Performance Evaluation 43 (2001) 39–62

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

Connection-wise end-to-end performance analysis of queueing networks with MMPP inputs

Huei-Wen Ferng*, Jin-Fu Chang

Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan, ROC

Received 3 March 1998; received in revised form 14 June 2000

Abstract

A systematic method for link/connection-wise end-to-end performance evaluation in queueing networks receiving heterogeneous Markov-modulated Poisson processes (H-MMPPs) is proposed. The method consists of (i) connection-wise nodal performance analysis; (ii) tagged departure process analysis; and (iii) moment matching. For the tagged departure process of an H-MMPPs/G/1 queue, we propose two decomposition schemes to approximate the output process of a tagged traffic stream which is mixed with other traffic streams. A moment matching method is further proposed to emulate the tagged output process as a two-state MMPP. The adjacent down-stream node along a reference connection can be then modeled as an H-MMPPs/G/1 queue. Recursively performing (i)–(iii), the end-to-end performance of a reference connection is obtained. The methodology developed in this paper can be applied to packet-switched high-speed networks, especially to asynchronous transfer mode (ATM) networks.

Keywords: MMPP; Queueing network; Output process characterization; End-to-end performance analysis

1. Introduction

Poisson processes have been previously extensively employed to model traffic flows in telephony/computer networks since Poissonian assumptions make performance analysis tractable. Well known results reported in the literature include, for examples, Jackson's theorem and the BCMP theorem [4,9]. The mean value analysis (MVA) technique has been widely applied to treat Poissonian queueing networks [9]. The M/G/1 queue as well as its variants have been thoroughly discussed by Takagi [33].

As transport technologies evolve, computer and communication networks are quickly moving towards a new era. Networks are now used to provide versatile services to meet different types of subscriber needs, including voice, video, data, etc. Multimedia traffic has characteristics very different from Poisson characteristics, generating high rate data at one time and low rate data at another. Strong correlation or long-range dependence has been observed in such traffic. Therefore, more sophisticated traffic models suitable for broadband networks are needed to capture the characteristics of multimedia traffic. In the literature, various mathematical models have been proposed for this purpose, including the continuous-time

* Corresponding author. Tel.: +886-2-236-352-51, ext: 533; fax: +886-2-236-382-47.
E-mail addresses: wen@santos.ee.ntu.edu.tw (H.-W. Ferng), jfchang@fcom.ee.ntu.edu.tw (J.-F. Chang).

0166-5316/01/$ – see front matter
PII: S0166-5316(00)00036-5

models such as the N-process [24], batch Markovian arrival process (BMAP) [20], Markovian arrival process (MAP) [19,20], Markov-modulated Poisson process (MMPP) [8], the discrete-time models such as the discrete-time batch Markovian arrival process (D-BMAP) [5], switched batch Bernoulli process (SBBP) [10], discrete-time Markovian arrival process (D-MAP) [5], Markov-modulated Bernoulli process (MMBP) [25]. Among these traffic models, the MMPP has received a great deal of attention, and an excellent survey of the MMPP is presented by Fischer and Meier-Hellstern in [8]. The MMPP is capable of capturing both time-varying arrival rates and correlations between interarrival times. It has been extensively used to model multimedia sources in broadband integrated services digital networks (B-ISDNs) [1,14,18,26,31]. Moreover, the MMPP model leads to tractable analysis as well as accurate results [12,26]. Although more sophisticated models such as the N-process (or its equivalent, the BMAP) have been proposed, the MMPP is still popular due to its considerably reduced complexity. Therefore, we use the MMPP to model a traffic flow in this paper.

Among various transport technologies, ATM has been widely adopted to support different services in B-ISDNs. It is capable of transporting constant bit rate (CBR), variable bit rate (VBR), available bit rate (ABR), and unspecified bit rate (UBR) services to meet different quality-of-service (QoS) requirements. Since ATM networks are connection-oriented, QoS requirements (such as cell delay, cell loss rate, and cell delay variation [3,21]) are defined on a per-connection basis. Although knowing the connection-wise end-to-end performance measures for a virtual connection (VC) is important, however, very few papers have addressed this issue. Refs. [2,6,16,23,28,30] are some of the previous works that have appeared in the literature to study end-to-end performance in ATM networks. Naser and Leon-Garcia [23] used a simulation approach to study the end-to-end performance of a CBR connection. Addie and Zukerman [2] studied the cell loss probability of a tree-type ATM network under discrete-time Gaussian arrivals. Kroner et al. [16] approximated the end-to-end delay jitter in a tandem ATM network under burst silence cell streams. Ren et al. [28] also studied the end-to-end performance of ATM networks under On–Off cell streams. Sohraby and Privalov [30] provided the end-to-end jitter analysis of feed-forward connection-oriented networks which support multiplexed CBR connections. In [6], Ferng and Chang obtained the connection-wise end-to-end cell sojourn delay time for a tagged VC in an ATM network using the H-MMPPs/Slotted D/1 queue to model an ATM multiplexer or an ATM switch output port. These estimates can be applied to facilitate call admission control, congestion control, routing algorithms, etc.

In ATM networks, data streams are broken into small fixed-size cells. A node in an ATM network is either a multiplexer or a switch output port. Following [18], each node can be modeled as an H-MMPPs/D/1 queue, where H says that a number of heterogeneous multiclass traffic streams are multiplexed/routed at the node, and D represents the deterministic service facility to reflect the fixed-size nature of cells. In this queueing model, infinite buffer size can be assumed without loss of generality if cell loss rate is in the order of 10^{-6} which holds in most situations [3]. Therefore, the queueing model for each VC can be regarded as a tandem configuration of H-MMPPs/D/1 queues. In this paper, we further generalize the server model from D to G for both theoretical and practical reasons. From a theoretical viewpoint, this extends the scope of open networks of queues from Poissonian to those receiving heterogeneous MMPPs. General service distribution fits other packet-switched networks as well. Park et al. [27] have considered a tandem configuration of queues with geometric servers and MMBP arrivals. From a practical perspective, we make this modification for two reasons. First, to account for the time consumed in a controller, e.g., leaky bucket which is a traffic shaper employed in ATM networks [3]: traffic inbound to an ATM network may first pass through a controller, then receives a service time of one cell. Second, to model the virtual server seen by a tagged traffic: in ATM networks, VCs are multiplexed/routed at a multiplexer/switch.

For a specific connection, we may describe its queueing behavior via a G server, whereas the general service time distribution is used to reflect the phenomenon that the tagged traffic is mixed with the other connections. This approach allows us to use decomposition schemes to simplify queueing analysis. Therefore, this paper can be regarded as an extension of our previous work [6].

Unlike analysis for open networks of queues with Poisson inputs, analysis for open networks fed by bursty and correlated processes seems to be intractable. A systematic and efficient (but may be approximate) method should be developed to understand the performance of such networks. A well-known approach, i.e., Kleinrock's independence approximation [2,4,6,22,27], was to first decompose a network into a set of isolated queues, and then analyze each queue in isolation. In this paper, we again employ this approach to analyze a network of queues with heterogeneous MMPP inputs. But we first need to solve the problem of how the input processes are fed into an intermediate queue: can these input processes be again assumed to be MMPPs? If yes, how should they be characterized? To solve this problem, we analyze the tagged output process of an H-MMPPs/G/1 queue and propose a moment matching scheme to emulate the tagged output process as an MMPP. These are the two main goals of this paper.

Although Stanford and Fischer [32], and Saito [29] have independently derived the tagged output process in the $\sum_i \mathrm{M}_i/\mathrm{G}_i/1$ queue, i.e., an M/G/1 queue with heterogeneous inputs and service distributions, analysis for the tagged output process of the H-MMPPs/G/1 is more involved and complicated. The complexity arises from the large number of states of the aggregated MMPPs to which the tagged MMPP belongs. For these reasons, we use the indirect method of employing decomposition instead of pursuing an exact analysis. We propose two decomposition schemes to derive the output process corresponding to a tagged stream in the H-MMPPs [1].

A moment matching scheme modified from that developed in [6] is next proposed to emulate the tagged output process as a two-state MMPP. By doing so, the adjacent down-stream node along a reference connection can be modeled again as an H-MMPPs/G/1 queue. This makes the analysis systematic and efficient. In the literature, moment matching has been used for different purposes, e.g., [6,11,12,17]. As in [6], we match the following four statistics of the tagged output process: the first moment of the interdeparture time, the squared coefficient of the variation of the interdeparture time, the third moment of the interdeparture time, and the lag 1 covariance of two adjacent interdeparture times. The reason why these four statistics are selected is given in [6]. Through connection-wise nodal performance of H-MMPPs/G/1 queue in [8], composite/tagged output process characterization, and moment matching, we can iteratively obtain the performance measures from the source node enroute to the destination node along a reference connection.

The rest of this paper is organized as follows. Section 2 describes the traffic and system models used in this paper. In Section 3, the approximate output process corresponding to a tagged MMPP of the H-MMPPs/G/1 queue is obtained via two decomposition schemes mentioned earlier. In Section 4, moment matching is applied to emulate the tagged output process as a two-state MMPP. Section 5 gives the systematic method for evaluating connection-wise end-to-end performance. Section 6 provides comprehensive numerical experiments and discussions. Computer simulations are given to demonstrate that the proposed approach indeed produces results of good accuracy. Section 7 concludes the paper and outlines potential future research.

[1] H-MMPPs represents a superposition of heterogeneous MMPPs and is also an MMPP. We use H-MMPPs instead of MMPP because per stream or connection-wise analysis is the major concern of this paper.

2. Traffic and system models

The queueing model considered is a network of queues receiving heterogeneous MMPP inputs. Both tandem and tree type configurations are considered in this paper, using a two-state MMPP to model each external input traffic. The following descriptions regarding the MMPP are taken from [8] for readers convenience. A two-state MMPP of type i is characterized by the following infinitesimal generator $\boldsymbol{Q}_i$ of the underlying Markov chain and rate matrix $\boldsymbol{\Lambda}_i$ with two Poisson rates λ_{1i} and λ_{2i}.

$$\boldsymbol{Q}_i = \begin{bmatrix} -\sigma_{1i} & \sigma_{1i} \\ \sigma_{2i} & -\sigma_{2i} \end{bmatrix}, \qquad \boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_{1i} & 0 \\ 0 & \lambda_{2i} \end{bmatrix}. \tag{1}$$

Assume a total of r two-state MMPPs, then the traffic descriptor $(\boldsymbol{Q}, \boldsymbol{\Lambda})$ of the aggregated process (which is again an MMPP with $m = 2^r$ possible states) can be written in terms of $(\boldsymbol{Q}_i, \boldsymbol{\Lambda}_i)$, $1 \leq i \leq r$, as follows [8]:

$$\boldsymbol{Q} = \boldsymbol{Q}_1 \oplus \boldsymbol{Q}_2 \oplus \cdots \oplus \boldsymbol{Q}_r, \tag{2}$$

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_1 \oplus \boldsymbol{\Lambda}_2 \oplus \cdots \oplus \boldsymbol{\Lambda}_r, \tag{3}$$

where $\oplus$ represents the Kronecker sum. Let λ denote the mean arrival rate of the aggregated process. Then $\lambda = \boldsymbol{\pi}\boldsymbol{\Lambda}\boldsymbol{e}$ with $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$ and $\boldsymbol{\pi}\boldsymbol{e} = 1$, where $\boldsymbol{e}$ and $\boldsymbol{0}$ are $m \times 1$ column vector of all ones and $1 \times m$ row vector of all zeros, respectively. The arrival rate from type i $(1 \leq i \leq r)$ traffic stream is independent of the other traffic streams and is given by

$$\boldsymbol{\Lambda}(i) = \boldsymbol{0} \oplus \cdots \oplus \boldsymbol{0} \oplus \boldsymbol{\Lambda}_i \oplus \boldsymbol{0} \oplus \cdots \oplus \boldsymbol{0}, \tag{4}$$

where $\boldsymbol{0}$ is a 2×2 zero matrix. The mean arrival rate from type i traffic stream is $\lambda_i^* = \boldsymbol{\pi}\boldsymbol{\Lambda}(i)\boldsymbol{e}$.

Service-type of queues in the network may be different and independently follow general service time distributions. For a specific queue, the service times of different customers are assumed independent and identically distributed (i.i.d.) with distribution function $\tilde{H}(\cdot)$ whose Laplace–Stieltjes transform (LST) and mean service time are $H(s)$ and h, respectively.

3. Departure processes of an H-MMPPs/G/1 queue

3.1. Composite/overall output process of an H-MMPPs/G/1 queue

3.1.1. Preliminaries

Let $\{\tau_n : n \geq 0\}$ be the successive epochs of departure with $\tau_0 = 0$. Further define L_n and J_n to be, respectively, the number of customers in the system and the state of the H-MMPPs (with m states) right after τ_n. Then $\{(L_n, J_n, \tau_{n+1} - \tau_n) : n \geq 0\}$ forms a semi-Markov sequence with the following transition matrix:

$$\tilde{\boldsymbol{Q}}(x) = \begin{bmatrix} \tilde{\boldsymbol{B}}_0(x) & \tilde{\boldsymbol{B}}_1(x) & \tilde{\boldsymbol{B}}_2(x) & \dots \\ \tilde{\boldsymbol{A}}_0(x) & \tilde{\boldsymbol{A}}_1(x) & \tilde{\boldsymbol{A}}_2(x) & \dots \\ \boldsymbol{0} & \tilde{\boldsymbol{A}}_0(x) & \tilde{\boldsymbol{A}}_1(x) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{5}$$

where $\tilde{\boldsymbol{A}}_n(x)$ and $\tilde{\boldsymbol{B}}_n(x)$ are $m \times m$ matrices of mass functions defined in [8] as follows:

$$\tilde{\boldsymbol{A}}_n(x) = \int_0^x \boldsymbol{P}(n,u)\,\mathrm{d}\tilde{H}(u), \quad n \geq 0,\ x \geq 0, \tag{6}$$

$$\tilde{\boldsymbol{B}}_n(x) = \tilde{\boldsymbol{U}}(x) * \tilde{\boldsymbol{A}}_n(x), \quad n \geq 0,\ x \geq 0, \tag{7}$$

$$\tilde{\boldsymbol{U}}(x) = \int_0^x \exp\{(\boldsymbol{Q} - \boldsymbol{\Lambda})t\}\boldsymbol{\Lambda}\,\mathrm{d}t, \tag{8}$$

where $\boldsymbol{P}(n,u)$ is an $m \times m$ matrix with entry $[\boldsymbol{P}(n,u)]_{ij}$ representing the probability that given the state of the arrival process is i at time 0, there are n customers arriving during a period of length u and the state of the arrival process is j at time u and the operator $*$ denotes matrix convolution. We use $\boldsymbol{Q}(s)$, $\boldsymbol{A}_n(s)$, $\boldsymbol{B}_n(s)$, and $\boldsymbol{U}(s)$ to denote the LSTs of $\tilde{\boldsymbol{Q}}(x)$, $\tilde{\boldsymbol{A}}_n(x)$, $\tilde{\boldsymbol{B}}_n(x)$, and $\tilde{\boldsymbol{U}}(x)$, respectively.

Assuming that (L_n, J_n) is stationary, we use $\boldsymbol{x} = (\boldsymbol{x}_0, \dots, \boldsymbol{x}_k, \dots)$ to denote the stationary probability of the number of customers in the system immediately after a departure, i.e., $\boldsymbol{x}\boldsymbol{Q}(0) = \boldsymbol{x}$. Here $\boldsymbol{x}_k = (x_{k,1}, \dots, x_{k,m})$ is the stationary probability vector with $x_{k,j} = \lim_{n\to\infty} \Pr\{L_n = k, J_n = j\}$.

The following results are taken from [7] because an MMPP is a special case of the N-process (or its equivalent, the BMAP).

3.1.2. Interdeparture time moments

Let $T_{\mathrm{D},i}$ denote the time between the ith and the $(i+1)$th departure of the output process. From [7], we have the following moments of interdeparture times:

$$E[T_{\mathrm{D},i}^n] = (-1)^n \left[\sum_{i=0}^{n-1} \frac{n!}{i!} H^{(i)}(0)\boldsymbol{x}_0\boldsymbol{R}^{-(n-i)}(0)\boldsymbol{e}\right] + (-1)^n H^{(n)}(0), \tag{9}$$

where $\boldsymbol{R}(0) = \boldsymbol{Q} - \boldsymbol{\Lambda}$. In particular, the first three moments are given by

$$E[T_{\mathrm{D},i}] = h - \boldsymbol{x}_0\boldsymbol{R}^{-1}(0)\boldsymbol{e}, \tag{10}$$

$$E[T_{\mathrm{D},i}^2] = H^{(2)}(0) - 2h\boldsymbol{x}_0\boldsymbol{R}^{-1}(0)\boldsymbol{e} + 2\boldsymbol{x}_0\boldsymbol{R}^{-2}(0)\boldsymbol{e}, \tag{11}$$

$$E[T_{\mathrm{D},i}^3] = -H^{(3)}(0) - 3H^{(2)}(0)\boldsymbol{x}_0\boldsymbol{R}^{-1}(0)\boldsymbol{e} + 6h\boldsymbol{x}_0\boldsymbol{R}^{-2}(0)\boldsymbol{e} - 6\boldsymbol{x}_0\boldsymbol{R}^{-3}(0)\boldsymbol{e}. \tag{12}$$

3.1.3. Covariance of two successive interdeparture times

From [7], we have

$$\begin{aligned} \mathrm{Cov}(T_{\mathrm{D},i}, T_{\mathrm{D},i+1}) = {} & h\boldsymbol{x}_0\boldsymbol{R}^{-1}(0)\boldsymbol{e} - [\boldsymbol{x}_0\boldsymbol{R}^{-1}(0)\boldsymbol{e}]^2 + \boldsymbol{x}_1\boldsymbol{A}_0'(0)\boldsymbol{R}^{-1}(0)\boldsymbol{e} \\ & + \boldsymbol{x}_0\boldsymbol{R}^{-1}(0)\boldsymbol{U}(0)\boldsymbol{A}_0(0)\boldsymbol{R}^{-1}(0)\boldsymbol{e} + \boldsymbol{x}_0\boldsymbol{U}(0)\boldsymbol{A}_0'(0)\boldsymbol{R}^{-1}(0)\boldsymbol{e}, \end{aligned} \tag{13}$$

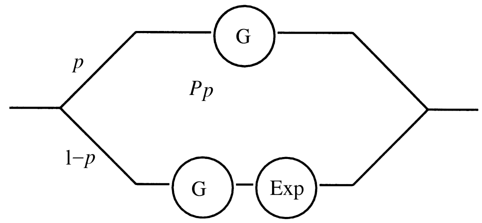


Fig. 1. Compound server model.

where $\boldsymbol{x}_0$ is calculated in [8] for the H-MMPPs/G/1 queue and $\boldsymbol{U}(0) = (\boldsymbol{\Lambda} - \boldsymbol{Q})^{-1}\boldsymbol{\Lambda}$. $\boldsymbol{A}_0(0)$ and $\boldsymbol{A}'_0(0)$ are given in Appendix A for several specific service distributions.

3.2. Tagged output process of an H-MMPPs/G/1 queue

Although Kleinrock's independence approximation allows us to decompose a network into a set of isolated queues, it is still necessary to determine the input processes to the intermediate nodes on an end-to-end path. Since the output process of a node may contribute in part to the input processes of the adjacent down-stream node, it becomes necessary to know the output process corresponding to a tagged MMPP at the input of the H-MMPPs/G/1 queue in order to facilitate the connection-wise end-to-end analysis of the tagged traffic stream.

For convenience, we denote the tagged MMPP by $\mathrm{MMPP_t}$ with traffic load ρ_{t} and the other interfering/cross-traffic sources as a whole by $\mathrm{MMPP_c}$ with traffic load ρ_{c}. Let ρ denote the total traffic load, i.e., $\rho = \lambda h = \rho_{\mathrm{t}} + \rho_{\mathrm{c}}$. Then the H-MMPPs arrivals are symbolically represented by $\mathrm{MMPP_t} + \mathrm{MMPP_c}$. The proposed decomposition is to replace the $\mathrm{MMPP_t} + \mathrm{MMPP_c}/\mathrm{G}/1$ queue by an $\mathrm{MMPP_t}/\mathrm{G_{eff}}/1$ queue. The modification is reflected in $\mathrm{G_{eff}}$ due to the interference arising from the cross-traffic. In this paper, we use the compound server model illustrated in Fig. 1 to model the modified effective server $\mathrm{G_{eff}}$ seen by the tagged $\mathrm{MMPP_t}$ stream only.

We now explain why this modified server model is employed. In the original $\mathrm{MMPP_t} + \mathrm{MMPP_c}/\mathrm{G}/1$ queue, the tagged $\mathrm{MMPP_t}$ stream and the cross-$\mathrm{MMPP_c}$ stream compete for service following the *first-come-first-serve* (FCFS) discipline. In other words, the tagged $\mathrm{MMPP_t}$ stream has only partial access to the service facility. Therefore, we modify the original server G to $\mathrm{G_{eff}}$ to reflect such partial usage. Note that when a customer from the $\mathrm{MMPP_t}$ stream is now at the head-of-line (HOL) in the decomposed queue, that customer may immediately receive service without extra waiting with probability $p = \rho_{\mathrm{t}}/\rho$ or may first wait if the server is busy with $\mathrm{MMPP_c}$ customers with probability $q = \rho_{\mathrm{c}}/\rho = 1 - \rho_{\mathrm{t}}/\rho = 1 - p$. This explains why these schemes use the two-branch compound server model depicted in Fig. 1. In this compound server model, we use an exponential random variable to capture the extra waiting time encountered by the $\mathrm{MMPP_t}$ customer. To use this model, an additional parameter describing the exponential server needs to be specified. We now describe two different decomposition schemes to fulfill different criteria.

Scheme I. This scheme determines the parameter pertinent to the exponential random variable in the compound server model via matching the mean service time of the compound server to the mean effective service time seen by the $\mathrm{MMPP_t}$ customers when competing with the $\mathrm{MMPP_c}$ customers. We first calculate the mean effective service time seen by the $\mathrm{MMPP_t}$ customer. Ignoring the effect of traffic burstiness [21], the server may serve a customer from the $\mathrm{MMPP_c}$ stream with probability ρ_{c} or become idle/serve a customer from the $\mathrm{MMPP_t}$ stream with probability $1-\rho_c$. Therefore, we have the following mean effective service time h_{eff}:

$$h_{\mathrm{eff}} = \sum_{i=0}^{\infty}(1-\rho_c)\rho_c^i(i+1)h = \frac{h}{1-\rho_c}. \tag{14}$$

Next, we calculate the mean service time of the compound server, first defining $h_{\exp}$ to be the mean time of the exponential random variable. Then the service time of the compound server is merely the weighted sum of service time of each branch, i.e., $ph + (1-p)(h+h_{\exp}) = h + (\rho_c/\rho)h_{\exp}$. Using (14) and equating the mean service time of the compound server to the mean effective service time seen by the $\mathrm{MMPP_t}$ customers, one obtains $h_{\exp} = [\rho/(1-\rho_c)]h$. This completes the specification of the compound server model. Via this compound server model, one can analyze the tagged departure process using the results of Section 3.1. We now summarize this scheme as follows:

- Set $h_{\exp} = [\rho/(1-\rho_c)]h$ for the compound server model sketched in Fig. 1.
- Calculate the departure process of $\mathrm{MMPP_t/G_{eff}/1}$ queue using (9) and (13) to obtain the departure process of the tagged $\mathrm{MMPP_t}$ traffic in the $\mathrm{MMPP_t + MMPP_c/G/1}$ queue.

Scheme II. Let us examine the relationship among interdeparture time, interarrival time, and sojourn time. We redefine $T_{D,i}$ to be the time between the ith and the $(i+1)$th departures of the tagged MMPP. Let $T_{A,i}$ denote the time between the ith and the $(i+1)$th arrivals of the tagged MMPP and $T_{S,i}$ stand for the sojourn time of the ith customer of the tagged MMPP, i.e., the time the ith customer of the tagged MMPP spent in the system. Then it can be easily shown that $T_{D,i} = T_{A,i} + (T_{S,i+1} - T_{S,i})$. This relation reveals that interdeparture time depends on $T_{A,i}$ as well as the sojourn time. Note that $T_{A,i}$ is related solely to the tagged MMPP while $T_{S,i}$ (or $T_{S,i+1}$) depends on both the service mechanism and cross-traffic (if that exists). Therefore, if the relation $T_{D,i} = T_{A,i} + (T_{S,i+1} - T_{S,i})$ for the $\mathrm{MMPP_t + MMPP_c/G/1}$ queue is still valid for the decomposed queue, we should keep $T_{S,i}$ (or $T_{S,i+1}$) unchanged. One viable way to maintain $T_{S,i}$ (or $T_{S,i+1}$) is to employ the compound server by matching the mean sojourn delay time that $\mathrm{MMPP_t}$ experiences in the $\mathrm{MMPP_t + MMPP_c/G/1}$ queue with the mean sojourn delay in the decomposed queue. This scheme is carried out as follows. First calculate the mean sojourn delay time $W_{q,i} + h$ for the $\mathrm{MMPP_t}$ in the undecomposed queue. Then recursively apply a search algorithm, say binary search, to find a mean service time of the compound server $h_m (= h + (1-p)h_{\exp})$ over a feasible region [2] such that the mean sojourn delay time $W_{q,a,\mathrm{eff}} + h_m$ of the decomposed queue satisfies $|W_{q,a,\mathrm{eff}} + h_m - W_{q,i} - h| < \epsilon$, where ϵ is a small number, specified to stop the recursion, e.g., $\epsilon = 10^{-8}$. Note that this scheme tries to maintain the random variable $T_{S,i}$ (or $T_{S,i+1}$) the same in both the original and the decomposed queue via keeping its first moment, i.e., its mean, unchanged. Using a binary search, a detailed description of this scheme is given in the following:

Step 1. Initialization: set a small number, say 10^{-8}, to ϵ; set the lower bound of $h_{\exp}$ to zero, i.e., $h_{\exp,l} = 0$, the upper bound of $h_{\exp}$ to $h^*_{\exp}$, i.e., $h_{\exp,u} = h^*_{\exp}$ [3]; and calculate $h_{\exp} = (h_{\exp,l} + h_{\exp,u})/2$, $h_{\mathrm{gap}} = h_{\exp,u} - h_{\exp,l}$.

Step 2. Calculate the mean sojourn delay time $W_{q,i} + h$ for the $\mathrm{MMPP_t}$ in the $\mathrm{MMPP_t + MMPP_c/G/1}$ queue using the results obtained in [8].

[2] Here, we let $h_{\exp}$ fall in the region $(0, h^*_{\exp})$, where $h^*_{\exp}$ represents an upper bound of $h_{\exp}$.

[3] In our calculation, we heuristically select $h^*_{\exp} = h/p$. Of course, one may choose other values.

Step 3. Let $h_\mathrm{m} = h + (1-p)h_\mathrm{exp}$ and calculate the mean sojourn delay time $W_{\mathrm{q},a,\mathrm{eff}} + h_\mathrm{m}$ for the $\mathrm{MMPP_t}$ in the decomposed $\mathrm{MMPP_t/G_{eff}/1}$ queue using the result obtained in [8].

Step 4. If $|W_{\mathrm{q},a,\mathrm{eff}} + h_\mathrm{m} - W_{\mathrm{q},i} - h| < \epsilon$, then go to Step 5; else $h_\mathrm{gap} = h_\mathrm{gap}/2$. If $W_{\mathrm{q},a,\mathrm{eff}} + h_\mathrm{m} - W_{\mathrm{q},i} - h < 0$, set $h_\mathrm{exp} = h_\mathrm{exp} + h_\mathrm{gap}$, and go to Step 3; else set $h_\mathrm{exp} = h_\mathrm{exp} - h_\mathrm{gap}$, and go to Step 3.

Step 5. Use h_exp as the mean time of the exponential random variable in the compound server model in Fig. 1.

Step 6. Calculate the departure process of the $\mathrm{MMPP_t/G_{eff}/1}$ queue using (9) and (13) to obtain the departure process of the $\mathrm{MMPP_t}$ traffic in the $\mathrm{MMPP_t + MMPP_c/G/1}$ queue.

Remark 1. Obviously, Scheme I saves computation time, whereas Scheme II consumes much more time in the search algorithm. For a real-time system, we prefer Scheme I.

Remark 2. The stoppage parameter ϵ affects not only the computation time but also the accuracy. Setting a smaller ϵ produces more accurate results but takes longer computation time.

4. Moment matching scheme

In order to approximate the overall or tagged output process as a two-state MMPP, we need to match selected statistics. As in [6], we select the following four statistics: (a) the first moment of the interdeparture time $E[T_{\mathrm{D},i}]$; (b) the squared coefficient of the variation of the interdeparture time $c^2(T_{\mathrm{D},i}) = \mathrm{Var}[T_{\mathrm{D},i}]/E^2[T_{\mathrm{D},i}]$; (c) the third moment of the interdeparture time $E[T^3_{\mathrm{D},i}]$; and (d) the lag 1 covariance of two adjacent interdeparture times $\mathrm{Cov}(T_{\mathrm{D},i}, T_{\mathrm{D},i+1})$.

In the following, we use $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21})$ to denote the traffic descriptor of the tagged traffic at the input end. The matching equations regarding these four selected statistics have been acquired in [6] and are reproduced here for readers convenience as follows:

$$\frac{\sigma_1^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})}}{\sigma_2^{(\mathrm{m})}\lambda_1^{(\mathrm{m})} + \sigma_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})}} = A_d, \tag{15}$$

$$1 + \frac{2\sigma_1^{(\mathrm{m})}\sigma_2^{(\mathrm{m})}(\lambda_1^{(\mathrm{m})} - \lambda_2^{(\mathrm{m})})^2}{[\sigma_2^{(\mathrm{m})}\lambda_1^{(\mathrm{m})} + \sigma_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})} + \lambda_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})}](\sigma_1^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})})^2} = B_d, \tag{16}$$

$$\begin{aligned}&\frac{3C_{d,1}(\sigma_1^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})})}{\sigma_2^{(\mathrm{m})}\lambda_1^{(\mathrm{m})} + \sigma_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})} + \lambda_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})}} \\ &\quad + \frac{6[(\sigma_1^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})})(\sigma_1^{(\mathrm{m})}\lambda_1^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})}\lambda_2^{(\mathrm{m})}) + \sigma_1^{(\mathrm{m})}\lambda_1^{(\mathrm{m})^2} + \sigma_2^{(\mathrm{m})}\lambda_2^{(\mathrm{m})^2}]}{(\sigma_2^{(\mathrm{m})}\lambda_1^{(\mathrm{m})} + \sigma_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})})(\sigma_2^{(\mathrm{m})}\lambda_1^{(\mathrm{m})} + \sigma_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})} + \lambda_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})})^2} = C_d,\end{aligned} \tag{17}$$

$$\frac{D_{d,1}\lambda_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})}}{\sigma_2^{(\mathrm{m})}\lambda_1^{(\mathrm{m})} + \sigma_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})} + \lambda_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})}} = D_d, \tag{18}$$

where $(\sigma_1^{(\mathrm{m})}, \sigma_2^{(\mathrm{m})}, \lambda_1^{(\mathrm{m})}, \lambda_2^{(\mathrm{m})})$ denotes the descriptor of the matched two-state MMPP, $A_d = E[T_{\mathrm{D},i}]$, $B_d = \mathrm{Var}[T_{\mathrm{D},i}]/E^2[T_{\mathrm{D},i}]$, $C_d = E[T^3_{\mathrm{D},i}]$, $C_{d,1} = E[T^2_{\mathrm{D},i}]$, $D_d = \mathrm{Cov}(T_{\mathrm{D},i}, T_{D,i+1})$, $D_{d,1} = (\mathrm{Var}[T_{\mathrm{D},i}] - E^2[T_{\mathrm{D},i}])/2$.

Ref. [6] has provided a procedure to determine $(\sigma_1^{(\mathrm{m})}, \sigma_2^{(\mathrm{m})}, \lambda_1^{(\mathrm{m})}, \lambda_2^{(\mathrm{m})})$ if (15)–(18) have a feasible solution set. The procedure is redescribed as follows:

- Introduce a new set of four variables $\alpha, \beta, \gamma, \delta$ as follows:

$$\alpha = \sigma_1^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})}, \quad \beta = \sigma_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})}\lambda_1^{(\mathrm{m})}, \quad \gamma = \lambda_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})}, \qquad \delta = \lambda_1^{(\mathrm{m})} + \lambda_2^{(\mathrm{m})}. \tag{19}$$

- Rewrite (15)–(18) in terms of $\alpha, \beta, \gamma, \delta$. After algebraic manipulations, we derive the following solutions:

$$\alpha = \frac{6(\zeta\xi - \eta)}{C_d\zeta(\zeta+\eta)^2 - 3C_{d,1}\zeta(\zeta+\eta) - 6(\xi+\xi^2)}, \quad \beta = \zeta\alpha, \quad \gamma = \eta\alpha, \quad \delta = \zeta + \xi\alpha \tag{20}$$

  with $\zeta = 1/A_d$, $\eta = D_d/[A_d(D_{d,1} - D_d)]$, $\xi = [(B_d - 1)(\zeta+\eta) + 2\eta]/(2\zeta)$.

- Solving (19), we further obtain

$$\lambda_1^{(\mathrm{m})} = \frac{\delta + \sqrt{\delta^2 - 4\gamma}}{2}, \qquad \lambda_2^{(\mathrm{m})} = \frac{\delta - \sqrt{\delta^2 - 4\gamma}}{2},$$
$$\sigma_1^{(\mathrm{m})} = \frac{\alpha\lambda_1^{(\mathrm{m})} - \beta}{\lambda_1^{(\mathrm{m})} - \lambda_2^{(\mathrm{m})}}, \qquad \sigma_2^{(\mathrm{m})} = \frac{\beta - \alpha\lambda_2^{(\mathrm{m})}}{\lambda_1^{(\mathrm{m})} - \lambda_2^{(\mathrm{m})}}. \tag{21}$$

  Here we assume $\lambda_1^{(\mathrm{m})} \geq \lambda_2^{(\mathrm{m})}$. For $\lambda_1^{(\mathrm{m})} < \lambda_2^{(\mathrm{m})}$, we switch the above solutions for $\lambda_1^{(\mathrm{m})}$ and $\lambda_2^{(\mathrm{m})}$.

When (15)–(18) do not have a feasible solution, the treatment developed in [6] for a discrete-time deterministic server is not satisfactory for the general server model treated in this paper and needs to be modified as follows.

If (15)–(18) do not have a feasible solution set due to the third moment failure (i.e., $\alpha < 0$), we set $\alpha = \sigma_{11} + \sigma_{21}$, and resolve (20) and (21) to obtain a feasible solution set. If (15)–(18) does not have a feasible solution set due to the conflict between $c^2(T_{\mathrm{D},i})$ and $\mathrm{Cov}(T_{\mathrm{D},i}, T_{\mathrm{D},i+1})$ (i.e., $\eta < 0$), we set $\lambda_1^{(\mathrm{m})} = \lambda_{11}$, $\lambda_2^{(\mathrm{m})} = \lambda_{21}$ and merely solve (15) and (16) via the first two moments. This can be treated as follows:

- Introduce a new set of two variables α, β as follows:

$$\alpha = \sigma_1^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})}, \qquad \beta = \sigma_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})} + \sigma_2^{(\mathrm{m})}\lambda_1^{(\mathrm{m})}. \tag{22}$$

- Rewrite (15) and (16) in terms of α and β. After algebraic manipulations, we have the following solutions:

$$\alpha = \frac{\beta}{\zeta}, \qquad \beta = \frac{2(\lambda_1^{(\mathrm{m})} - \zeta)(\zeta - \lambda_2^{(\mathrm{m})})}{B_d - 1} - \lambda_1^{(\mathrm{m})}\lambda_2^{(\mathrm{m})}. \tag{23}$$

- Solving (22), we further obtain

$$\sigma_1^{(\mathrm{m})} = \frac{\alpha\lambda_1^{(\mathrm{m})} - \beta}{\lambda_1^{(\mathrm{m})} - \lambda_2^{(\mathrm{m})}}, \qquad \sigma_2^{(\mathrm{m})} = \frac{\beta - \alpha\lambda_2^{(\mathrm{m})}}{\lambda_1^{(\mathrm{m})} - \lambda_2^{(\mathrm{m})}}. \tag{24}$$

If (15)–(18) do not have a feasible solution set because $c^2(T_{\mathrm{D},i}) < 1$, we then approximate the output process as a Poisson process by setting $\lambda_1^{(\mathrm{m})} = \lambda_2^{(\mathrm{m})} = \zeta$, $\sigma_1^{(\mathrm{m})} = \sigma_{11}$, $\sigma_2^{(\mathrm{m})} = \sigma_{21}$. This is because $c^2(T_{\mathrm{D},i}) > 1$ for two-state MMPPs, while $c^2(T_{\mathrm{D},i}) = 1$ for a Poisson process.

The detailed description of the moment matching algorithm is now summarized as follows:

Step 1. Input departure statistics of the tagged traffic (say $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21})$).

Step 2. If $c^2(T_{\mathrm{D},i}) < 1$ then use a Poisson process to emulate the departure process by setting $\lambda_1^{(\mathrm{m})} = \lambda_2^{(\mathrm{m})} = \zeta$, $\sigma_1^{(\mathrm{m})} = \sigma_{11}$, $\sigma_2^{(\mathrm{m})} = \sigma_{21}$ and go to Step 5; else go to the next step.

Step 3. If $c^2(T_{\mathrm{D},i})$ and $\mathrm{Cov}(T_{\mathrm{D},i}, T_{\mathrm{D},i+1})$ conflict, i.e., $\eta < 0$, then set $\lambda_1^{(\mathrm{m})} = \lambda_{11}$, $\lambda_2^{(\mathrm{m})} = \lambda_{21}$, use (23) and (24) to get $\sigma_1^{(\mathrm{m})}$ and $\sigma_2^{(\mathrm{m})}$, and go to Step 5; else go to the next step.

Step 4. If $\alpha > 0$, directly use (20) and (21) else set $\alpha = \sigma_{11} + \sigma_{21}$ then use (20) and (21) to get $(\sigma_1^{(\mathrm{m})}, \sigma_2^{(\mathrm{m})}, \lambda_1^{(\mathrm{m})}, \lambda_2^{(\mathrm{m})})$. Go to Step 5.

Step 5. Output the descriptor $(\sigma_1^{(\mathrm{m})}, \sigma_2^{(\mathrm{m})}, \lambda_1^{(\mathrm{m})}, \lambda_2^{(\mathrm{m})})$ as the parameters of the departure process.

The effectiveness and accuracy of the moment matching algorithm are examined through numerical experiments in Section 6.2.

5. Connection-wise end-to-end performance analysis

We first consider a tandem network because a reference connection in a network can be configured into such a topology. Since a superposition of MMPPs yields another MMPP, we can divide the traffic streams inbound to a node into two types without loss of generality, i.e., the tagged MMPP and the aggregation of other interfering/cross-MMPPs. This simplified tandem configuration of queues is shown in Fig. 2. Next, we apply Kleinrock's independence approximation to decompose a tandem network into a set of isolated queues. Since the source node receives a collection of MMPPs, it is clearly an H-MMPPs/G/1 queue. As for other nodes, the tagged departure process analysis together with moment matching to emulate the output process as a two-state MMPP make the H-MMPPs/G/1 queue still suitable to model the intermediate nodes. Therefore, a systematic and efficient method can be developed by recursively performing nodal performance analysis of the H-MMPPs/G/1 queue in [8] and the tagged departure process analysis, together with moment matching from the source node en route to the destination node along the reference connection. The connection-wise end-to-end evaluation algorithm is described as follows.

Step 1. Identify the traffic sources inbound to all the nodes along the reference connection. Remove unrelated traffic sources and nodes from the network. The remained reference connection can then be regarded as a tandem configuration of queues.

Step 2. Initialize the node number counter NC, i.e., set NC $= 1$.

Step 3. Perform nodal performance analysis for node NC using the results in [8] to gather the tagged performance measures.

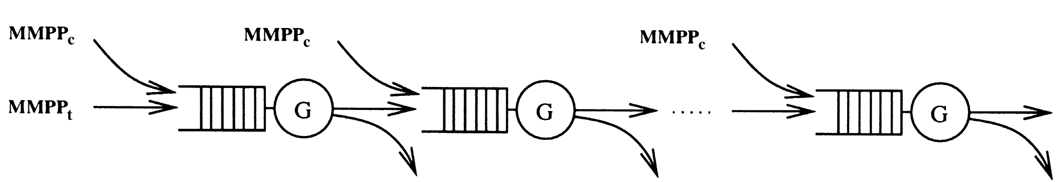


Fig. 2. Simplified tandem configuration of H-MMPPs/G/1 queues.

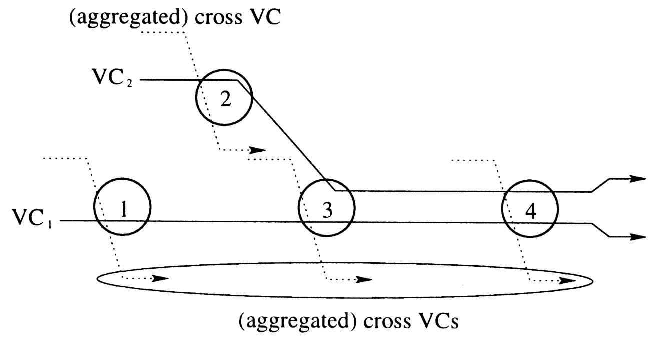


Fig. 3. Tree type network: Scenario-I.

Step 4. Apply the decomposition scheme (either Scheme I or Scheme II) to collect the tagged departure statistics.

Step 5. Use moment matching to emulate an output process as a two-state MMPP. Do this for each individual output process which is routed to node NC + 1.

Step 6. Increase NC by one, i.e., NC $\leftarrow$ NC + 1. If NC $\neq N$ (N represents the total number of nodes along the reference connection), go to Step 3; else redo Steps 3–5, then go to Step 7.

Step 7. Gather nodal performance measures from node 1 to N and output the connection-wise end-to-end performance measures.

We now examine whether the above algorithm works for a tree type network such as those depicted in Figs. 3 and 4. Applying Kleinrock's independence approximation to a tree type network, one can decompose a tree type network into several connections. For a reference connection, traffic streams inbound to a node may now include external sources and output processes of the upstream nodes as well. Output processes from other connections may also contribute to a node on the reference connection. This requires a few modifications when using the above algorithm. The modifications are described as follows:

- First, perform the above algorithm for other cross-connections to determine the output processes contributing to a node along the reference connection, then use moment matching to emulate these processes as two-state MMPPs.
- Next, treat these processes as external sources by ignoring the dependence among the traffic streams inbound to a node on the reference connection.
- Finally, perform again the above algorithm for the reference connection to obtain the connection-wise end-to-end performance measures.

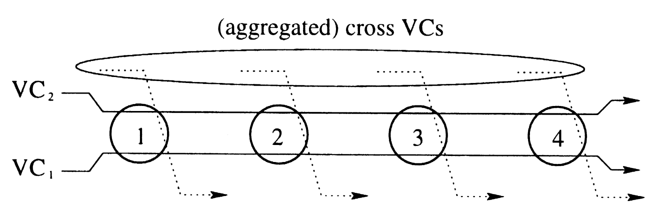


Fig. 4. Tree type network: Scenario-II.

6. Numerical examples and discussions

In this section, we first examine the effectiveness and accuracy of the decomposition schemes proposed to extract a tagged output process from the H-MMPPs/G/1 queue and moment matching scheme under the following different service distributions: deterministic (D), exponential (Exp), and k-stage Erlangian (E_k). Next, we apply the systematic approach proposed in Section 5 to estimate the connection-wise end-to-end sojourn delay for both tandem and tree type configurations. In these examples, computer simulations are provided (with 95% confidence interval) to validate the accuracy of approximation. In the following examples, for simplicity, we set the mean service time as $h = 1$.

6.1. The tagged departure statistics of an H-MMPPs/G/1 queue

This subsection shows the effectiveness and accuracy of the decomposition schemes. Consider the H-MMPPs/G/1 queue. As mentioned in Section 5, we may split traffic streams into the tagged MMPP and the aggregated interfering/cross-MMPP. For numerical simplicity, we assume both the tagged and aggregated MMPPs to be two-state MMPPs and consider two different examples: (I) the tagged traffic specified by $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.01, 0.12, 0.04)$ with traffic load $\rho_1 = 0.08$, and the cross-traffic specified by $\sigma_{12} = \sigma_{22} = 0.1$, $\lambda_{12}/\lambda_{22} = 2$; (II) the tagged traffic specified by $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.01, 0.36, 0.12)$ with traffic load $\rho_1 = 0.24$, and the same cross-traffic as (I). Varying the cross-traffic load, we obtain the tagged output statistics in Figs. 5(a)–(d) for (I) and Figs. 6(a)–(d) for (II) via characterizing the composite departure statistics using the decomposed $\mathrm{MMPP_t}/\mathrm{G_{eff}}/1$ model. From these figures, we note that (i) $E[T_{\mathrm{D},i}]$ is correctly characterized since flow conservation is not violated in decomposition; (ii) other statistics have good accuracy except at heavy traffic load (i.e., when the cross-traffic load goes beyond 0.8 for (I) and 0.6 for (II)); (iii) differences resulting from these two schemes are negligible except at heavy traffic load (In fact, Scheme II provides more accurate results than Scheme I at heavy traffic load.) These phenomena strongly indicate that the proposed decomposition schemes work well in practical situations.

According to the phenomena exhibited by the tagged departure statistics, we give a summary in the following based on numerical experiments. First, mean interdeparture time is flat due to flow conservation under fixed rate of the tagged input traffic. Second, $c^2(T_{\mathrm{D},i})$ and $E[T_{\mathrm{D},i}^3]$ decay slowly as the cross-traffic load increases, and both of them fall below the corresponding statistic of the input traffic. Third, $\mathrm{Cov}(T_{\mathrm{D},i}, T_{\mathrm{D},i+1})$ is always smaller than the input traffic for Exp, while for D and E_4 it is at first larger and then becomes smaller than the input traffic as the cross-traffic load increases.

6.2. Connection-wise end-to-end sojourn delay

In the following, we present analytical results using Schemes I and II. For clarity, we use Analysis-I and Analysis-II to distinguish the results of these two schemes.

We now examine the effectiveness and accuracy of the moment matching scheme by observing the tagged sojourn delay at the second node of a tandem queue. The queueing model employed is a two-node tandem queue with two heterogeneous MMPP inputs (Fig. 2) under different service assignments, for example, $(\mathrm{D}, \mathrm{E}_4)$ says that D is used by the first node while E_4 is employed by the second node. We consider two sets of traffic parameters: Set (I), the tagged MMPP at the first node specified by $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) =$

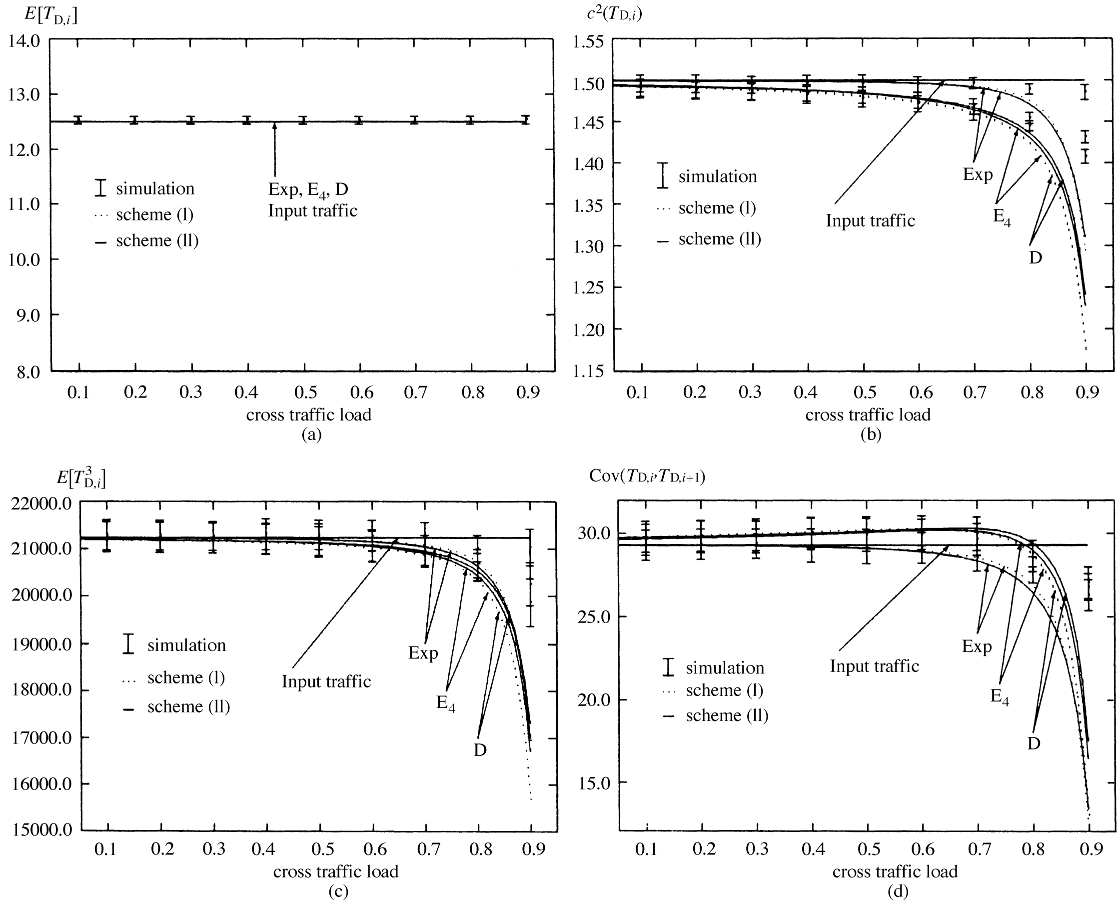


Fig. 5. Tagged departure statistics vs. cross-traffic load of an H-MMPPs/G/1 queue with two input sources: tagged traffic $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.01, 0.12, 0.04)$, cross-traffic $\sigma_{12} = \sigma_{22} = 0.1$, $\lambda_{12}/\lambda_{22} = 2$ under three types of servers: D, E_4, Exp with mean service time $h = 1$.

$(0.01, 0.01, 0.08, 0.04)$ with traffic load $\rho_1 = 0.06$; the cross-MMPP at the first node specified by $(\sigma_{12}, \sigma_{22}, \lambda_{12}, \lambda_{22}) = (0.02, 0.02, 4\rho_2/3, 2\rho_2/3)$ with traffic load ρ_2; and the cross-MMPP at the second node specified by $(\sigma_{13}, \sigma_{23}, \lambda_{13}, \lambda_{23}) = (0.01, 0.04, 5(0.1 + 2\rho_2)/9, 5(0.1 + 2\rho_2)/18)$ with traffic load $\rho_3 = \rho_2 + 0.05$. Set (II) is the same traffic parameters as (I) except the tagged traffic is now specified by $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.02, 1.8/7, 0.6/7)$ with traffic load $\rho_1 = 0.2$. Varying ρ_3 among 0.25, 0.45, and 0.65 (i.e., ρ_2 among 0.2, 0.4, and 0.6), we get the sojourn delay at the second node for the tagged MMPP in Tables 1–3.

For parameter set (I), the approximate analysis results show very good accuracy when compared with simulation results. The error is at most 2.2%. Analysis-I and Analysis-II show little difference in these estimates. But in the case of parameter set (II), the worst case error reaches 8.2% for both Analysis-I and Analysis-II with one exception; approximately 17.3% error for Analysis-II at heavy traffic load (i.e., $\rho_1 = 0.2$ and $\rho_3 = 0.65$). This ambiguous error for Analysis-II results largely from the setting of the

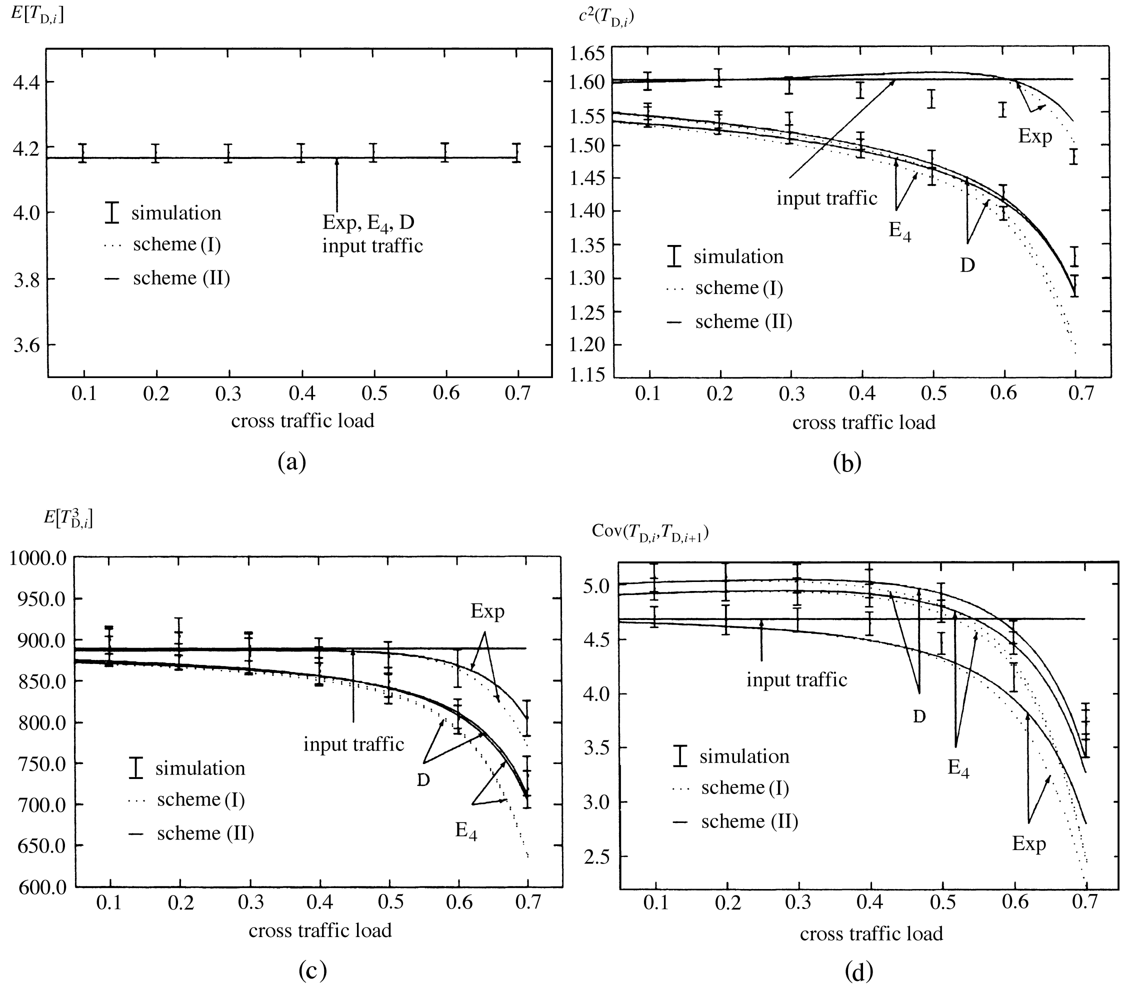


Fig. 6. Tagged departure statistics vs. cross-traffic load of an H-MMPPs/G/1 queue with two input sources: tagged traffic $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.01, 0.36, 0.12)$, cross-traffic $\sigma_{12} = \sigma_{22} = 0.1$, $\lambda_{12}/\lambda_{22} = 2$ under three types of servers: D, E_4, Exp with mean service time $h = 1$.

stoppage parameter ϵ. In order to have a quick calculation, we set a larger value to ϵ which degrades the accuracy.

We now investigate the connection-wise end-to-end performance analysis for both tandem and tree type configurations, first considering the four-node tandem configuration shown in Fig. 2. Assume that all external cross-traffic streams are identical and mutually independent, and servers are homogeneous at different nodes. Using the input traffic specified in (I) of Section 6.1, we obtain sojourn delay at each node for three different service mechanisms D, Exp, and E_4 in Tables 4–6. These tables reflect that the error is at most 4%, In networks, such as ATM, it is possible for different VCs to partly or entirely share the same route. Figs. 3 and 4 give two simple scenarios of tree type topology in which two tagged VCs

Table 1
Delay at the second node of a two-node tandem queue with (D, D), (D, E_4), and (D, Exp) service assignments

ρ_c		Tagged traffic (0.01, 0.01, 0.08, 0.04)			Tagged traffic (0.01, 0.02, 1.8/7, 0.6/7)		
		0.25	0.45	0.65	0.25	0.45	0.65
D	Simulation	1.209±0.2%	1.542 ± 0.3%	2.476 ± 0.6%	1.352 ± 0.5%	2.064 ± 0.6%	5.782 ± 2.4%
	Analysis-I	1.2355	1.5650	2.4928	1.4636	2.1580	5.3061
	Analysis-II	1.2356	1.5654	2.4935	1.4649	2.0010	4.7867
E_4	Simulation	1.273 ± 0.3%	1.688 ± 0.6%	2.826 ± 0.9%	1.500 ± 0.7%	2.365 ± 1.1%	6.691 ± 2.8%
	Analysis-I	1.2943	1.7050	2.8481	1.5791	2.4292	6.1446
	Analysis-II	1.2944	1.7054	2.8489	1.5807	2.2472	5.6064
Exp	Simulation	1.451 ± 0.6%	2.113 ± 0.8%	3.854 ± 1.3%	1.880 ± 0.9%	3.234 ± 1.5%	9.215 ± 3.4%
	Analysis-I	1.4702	2.1225	3.8900	1.9251	3.2210	8.5529
	Analysis-II	1.4704	2.1231	3.8907	1.9273	2.9797	7.9741

Table 2
Delay at the second node of a two-node tandem queue with (E_4, D), (E_4, E_4), and (E_4, Exp) service assignments

ρ_c		Tagged traffic (0.01, 0.01, 0.08, 0.04)			Tagged traffic (0.01, 0.02, 1.8/7, 0.6/7)		
		0.25	0.45	0.65	0.25	0.45	0.65
D	Simulation	1.221 ± 0.1%	1.549 ± 0.3%	2.481 ± 0.7%	1.402 ± 0.5%	2.101 ± 0.7%	5.750 ± 3.1%
	Analysis-I	1.2356	1.5652	2.4931	1.4665	2.1608	5.3293
	Analysis-II	1.2357	1.5655	2.4932	1.4672	2.0010	4.7867
E_4	Simulation	1.281 ± 0.3%	1.693 ± 0.5%	2.841 ± 0.8%	1.530 ± 0.9%	2.400 ± 1.6%	6.723 ± 2.0%
	Analysis-I	1.2944	1.7052	2.8485	1.5825	2.4329	6.1694
	Analysis-II	1.2945	1.7055	2.8486	1.5835	2.2472	5.6064
Exp	Simulation	1.461 ± 0.6%	2.108 ± 0.4%	3.907 ± 1.0%	1.900 ± 1.2%	3.249 ± 2.3%	9.272 ± 3.8%
	Analysis-I	1.4704	2.1228	3.8902	1.9296	3.2276	8.5813
	Analysis-II	1.4705	2.1232	3.8903	1.9309	2.9797	7.9741

Table 3
Delay at the second node of a two-node tandem queue with (Exp, D), (Exp, E_4), and (Exp, Exp) service assignments

ρ_c		Tagged traffic (0.01, 0.01, 0.08, 0.04)			Tagged traffic (0.01, 0.02, 1.8/7, 0.6/7)		
		0.25	0.45	0.65	0.25	0.45	0.65
D	Simulation	1.237 ± 0.2%	1.567 ± 0.3%	2.500 ± 0.6%	1.480 ± 0.4%	2.189 ± 0.8%	5.879 ± 2.6%
	Analysis-I	1.2362	1.5670	2.4995	1.4827	2.1871	5.6749
	Analysis-II	1.2362	1.5670	2.4997	1.4827	2.1882	5.6414
E_4	Simulation	1.297 ± 0.3%	1.709 ± 0.6%	2.850 ± 0.8%	1.597 ± 0.9%	2.464 ± 0.5%	6.815 ± 3.2%
	Analysis-I	1.2951	1.7073	2.8549	1.6016	2.4670	6.5374
	Analysis-II	1.2951	1.7073	2.8551	1.6016	2.4682	6.4983
Exp	Simulation	1.476 ± 0.4%	2.121 ± 0.7%	3.877 ± 0.5%	1.940 ± 0.9%	3.302 ± 1.9%	9.287 ± 3.2%
	Analysis-I	1.4715	2.1256	3.8955	1.9548	3.2846	8.9996
	Analysis-II	1.4715	2.1256	3.8956	1.9548	3.2854	8.9480

Table 4
Delay at each node of the four-node tandem configuration with (D, D, D, D) service assignment

ρ_c		Node 1	Node 2	Node 3	Node 4	Total
0.2	Simulation	$1.223 \pm 0.3\%$	$1.173 \pm 0.2\%$	$1.172 \pm 0.2\%$	$1.167 \pm 0.2\%$	$4.735 \pm 0.1\%$
	Analysis-I	1.2209	1.2195	1.2183	1.2171	4.8758
	Analysis-II	1.2209	1.2197	1.2186	1.2175	4.8767
0.4	Simulation	$1.550 \pm 0.5\%$	$1.511 \pm 0.3\%$	$1.502 \pm 0.4\%$	$1.493 \pm 0.5\%$	$6.056 \pm 0.2\%$
	Analysis-I	1.5513	1.5473	1.5437	1.5405	6.1829
	Analysis-II	1.5513	1.5481	1.5451	1.5423	6.1869
0.6	Simulation	$2.486 \pm 0.7\%$	$2.456 \pm 1.0\%$	$2.440 \pm 1.0\%$	$2.408 \pm 0.7\%$	$9.790 \pm 0.5\%$
	Analysis-I	2.4808	2.4663	2.4538	2.4429	9.8438
	Analysis-II	2.4808	2.4691	2.4587	2.4492	9.8578
0.8	Simulation	$7.694 \pm 3.7\%$	$7.653 \pm 2.9\%$	$7.724 \pm 4.4\%$	$7.619 \pm 2.1\%$	$30.691 \pm 2.0\%$
	Analysis-I	7.6556	7.6640	7.4537	7.5040	30.2773
	Analysis-II	7.6556	7.6849	7.8777	6.9516	30.1697

Table 5
Delay at each node of the four-node tandem configuration with (E_4, E_4, E_4, E_4) service assignment

ρ_c		Node 1	Node 2	Node 3	Node 4	Total
0.2	Simulation	$1.277 \pm 0.5\%$	$1.251 \pm 0.4\%$	$1.241 \pm 0.4\%$	$1.238 \pm 0.2\%$	$5.007 \pm 0.2\%$
	Analysis-I	1.2755	1.2741	1.2728	1.2716	5.0941
	Analysis-II	1.2755	1.2743	1.2731	1.2719	5.0948
0.4	Simulation	$1.690 \pm 0.6\%$	$1.658 \pm 0.5\%$	$1.648 \pm 0.5\%$	$1.653 \pm 0.7\%$	$6.649 \pm 0.2\%$
	Analysis-I	1.6827	1.6785	1.6746	1.6711	6.7069
	Analysis-II	1.6827	1.6790	1.6756	1.6725	6.7098
0.6	Simulation	$2.797 \pm 1.0\%$	$2.768 \pm 1.0\%$	$2.771 \pm 1.2\%$	$2.745 \pm 1.0\%$	$11.081 \pm 0.6\%$
	Analysis-I	2.7897	2.7752	2.7626	2.7515	11.0790
	Analysis-II	2.7897	2.7768	2.7654	2.7551	11.0871
0.8	Simulation	$8.675 \pm 3.0\%$	$8.698 \pm 2.4\%$	$8.834 \pm 4.0\%$	$8.666 \pm 2.3\%$	$34.872 \pm 2.0\%$
	Analysis-I	8.6603	8.6799	8.9099	8.5453	34.7953
	Analysis-II	8.6603	8.6904	8.9132	7.9156	34.1794

Table 6
Delay at each node of the four-node tandem configuration with (Exp, Exp, Exp, Exp) service assignment

ρ_c		Node 1	Node 2	Node 3	Node 4	Total
0.2	Simulation	$1.440 \pm 0.4\%$	$1.443 \pm 0.6\%$	$1.437 \pm 0.6\%$	$1.437 \pm 0.5\%$	$5.756 \pm 0.3\%$
	Analysis-I	1.4385	1.4383	1.4382	1.4380	5.7530
	Analysis-II	1.4385	1.4383	1.4382	1.4380	5.7530
0.4	Simulation	$2.069 \pm 1.2\%$	$2.077 \pm 1.6\%$	$2.079 \pm 1.6\%$	$2.075 \pm 0.5\%$	$8.299 \pm 0.2\%$
	Analysis-I	2.0703	2.0699	2.0695	2.0691	8.2787
	Analysis-II	2.0703	2.0699	2.0695	2.0691	8.2787
0.6	Simulation	$3.700 \pm 0.7\%$	$3.741 \pm 1.0\%$	$3.705 \pm 1.0\%$	$3.705 \pm 1.9\%$	$14.851 \pm 0.5\%$
	Analysis-I	3.6887	3.6865	3.6844	3.6822	14.7418
	Analysis-II	3.6887	3.6863	3.6839	3.6814	14.7402
0.8	Simulation	$11.846 \pm 4.7\%$	$11.539 \pm 4.0\%$	$11.524 \pm 3.2\%$	$11.892 \pm 2.8\%$	$46.803 \pm 1.9\%$
	Analysis-I	11.6057	11.5118	11.4303	11.3594	45.9072
	Analysis-II	11.6057	11.4966	11.4031	11.3227	45.8281

Table 7
Traffic parameters employed in the tree type network of Fig. 3 (Scenario-I)

	Node 1	Node 2	Node 3	Node 4
Parameter set I				
VC_1 ($\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}$)	(0.01,0.01,0.08,0.04)	–	–	–
VC_2 ($\sigma_{12}, \sigma_{22}, \lambda_{12}, \lambda_{22}$)	–	(0.01,0.04,0.15,0)	–	–
Cross-VCs	(0.02,0.02,0.8/3,0.4/3)	(0.01,0.02,4.5/14,1.5/14)	(0.01,0.01,0.4,0.2)	(0.01,0.04,5/18,2.5/18)
Parameter set II				
VC_1 ($\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}$)	(0.01,0.01,0.08,0.04)	–	–	–
VC_2 ($\sigma_{12}, \sigma_{22}, \lambda_{12}, \lambda_{22}$)	–	(0.01,0.04,0.15,0)	–	–
Cross-VCs	(0.02,0.02,1.6/3,0.8/3)	(0.01,0.02,8.1/14,2.7/14)	(0.01,0.01,2/3,1/3)	(0.01,0.04,0.5,0.25)
Parameter set III				
VC_1 ($\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}$)	(0.01,0.01,0.08,0.04)	–	–	–
VC_2 ($\sigma_{12}, \sigma_{22}, \lambda_{12}, \lambda_{22}$)	–	(0.01,0.04,0.15, 0)	–	–
Cross-VCs	(0.02,0.02,0.8,0.4)	(0.01,0.02,11.7/14,3.9/14)	(0.01,0.01,2.8/3,1.4/3)	(0.01,0.04,13/18,6.5/18)

partially share the same route in Fig. 3 while two tagged VCs share entirely the same route in Fig. 4. For convenience, the scenarios in Figs. 3 and 4 are referred as Scenario-I and Scenario-II, respectively. Using the traffic parameters specified in Table 7 for Scenario-I under (D, D, D, D), (E_4, E_4, E_4, E_4), (Exp, Exp, Exp, Exp), and (D, Exp, E_4, D) service assignments, we obtain the sojourn delay at each node for VC_1 and VC_2 in Tables 8–11. Also applying the traffic parameters given in Table 12 for Scenario-II under these service assignments, we obtain the sojourn delay at each node for VC_1 and VC_2 in Tables 13–16.

Table 8
Delay at each node for each connection in Scenario-I with (D, D, D, D) service assignment

Sets			Node 1	Node 2	Node 3	Node 4	Total
I	VC_1	Simulation	1.187 ± 0.1%	–	1.488 ± 0.4%	1.299 ± 0.3%	3.974 ± 0.2%
		Analysis-I	1.1876	–	1.5213	1.4000	4.1090
		Analysis-II	1.1876	–	1.5216	1.4006	4.1098
	VC_2	Simulation	–	1.356 ± 0.5%	1.485 ± 0.8%	1.307 ± 0.4%	4.148 ± 0.4%
		Analysis-I	–	1.3514	1.5535	1.4250	4.3298
		Analysis-II	–	1.3514	1.5539	1.4259	4.3313
II	VC_1	Simulation	1.500 ± 0.4%	–	2.613 ± 1.5%	1.853 ± 0.3%	5.966 ± 0.6%
		Analysis-I	1.4947	–	2.6349	1.9581	6.0877
		Analysis-II	1.4947	–	2.6363	1.9598	6.0908
	VC_2	Simulation	–	1.939 ± 0.6%	2.666 ± 1.5%	1.879 ± 0.4%	6.484 ± 0.7%
		Analysis-I	–	1.9310	2.7082	2.0046	6.6439
		Analysis-II	–	1.9310	2.7106	2.0074	6.6490
III	VC_1	Simulation	2.541 ± 0.8%	–	18.224 ± 5.5%	4.487 ± 1.0%	25.254 ± 4.0%
		Analysis-I	2.5242	–	18.2982	4.4035	25.2259
		Analysis-II	2.5242	–	18.2754	4.2795	25.0791
	VC_2	Simulation	–	5.375 ± 2.8%	18.553 ± 6.4%	4.577 ± 2.4%	28.505 ± 4.4%
		Analysis-I	–	5.3450	18.3932	4.4887	28.2269
		Analysis-II	–	5.3450	18.3727	4.3994	28.1171

Table 9
Delay at each node for each connection in Scenario-I with (E_4, E_4, E_4, E_4) service assignment

Sets			Node 1	Node 2	Node 3	Node 4	Total
I	VC_1	Simulation	$1.236 \pm 0.3\%$	–	$1.631 \pm 0.6\%$	$1.443 \pm 0.5\%$	$4.310 \pm 0.4\%$
		Analysis-I	1.2344	–	1.6505	1.5000	4.3849
		Analysis-II	1.2344	–	1.6507	1.5005	4.3856
	VC_2	Simulation	–	$1.442 \pm 0.6\%$	$1.644 \pm 0.9\%$	$1.455 \pm 0.6\%$	$4.541 \pm 0.4\%$
		Analysis-I	–	1.4379	1.6894	1.5307	4.6580
		Analysis-II	–	1.4379	1.6898	1.5314	4.6590
II	VC_1	Simulation	$1.621 \pm 0.6\%$	–	$2.980 \pm 1.5\%$	$2.134 \pm 0.5\%$	$6.735 \pm 0.7\%$
		Analysis-I	1.6162	–	2.9993	2.1934	6.8089
		Analysis-II	1.6162	–	3.0000	2.1940	6.8102
	VC_2	Simulation	–	$2.142 \pm 0.6\%$	$3.057 \pm 1.7\%$	$2.175 \pm 0.7\%$	$7.374 \pm 0.8\%$
		Analysis-I	–	2.1507	3.0824	2.2487	7.4817
		Analysis-II	–	2.1507	3.0834	2.2494	7.4835
III	VC_1	Simulation	$2.852 \pm 1.4\%$	–	$19.437 \pm 5.0\%$	$5.250 \pm 1.2\%$	$27.538 \pm 3.6\%$
		Analysis-I	2.8514	–	19.4908	5.1297	27.4718
		Analysis-II	2.8514	–	19.4748	4.9907	27.3168
	VC_2	Simulation	–	$6.101 \pm 3.0\%$	$19.556 \pm 4.5\%$	$5.417 \pm 2.7\%$	$31.075 \pm 2.7\%$
		Analysis-I	–	6.0495	19.5975	5.2259	30.8729
		Analysis-II	–	6.0495	19.5696	5.1200	30.7391

Table 10
Delay at each node for each connection in Scenario-I with (Exp, Exp, Exp, Exp) service assignment

Sets			Node 1	Node 2	Node 3	Node 4	Total
I	VC_1	Simulation	$1.377 \pm 0.6\%$	–	$2.044 \pm 0.9\%$	$1.813 \pm 0.8\%$	$5.234 \pm 0.6\%$
		Analysis-I	1.3743	–	2.0368	1.8027	5.2139
		Analysis-II	1.3743	–	2.0368	1.8028	5.2139
	VC_2	Simulation	–	$1.701 \pm 1.1\%$	$2.092 \pm 0.7\%$	$1.842 \pm 0.9\%$	$5.635 \pm 0.5\%$
		Analysis-I	–	1.6947	2.0947	1.8506	5.6400
		Analysis-II	–	1.6947	2.0947	1.8506	5.6399
II	VC_1	Simulation	$1.983 \pm 1.1\%$	–	$4.054 \pm 1.8\%$	$2.904 \pm 1.0\%$	$8.941 \pm 0.6\%$
		Analysis-I	1.9764	–	4.0480	2.9037	8.9281
		Analysis-II	1.9764	–	4.0478	2.9035	8.9277
	VC_2	Simulation	–	$2.792 \pm 1.3\%$	$4.130 \pm 2.3\%$	$2.994 \pm 1.5\%$	$9.916 \pm 0.8\%$
		Analysis-I	–	2.7903	4.1587	2.9864	9.9354
		Analysis-II	–	2.7903	4.1581	2.9847	9.9331
III	VC_1	Simulation	$3.773 \pm 1.1\%$	–	$23.174 \pm 5.8\%$	$7.587 \pm 2.0\%$	$34.533 \pm 4.0\%$
		Analysis-I	3.7821	–	22.8746	7.2835	33.9402
		Analysis-II	3.7821	–	22.8538	7.2244	33.8604
	VC_2	Simulation	–	$8.031 \pm 3.9\%$	$23.548 \pm 7.3\%$	$7.648 \pm 1.8\%$	$39.229 \pm 4.8\%$
		Analysis-I	–	7.9724	23.0699	7.4199	38.4622
		Analysis-II	–	7.9724	23.0382	7.3356	38.3461

Table 11
Delay at each node for each connection in Scenario-I with (D, Exp, E_4, D) service assignment

Sets			Node 1	Node 2	Node 3	Node 4	Total
I	VC_1	Simulation	$1.187 \pm 0.1\%$	–	$1.636 \pm 0.7\%$	$1.336 \pm 0.3\%$	$4.160 \pm 0.4\%$
		Analysis-I	1.1876	–	1.6508	1.4006	4.2390
		Analysis-II	1.1876	–	1.6510	1.4009	4.2396
	VC_2	Simulation	–	$1.696 \pm 1.0\%$	$1.684 \pm 0.9\%$	$1.359 \pm 0.2\%$	$4.738 \pm 0.6\%$
		Analysis-I	–	1.6947	1.6920	1.4272	4.8139
		Analysis-II	–	1.6947	1.6920	1.4276	4.8143
II	VC_1	Simulation	$1.500 \pm 0.4\%$	–	$2.984 \pm 1.9\%$	$1.892 \pm 0.3\%$	$6.377 \pm 0.9\%$
		Analysis-I	1.4947	–	3.0026	1.9601	6.4574
		Analysis-II	1.4947	–	3.0029	1.9606	6.4583
	VC_2	Simulation	–	$2.808 \pm 1.2\%$	$3.058 \pm 1.2\%$	$1.939 \pm 0.6\%$	$7.805 \pm 0.7\%$
		Analysis-I	–	2.7903	3.0924	2.0112	7.8939
		Analysis-II	–	2.7903	3.0920	2.0111	7.8934
III	VC_1	Simulation	$2.537 \pm 0.9\%$	–	$19.273 \pm 4.4\%$	$4.518 \pm 1.7\%$	$26.328 \pm 3.3\%$
		Analysis-I	2.5242	–	19.5233	4.4146	26.4620
		Analysis-II	2.5242	–	19.4928	4.2819	26.2989
	VC_2	Simulation	–	$8.061 \pm 2.4\%$	$19.384 \pm 7.3\%$	$4.622 \pm 1.6\%$	$32.066 \pm 4.8\%$
		Analysis-I	–	7.9724	19.6370	4.5112	32.1206
		Analysis-II	–	7.9724	19.6027	4.4046	31.9797

From these tables, approximately 7.7% and 9% of the worst case error are observed for Scenario-I and Scenario-II, respectively.

Through these numerical experiments, we observe that the systematic connection-wise end-to-end performance analysis algorithm performs satisfactorily.

Table 12
Traffic parameters employed in the tree type network of Fig. 4 (Scenario-II)

	Node 1	Node 2	Node 3	Node 4
Parameter set I				
VC_1 $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21})$	(0.01,0.01,0.08,0.04)	–	–	–
VC_2 $(\sigma_{12}, \sigma_{22}, \lambda_{12}, \lambda_{22})$	(0.01,0.04,0.15,0)	–	–	–
Cross-VCs	(0.02,0.02,0.8/3,0.4/3)	(0.01,0.02,4.5/14,1.5/14)	(0.01,0.01,0.4,0.2)	(0.01,0.04,5/18,2.5/18)
Parameter set II				
VC_1 $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21})$	(0.01,0.01,0.08,0.04)	–	–	–
VC_2 $(\sigma_{12}, \sigma_{22}, \lambda_{12}, \lambda_{22})$	(0.01,0.04,0.15,0)	–	–	–
Cross-VCs	(0.02,0.02,1.6/3,0.8/3)	(0.01,0.02,8.1/14,2.7/14)	(0.01,0.01,2/3,1/3)	(0.01,0.04,0.5,0.25)
Parameter set III				
VC_1 $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21})$	(0.01,0.01,0.08,0.04)	–	–	–
VC_2 $(\sigma_{12}, \sigma_{22}, \lambda_{12}, \lambda_{22})$	(0.01,0.04,0.15,0)	–	–	–
Cross-VCs	(0.02,0.02,0.8,0.4)	(0.01,0.02,11.7/14,3.9/14)	(0.01,0.01,2.8/3,1.4/3)	(0.01,0.04,13/18,6.5/18)

Table 13
Delay at each node for each connection in Scenario-II with (D, D, D, D) service assignment

Sets			Node 1	Node 2	Node 3	Node 4	Total
I	VC_1	Simulation	1.332 ± 0.3%	1.329 ± 0.3%	1.420 ± 0.5%	1.283 ± 0.4%	5.364 ± 0.2%
		Analysis-I	1.3316	1.4202	1.5186	1.3982	5.6687
		Analysis-II	1.3316	1.4205	1.5194	1.3991	5.6707
	VC_2	Simulation	1.358 ± 0.4%	1.343 ± 0.5%	1.435 ± 0.6%	1.297 ± 0.3%	5.432 ± 0.4%
		Analysis-I	1.3560	1.4478	1.5494	1.4220	5.7752
		Analysis-II	1.3560	1.4482	1.5506	1.4234	5.7783
II	VC_2	Simulation	1.859 ± 0.7%	2.095 ± 0.7%	2.521 ± 1.5%	1.815 ± 0.5%	8.289 ± 0.4%
		Analysis-I	1.8511	2.1779	2.6243	1.9547	8.6079
		Analysis-II	1.8511	2.1795	2.6275	1.9354	8.5935
	VC_2	Simulation	1.906 ± 0.6%	2.139 ± 1.2%	2.584 ± 1.0%	1.851 ± 0.8%	8.480 ± 0.6%
		Analysis-I	1.9008	2.2383	2.6908	1.9953	8.8252
		Analysis-II	1.9008	2.2407	2.6962	1.9959	8.8336
III	VC_1	Simulation	4.604 ± 2.5%	8.903 ± 2.4%	17.971 ± 5.5%	4.400 ± 1.5%	35.879 ± 2.9%
		Analysis-I	4.5584	8.7728	18.0957	4.3547	35.7816
		Analysis-II	4.5584	8.7738	17.8283	4.2614	35.4219
	VC_2	Simulation	4.765 ± 1.9%	8.630 ± 3.3%	18.407 ± 6.3%	4.521 ± 2.1%	36.322 ± 3.3%
		Analysis-I	4.6843	8.8619	18.2501	4.4343	36.2305
		Analysis-II	4.6843	8.8621	18.1217	4.3591	36.0272

Table 14
Delay at each node for each connection in Scenario-II with (E_4, E_4, E_4, E_4) service assignment

Sets			Node 1	Node 2	Node 3	Node 4	Total
I	VC_1	Simulation	1.415 ± 0.5%	1.482 ± 0.6%	1.587 ± 0.9%	1.421 ± 0.4%	5.906 ± 0.4%
		Analysis-I	1.4140	1.5246	1.6476	1.4980	6.0842
		Analysis-II	1.4140	1.5248	1.6482	1.4987	6.0858
	VC_2	Simulation	1.444 ± 0.6%	1.500 ± 0.8%	1.608 ± 0.4%	1.441 ± 0.5%	5.993 ± 0.4%
		Analysis-I	1.4438	1.5582	1.6851	1.5274	6.2144
		Analysis-II	1.4438	1.5585	1.6859	1.5285	6.2167
II	VC_2	Simulation	2.062 ± 0.5%	2.413 ± 1.1%	2.928 ± 1.9%	2.092 ± 0.5%	9.496 ± 0.7%
		Analysis-I	2.0542	2.4554	2.9895	2.1894	9.6885
		Analysis-II	2.0542	2.4563	2.9910	2.1644	9.6658
	VC_2	Simulation	2.123 ± 1.0%	2.467 ± 1.1%	3.015 ± 1.4%	2.130 ± 0.9%	9.734 ± 0.6%
		Analysis-I	2.1127	2.5252	3.0648	2.2383	9.9411
		Analysis-II	2.1127	2.5266	3.0673	2.2358	9.9425
III	VC_1	Simulation	5.169 ± 1.9%	9.955 ± 2.4%	19.389 ± 4.9%	5.241 ± 2.1%	39.755 ± 2.6%
		Analysis-I	5.1468	9.7282	19.2994	5.0783	39.2527
		Analysis-II	5.1468	9.7313	19.0171	4.9713	38.8665
	VC_2	Simulation	5.332 ± 1.6%	9.683 ± 2.9%	19.578 ± 5.7%	5.342 ± 1.7%	39.934 ± 2.6%
		Analysis-I	5.2816	9.8254	19.4582	5.1696	39.7348
		Analysis-II	5.2816	9.8190	19.3079	5.0762	39.4846

Table 15
[]Delay at each node for each connection in Scenario-II with (Exp, Exp, Exp, Exp) service assignment

Sets			Node 1	Node 2	Node 3	Node 4	Total
I	VC_1	Simulation	$1.657 \pm 0.7\%$	$1.845 \pm 1.0\%$	$2.039 \pm 1.1\%$	$1.808 \pm 0.6\%$	$7.350 \pm 0.5\%$
		Analysis-I	1.6600	1.8377	2.0365	1.8025	7.3367
		Analysis-II	1.6600	1.8377	2.0365	1.8025	7.3368
	VC_2	Simulation	$1.715 \pm 1.0\%$	$1.894 \pm 0.8\%$	$2.094 \pm 0.8\%$	$1.840 \pm 1.4\%$	$7.543 \pm 0.5\%$
		Analysis-I	1.7045	1.8881	2.0930	1.8493	7.5349
		Analysis-II	1.7045	1.8881	2.0930	1.8492	7.5348
II	VC_2	Simulation	$2.641 \pm 1.1\%$	$3.289 \pm 1.4\%$	$4.088 \pm 2.4\%$	$2.924 \pm 0.9\%$	$12.942 \pm 1.0\%$
		Analysis-I	2.6505	3.2682	4.0456	2.9026	12.8668
		Analysis-II	2.6505	3.2681	4.0451	2.9024	12.8661
	VC_2	Simulation	$2.724 \pm 1.2\%$	$3.345 \pm 1.6\%$	$4.154 \pm 2.2\%$	$2.974 \pm 1.5\%$	$13.196 \pm 0.8\%$
		Analysis-I	2.7308	3.3632	4.1524	2.9820	13.2284
		Analysis-II	2.7308	3.3630	4.1511	2.9801	13.2249
III	VC_1	Simulation	$6.943 \pm 2.0\%$	$12.680 \pm 2.7\%$	$22.616 \pm 4.7\%$	$7.538 \pm 2.1\%$	$49.777 \pm 2.3\%$
		Analysis-I	6.8057	12.3067	22.8181	7.2633	49.1939
		Analysis-II	6.8057	12.2925	22.7407	7.1919	49.0307
	VC_2	Simulation	$6.916 \pm 3.0\%$	$12.540 \pm 3.9\%$	$23.401 \pm 5.3\%$	$7.771 \pm 2.6\%$	$50.628 \pm 2.5\%$
		Analysis-I	6.9607	12.4864	23.0084	7.3917	49.8471
		Analysis-II	6.9607	12.4707	22.9262	7.2987	49.6563

Table 16
Delay at each node for each connection in Scenario-II with (D, Exp, E_4, D) service assignment

Sets			Node 1	Node 2	Node 3	Node 4	Total
I	VC_1	Simulation	$1.332 \pm 0.3\%$	$1.783 \pm 0.7\%$	$1.624 \pm 0.6\%$	$1.329 \pm 0.3\%$	$6.069 \pm 0.3\%$
		Analysis-I	1.3316	1.8349	1.6496	1.3999	6.2160
		Analysis-II	1.3316	1.8354	1.6501	1.4004	6.2175
	VC_2	Simulation	$1.358 \pm 0.4\%$	$1.831 \pm 0.9\%$	$1.659 \pm 1.0\%$	$1.346 \pm 0.4\%$	$6.194 \pm 0.5\%$
		Analysis-I	1.3560	1.8842	1.6878	1.4245	6.3525
		Analysis-II	1.3560	1.8850	1.6884	1.4253	6.3547
II	VC_2	Simulation	$1.858 \pm 0.7\%$	$3.227 \pm 1.3\%$	$2.957 \pm 1.7\%$	$1.880 \pm 0.5\%$	$9.923 \pm 0.8\%$
		Analysis-I	1.8511	3.2605	2.9962	1.9572	10.0650
		Analysis-II	1.8511	3.2624	2.9979	1.9584	10.0698
	VC_2	Simulation	$1.908 \pm 0.6\%$	$3.278 \pm 1.4\%$	$3.068 \pm 1.0\%$	$1.909 \pm 0.7\%$	$10.163 \pm 0.5\%$
		Analysis-I	1.9008	3.3514	3.0763	2.0026	10.3311
		Analysis-II	1.9008	3.3546	3.0787	2.0040	10.3381
III	VC_1	Simulation	$4.597 \pm 2.3\%$	$12.532 \pm 1.8\%$	$19.259 \pm 4.1\%$	$4.554 \pm 1.8\%$	$40.942 \pm 2.5\%$
		Analysis-I	4.5584	12.3837	19.4701	4.3938	40.8060
		Analysis-II	4.5584	12.3850	19.3690	4.2677	40.5801
	VC_2	Simulation	$4.770 \pm 1.9\%$	$12.194 \pm 3.5\%$	$19.395 \pm 7.5\%$	$4.667 \pm 2.1\%$	$41.025 \pm 3.8\%$
		Analysis-I	4.6843	12.4766	19.5551	4.4713	41.1873
		Analysis-II	4.6843	12.4769	19.4834	4.3731	41.0177

7. Conclusions

Due to the failure of Poisson processes to capture the statistical characteristics of real traffic in both contemporary and future communication networks, the connection-wise end-to-end performance analysis needs major revision. Unfortunately, correlated arrival models such as the MMPP make the connection-wise end-to-end analysis intractable. Instead of pursuing an exact analysis, we propose an approximate but systematic and efficient method for the connection-wise end-to-end performance evaluation in a network of queues receiving heterogeneous MMPPs. The pivotal points of this method are the tagged departure process analysis and the moment matching scheme/algorithm to approximate an output process as a two-state MMPP. Analysis of a tandem configuration of queues becomes simple and efficient. For tree type networks, applying Kleinrock's independence approximation together with these two techniques enable us to decompose the network into several connections of tandem configuration, so the proposed methodology is also applicable. Through comprehensive numerical experiments, we show that this method provides good accuracy. This method has potential applications to packet-switched high-speed networks, especially for connection-wise end-to-end performance of a VC in ATM networks.

The current work can be extended to systems with priority service discipline, and it can also be extended to incorporate server vacations.

Appendix A. Relation between $\boldsymbol{A}_0(0)$ and $\boldsymbol{A}_0'(0)$

In this Appendix, we establish the relation between $\boldsymbol{A}_0(0)$ and $\boldsymbol{A}_0'(0)$ under a number of service distributions for the model sketched in Fig. 1. $\boldsymbol{A}_0(0)$ and $\boldsymbol{A}_0'(0)$ are defined as follows:

$$\boldsymbol{A}_0(0) = \int_0^\infty \mathrm{d}\tilde{\boldsymbol{A}}_0(x), \tag{A.1}$$

$$\boldsymbol{A}_0'(0) = -\int_0^\infty x\, \mathrm{d}\tilde{\boldsymbol{A}}_0(x). \tag{A.2}$$

1. Deterministic service distribution, i.e., $\mathrm{d}\tilde{H}(x)/\mathrm{d}x = \delta(x-h)$:

$$\boldsymbol{A}_0(0) = \exp[\boldsymbol{R}(0)h], \tag{A.3}$$

$$\boldsymbol{A}_0'(0) = -h\boldsymbol{A}_0(0). \tag{A.4}$$

2. k-Stage Erlangian service distribution, i.e., $\mathrm{d}\tilde{H}(x)/\mathrm{d}x = \Gamma(x;k,\mu)$ [13,15]:

$$\boldsymbol{A}_0(0) = \mu^k[\mu\boldsymbol{I} - \boldsymbol{R}(0)]^{-k}, \tag{A.5}$$

$$\boldsymbol{A}_0'(0) = -k\boldsymbol{A}_0(0)[\mu\boldsymbol{I} - \boldsymbol{R}(0)]^{-1}. \tag{A.6}$$

3. Serial connection of deterministic and k-stage Erlangian server, i.e., $\mathrm{d}\tilde{H}(x)/\mathrm{d}x = \Gamma(x-h;k,\mu)$:

$$\boldsymbol{A}_0(0) = \mu^k \exp[\boldsymbol{R}(0)h][\mu\boldsymbol{I} - \boldsymbol{R}(0)]^{-k}, \tag{A.7}$$

$$\boldsymbol{A}_0'(0) = -h\boldsymbol{A}_0(0) - k\boldsymbol{A}_0(0)[\mu\boldsymbol{I} - \boldsymbol{R}(0)]^{-1}. \tag{A.8}$$

4. Serial connection of exponential and k-stage Erlangian server, i.e., $\mathrm{d}\tilde{H}(x)/\mathrm{d}x = \mu_2 \mathrm{e}^{-\mu_2 x} \star \Gamma(x; k, \mu_1)$ where $\star$ denotes the functional convolution:

$$\boldsymbol{A}_0(0) = \mu_1^k \mu_2 [\mu_1 \boldsymbol{I} - \boldsymbol{R}(0)]^{-k} [\mu_2 \boldsymbol{I} - \boldsymbol{R}(0)]^{-1}, \tag{A.9}$$

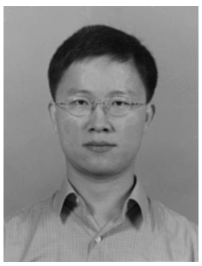
$$\boldsymbol{A}_0'(0) = -\boldsymbol{A}_0(0)[\mu_2 \boldsymbol{I} - \boldsymbol{R}(0)]^{-1} - k\boldsymbol{A}_0(0)[\mu_1 \boldsymbol{I} - \boldsymbol{R}(0)]^{-1}. \tag{A.10}$$

5. Compound model: the resultant $\boldsymbol{A}_0(0)$ and $\boldsymbol{A}_0'(0)$ are weighted sum of the individual $\boldsymbol{A}_0(0)$ and $\boldsymbol{A}_0'(0)$.

References

[1] A. Adas, Traffic models in broadband networks, IEEE Commun. Mag. 35 (7) (1997) 82–89.
[2] R.G. Addie, M. Zukerman, Queueing performance of a tree type ATM network, in: Proceedings of the IEEE INFOCOM'94, Toronto, Canada, June 1994, pp. 48–55.
[3] ATM User-network Interface Specification, Version 3.0, The ATM Forum, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[4] D. Bertsekas, R. Gallager, Data Networks, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1987.
[5] C. Blondia, A discrete-time batch Markovian arrival process as B-ISDN traffic model, Belg. J. Oper. Res. Statist. Comput. Sci. 32 (3–4) (1994) 3–23.
[6] H.W. Ferng, J.F. Chang, Connection-wise end-to-end delay analysis in ATM networks, IEICE Trans. Commun. E83-B (3) (2000) 659–671.
[7] H.W. Ferng, J.F. Chang, Departure processes of BMAP/G/1 queues, submitted for publication.
[8] W. Fischer, K.S. Meier-Hellstern, The Markov-modulated Poisson process (MMPP) cookbook, Performance Evaluation 18 (1993) 149–171.
[9] P. Harrison, N. Pated, Performance Modeling of Communication Networks and Computer Architectures, Addison-Wesley, New York, 1993.
[10] O. Hashida, Y. Takahashi, S. Shimogawa, Switched batch Bernoulli process (SBBP) and the discrete-time SBBP/G/1 queue with application to statistical multiplexer performance, IEEE J. Select. Areas Commun. 9 (3) (1991) 394–401.
[11] H. Heffes, A class of data traffic processes–covariance function characterization and related queueing results, Bell Syst. Tech. J. 59 (6) (1980) 897–929.
[12] H. Heffes, D.M. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, IEEE J. Select. Areas Commun. SAC-4 (6) (1986) 856–868.
[13] P.G. Hoel, S.C. Port, C.J. Stone, Introduction to Probability Theory, Houghton Mifflin, Boston, MA, 1971.
[14] Y.H. Kim, C.K. Un, Performance analysis of statistical multiplexing for heterogeneous bursty traffic in an ATM network, IEEE Trans. Commun. 42 (2–4) (1994) 745–753.
[15] L. Kleinrock, Queueing Systems: Theory, Vol. I, Wiley, New York, 1975.
[16] H. Kroner, M. Eberspacher, T.H. Theimer, P.J. Kuhn, U. Briem, Approximate analysis of the end-to-end delay in ATM networks, in: Proceedings of the IEEE INFOCOM'92, Florence, Italy, May 1992, pp. 978–986.
[17] A. Kuczura, The interrupted Poisson process as an overflow process, Bell Syst. Tech. J. 52 (3) (1973) 437–448.
[18] P.J. Kuehn, Remainder on queueing theory for ATM networks, Telecommun. Systems 5 (1996) 1–24.
[19] D.M. Lucantoni, K.S. Meier-Hellstern, M.F. Neuts, A single-server queue with server vacations and a class of non-renewal arrival processes, Adv. Appl. Probab. 22 (1990) 676–705.
[20] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Commun. Statist. Stochastic Models 7 (1) (1991) 1–46.
[21] D.E. McDysan, D.L. Spohn, ATM Theory and Application, McGraw-Hill, New York, 1995.
[22] E. Modiano, J.E. Wieselthier, A simple analysis of average queueing delay in tree networks, IEEE Trans. Inform. Theory 42 (2) (1996) 660–664.
[23] H. Naser, A. Leon-Garcia, A simulation study of delay and delay variation in ATM networks, in: Proceedings of the IEEE INFOCOM'96, San Francisco, CA, March 1996, pp. 393–400.
[24] M.F. Neuts, A versatile Markovian point process, J. Appl. Probab. 16 (1979) 746–779.
[25] C.H. Ng, L. Bai, B.H. Soong, Modelling multimedia traffic over ATM using MMBP, IEE Proc.-Commun. 144 (5) (1997) 307–310.

[26] R.O. Onvural, Asynchronous Transfer Mode Networks: Performance Issues, Artech House, Boston, 1994.
[27] D. Park, H.G. Perros, H. Yamashita, Approximate analysis of discrete-time tandem queueing networks with bursty and correlated input traffic and customer loss, Oper. Res. Lett. 15 (1994) 95–104.
[28] J.F. Ren, J.W. Mark, J.W. Wong, End-to-end performance in ATM networks, in: Proceedings of the IEEE ICC'94, New Orleans, LA, May 1994, pp. 996–1002.
[29] H. Saito, Interdeparture times from an M/G/1 queueing system with heterogeneous inputs, in: Proceedings of the National Spring Conference, Operations Research Society of Japan, 1-E-4, 1989.
[30] K. Sohraby, A. Privalov, End-to-end jitter analysis in networks of periodic flows, in: Proceedings of the IEEE INFOCOM'99, New York, March 1999.
[31] G.D. Stamoulis, M.E. Anagnostou, A.D. Georgantas, Traffic source models for ATM networks: a survey, Comput. Commun. 17 (6) (1994) 428–438.
[32] D. Stanford, W. Fischer, The interdeparture-time distribution for each class in the $\sum_i \mathrm{M}_i/\mathrm{G}_i/1$ queue, Queueing Systems 4 (1989) 179–191.
[33] H. Takagi, Queueing Analysis: A Foundation of Performance Evaluation: Vacation and Priority Systems, Vol. 1, Elsevier, Amsterdam, 1991.



Huei-Wen Ferng was born in Taiwan in 1970. He received his B.S.E.E. degree from the National Tsing Hwa University, Hsinchu, Taiwan, in 1993 and Ph.D. in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 2000. His research interests include queueing theory, teletraffic modeling, and performance analysis in ATM networks.

Jin-Fu Chang was born in Taiwan in 1948. He received his B.S.E.E. degree from the National Taiwan University, Taipei, Taiwan, in 1970 and Ph.D. in electrical engineering and computer sciences from the University of California, Berkeley, in 1977. He is currently Professor of Electrical Engineering at the National Taiwan University where he spent most of his past career time as a professor since August 1982. He was also the Department Chairman from August 1985 to July 1987 and then was on leave at the Ministry of Education as Director of Science and Technology Advisory Office from July 1987 to June 1990. He was an Adjunct Professor (from February to July 1982) and an Associate Professor (from September 1984 to September 1985) at the Electrical and Computer Engineering Department, Naval Postgraduate School, Monterey, CA. He spent the summer of 1989 visiting AT&T Bell Labs, Holmdel, NJ, and the summer of 1995 visiting the Computer Laboratory, Cambridge University, UK. From August 1991 to July 1994, he was at the National Central University as Dean of Academic Affairs. He was Vice Chairman of the National Science Council, Taiwan from June 1996 to May 2000. He was also a Research Fellow at the Institute of Information Science, Academia Sinica, for a number of years. His research interests include computer communications, high-speed networks, performance analysis, wireless communications, and cryptography. Dr. Chang was elected one of the 10 Outstanding Young Engineers by the Chinese Institute of Engineers in 1982. He received a publication award from the Sun Yet-Sen Culture Foundation in 1983, a two-year Research Award from the Ministry of Education in 1983, Distinguished Research Awards from the National Science Council from 1986 to 1996, a Paper Award from the Chinese Institute of Engineering in 1986, and the Academic Achievement Award in Engineering from the Ministry of Education in 1990. Dr. Chang is a member of the ACM and a Fellow of the IEEE. He was the Secretary General of the Chinese Computer Society (from 1989 to 1993), on the Board of Trustees of the Taiwan Power Company from September 1988 to September 1990, and a Board Member of the Chinese Institute of Electrical Engineering from 1986 to 1988. He was President of the Institute of Information and Computing Machinery, Taiwan, a sister society of the ACM, from March 1998 to March 2000. He also held the chair of ACM, Taipei, Taiwan Chapter from January 1996 to January 1998.