PAPER

Connection-Wise End-to-End Delay Analysis in ATM Networks

Huei-Wen FERNG[†] *and* Jin-Fu CHANG[††], *Nonmembers*

SUMMARY A systematic method for connection-wise end-to-end delay analysis in asynchronous transfer mode (ATM) networks is proposed. This method consists of the followings: (i) per-stream nodal analysis; (ii) output processes characterization; and (iii) moment matching scheme. Following our previous work [1], we employ H-MMPPs/Slotted D/1 to model ATM queues. Each virtual connection (VC) in ATM networks can be regarded as a tandem configuration of such queues. In [1], the per-stream analytical results for such an H-MMPPs/Slotted D/1 queue have been provided. In this paper, not only the composite output process is exactly characterized, but also the component in an output process that corresponds to a specific traffic stream is approximated via a decomposition scheme. A moment matching scheme to emulate the per-stream output process as a two-state MMPP is further proposed. Through moment matching, we can then approximate the connection-wise end-to-end delay by recursively performing the nodal performance analysis. The connection-wise end-to-end delay is crucial to network resource decision or control problems such as call admission control (CAC) and routing.

key words: *ATM, MMPP, H-MMPPs/Slotted D/1 queue, tandem queues, end-to-end performance*

1. Introduction

Broadband integrated services digital networks (B-ISDNs) [2] provide a means of exchanging multimedia information such as voice, video, image, data, and etc. Among various transport technologies, ATM is targeted to provide such diverse services in one network. It provides different quality of service (QoS) according to service requirements. Since the QoS requirements are defined on a per-connection/stream basis, performance evaluation in connection-oriented ATM networks has the need to be done for a specific connection which is complicated by interference from other traffic streams. Past researches mostly focused on the performance analysis for an isolated node. Recently, end-to-end QoS, e.g., connection-wise end-to-end sojourn delay time, has become more and more attentioned. Knowledge of end-to-end QoS helps us to understand the interplay between traffic descriptors and QoS requirements. [3]–[5] are some of the works that have appeared in the literature to study the end-to-end performance in ATM networks. Addie and Zukerman [3] studied the performance of a tree type ATM network with discrete-time Gaussian arrivals. Kroner et al. [4] approximated the end-to-end delay jitter in ATM networks with burst silence cell streams. Ren et al. [5] also studied the end-to-end performance of ATM networks with On-Off cell streams. For a specific connection or tagged VC, we shall in this paper examine the end-to-end cell sojourn delay time. We propose an analytical method to estimate these delays. These estimates can be applied to facilitate call admission control, congestion control, routing algorithms and etc.

The output stream of a switch output port or a multiplexer becomes one of the input streams to the next stage. Research in this regard can be found in [6]–[10]. Ohba et al. [6] studied the departure process of a designated GI-stream interfered by other batch Bernoulli and IPP arrivals. Park et al. [7] studied the departure process of the lossy geometric server receiving an MMBP input. Saito [8] studied the departure process of an N/G/1 queue based on the embedded Markov chain at departure epochs. Takine et al. [9] studied the departure process of a discrete-time lossy deterministic server with correlated arrivals. Kleinrock [10] discussed the output process of queues with server(s) and input of exponential type. Among these past works, [7]–[10] investigated the composite departure process only. Although [6] dealt with the per-stream departure process, the results obtained are not in closed-form and the model employed is perhaps too simplified. In our work, part of the accomplishment is to analyze the exact composite departure statistics of the H-MMPPs/Slotted D/1 queue. In addition, we obtain the approximate per-stream departure statistics corresponding to a specific traffic stream via a decomposition scheme. Along with a method of moment matching, we succeed in obtaining the end-to-end sojourn delay time of a tagged VC by matching the corresponding per-stream output process to a two-state MMPP at each node en route to the destination node. There are several existing moment matching techniques, e.g., [11], [12]. In [11], Heffes and Lucantoni modeled a superposition of packet arrival processes by an MMPP so that the following four statistics are matched: mean cell arrival rate, short term and long term variance-to-mean ratio of the number of arrivals,

Manuscript received April 30, 1999.

Manuscript revised September 29, 1999.

[†]The author is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan 10617 R.O.C.

[††]The author was with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan 10617 R.O.C. He is now with the National Science Council, Taipei, Taiwan 10622 R.O.C.

and the third moment of the number of arrivals in a period. In this work, the following four departure statistics are matched: the first moment of interdeparture times, squared coefficient of variation of interdeparture times, the third moment of interdeparture times, and lag 1 covariance of interdeparture times. These four statistics are selected for the purpose of characterizing the output process. A detailed discussion about why these four statistics are selected is given in Sect. 4. We approximate the per-stream output process by a two-state MMPP if the matching scheme works; or the per-stream output process is taken to be Poisson. The failure of moment matching is due to that burstiness of the (estimated) output process has been found to be reduced to beyond the scope of MMPPs.

The rest of this paper is organized as follows. Preliminaries are given in Sect. 2. In Sect. 3, we analyze the composite and per-stream departure processes for the H-MMPPs/Slotted D/1 queue. In Sect. 4, we propose a moment matching scheme to emulate the per-stream output process as a two-state MMPP. Section 5 gives detailed description of the systematic method to evaluate connection-wise end-to-end performance measures. In Sect. 6, numerical examples are given. Computer simulations are also provided to demonstrate the accuracy of approximation. Section 7 concludes the paper.

2. Preliminaries

Following [1], we use the H-MMPPs/Slotted D/1 queue, a discrete-time deterministic server receiving heterogeneous MMPPs, to model ATM queues, e.g., a multiplexer or an output port of a switch. For the convenience of subsequent sections, we here present notation related to the MMPP [13] and results of an H-MMPPs/Slotted D/1 queue previously obtained in [1].

2.1 MMPP Traffic

The following discussions can be found in [13].

A two-state MMPP of type j is characterized by the following infinitesimal generator $\boldsymbol{Q}_j$ of the underlying Markov chain and rate matrix $\boldsymbol{\Lambda}_j$ with two Poisson arrival rates λ_{1j} and λ_{2j}, i.e.,

$$\boldsymbol{Q}_j = \begin{bmatrix} -\sigma_{1j} & \sigma_{1j} \\ \sigma_{2j} & -\sigma_{2j} \end{bmatrix}, \quad \boldsymbol{\Lambda}_j = \begin{bmatrix} \lambda_{1j} & 0 \\ 0 & \lambda_{2j} \end{bmatrix}. \tag{1}$$

Because of the property that superposition of independent MMPPs yields again an MMPP, the superposition of r different types of two-state MMPP is also MMPP with $m = 2^r$ possible states. The state space of the superposed MMPP can be described using Kronecker product and sum of matrices. In other words, the superposition of r independent two-state MMPPs parameterized by the descriptor $(\boldsymbol{Q}_j, \boldsymbol{\Lambda}_j)$, $1 \leq j \leq r$, can be represented by the m-state MMPP parameterized by the descriptor $(\boldsymbol{Q}, \boldsymbol{\Lambda})$ with

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{Q}_1 \oplus \boldsymbol{Q}_2 \oplus \ldots \oplus \boldsymbol{Q}_r, \\ \boldsymbol{\Lambda} &= \boldsymbol{\Lambda}_1 \oplus \boldsymbol{\Lambda}_2 \oplus \ldots \oplus \boldsymbol{\Lambda}_r \end{aligned} \tag{2}$$

where $\oplus$ denotes the Kronecker sum operator. The rate matrix $\boldsymbol{\Lambda}$ is a diagonal matrix with $\lambda_1,\ldots,\lambda_m$ on the diagonal, i.e., $\boldsymbol{\Lambda} = Diag(\lambda_1, \lambda_2, \ldots, \lambda_m)$.

For the superposed m-state MMPP $(\boldsymbol{Q}, \boldsymbol{\Lambda})$, the steady-state probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ of the underlying Markov chain is given by solving the equations $\boldsymbol{\pi Q} = \boldsymbol{0}$ and $\boldsymbol{\pi e} = 1$ 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.

Define the conditional probability $P_{i,j}(n,t) \overset{\triangle}{=} Pr\{N(t) = n, J(t) = j \mid N(0) = 0, J(0) = i\}$ where $N(t)$ and $J(t)$ denote respectively the number of arrivals during the interval $(0,t)$ and the state of the underlying Markov process at time t. From [13], the $m \times m$ matrix of probabilities $\boldsymbol{P}(n,t) \overset{\triangle}{=} [P_{i,j}(n,t)]_{1 \leq i,j \leq m}$ has the probability generating function

$$\boldsymbol{P}^*(z,t) = e^{\boldsymbol{R}(z)t}\ , \quad |\ z\ | \leq 1 \tag{3}$$

with

$$\boldsymbol{R}(z) = \boldsymbol{Q} + (z-1)\boldsymbol{\Lambda}. \tag{4}$$

And the effective arrival rate of the superposed MMPP $(\boldsymbol{Q}, \boldsymbol{\Lambda})$ is $\lambda^* = \boldsymbol{\pi} \frac{d}{dz} \boldsymbol{R}(z) \mid_{z=1} \boldsymbol{e} = \boldsymbol{\pi \lambda}$ where $\boldsymbol{\lambda} = \boldsymbol{\Lambda e}$. The arrival rate from type j $(1 \leq j \leq r)$ traffic stream is independent of the other traffic streams and is given by

$$\boldsymbol{\Lambda}(j) = \boldsymbol{0} \oplus \ldots \oplus \boldsymbol{0} \oplus \boldsymbol{\Lambda}_j \oplus \boldsymbol{0} \oplus \ldots \oplus \boldsymbol{0} \tag{5}$$

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

2.2 System Size of the H-MMPPs/Slotted D/1 Queue at Slot Boundaries

The followings are taken from [1].

Let L_n^s denote the number of cells in the system immediately after the end of the nth slot, i.e., L_n^s includes cells arrived and accommodated into the system in the nth slot but excludes the cell departed at the end of the nth slot. Then $L_{n+1}^s = (L_n^s - 1)^+ + \tilde{A}_{n+1}$ where $\tilde{A}_n$ denotes the number of cells arrive in the nth slot and $(x)^+ = max(x, 0)$ (see the system time diagram shown in Fig. 1). Let J_n^s denote the state of the input process

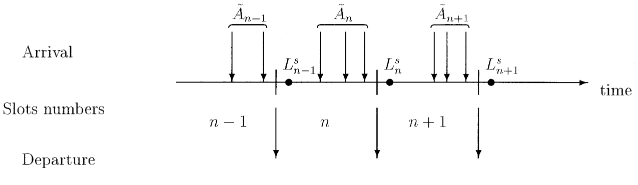


Fig. 1 System time diagram.

at the end of the nth slot. And let $\boldsymbol{A}_i \overset{\triangle}{=} \boldsymbol{P}(i,T)$ $(i \geq 0)$ where T denotes the duration of a slot and is used as the time unit of the time axis. $\boldsymbol{A}_i$ may be computed using the method in [13] or [14]. Then $\{(L_n^s, J_n^s) : n \geq 0\}$ forms the infinite-state Markov chain with state space $\{0,1,2,\ldots\} \times \{1,2,\ldots,m\}$ and has transition probability matrix in the following form:

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{B}_0^s & \boldsymbol{B}_1^s & \boldsymbol{B}_2^s & \cdots \\ \boldsymbol{A}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots \\ \boldsymbol{0} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{6}$$

where $\boldsymbol{B}_i^s = \boldsymbol{A}_i$, $i \geq 0$ and $\boldsymbol{0}$ is an $m \times m$ zero matrix.

Let $\boldsymbol{x}^s = (\boldsymbol{x}_0^s, \boldsymbol{x}_1^s, \ldots, \boldsymbol{x}_n^s, \ldots)$ denote the stationary probability vector of $\boldsymbol{U}$, where $\boldsymbol{x}_i^s$ is a $1 \times m$ vector whose jth element $x_{i,j}^s = \lim_{n\to\infty} Pr\{L_n^s = i, J_n^s = j\}$. $\boldsymbol{x}^s$ satisfies $\sum_{n=0}^{\infty} \boldsymbol{x}_n^s \boldsymbol{e} = 1$ and $\boldsymbol{x}^s = \boldsymbol{x}^s \boldsymbol{U}$ or equivalently,

$$\boldsymbol{x}_i^s = \boldsymbol{x}_0^s \boldsymbol{B}_i^s + \sum_{\nu=1}^{i+1} \boldsymbol{x}_\nu^s \boldsymbol{A}_{i+1-\nu}\ , \quad i \geq 0 \tag{7}$$

where $\boldsymbol{x}_0^s$ can be calculated via resorting to [15] and [16]. Then use (7), we can obtain the system size distribution at slot boundaries.

3. Analysis of Departure Processes

3.1 Interdeparture Time Distribution of the Composite Departure Process

Following a similar philosophy done for a finite queue accepting correlated arrivals in [9], we now derive interdeparture time distribution for the composite output stream of an H-MMPPs/Slotted D/1 queue. Comparing with [9], we shall not only derive neater but more results. For example, the lag 1 covariance is not treated in [9]. Let $\boldsymbol{u}_k$ be a $1 \times m$ vector whose jth element $u_{k,j}$ is the joint stationary probability that the number of cells in the system is k and the state of the underlying Markov chain immediately after a departure is j, then $u_{k,j} = Pr\{L_n^s = k, J_n^s = j \mid L_{n-1}^s \geq 1\}$. The condition $\{L_{n-1}^s \geq 1\}$ is to ensure that the number of cells in the system immediately after the $(n-1)$st slot is at least one and there is a cell departing from the system at the end of the nth slot. Rewrite $u_{k,j}$ as follows:

$$\begin{aligned} u_{k,j} &= \frac{Pr\{L_{n-1}^s \geq 1, L_n^s = k, J_n^s = j\}}{Pr\{L_{n-1}^s \geq 1\}} \\ &= \frac{\sum_{i=1}^{k+1} Pr\{L_{n-1}^s = i, L_n^s = k, J_n^s = j\}}{1 - Pr\{L_{n-1}^s = 0\}}. \end{aligned} \tag{8}$$

Then from (7), (8), and the definition of $\boldsymbol{u}_k$,

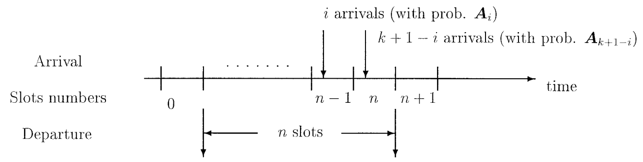

i arrivals (with prob. $\boldsymbol{A}_i$)
$k+1-i$ arrivals (with prob. $\boldsymbol{A}_{k+1-i}$)
Arrival
. . . .
time
Slots numbers
0
$n-1$
n
$n+1$
Departure
n slots

Fig. 2 Illustration of the time diagram of $\boldsymbol{d}_k^{(1)}(n)$.

$$\boldsymbol{u}_k = \frac{\sum_{i=1}^{k+1} \boldsymbol{x}_i^s \boldsymbol{A}_{k+1-i}}{1 - \boldsymbol{x}_0^s \boldsymbol{e}} = \frac{\boldsymbol{x}_k^s - \boldsymbol{x}_0^s \boldsymbol{B}_k^s}{1 - \boldsymbol{x}_0^s \boldsymbol{e}}, \quad k \geq 0. \tag{9}$$

For convenience, we number the slot that the first cell departure occurs as the 0th slot. Let D_m denote the mth interdeparture time. Let $\boldsymbol{d}_k^{(1)}(n)$ be a $1 \times m$ vector whose jth element $d_{k,j}^{(1)}(n)$ is the joint stationary probability that D_1 equals n (slots), the number of cells in the system after the second departure is k and the state of the underlying Markov chain immediately after the second departure is j, i.e., $d_{k,j}^{(1)}(n) = Pr\{D_1 = n, L_n^s = k, J_n^s = j\}$. We use the superscript $^{(1)}$ to indicate a single interdeparture time. We now derive $\boldsymbol{d}_k^{(1)}(n)$ $(n \geq 1)$ as follows. First, we consider $\boldsymbol{d}_k^{(1)}(1)$. Only when the number of cells left behind the first departure at the end of the 0th slot is at least one, say i cells, the interdeparture time becomes one. Thus, $\boldsymbol{d}_k^{(1)}(1) = \sum_{i=1}^{k+1} \boldsymbol{u}_i \boldsymbol{A}_{k+1-i}$ $(k \geq 0)$. We next consider $\boldsymbol{d}_k^{(1)}(n)$ $(n \geq 2)$. When the number of cells left behind the first departure is zero, the interdeparture time is then longer than one. If the first interdeparture time is n, there must be no arrivals between the first and the $(n-2)$nd slot since the first departure. And there must be at least one cell arriving in the $(n-1)$st slot. Thus, $\boldsymbol{d}_k^{(1)}(n) = \boldsymbol{u}_0 \boldsymbol{A}_0^{n-2} \sum_{i=1}^{k+1} \boldsymbol{A}_i \boldsymbol{A}_{k+1-i}$ $(k \geq 0,\ n \geq 2)$. The detail of $\boldsymbol{d}_k^{(1)}(n)$ $(n \geq 2)$ is illustrated in Fig. 2. Let $d^{(1)}(n)$ denote the probability that interdeparture time D_1 is of length n (slots), i.e., $d^{(1)}(n) = \sum_{k=0}^{\infty} \boldsymbol{d}_k^{(1)}(n)\boldsymbol{e}$. Then

$$\begin{aligned} d^{(1)}(1) &= 1 - \boldsymbol{u}_0 \boldsymbol{e}, \\ d^{(1)}(n) &= \boldsymbol{u}_0 \boldsymbol{A}_0^{n-2}(\boldsymbol{I} - \boldsymbol{A}_0)\boldsymbol{e}, \quad n \geq 2. \end{aligned} \tag{10}$$

Let $\tilde{D}^{(1)}(z)$ denote the probability generating function of the interdeparture time distribution. Then we have

$$\begin{aligned} \tilde{D}^{(1)}(z) &= \sum_{n=1}^{\infty} d^{(1)}(n) z^n \\ &= z(1 - \boldsymbol{u}_0 \boldsymbol{e}) \\ &\quad + z^2 \boldsymbol{u}_0 (\boldsymbol{I} - z\boldsymbol{A}_0)^{-1} (\boldsymbol{I} - \boldsymbol{A}_0)\boldsymbol{e}. \end{aligned} \tag{11}$$

Let $D^{(m)}$ denote the mth factorial moment of the interdeparture time distribution, that is, $D^{(m)} =$

$\frac{d^m}{dz^m}\tilde{D}^{(1)}(z)\mid_{z=1}$. Then

$$\begin{aligned} D^{(1)} &= 1+\boldsymbol{u}_0(\boldsymbol{I}-\boldsymbol{A}_0)^{-1}\boldsymbol{e}, \\ D^{(m)} &= m!\boldsymbol{u}_0(\boldsymbol{I}-\boldsymbol{A}_0)^{-m}\boldsymbol{A}_0^{m-2}\boldsymbol{e}, \\ & \quad m \geq 2. \end{aligned} \tag{12}$$

From (9) and (12), $D^{(1)}$ can be further reduced to $1/(1-\boldsymbol{x}_0^s\boldsymbol{e}) = 1/\rho = 1/\lambda^*$ where ρ denotes the utilization factor. Therefore, the mean input rate equals the mean output rate that can also be deduced from flow conservation. From the definition of factorial moment of the interdeparture time, we can obtain moments of D_1 as follows:

$$\begin{aligned} &E[D_1] = D^{(1)}, \quad E[D_1^2] = D^{(2)} + D^{(1)}, \\ &E[D_1^3] = D^{(3)} + 3D^{(2)} + D^{(1)}. \end{aligned} \tag{13}$$

3.2 Joint Distribution of Successive Interdeparture Times and Their Correlation for the Composite Departure Process

As in Sect. 3.1, let $\boldsymbol{u}_k$ denote a $1 \times m$ vector whose jth element is the joint stationary probability that the number of cells in the system is k and the state of the underlying Markov chain immediately after the departure is j. Also D_m denotes the mth interdeparture time.

Let $\boldsymbol{d}_k^{(2)}(n_1,n_2)$ denote a $1 \times m$ vector whose jth element $d_{k,j}^{(2)}(n_1,n_2)$ is the joint stationary probability of the two successive interdeparture times of length n_1 and n_2, respectively, the number of cells in the system is k and the state of the underlying Markov chain at the end of the second departure point is j, i.e., $d_{k,j}^{(2)}(n_1,n_2) = Pr\{D_1 = n_1, D_2 = n_2, L_{n_1+n_2}^s = k, J_{n_1+n_2}^s = j\}$. The derivation of $\boldsymbol{d}_k^{(2)}(n_1,n_2)$ for various n_1 and n_2 can be done for the following situations: (a) when $n_1 \geq 1$ and $n_2 = 1$: $\boldsymbol{d}_k^{(2)}(n_1,1) = \sum_{i=1}^{k+1}\boldsymbol{d}_i^{(1)}(n_1)\boldsymbol{A}_{k+1-i}$ $(k \geq 0)$; (b) when $n_1 \geq 1$ and $n_2 \geq 2$: $\boldsymbol{d}_k^{(2)}(n_1,n_2) = \boldsymbol{d}_0^{(1)}(n_1)\boldsymbol{A}_0^{n_2-2}\sum_{i=1}^{k+1}\boldsymbol{A}_i\boldsymbol{A}_{k+1-i}$ $(k \geq 0)$. Let $d^{(2)}(n_1,n_2) \overset{\triangle}{=} \sum_{k=0}^{\infty}\boldsymbol{d}_k^{(2)}(n_1,n_2)\boldsymbol{e}$, then we can obtain $d^{(2)}(n_1,n_2)$ for each case of n_1 and n_2 as follows:

$$\begin{aligned} d^{(2)}(1,1) &= 1-\boldsymbol{u}_0\boldsymbol{e}-\boldsymbol{u}_1\boldsymbol{A}_0\boldsymbol{e} \\ &\quad \text{for } n_1 = n_2 = 1, \\ d^{(2)}(n_1,1) &= \boldsymbol{u}_0\boldsymbol{A}_0^{n_1-2}(\boldsymbol{I}-\boldsymbol{A}_0-\boldsymbol{A}_1\boldsymbol{A}_0)\boldsymbol{e} \\ &\quad \text{for } n_1 \geq 2 \text{ and } n_2 = 1, \\ d^{(2)}(1,n_2) &= \boldsymbol{u}_1\boldsymbol{A}_0^{n_2-1}(\boldsymbol{I}-\boldsymbol{A}_0)\boldsymbol{e} \\ &\quad \text{for } n_1 = 1 \text{ and } n_2 \geq 2, \\ d^{(2)}(n_1,n_2) &= \boldsymbol{u}_0\boldsymbol{A}_0^{n_1-2}\boldsymbol{A}_1\boldsymbol{A}_0^{n_2-1}(\boldsymbol{I}-\boldsymbol{A}_0)\boldsymbol{e} \\ &\quad \text{for } n_1 \geq 2 \text{ and } n_2 \geq 2. \end{aligned} \tag{14}$$

Define $\tilde{D}_c(y,z) \overset{\triangle}{=} \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty} d^{(2)}(n_1,n_2)y^{n_1}z^{n_2}$ and use (14), we have

$$\begin{aligned} &\tilde{D}_c(y,z) \\ &\quad = (1-\boldsymbol{u}_0\boldsymbol{e}-\boldsymbol{u}_1\boldsymbol{A}_0\boldsymbol{e})yz + \boldsymbol{u}_0(\boldsymbol{I}-y\boldsymbol{A}_0)^{-1} \\ &\qquad \times (\boldsymbol{I}-\boldsymbol{A}_0-\boldsymbol{A}_1\boldsymbol{A}_0)\boldsymbol{e}y^2z + \boldsymbol{u}_1\boldsymbol{A}_0 \\ &\qquad \times (\boldsymbol{I}-z\boldsymbol{A}_0)^{-1}(\boldsymbol{I}-\boldsymbol{A}_0)\boldsymbol{e}yz^2 \\ &\qquad + \boldsymbol{u}_0(\boldsymbol{I}-y\boldsymbol{A}_0)^{-1}\boldsymbol{A}_1\boldsymbol{A}_0 \\ &\qquad \times (\boldsymbol{I}-z\boldsymbol{A}_0)^{-1}(\boldsymbol{I}-\boldsymbol{A}_0)\boldsymbol{e}y^2z^2. \end{aligned} \tag{15}$$

Then we can derive $E[D_{k-1}D_k]$ from (15) by differentiating it with respect to y and z, respectively. After tedious manipulations, we obtain

$$\begin{aligned} &E[D_{k-1}D_k] \\ &\quad = E[D_1D_2] \\ &\quad = \frac{\partial^2}{\partial z \partial y}\tilde{D}_c(y,z)\mid_{y=1,z=1} \\ &\quad = 1 + (\boldsymbol{u}_0+\boldsymbol{u}_1\boldsymbol{A}_0)(\boldsymbol{I}-\boldsymbol{A}_0)^{-1}\boldsymbol{e} \\ &\qquad + \boldsymbol{u}_0(\boldsymbol{I}-\boldsymbol{A}_0)^{-2}(2\boldsymbol{I}-\boldsymbol{A}_0)\boldsymbol{A}_1\boldsymbol{A}_0 \\ &\qquad \times (\boldsymbol{I}-\boldsymbol{A}_0)^{-1}\boldsymbol{e}. \end{aligned} \tag{16}$$

The covariance of D_{k-1} and D_k can be obtained via the following relation

$$\begin{aligned} Cov(D_{k-1},D_k) &= Cov(D_1,D_2) \\ &= E[D_1D_2] - E^2[D_1]. \end{aligned} \tag{17}$$

From (12) and (17), we now have the departure statistics to characterize the composite output process.

3.3 Per-Stream Departure Process of a Tagged Traffic Stream

In order to facilitate end-to-end performance analysis in ATM networks, it is necessary to figure out the per-stream output process corresponding to a tagged cell stream. However, it is much more difficult to extract the per-stream output process corresponding to a tagged cell stream from the composite output process of the H-MMPPs/Slotted D/1 queue. Instead, we propose a heuristic decomposition scheme to obtain the approximate statistics. The usefulness and effectiveness of the decomposition scheme are checked in Sect. 6.2 through numerical examples.

We now describe the decomposition scheme as follows. Denote the tagged MMPP by $\mathrm{MMPP_t}$ and the other $r-1$ cross traffic streams as a whole by $\mathrm{MMPP_c}$. Then the H-MMPPs arrivals are now replaced by $\mathrm{MMPP_t}$+$\mathrm{MMPP_c}$. The decomposition scheme is to substitute the $\mathrm{MMPP_t}$+$\mathrm{MMPP_c}$/Slotted D/1 queue by a decomposed $\mathrm{MMPP_t}$/Slotted $\mathrm{D_{eff}}$/1 queue† with a

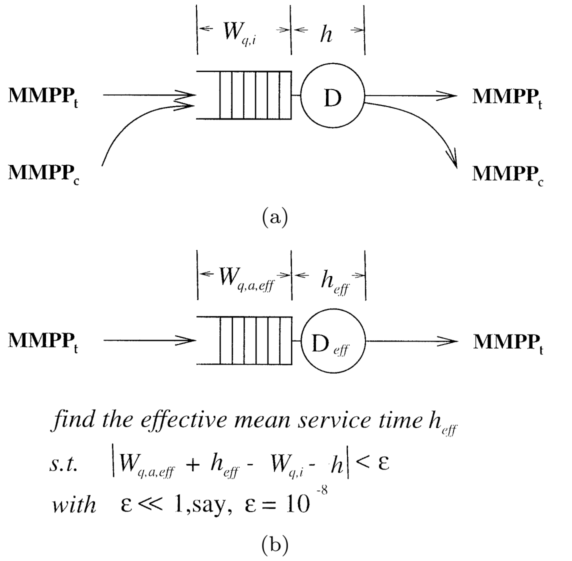

$W_{q,i}$
h
MMPP$_t$
D
MMPP$_t$
MMPP$_c$
MMPP$_c$
(a)
$W_{q,a,eff}$
h_{eff}
MMPP$_t$
D$_{eff}$
MMPP$_t$
find the effective mean service time h_{eff}
s.t. $|W_{q,a,eff} + h_{eff} - W_{q,i} - h| < \varepsilon$
with $\varepsilon \ll 1$,say, $\varepsilon = 10^{-8}$
(b)

Fig. 3 The decomposition scheme: (a) the undecomposed original queueing model; (b) the replaced queueing model.

modified effective deterministic server D$_{\text{eff}}$. The modification of server is to reflect the interference from the cross cell streams. Since we still employ a deterministic server, it is merely necessary to determine its service time. The effective service time h_{eff} is determined via matching the per-stream sojourn delay time that MMPP$_t$ experiences in MMPP$_t$+MMPP$_c$/Slotted D/1 queue with that in the decomposed queue. That is, we first calculate the per-stream sojourn delay time $W_{q,i} + h$ ($W_{q,i}$ can be calculated using the result in [1]) for the MMPP$_t$ in the undecomposed queue. Then recursively apply binary search to find an h_{eff} over a feasible region such that the sojourn delay time $W_{q,a,eff} + h_{eff}$ ($W_{q,a,eff}$ can also be calculated using the result in [1]) of the decomposed queue satisfies $|W_{q,a,eff} + h_{eff} - W_{q,i} - h| < \epsilon$ where ϵ is a small number, say $\epsilon = 10^{-8}$, specified for stoppage of the recursion. Using the decomposition scheme, we obtain the approximate per-stream departure statistics corresponding to a tagged cell stream. This scheme is also illustrated in Fig. 3.

In this work, we again approximate the per-stream output process by a two-state MMPP through moment matching technique stated latter. Of course, we can approximate the per-stream output process by a multi-state (higher than two) MMPP to gain more accurate results, but this causes more complexities on moment matching scheme.

4. Moment Matching Scheme

For more general description, we introduce the following notation:

- $T_{A,i}$: the time between the ith and the $(i+1)$st arrivals of the matched two-state MMPP parameterized by the descriptor $(\sigma_1^{(m)}, \sigma_2^{(m)}, \lambda_1^{(m)}, \lambda_2^{(m)})$.
- $T_{D,i}$: the time between the ith and the $(i+1)$st departures of the output process.

In order to match the composite or per-stream output process as a two-state MMPP traffic source, we need to select four statistics of the output process to match. In this work, we select the following four statistics: (a) the first moment of interdeparture times, i.e., $E[T_{D,i}]$; (b) squared coefficient of variation of interdeparture times, i.e., $c^2(T_{D,i}) = Var[T_{D,i}]/E^2[T_{D,i}]$; (c) the third moment of interdeparture times, i.e., $E[T_{D,i}^3]$; (d) lag 1 covariance of interdeparture times, i.e., $Cov(T_{D,i}, T_{D,i+1})$.

We now explain why these four statistics are selected. From [17]: (i) the mean $E[T_{D,i}]$ is a measure of central tendency of the random variable $T_{D,i}$; (ii) the variance $Var[T_{D,i}]$ is a measure of the dispersion of the random variable $T_{D,i}$ with respect to (w.r.t.) the mean. Equivalently, $c^2(T_{D,i})$ measures the dispersion; (iii) the skewness $\nu = E[(T_{D,i} - E[T_{D,i}])^3]/(Var[T_{D,i}])^{\frac{3}{2}}$ is a measure of symmetry of the random variable $T_{D,i}$ w.r.t. the mean; (iv) the covariance $Cov(T_{D,i}, T_{D,i+1})$ is a measure of dependence between $T_{D,i}$ and $T_{D,i+1}$. We note that (i)–(iii) concern the shape of the distribution of the random variable $T_{D,i}$ and (iv) describes the correlation/dependence between $T_{D,i}$ and $T_{D,i+1}$. These four statistics well reflect the output process.

We obtain the four statistics after tedious algebraic manipulations as follows:

$$E[T_{A,i}] = \frac{\sigma_1^{(m)} + \sigma_2^{(m)}}{\sigma_2^{(m)}\lambda_1^{(m)} + \sigma_1^{(m)}\lambda_2^{(m)}}, \tag{18}$$

$$\begin{aligned} c^2(T_{A,i}) &= 1 + 2\sigma_1^{(m)}\sigma_2^{(m)}(\lambda_1^{(m)} - \lambda_2^{(m)})^2 \\ &\quad /\{[\sigma_2^{(m)}\lambda_1^{(m)} + \sigma_1^{(m)}\lambda_2^{(m)} + \lambda_1^{(m)}\lambda_2^{(m)}] \\ &\quad \times \sigma_1^{(m)} + \sigma_2^{(m)})^2\}, \end{aligned} \tag{19}$$

$$\begin{aligned} E[T_{A,i}^3] &= 3E[T_{A,i}^2](\sigma_1^{(m)} + \sigma_2^{(m)})/[\sigma_2^{(m)}\lambda_1^{(m)} \\ &\quad + \sigma_1^{(m)}\lambda_2^{(m)} + \lambda_1^{(m)}\lambda_2^{(m)}] + 6[(\sigma_1^{(m)} \\ &\quad + \sigma_2^{(m)})(\sigma_1^{(m)}\lambda_1^{(m)} + \sigma_2^{(m)}\lambda_2^{(m)}) \\ &\quad + \sigma_1^{(m)}\lambda_1^{(m)^2} + \sigma_2^{(m)}\lambda_2^{(m)^2}]/\{(\sigma_2^{(m)} \\ &\quad \times \lambda_1^{(m)} + \sigma_1^{(m)}\lambda_2^{(m)})(\sigma_2^{(m)}\lambda_1^{(m)} \\ &\quad + \sigma_1^{(m)}\lambda_2^{(m)} + \lambda_1^{(m)}\lambda_2^{(m)})^2\}, \end{aligned} \tag{20}$$

[†] From flow conservation, the mean output rates of MMPP$_t$ in both MMPP$_t$+MMPP$_c$/Slotted D/1 and MMPP$_t$/Slotted D$_{\text{eff}}$/1 queues are equal. Therefore, the mean interdeparture time is obtainable without approximation using this decomposition scheme.

$$\begin{aligned} Cov(T_{A,i}, T_{A,i+1}) = {} & \{Var[T_{A,i}] - E^2[T_{A,i}]\} \\ & \times \lambda_1^{(m)}\lambda_2^{(m)}/[2(\sigma_2^{(m)}\lambda_1^{(m)} \\ & + \sigma_1^{(m)}\lambda_2^{(m)} + \lambda_1^{(m)}\lambda_2^{(m)})]. \end{aligned} \tag{21}$$

Therefore, we now work with the following four matching equations:

$$\frac{\sigma_1^{(m)} + \sigma_2^{(m)}}{\sigma_2^{(m)}\lambda_1^{(m)} + \sigma_1^{(m)}\lambda_2^{(m)}} = A_d, \tag{22}$$

$$\begin{aligned} & 1 + 2\sigma_1^{(m)}\sigma_2^{(m)}(\lambda_1^{(m)} - \lambda_2^{(m)})^2/\{[\sigma_2^{(m)}\lambda_1^{(m)} \\ & +\sigma_1^{(m)}\lambda_2^{(m)} + \lambda_1^{(m)}\lambda_2^{(m)}](\sigma_1^{(m)} + \sigma_2^{(m)})^2\} \\ & \ = B_d, \end{aligned} \tag{23}$$

$$\begin{aligned} & \frac{3C_{d,1}(\sigma_1^{(m)} + \sigma_2^{(m)})}{\sigma_2^{(m)}\lambda_1^{(m)} + \sigma_1^{(m)}\lambda_2^{(m)} + \lambda_1^{(m)}\lambda_2^{(m)}} \\ & \quad + 6[(\sigma_1^{(m)} + \sigma_2^{(m)})(\sigma_1^{(m)}\lambda_1^{(m)} + \sigma_2^{(m)}\lambda_2^{(m)}) \\ & \quad + \sigma_1^{(m)}{\lambda_1^{(m)}}^2 + \sigma_2^{(m)}{\lambda_2^{(m)}}^2]/\{(\sigma_2^{(m)}\lambda_1^{(m)} \\ & \quad + \sigma_1^{(m)}\lambda_2^{(m)})(\sigma_2^{(m)}\lambda_1^{(m)} + \sigma_1^{(m)}\lambda_2^{(m)} \\ & \quad + \lambda_1^{(m)}\lambda_2^{(m)})^2\} = C_d, \end{aligned} \tag{24}$$

$$\frac{D_{d,1}\lambda_1^{(m)}\lambda_2^{(m)}}{\sigma_2^{(m)}\lambda_1^{(m)} + \sigma_1^{(m)}\lambda_2^{(m)} + \lambda_1^{(m)}\lambda_2^{(m)}} = D_d \tag{25}$$

where $A_d = E[T_{D,i}]$, $B_d = Var[T_{D,i}]/E^2[T_{D,i}]$, $C_d = E[T_{D,i}^3]$, $C_{d,1} = E[T_{D,i}^2]$, $D_d = Cov(T_{D,i}, T_{D,i+1})$, $D_{d,1} = (Var[T_{D,i}] - E^2[T_{D,i}])/2$.

The system of equations in (22)–(25) can be solved by the following procedure:

- Introduce another set of four variables α, β, γ, δ as follows:

$$\begin{aligned} \alpha &= \sigma_1^{(m)} + \sigma_2^{(m)}, \\ \beta &= \sigma_1^{(m)}\lambda_2^{(m)} + \sigma_2^{(m)}\lambda_1^{(m)}, \\ \gamma &= \lambda_1^{(m)}\lambda_2^{(m)}, \\ \delta &= \lambda_1^{(m)} + \lambda_2^{(m)}. \end{aligned} \tag{26}$$

- Rewrite (22)–(25) in terms of variables α, β, γ, δ. After algebraic manipulations, we have the following solution:

$$\begin{aligned} \alpha &= 6(\zeta\xi - \eta)/[C_d\zeta(\zeta+\eta)^2 - 3C_{d,1}\zeta(\zeta+\eta) \\ &\quad -6(\xi + \xi^2)], \\ \beta &= \zeta\alpha, \\ \gamma &= \eta\alpha, \\ \delta &= \zeta + \xi\alpha \end{aligned} \tag{27}$$

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 (27), we further obtain

$$\begin{aligned} \lambda_1^{(m)} &= \frac{\delta + \sqrt{\delta^2 - 4\gamma}}{2}, \\ \lambda_2^{(m)} &= \frac{\delta - \sqrt{\delta^2 - 4\gamma}}{2}, \\ \sigma_1^{(m)} &= \frac{\alpha\lambda_1^{(m)} - \beta}{\lambda_1^{(m)} - \lambda_2^{(m)}}, \\ \sigma_2^{(m)} &= \frac{\beta - \alpha\lambda_2^{(m)}}{\lambda_1^{(m)} - \lambda_2^{(m)}}. \end{aligned} \tag{28}$$

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

In order to apply the above procedure to match the output process of the H-MMPPs/Slotted D/1 queue, we note that

$$\begin{aligned} & E[T_{D,i}] = TE[D_1], \\ & E[T_{D,i}^2] = T^2E[D_1^2], \\ & E[T_{D,i}^3] = T^3E[D_1^3], \\ & Cov(T_{D,i}, T_{D,i+1}) = T^2Cov(D_{k-1}, D_k). \end{aligned} \tag{29}$$

Then using (12), (17), (27)–(29), we are able to match the output process as a two-state MMPP $(\sigma_1^{(m)}, \sigma_2^{(m)}, \lambda_1^{(m)}, \lambda_2^{(m)})$. Note that (28) may not have a feasible solution set. For such case, we then approximate the output process as a Poisson process by setting $\lambda_1^{(m)} = \lambda_2^{(m)} = \zeta$, $\sigma_1^{(m)}$ and $\sigma_2^{(m)}$ are freely specified. The reason why we select the Poisson process is that only its arrival rate is needed to specify and this parameter can be correctly obtained as mentioned earlier in Sect. 3.3, while the other output statistics are underestimated (at heavy traffic load) using the decomposition scheme (see Sect. 6.2).

5. Connection-Wise End-to-End Performance Analysis in ATM Networks

As we mentioned earlier, each VC in ATM networks can be regarded as a tandem configuration of H-MMPPs/Slotted D/1 queues. Applying the results we obtained in the previous sections, we now propose a recursive scheme outlined in Fig. 4 to evaluate the connection-wise performance for a specific VC. This scheme is done for the tandem queues shown in Fig. 5 in which all external arrival sources are assumed to be two-state MMPPs and mutually independent.

When applying our heuristic recursive calculation of the per-stream end-to-end sojourn delay time, we must modify the sojourn delay time after the first node. This is because the per-stream output processes of the first and the descendent nodes possess the discrete-time nature. Thus, if we use two-state MMPP to match the

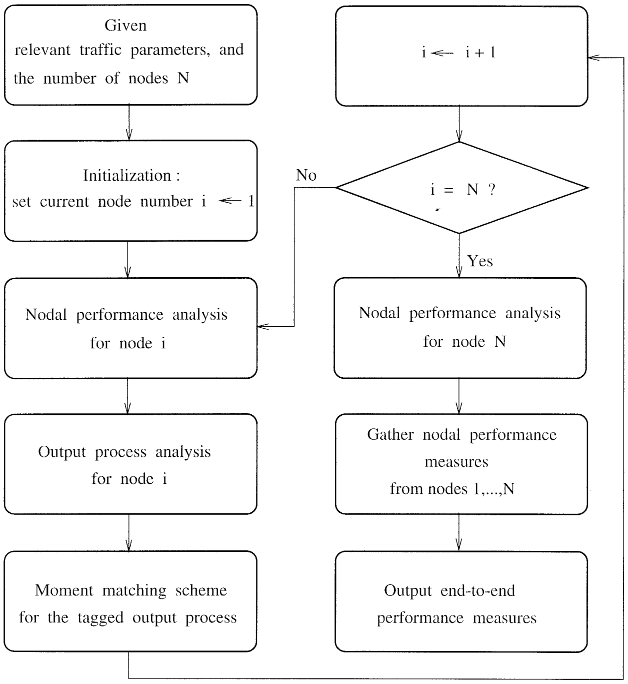


Fig. 4 A recursive scheme applied to the end-to-end performance analysis.

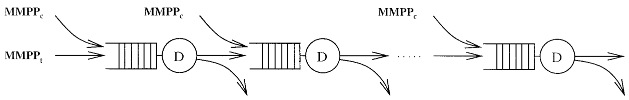


Fig. 5 The tandem queueing configuration.

per-stream output process, we must modify the results by the following interpolation scheme:

- Let the cross traffic be zero, we calculate its per-stream mean delay time drift $\triangle W_{q,i,0} = -W_{q,i,0}$. The subscript 0 denotes zero cross traffic and $W_{q,i,0}$ is the per-stream mean delay time. Note that the tagged traffic coming from the upstream node exhibits interarrival time of multiple slots. Then per-stream mean delay time should be 0. However, $W_{q,i,0}$ is not equal to 0 because we employ MMPP to match the per-stream output process. Hence we need to set $\triangle W_{q,i,0} = -W_{q,i,0}$ to compensate the per-stream mean delay time.
- Assume that the tagged traffic source has effective load ρ_i, we heuristically set $\triangle W_{q,i,1-\rho_i} = 0$, i.e., the overall traffic load equals one. Under the full load condition, the per-stream mean delay time drift becomes negligible. Hence we set $\triangle W_{q,i,1-\rho_i} = 0$.
- We use $\triangle W_{q,i,\rho_c} = -(1-\rho_i-\rho_c)W_{q,i,0}/(1-\rho_i)$ when the tagged traffic load and cross traffic load are ρ_i and ρ_c, respectively. Note that $\triangle W_{q,i,\rho_c}$ can be easily obtained via internal interpolation between $(0, \triangle W_{q,i,0})$ and $(1-\rho_i, \triangle W_{q,i,1-\rho_i})$. This is why we call this scheme an "interpolation" scheme.

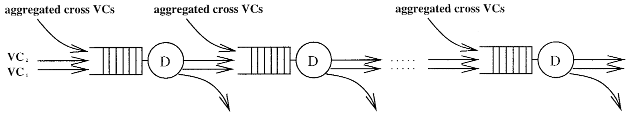


Fig. 6 An example of two VCs sharing the same route.

Using the recursive calculation together with the interpolation scheme, we can obtain a more accurate estimate of the per-stream sojourn delay time incurred at the intermediate nodes.

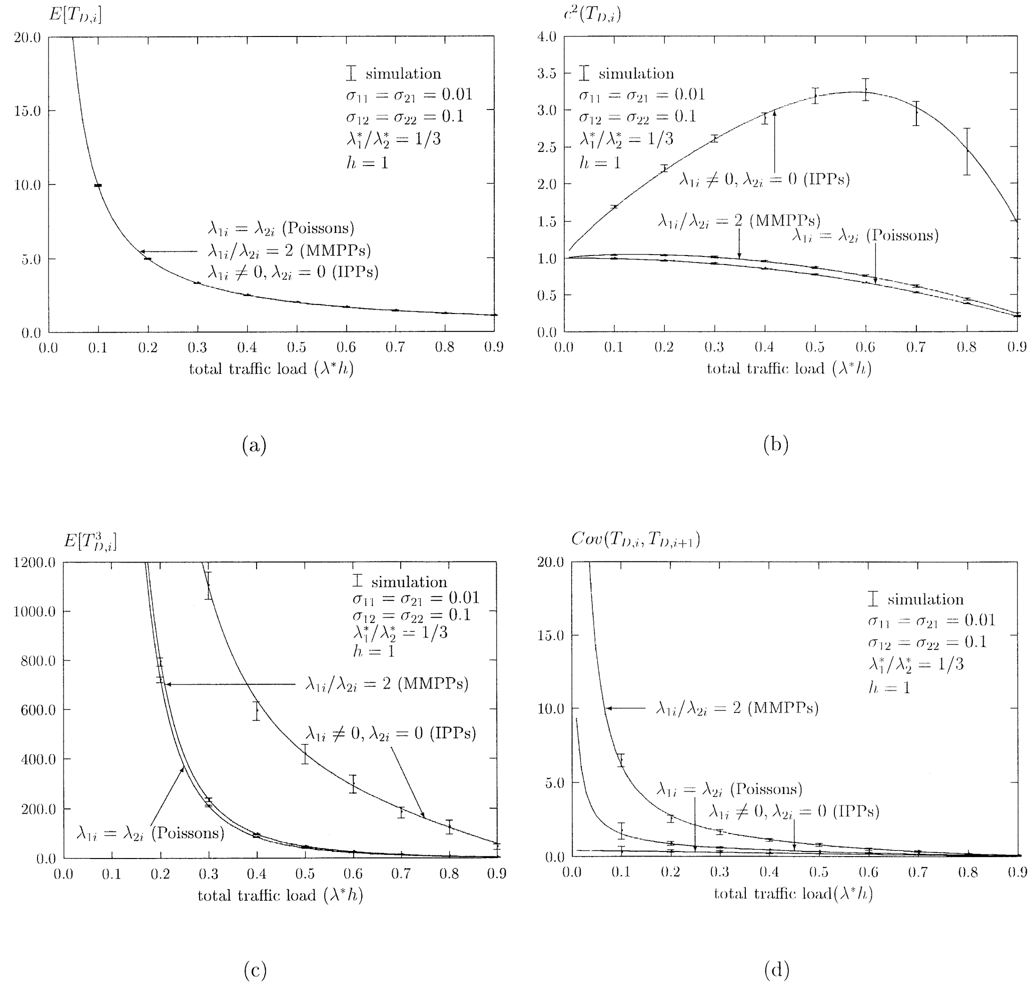
Note that it is possible for different VCs to use partly/entirely the same route, e.g., Fig. 6. The tandem configuration in Fig. 5 can still be employed for such case in an approximate manner if we neglect the correlation between these VCs when performing the recursion to obtain the connection-wise end-to-end performance measures. In Sect. 6.3, we also demonstrate the effectiveness and good accuracy of the approximation through numerical experiments (see Table 3).

6. Numerical Examples and Discussions

In this section, we investigate both the departure statistics and the connection-wise end-to-end performance through numerical experiments. Computer simulations are provided (with 95% confidence interval) to show the accuracy of approximation. For convenience, we use the definition of burstiness in [2], i.e., burstiness $B =$ *Peak rate/ Average rate* in the following discussions. In this work, burstiness $B = \lambda_1(\sigma_1+\sigma_2)/(\sigma_2\lambda_1+\sigma_1\lambda_2)$ for a two-state MMPP with descriptor $(\sigma_1, \sigma_2, \lambda_1, \lambda_2)$ (we assume $\lambda_1 \geq \lambda_2$). It is easy to show that the mean burst duration is in proportion to $1/\sigma_1$ (or $1/\sigma_2$). In the following examples, we set service time $h = T = 1$.

6.1 Composite Departure Statistics

We consider an example of $r = 2$ under $\lambda_1^*/\lambda_2^* = 1/3$, i.e., we feed two sources which are both represented by two-state MMPPs into the slotted deterministic server. Varying total traffic load from 0.01 to 0.9, we obtain the composite departure statistics in Figs. 7(a)–(d) for three different values of burstiness (under $\sigma_{11} = \sigma_{21} = 0.01$ and $\sigma_{12} = \sigma_{22} = 0.1$). In Fig. 7(a), $E[T_{D,i}]$ for MMPPs, Poissons, and IPPs are all equal. This follows the flow conservation we mentioned earlier. In Fig. 7(b), $c^2(T_{D,i})$ decreases in a smooth manner for MMPPs and Poissons as total traffic load increases and $c^2(T_{D,i})$ for MMPPs is slightly above Poissons. For IPPs, $c^2(T_{D,i})$ increases as total traffic load goes from 0.01 to 0.58 and then decreases. We also note that if input traffic is bursty, then $c^2(T_{D,i})$ is large (as shown in Fig. 7(b): $B = 1$ for Poissons, $B = 4/3$ for MMPPs, $B = 2$ for IPPs). The difference between $c^2(T_{D,i})$ of IPPs and MMPPs (or Poissons) is considerably large. In Fig. 7(c), $E[T_{D,i}^3]$



(a) (b)

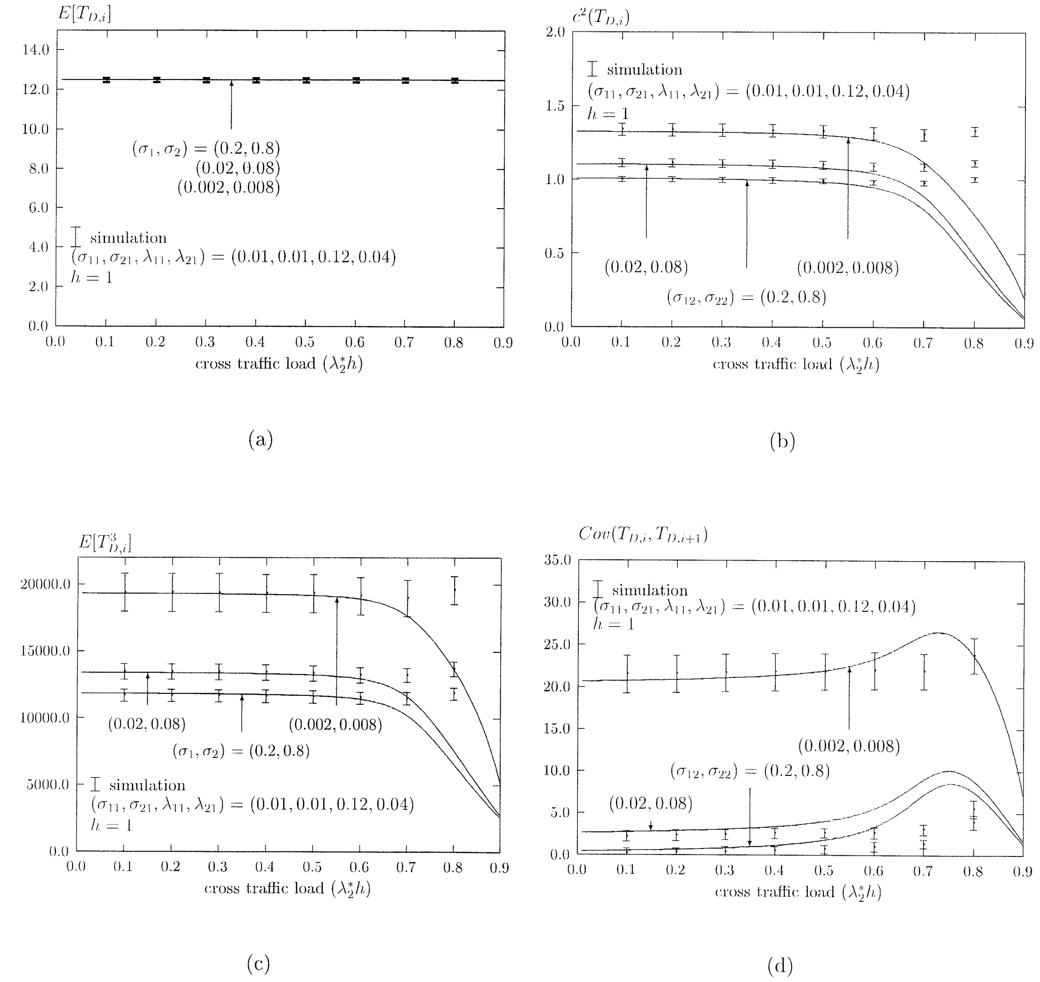
(c) (d)

Fig. 7 Composite departure statistics under three different sets of $(\lambda_{1i}, \lambda_{2i})$ $(i = 1, 2)$: (a) $E[T_{D,i}]$; (b) $c^2(T_{D,i})$; (c) $E[T^3_{D,i}]$;(d) $Cov(T_{D,i}, T_{D,i+1})$ versus total traffic load.

decreases rather sharply for MMPPs, IPPs, and Poissons. We notice that the difference between MMPPs and Poissons is very small while the difference between IPPs and MMPPs (or Poissons) is considerably large. Both $c^2(T_{D,i})$ and $E[T^3_{D,i}]$ become larger if input traffic gets more bursty ($c^2(T_{D,i})$ and $E[T^3_{D,i}]$ for IPPs are the largest while for Poissons they are the smallest). From (21), we know that both IPP and Poisson are renewal while MMPP is not. After going through a queue, we see from Fig. 7(d) that composite output process becomes non-renewal (i.e., $Cov(T_{D,i}, T_{D,i+1}) \neq 0$) due to the discreteness of the queue. $Cov(T_{D,i}, T_{D,i+1})$ decreases as total traffic load increases (and finally behaves more like the deterministic process at heavy total traffic load). $Cov(T_{D,i}, T_{D,i+1})$ decreases almost linearly for Poissons as total traffic load increases and is below that of MMPPs. $Cov(T_{D,i}, T_{D,i+1})$ decreases rather sharply for IPPs and MMPPs.

6.2 Per-Stream Departure Statistics

In many situations, it is important to know the per-stream output process corresponding to a tagged source multiplexed with other cross traffic at the input. We fix the descriptor of the tagged traffic at $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.01, 0.12, 0.04)$ and apply our proposed heuristic decomposition scheme to find the aper-stream departure statistics corresponding to the tagged source. In Fig. 8, we compare the per-stream departure statistics corresponding to the tagged source under three different interfering MMPPs represented by $(\sigma_{12}, \sigma_{22})$ = (0.002,0.008), (0.02,0.08), and (0.2,0.8). We notice the followings according to simulation. First, $E[T_{D,i}]$ maintains a constant level since we feed a fixed tagged traffic source. Second, $c^2(T_{D,i})$, $E[T^3_{D,i}]$, and $Cov(T_{D,i}, T_{D,i+1})$ almost maintain at a constant level regardless of the variation in cross traffic except they go up slightly when the cross traffic load becomes large. Third, $c^2(T_{D,i}) \geq 1$, $Cov(T_{D,i}, T_{D,i+1}) > 0$, i.e., (22)–



(a)

(b)

(c)

(d)

Fig. 8 Per-stream departure statistics: (a) $E[T_{D,i}]$; (b) $c^2(T_{D,i})$; (c) $E[T_{D,i}^3]$;(d) $Cov(T_{D,i}, T_{D,i+1})$ versus cross traffic load.

(25) are satisfied. Hence we can approximate the per-stream output process of a tagged traffic as a two-state MMPP. Fourth, the decomposition scheme works well except for higher utilization. It underestimates the departure statistics except the mean interdeparture time at heavy traffic load. This makes the moment matching scheme fail to work. Since only the mean interdeparture time is valid, the estimated departure process is replaced by a Poisson process.

6.3 Per-Stream Mean End-to-End Sojourn Delay Time

Due to the discrete nature of the H-MMPPs/Slotted D/1 queue, we need the interpolation scheme previously described in Sect. 5 to compensate the sojourn delay time after the first node when applying the recursive calculation of the per-stream end-to-end sojourn delay time. In the following, we show how this scheme works at the second node. Consider a two-node tandem queueing configuration in Fig. 5 in which each node is modeled by a deterministic server fed by a fixed tagged MMPP at the first node described by $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.01, 0.12, 0.04)$ and a cross MMPP at the first and second node, respectively, with traffic load varying from 0.01 to 0.9 under the fixed ratio of $\lambda_{12}/\lambda_{22} (= 2)$ (here, two different sets of cross MMPPs are arranged for comparison by setting $(\sigma_{12}, \sigma_{22})$ at (0.2,0.8) and (0.002,0.008), respectively). In Fig. 9, we show the uncompensated sojourn delay $W_{q,i} + h$ and the mean delay time drift $\triangle W_{q,i,\rho_c}$ s the cross traffic load varies from 0.01 to 0.9. Comparing with the simulation result, the uncompensated sojourn delay time indeed is overestimated except at very heavy traffic load and the overestimated value decreases as the (cross) traffic load increases. As shown in Fig. 9, $\triangle W_{q,i,\rho_c} < 0$ and $-\triangle W_{q,i,\rho_c}$ is roughly equal to the deviation of the overestimated sojourn delay from the simulated value so that the compensated sojourn delay $(W_{q,i} + h + \triangle W_{q,i,\rho_c})$ provides a more accurate result than the uncompensated one. We also note that the compensation effect contributes more at light traffic load than at heavy traffic load. This is equivalent to say the interpolation scheme contributes little when traffic is heavy. One may wonder why the sojourn delay time at heavy traffic load maintains comparatively accurate? This is due to that the decomposition scheme underestimates the per-stream departure statistics (except the mean interdeparture time) and the interpolation scheme does little help. We note that the estimated output process well matches an MMPP when the decomposition scheme works, while a Poisson process is used to replace the estimated output process when moment matching fails to work. It is the use of Poisson process which makes the estimated sojourn delay time (at an intermediate node) remains accurate, not the interpolation scheme. However, accuracy of the (compensated) sojourn delay is better at light traffic load than at heavy traffic load (see Table 1). The error of approximation reaches 5–10% at high traffic load. In the following examples, we shall show the compensated results only.

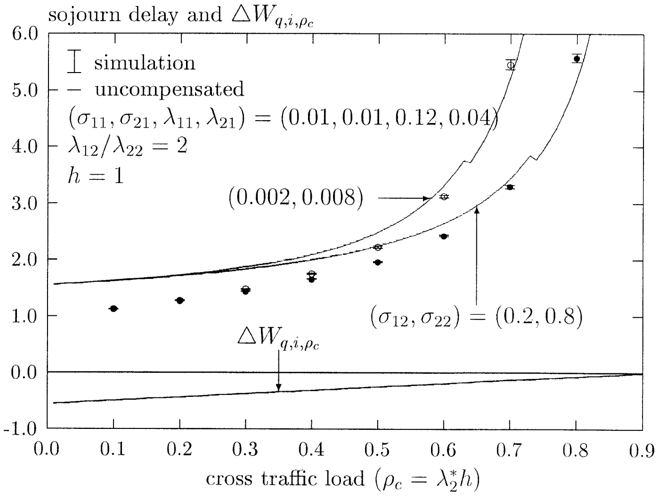


Fig. 9 Uncompensated sojourn delay $W_{q,i} + h$ and mean delay time drift $\triangle W_{q,i,\rho_c}$ at the second node.

Now we consider a tandem queueing configuration of four nodes (see Fig. 5). Each node is modeled as an H-MMPPs/Slotted D/1 queue fed by two sources (one is the tagged traffic and the other denotes the cross traffic). We assume all nodes are synchronized and all cross traffic streams are identical and mutually independent. Consider this model for the scenario of VCs in ATM networks, with tagged information flow parameterized by $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.01, 0.12, 0.04)$. Then we see how the per-stream mean end-to-end sojourn delay time $S_i^{ee} = \sum_{j=1}^{N=4} S_{i,j}$, ($N$: the number of nodes, i: connection identity, and j: intermediate node identity, $S_{i,j}$: per-stream mean sojourn delay time corresponding to connection i at intermediate node j) is affected by the cross traffic in Figs. 10(a)–(b) (when cross traffic load varies from 0.01 to 0.9). In Fig. 10(a), we change the mean burst duration of the cross MMPP by varying $\sigma_{12} = \sigma_{22}$ from 0.01 to 0.1 under $\lambda_{12}/\lambda_{22} = 2$ and $B = 4/3$. The result corresponding to Poissonian cross traffic is also included. The cross MMPP causes the per-stream mean end-to-end sojourn delay time to become larger than Poisson. Cross traffic with long mean burst duration makes the per-stream mean end-to-end sojourn delay time large. Table 1 gives the results for the case of $\sigma_{12} = \sigma_{22} = 0.01$ for detailed numerical comparison. From this table, we note that the approximation errors are about 0.8% (0.6%) for $\rho_c = 0.2$ and −5.4% (−5.0%) for $\rho_c = 0.8$ at the second node (for the total sojourn delay). In Fig. 10(b), we change the mean burst duration of the cross MMPP with $B = 10/9$ by varying $(\sigma_{12}, \sigma_{22})$ from $(0.002, 0.008)$ to $(0.2, 0.8)$ under $\lambda_{12}/\lambda_{22} = 2$. Results corresponding to Poissonian and IPP (with $B = 5/4$) cross traffic under $(\sigma_{12}, \sigma_{22}) = (0.02, 0.08)$ are also included. We observe that under fixed σ_{12}/σ_{22}, IPP and Poisson respectively serve as

Note that it is possible for different VCs to use

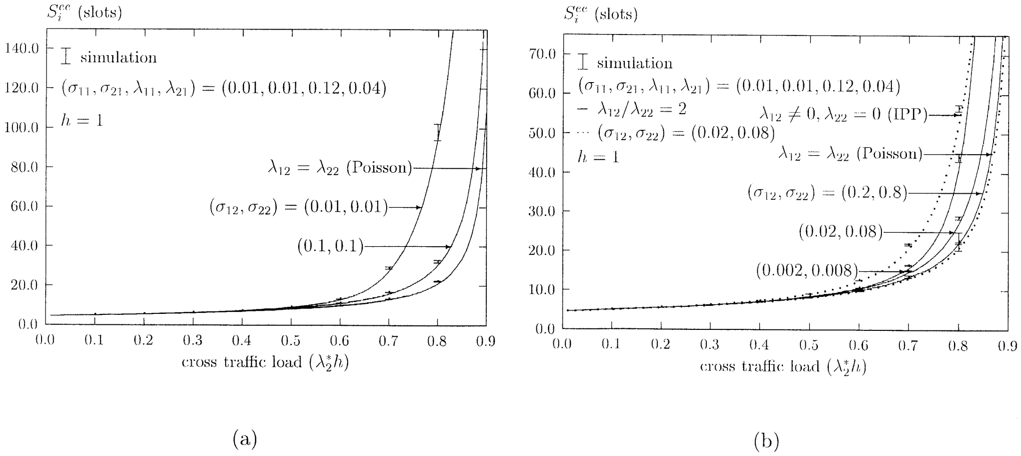


Fig. 10 Per-stream mean end-to-end sojourn delay time S_i^{ee} versus cross traffic load of a four-node tandem configuration accepting two sources: (a) cross traffic with $\sigma_{12} = \sigma_{22}$; (b) cross traffic with $\sigma_{12}/\sigma_{22} = 0.25$.

Table 1 Delay at each node of the four-node tandem queues in Fig. 5.

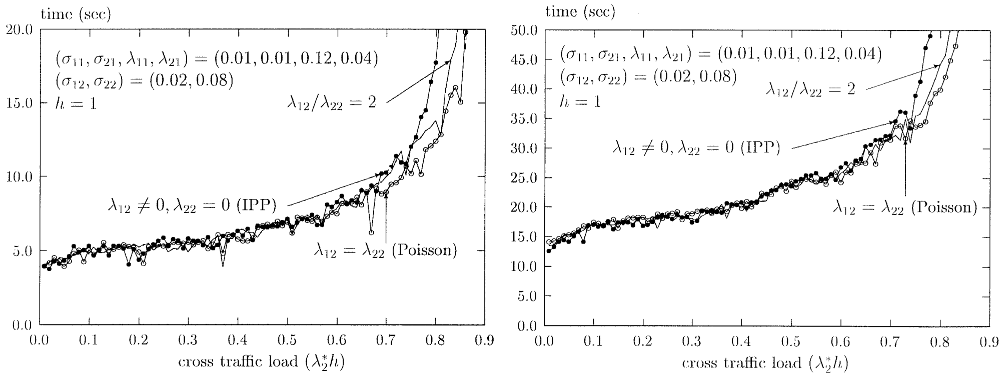
ρ_c		Node 1	Node 2	Node 3	Node 4	Total
0.2	Simulation	1.724±0.2%	1.276±0.2%	1.273±0.5%	1.270±0.4%	5.544±0.2%
	Analysis	1.7219	1.2860	1.2853	1.2847	5.5779
0.4	Simulation	2.079±0.3%	1.743±0.8%	1.728±0.6%	1.718±0.5%	7.269±0.3%
	Analysis	2.0746	1.7559	1.7521	1.7486	7.3312
0.6	Simulation	3.552±1.5%	3.301±2.0%	3.248±1.7%	3.204±1.0%	13.305±0.9%
	Analysis	3.5062	3.2873	3.0986	3.0986	12.9908
0.8	Simulation	24.946±7.0%	24.316±5.3%	24.648±5.4%	24.095±7.1%	98.005±4.2%
	Analysis	24.1057	22.9920	22.9920	22.9920	93.0816

Table 2 Traffic parameters employed in the example of Fig. 6.

	Node 1	Node 2	Node 3	Node 4
VC$_1$ $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21})$	(0.01,0.01,0.12,0.04)	—	—	—
VC$_2$ $(\sigma_{12}, \sigma_{22}, \lambda_{12}, \lambda_{22})$	(0.01,0.01,0.10,0.06)	—	—	—
cross VCs $(\sigma_{1c}, \sigma_{2c}, \lambda_{1c}, \lambda_{2c})$	(0.01,0.01,$\frac{4\rho_c}{3}$,$\frac{2\rho_c}{3}$)	(0.01,0.01,$\frac{4\rho_c}{3}$,$\frac{2\rho_c}{3}$)	(0.01,0.01,$\frac{4\rho_c}{3}$,$\frac{2\rho_c}{3}$)	(0.01,0.01,$\frac{4\rho_c}{3}$,$\frac{2\rho_c}{3}$)

Table 3 Delay at each node of the four-node tandem queues in Fig. 6.

ρ_c			Node 1	Node 2	Node 3	Node 4	Total
0.2	VC$_1$	Simulation	1.821±0.2%	1.325±0.3%	1.317±0.3%	1.314±0.4%	5.776±0.2%
		Analysis	1.8181	1.4296	1.4277	1.4266	6.1020
	VC$_2$	Simulation	1.800±0.1%	1.314±0.3%	1.308±0.2%	1.306±0.2%	5.729±0.1%
		Analysis	1.7997	1.4179	1.4142	1.4140	6.0459
0.4	VC$_1$	Simulation	2.318±0.4%	1.931±0.5%	1.907±0.5%	1.900±0.6%	8.055±0.4%
		Analysis	2.3178	2.0395	2.0321	1.9819	8.3714
	VC$_2$	Simulation	2.278±0.4%	1.897±0.6%	1.876±0.3%	1.874±0.5%	7.925±0.2%
		Analysis	2.2732	1.9932	1.9917	1.9819	8.2400
0.6	VC$_1$	Simulation	5.320±2.2%	4.990±2.0%	4.896±1.0%	4.86±2.0%5	20.071±0.7%
		Analysis	5.2587	4.6809	4.6809	4.6809	19.3015
	VC$_2$	Simulation	5.118±1.5%	4.832±1.8%	4.7370±1.8%	4.734±1.8%	19.422±0.8%
		Analysis	5.0606	4.6809	4.6809	4.6809	19.1034
0.8	VC$_1$	Simulation	101.6±15%	96.93±6.9%	88.28±16%	91.05±12%	377.8±6.2%
		Analysis	93.293	90.295	90.295	90.295	364.18
	VC$_2$	Simulation	99.43±13%	94.99±6.4%	87.25±13%	93.55±13%	375.2±5.3%
		Analysis	92.663	90.295	90.295	90.295	363.55



(a)

(b)

Fig. 11 Computation time: (a) for the decomposition scheme; (b) for the calculation of a four-node tandem queueing configuration.

partly/entirely the same route as mentioned earlier. Hence the tandem configuration in Fig. 5 can be employed for such case in an approximate manner if we may neglect the correlation between these VCs when performing the recursion to obtain the connection-wise end-to-end performance measures. We now consider a case in Fig. 6 where two VCs share entirely the same route and all other cross VCs are assumed to be still identical and mutually independent. For this configuration with the traffic parameter sets in Table 2, we obtain delays along both routes VC$_1$ and VC$_2$ in Table 3. Table 3 says that the approximation is still good since there is merely about 8.6% error for the worst case.

The above examples have demonstrated the accuracy of our proposed method in conducting end-to-end performance analysis. Let us further examine the efficiency regarding the computation time consumed. The following examples are run on a *Sun SPARC-20E* workstation using Matlab Ver.5.1. Shown in Fig. 11(a) is the CPU time used to find the value of h_{eff} under the following parameters: the tagged traffic fixed at $(\sigma_{11}, \sigma_{21}, \lambda_{11}, \lambda_{21}) = (0.01, 0.01, 0.12, 0.04)$, at every node the cross traffic fixed only for $(\sigma_{12}, \sigma_{22}) = (0.02, 0.08)$ but the traffic load varies from 0.01 to 0.9 under three different traffic sources, i.e., MMPP ($\lambda_{21}/\lambda_{22} = 2$), IPP ($\lambda_{12} \neq 0$, $\lambda_{22} = 0$), and Poisson ($\lambda_{12} = \lambda_{22}$). Note that the above traffic arrangement has been employed in previous examples (i.e., in Fig. 10). The decomposition scheme needs about 4–10 seconds when the traffic load is lower than 0.7 and up to 10–50 seconds when the traffic load goes beyond 0.7. Using the above traffic arrangement, Fig. 11(b) shows the CPU time consumed for a four-node end-to-end calculation. The CPU time falls in the range 12–30 seconds when the traffic load varies from 0.01 to 0.7. At extremely heavy traffic load, the calculation even consumes more than one minute. This seems unsatisfactory for practical CAC need. However, if faster processors such as *Sun UltraSPARC* are used, the computation time may be reduced by a factor (at least five to ten) to make the proposed method usable on-line.

7. Conclusion

In this paper, we have investigated the composite output process of the H-MMPPs/Slotted D/1 queueing system and proposed a simple heuristic decomposition scheme to obtain the per-stream output process corresponding to a tagged traffic stream. This decomposition scheme can also be used in other queueing systems. A moment matching scheme used to emulate the output process as a two-state MMPP is also provided. These results can be easily extended to obtain the composite end-to-end sojourn delay time or per-stream end-to-end sojourn delay time.

Our results enable us to conduct end-to-end performance analysis for each VC in ATM networks. It can be of course extended to evaluate the overall network performance since a network consists of a collection of individual VCs. In other words, we offer a powerful analytical method to call admission control, congestion control and routing algorithms in ATM networks.

References

[1] H.W. Ferng and J.F. Chang, "Per-stream nodal delay analysis in ATM networks," Research Note, Dept. Elect. Eng., National Taiwan University, 1997.

[2] D.E. McDysan and D.L. Spohn, ATM Theory and Application, McGraw-Hill, New York, 1995.

[3] R.G. Addie and M. Zukerman, "Queueing performance of a tree type ATM network," IEEE INFOCOM'94, Toronto, Canada, pp.48–55, June 1994.

[4] H. Kroner, M. Eberspacher, T.H. Theimer, P.J. Kuhn, and U. Briem, "Approximate analysis of the end-to-end delay in ATM networks," IEEE INFOCOM'92, Florence, Italy, pp.978–986, May 1992.

[5] J.F. Ren, J.W. Mark, and J.W. Wong, "End-to-end performance in ATM networks," IEEE ICC'94, New Orleans, USA, pp.996–1002, May 1994.

[6] Y. Ohba, M. Murata, and H. Miyahara, "Analysis of interdeparture processes for bursty traffic in ATM networks," IEEE J. Sel. Areas Commun., vol.9, no.3, pp.468–476, April 1991.

[7] D. Park, H.G. Perros, and H. Yamashita, "Approximate analysis of discrete-time tandem queueing networks with bursty and correlated input traffic and customer loss," Oper. Res. Lett., vol.15, pp.95–104, 1994.

[8] H. Saito, "The departure process of an N/G/1 queue," Performance Evaluation, vol.11, pp.241–251, 1990.

[9] T. Takine, T. Suda, and T. Hasegawa, "Cell loss and output process analyses of a finite-buffer discrete-time ATM queueing system with correlated arrivals," IEEE Trans. Commun., vol.43, no.2/3/4, pp.1022–1037, Feb./March/April 1995.

[10] L. Kleinrock, Queueing Systems, vol.I: Theory, Wiley, New York, 1975.

[11] H. Heffes and D.M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," IEEE J. Sel. Areas Commun., vol.SAC-4, no.6, pp.856–868, Sept. 1986.

[12] H. Heffes, "A class of data traffic processes—Covariance function characterization and related queuing results," Bell Syst. Tech. J., vol.59, no.6, pp.897–929, July-Aug. 1980.

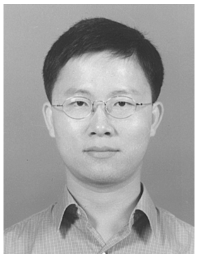
[13] W. Fischer and K.S. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," Performance Evaluation, vol.18, pp.149–171, 1993.

[14] D.M. Lucantoni and V. Ramaswami, "Efficient algorithms for solving the non-linear matrix equations arising in phase type queues," Commun. Statist.–Stochastic Models, vol.1, no.1, pp.29–51, 1985.

[15] M.F. Neuts, Structured Stochastic Matrices of $M/G/1$ Type and Their Applications, Dekker, New York, 1989.

[16] V. Ramaswami, "Nonlinear matrix equations in applied probability—Solution techniques and open problems," SIAM Rev., vol.30, no.2, pp.256–263, June 1988.

[17] A.M. Law and W.D. Kelton, Simulation Modeling & Analysis, 2 ed., McGraw-Hill, New York, 1991.



Huei-Wen Ferng was born in Taiwan, R.O.C., on September 5, 1970. He received the B.S.E.E. degree from the National Tsing Hwa University, Hsinchu, Taiwan in 1993. He is currently working toward the Ph.D. degree in electrical engineering at the National Taiwan University, Taipei, Taiwan. His research interests include queueing theory, teletraffic modeling, performance analysis in ATM networks.

Jin-Fu Chang was born in Taiwan, R.O.C., on March 9, 1948. He received the B.S.E.E. degree from the National Taiwan University, Taipei, Taiwan, in 1970 and the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1977. He is currently Vice Chairman of the National Science Council, Taiwan. He was a Professor of Electrical Engineering at the National Taiwan University for a number of years 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, U.K. From August 1991 to July 1994, he was at the National Central University as Dean of Academic Affairs. 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 Ten 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. He is also a Member of the Association of Computing Machinery. 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. Dr. Chang is a Fellow of IEEE.