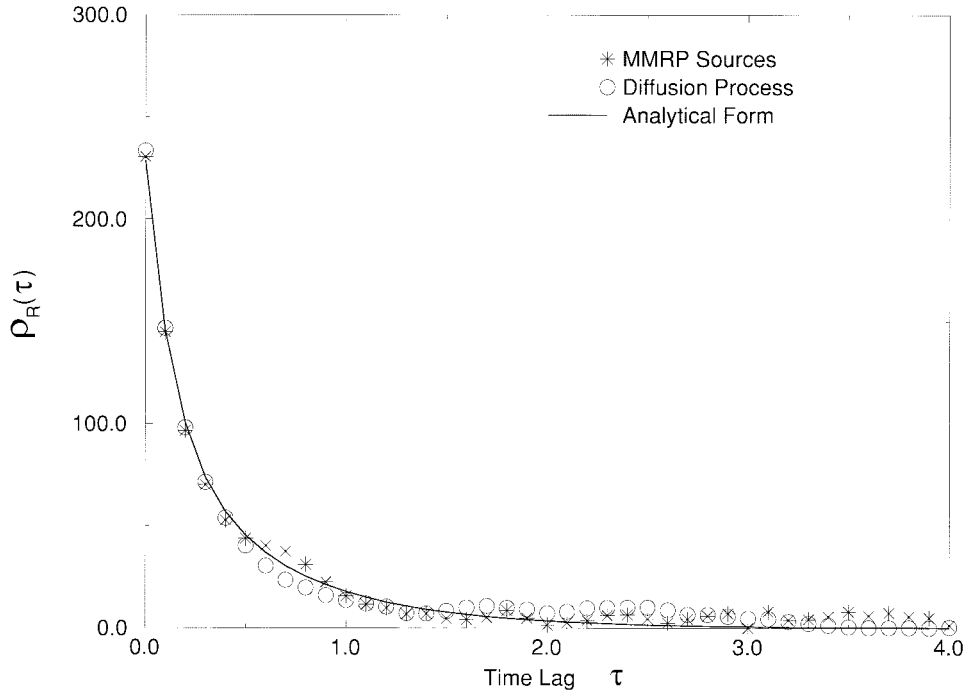


Fig. 8. Autocorrelation functions of $R(t)$ (shown by $*$), $\tilde{R}(t)$ (shown by circles), and the analytical expression $\rho_{\tilde{R}}(t)$. The state durations are exponentially distributed.

shows a simulated sample path of the *superposed* traffic stream from 100 independent sources $(K = 100)$, each of which generates the four-level bursty traffic $(M = 4)$ similar to Fig. 4. Fig. 6 shows a simulated sample path of the diffusion process $\tilde{R}(t)$ (as defined in (23) with corresponding matrices $\boldsymbol{B}$ and $\boldsymbol{A}$), which is an approximation of the superposed traffic in Fig. 5. We can see from these two figures that the sample path of the aggregate rate process $R(t)$ of (2) and its diffusion approximation process $\tilde{R}(t)$ look statistically almost identical.

The cumulative distributions of the superposed traffic $R(t)$ from the Markov modulated sources and its diffusion process representation $\tilde{R}(t)$ are plotted in Fig. 7 in the form of *fractile* diagram ("Q-Q plot") to show how close they are to the Gaussian distribution with $\mu_{\tilde{R}} = 260.06$ and $\sigma^2_{\tilde{R}} = 228.65$ obtained from formulas in (25). In the fractile diagram, the $x$-axis represents the aggregate traffic rate $r$ and the $y$-axis represents $H^{-1}(\Phi(r))$, where $\Phi(\cdot)$ is analytical cumulative Gaussian distribution function and $H^{-1}(\cdot)$ is the inverse of empirical histogram function. Thus, any Gaussian distribution is represented by a 45-degree straight line. Fig. 8 shows the autocorrelation functions of the superposed traffic $R(t)$ from the MMRP sources, its diffusion approximation process $\tilde{R}(t)$, and its analytical form of (25) . These figures therefore validate that the diffusion process $\tilde{R}(t)$ generated by (15) and (23) captures the marginal distribution and autocorrelation of the original aggregate rate process.

Fig. 9 shows $\log_{10}(P[Q > x])$ versus $x$ when $R(t)$ enters a multiplexer with infinite buffer and bandwidth $C = 285$ [cells/s], which corresponds to link utilization $\rho \approx 0.9$. The star line is obtained from the simulation by feeding $R(t)$ of (2) into the multiplexer, whereas the circle line is obtained by feeding its diffusion approximation process $\tilde{R}(t)$ of (23) into the multiplexer. The solid line is the analytical approximation by (35). The figure shows that the diffusion process $\tilde{R}(t)$ approximates the original aggregate MMRP process accurately in terms of the queueing behavior as well, and the analytical formula (35) of asymptotic queue length distribution, derived from the diffusion approximation $\tilde{Q}(t)$, captures the decay rate of complementary queue length distribution and provides a very good estimate for buffer overflow probabilities.

In Fig. 10, we fix $K = 25$ and vary the capacity $C$ to see how the diffusion approximations $\tilde{R}(t)$ and $\tilde{Q}(t)$ perform under different utilizations. For $C = 71.25$, 81.25, and 92.85 [cells/s] (which corresponds to the link utilization $\rho = 0.9, 0.8$ and 0.7, respectively), we plot $\log_{10}(P[Q > x])$ versus $x$ as in Fig. 9. We can see that while the diffusion process $\tilde{R}(t)$ approximates the $R(t)$ in all these cases, the accuracy of analytical asymptotic estimate in (35) deviates as the link utilization becomes lower. Nevertheless, the approximation formula (35) seems to serve as an upper-bound in all these cases.

In the above example, we have assumed that the holding time at each state is exponentially distributed. Now we make the holding times other than exponentially distributed while keeping other parameters unchanged.

In Figs. 11 and 12, we plot the quantities similar to Figs. 7 and 8 when we make the holding times constant. While the distribution functions fit very well to a Gaussian distribution, there are some discrepancies in their autocorrelation functions $\rho_{\tilde{R}}(t)$. The process $R(t)$ has a somewhat shorter correlation

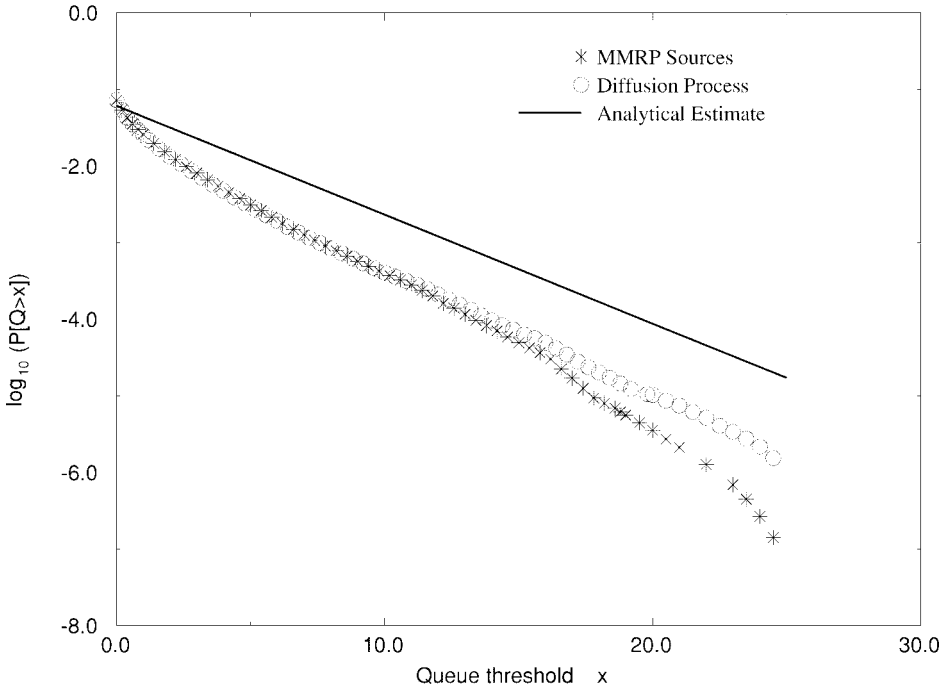


Fig. 9. Complementary queue length distributions $\log_{10}(P[Q > x])$: simulation inputs are $K = 100$ MMRP sources (shown in by) and the diffusion process approximation $\bar{R}(t)$ (shown in circles). Analytical asymptotic estimate of (35) is shown in the solid curve. The output link utilization $\rho = 0.9$ is assumed.

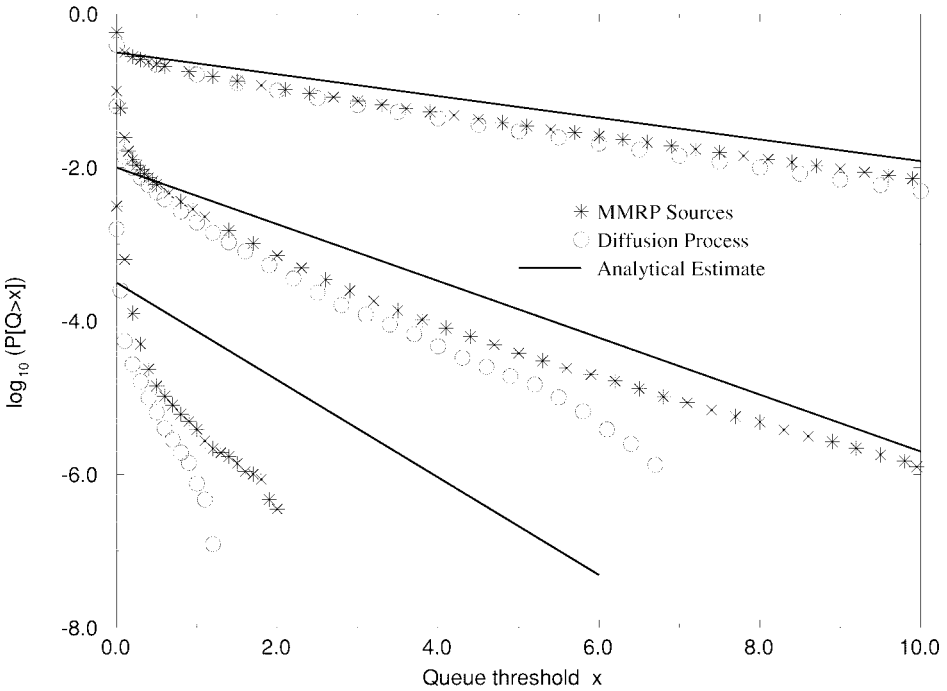


Fig. 10. $\log_{10}(P[Q > x])$: comparison of simulation results with MMRP sources and its diffusion process approximation, and analytical asymptotic estimate of (35). The three groups of the curves are for the output link utilization $\rho = 0.9, 0.8,$ and $0.7$ (from top to bottom). The number of sources $K = 25$.

span compared to its diffusion approximation process $\tilde{R}(t)$. As a result, we can see from Fig. 13 that the $P[Q > x]$ has a sharper decay compared to $P[\tilde{Q} > x]$ of its diffusion approximation. The analytical asymptotic estimate by (35) is shown to serve as an upper-bound.

Next, we consider that each source has a two-stage hyperexponential $(H_2)$ holding time distribution with mean $\alpha_i^{-1}$ when it is in $i$-state, $i = 0, 1, 2, 3$. The hyperexponential distributions of holding times for the four different states have the following probability density functions:

$$f_0(x) = \frac{5}{3}\, e^{-5x} + \frac{35}{6}\, e^{-8(3/4)x}$$
$$f_1(x) = \frac{5}{2}\, e^{-5x} + \frac{5}{18}\, e^{-(5/9)x}$$
$$f_2(x) = \frac{1}{5}\, e^{-x} + \frac{24}{5}\, e^{-6x}$$
$$f_3(x) = \frac{3}{10}\, e^{-x} + 2\,\frac{9}{20}\, e^{-(7/2)x}$$

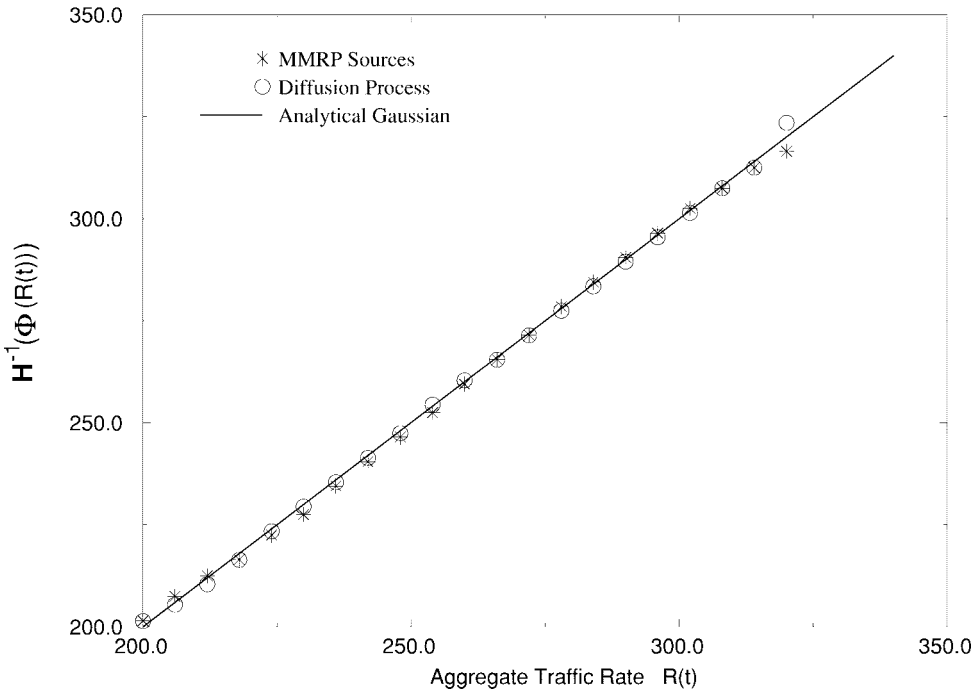


Fig. 11. Fractile diagram for the distribution functions of the aggregate traffic process $R(t)$, its diffusion process representation $\bar{R}(t)$, and the analytically derived Gaussian process. The state durations are constants.

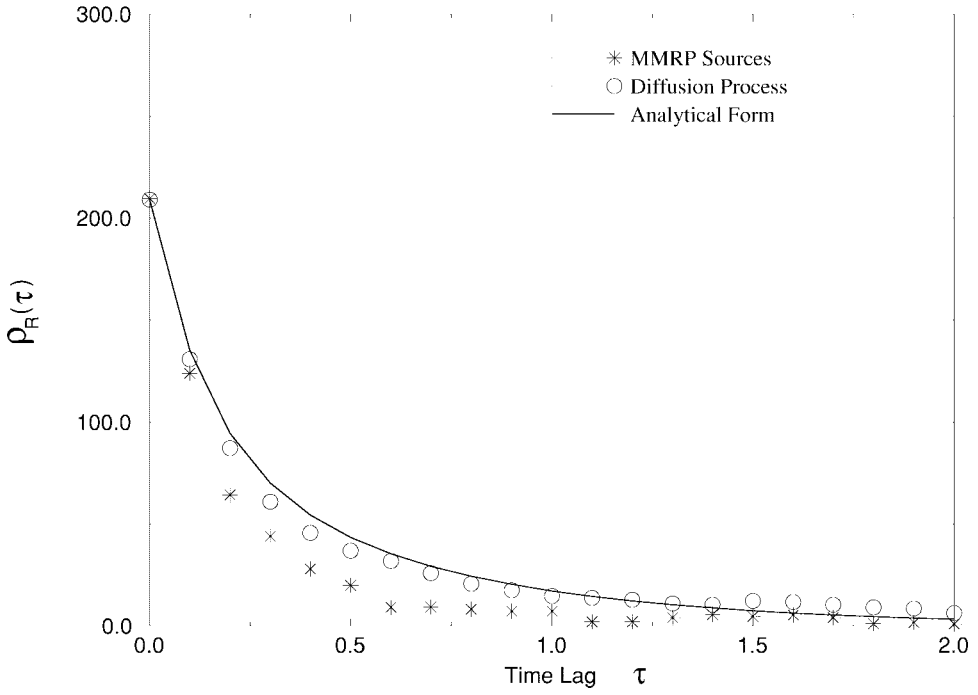


Fig. 12. Autocorrelation functions of $R(t)$, $\bar{R}(t)$, and the analytically derived form $\rho_{\tilde{R}}(\tau)$. The state durations are constants.

which have the squared coefficients of variation 1.2, 2.3, 3, and 1.9, respectively. In Fig. 14, we show the case with $K = 25$ sources. The link utilization is $\rho = 0.8$ with $C = 81.25$ [cells/s]. We see that $P[Q > x]$ decays at a slower rate compared to that in Fig. 10. This is expected because the hyperexponential distribution has a larger variance than the exponential or constant case. The process $\tilde{R}(t)$ approximates $R(t)$ well in terms of the queue length distribution. The analytical asymptotic estimate in (35) captures the decay rate reasonably well in this case also, although it gives an overestimate of an order of magnitude.

From what we have shown above, the accuracy and application of diffusion approximations depend on the traffic characteristics and system utilizations. The process $\tilde{R}(t)$ based on the diffusion process $\boldsymbol{X}(t)$ of (15) approximates the aggregate traffic $R(t)$ very well in terms of its marginal

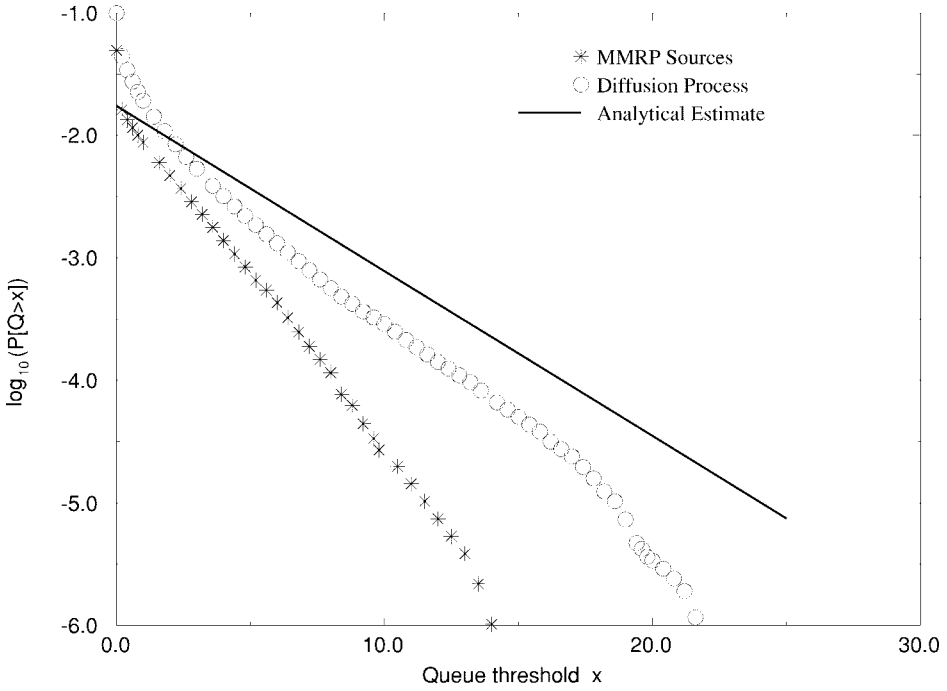


Fig. 13. $\log_{10}(P[Q > x])$ versus $x$. The simulation inputs parameters are the same as the cases for Fig. 9, except that the state durations are assumed constants.

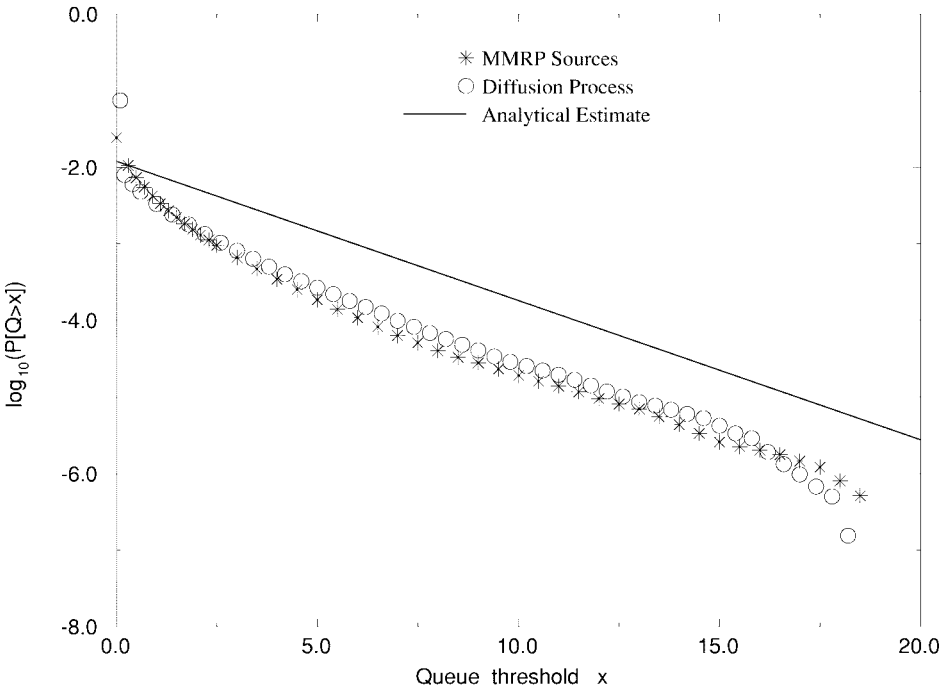


Fig. 14. $\log_{10}(P[Q > x])$ versus $x$. Simulations are done with $K = 25$ sources. The output link utilization $\rho = 0.8$ and the distributions of state durations are $H_2$ (two-stage hyperexponential).

distribution, auto-correlation, and queueing behaviors when the holding times are exponentially distributed. This is not unexpected because the original process $\boldsymbol{N}(t)$ is a Markov process for the exponentially distributed holding times. As holding times deviate from the exponential distribution, $\boldsymbol{N}(t)$ is no longer Markovian. While the Markov process $\boldsymbol{X}(t)$ can still capture the first- and second order statistics of $\boldsymbol{N}(t)$, it may not represent its queueing behavior accurately, which may depend on higher orders statistics. The analytical asymptotic formula in (35) provides a very good estimate for the complementary queue length distribution under relatively heavy traffic (say, $\rho \geq 0.8$). When $\rho$ is moderate or low, (35) tends to overestimate $P[Q > B]$. Thus, under moderate to light traffic, a technique based on large deviation theory may

be more appropriate, but that subject is beyond the scope of this paper. For the theory and applications of large deviations to ATM networks, readers are referred to Shwartz and Weiss [29].

## VIII. MMRP Sources with Long-Range Dependence

So far in this paper, we have considered the traditional MMRP sources whose holding-times have finite variances and whose aggregate traffic process has exponentially decaying autocorrelation functions. Recent empirical studies (e.g., [4], [20], [25], [33]) on actual packet network traffic indicate that aggregate packet streams often exhibit statistically *self-similar* characteristics. Such traffic streams possess the property of *long-range dependence* (LRD), i.e., they have power-law decaying autocorrelation functions rather than the exponentially decaying ones. As a result, the asymptotic complementary queue length distributions for a queue fed by such traffic streams will have subexponential decays. For more detailed discussions and analysis on self-similar and LRD traffic, the reader is referred to [28] and references therein.

To relate the LRD traffic to the MMRP models treated in this paper, we show below that the process $R(t)$ will have the LRD property if the holding times for individual MMRP sources have power-tail distributions (or called *heavy-tailed*)

$$P[T_i > t] \sim \frac{1}{t^{\beta_i}}, \qquad 0 < \beta_i \le 2 \quad \text{and} \quad i = 1, \cdots, M \tag{40}$$

which implies infinite variance and moments of higher orders.

Following the work by Norros [24], we characterize the integrated arrival process $I(t)$ by the stochastic differential equation

$$dI(t) = \mu_{\tilde{R}} \cdot dt + \sqrt{a} \cdot dZ^H(t) \tag{41}$$

in which we replace $W(t)$ in (29) by a *fractional* Brownian motion $Z^H(t)$ with a *Hurst* parameter $H \in [\frac{1}{2}, 1)$. In case $H = \frac{1}{2}, Z^H(t) = W(t)$.

If the traffic characterized by (41) is fed to a multiplexer with an output link capacity $C, P[Q > x]$ can be approximated by a *Weibull* distribution [24]

$$P[Q > x] \sim \exp\left(-\frac{(C - \mu_{\tilde{R}})^{2H}}{2\kappa^2 a}\, x^{2-2H}\right) \tag{42}$$

where $\kappa = H^H(1-H)^{1-H}$. Note that when $H = \frac{1}{2}$, the above formula agrees with the exponent in (35). Then it remains to show how $a$ and $H$ of (41) can be determined from the MMRP sources with heavy-tailed holding times characterized by (40).

The derivation of $a$ using (28) requires the squared coefficients of variation $c_m(x_m)$ as in (7). They must be finite in our modeling. However, if we should adopt the power-tail distributions (40), we would have infinite $c_m(x_m)$. Therefore, we proceed to approximate a power-tail distributions $1/t^{\beta}$ by an $L$-stage hyper-exponential distribution $H_L$ (called *truncated power-tails* in Greiner [11]):

$$P_L(t) = \frac{1-s}{1-s^L} \sum_{n=0}^{L-1} s^n \exp\left(-\frac{t}{r^L}\right), \quad \text{where } 0 < s < 1 \text{ and } r > 1. \tag{43}$$

Greiner has shown that $P_L(t)$ satisfies (40) as $L \to \infty$, and $\beta = -\log(s)/\log(r)$. Thus, for any finite $L$, (43) has finite variance and can serve as an approximation of the power-tail distribution in (40), and the $c_m(x_m)$ are then calculated by the method presented in Section II.

The Hurst parameter $H$ in (41) is chosen as follows:

$$H = \max_{1 \le i \le M-1, R_i > (C/K)} \left(\frac{3 - \beta_i}{2}\right) \tag{44}$$

i.e., the maximum possible $H$ in overload states. The Hurst parameters in underload states take no effects in queueing performance, as suggested in [6] by Choudhury and Whitt.

## IX. Conclusion

We have formulated a diffusion process approximation model to characterize the superposed traffic stream from many MMRP sources and then analyze its queueing behavior in a statistical multiplexer. The traffic from $K$ MMRP sources with $M$ states can be represented as a closed queueing network with $K$ customers and $M$ infinite-server nodes. Our source model is more general than those assumed by many previous studies: each source can have an arbitrary number of states, and its duration in a given state can have an arbitrary distribution and the cell generation rate can be state-dependent.

The diffusion processes developed in our paper can adequately capture the first- and second-order statistics of the multiplexed input and the queue process (i.e., long-term infinitesimal mean and variance). The processes $R(t)$ and $Q(t)$ are non-Markovian and are very difficult to analyze. Our diffusion approximation analysis provides simple and closed-form formulas for such performance measures as the overload probability $P[R(t) > C]$ and the buffer overflow probability $P[Q(t) > B]$, and can avoid computational complexities associated with a large number for $K$. The simulation results show the diffusion processes give reasonably accurate approximations, unless the durations of states are highly skewed from exponential distributions. We also showed how our diffusion approximation analysis results can be used to estimate an equivalent bandwidth for given CLR.

Although we have treated homogeneous sources only, our approach can be easily generalized to a system with multiple types of traffic—each traffic type should be modeled as in this paper, and the overall process is then simply a sum of these components. In other words, the quantities defined in (25) and (28) should be represented as the sums of the corresponding quantities of individual types, as we have shown in [17].

## References

[1] R. Addie and M. Zukerman, "An approximation for performance evaluation of stationary single server queues," *IEEE Trans. Commun.*, vol. 42, pp. 3150–3160, Dec. 1994.

[2] D. Anick, D. Mitra, and M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 61, no. 8, pp. 1871–1894, 1982.

[3] R. Bellman, *Introduction to Matrix Analysis*. New York: McGraw-Hill, 1970.

[4] J. Beran, R. Sherman, M. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol. 43, pp. 1566–1579, 1995.

[5] G. Choudhury, D. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, pp. 203–217, 1996.
[6] G. Choudhury and W. Whitt, "Long-tail buffer-content distributions in broadband networks," *Performace Evaluation*, vol. 30, no. 3, pp. 177–190, 1997.
[7] D. R. Cox and H. D. Miller, *Theory of Stochastic Process*. Methuen, 1965.
[8] D. R. Cox, *Renewal Theory*. London, U.K.: Methuen, 1972.
[9] A. Elwalid, D. Mitra, and T. E. Stern, "Statistical multiplexing of Markov modulated sources: Theory of computational algorithms," in *Proc. 13th Int. Teletraffic Congr. (ITC-13)*, 1991, pp. 495–500.
[10] E. Gelenbe, "On approximate computer system models," *J. Ass. Comput. Mach.*, vol. 22, pp. 261–263, 1975.
[11] M. Greiner, M. Jobmann, and L. Lipsky, "The importance of power-tail distributions for telecommunications traffic models," Tech. Rep., Inst. Inform., Technol. Univ. München, München, Germany, 1995.
[12] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 856–868, 1986.
[13] D. Iglehart, "Limiting diffusion approximation for the many server queue and the repairman problem," *J. Appl. Prob.*, vol. 2, pp. 429–441, 1965.
[14] C. Knessl and J. Morrison, "Heavy traffic analysis of a data handling system with multiple sources," *SIAM J. Appl. Math.*, vol. 51, pp. 187–213, 1991.
[15] H. Kobayashi, "Application of the diffusion approximation to queueing networks, Part I: Equilibrium queue distribution," *J. Ass. Comput. Mach.*, vol. 21, no. 2, pp. 316–328, 1974.
[16] H. Kobayashi, Y. Onozato, and D. Huynh, "An approximation method for design and analysis of an ALOHA system," *IEEE Trans. Commun.*, vol. COM-25, no. 1, pp. 148–157, 1977.
[17] H. Kobayashi and Q. Ren, "A diffusion approximation analysis of an ATM statistical multiplexer with multiple types of traffic, Part I: Equilibrium state solutions," in *Proc. ICC'93*, vol. 2, 1993, pp. 1047–1053.
[18] K. Kobayashi, "Steady state approximation analysis for ATM multiplexer By diffusion process without reflection barrier," in *Proc. Symp. Performance Models Inform. Commun. Networks*, Hakone, Japan, Jan. 19–21, 1994, pp. 309–320.
[19] L. Kosten, "Stochastic theory of data handling systems with groups of multiple sources," *in Performance of Computer-Communication Systems*, H. Rudin and W. Bux, Eds. Amsterdam, The Netherlands: North-Holland, 1984, pp. 321–331.
[20] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol 2, pp. 1–15, 1994.
[21] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, pp. 834–844, 1988.
[22] B. Melamed and B. Sengupta, "TES modeling of video traffic," *IEICE Trans. Commun.*, vol. E75-B, no. 12, pp. 1292–1300, 1992.
[23] I. Norros, J. Roberts, A. Simonian, and J. Virtamo, "The superposition of variable bit rate sources in an ATM multiplexer," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 378–387, 1991.
[24] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 953–962, 1995.
[25] V. Paxon and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226–244, 1995.
[26] G. Ramamurthy and Q. Ren, "Multicast connection admission control policy for ATM switches," in *Proc. IEEE INFOCOM'97*, Apr. 1997.
[27] Q. Ren and H. Kobayashi, "Diffusion process approximations of a statistical multiplexer with Markov modulated bursty traffic sources," in *Proc. GLOBECOM'94*, pp. 1100–1106.
[28] J. Roberts, U. Mocci, and J. Virtamo, Eds., "Broadband network teletraffic," Final Rep. Action 242, 1997.
[29] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communications, and Computing (Stochastic Modeling)*. New York: Book News, 1995.
[30] A. Simonian, "Stationary analysis of a fluid queue with input rate varying as an Ornstein–Uhlenbeck process," *SIAM J. Appl. Math.*, vol. 51, pp. 828–842, 1991.
[31] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, vol. 4, no. 6, pp. 1124–1132, 1986.
[32] W. Whitt, "Approximating a point process by a renewal process: Two basic methods," *Oper. Res.*, vol. 30, no. 1, pp. 125–147, 1982.
[33] W. Willinger, M. Taqqu, W. Leland, and D. Wilson, "Self-similarity in high speed packet traffic: Analysis and modeling of Ethernet traffic measurement," *Stat. Sci.*, vol. 10, pp. 67–85, 1995.

**Qiang Ren** received the B.A. degree in mathematics from Beijing University, China, in 1989, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, in 1991 and 1994, respectively.

He joined C&C Research Laboratories of NEC USA, Inc., Princeton, as a Research Staff Member in February 1994. His research interests include design, control, and performance analysis of broadband communication networks, queueing theory, and its applications to computer communications.

**Hisashi Kobayashi** received the B.S.E. and M.S.E degrees from the University of Tokyo, in 1961 and 1963, respectively, and the Ph.D. degree from Princeton University, Princeton, NJ, in 1967 all in electrical engineering.

He was a Radar Engineer at Toshiba Electric Company, Kawasaki, Japan (1963–1965), prior to coming to Princeton as an Orson Desaix Munn Fellow. From 1967 to 1982 he was with IBM T. J. Watson Research Center, Yorktown Heights, NY, where he worked on data transmission, seismic signal processing, digital magnetic recording, image compression, performance modeling of computers and communication systems, and queueing network theory. He is the inventor (1971) of a high-density magnetic recording method, now widely known as partial-response, maximum-likelihood decoding (PRML) scheme, a co-inventor (1974) of *relative address coding* for image compression. He also developed the convolutional algorithm (1975) for a multiclass queueing network, and the diffusion approximation method for a general queueing network (1974). He served as Senior Manager of Systems Analysis and Algorithms (1974–1980), and Department Manager of VLSI Design (1981–1982). From 1982 to 1986, he was the Founding Director of the IBM Tokyo Research Laboratory. In 1986, he joined Princeton University as Dean of Engineering and Applied Science (1986–1991), and as the Sherman Fairchild University Professor of Electrical Engineering and Computer Science. His current research interests include coding and modulation for wireless communications and digital recording, performance analysis of ATM and optical networks, and teletraffic theory. He is the author of *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology* (Reading, MA: Addison-Wesley, 1978), and has authored more than 120 technical papers.

Dr. Kobayashi is the recipient of the *Humboldt Prize* from Germany (1979); IFIP (International Federation of Information Processing)'s Silver Core Award (1980); IBM Outstanding Contribution Awards (1975 and 1984); and IBM Invention Achievement Awards (1971 and 1973). He was elected a member of *Engineering Academy of Japan* (1992).