Delay Distributions in Discrete Time Multiclass Tandem Communication Network Models

Original Scientific Paper

Nandyala D. Gangadhar
Department of Computer Science and Engineering, Faculty of Engineering and Technology,
M. S. Ramaiah University of Applied Sciences
470-P Phase IV, Peenya, Bengaluru 560058, India
n_d_gangadhar@ieee.org; gangadhar.cs.et@msruas.ac.in

Govind R. Kadambi
Department of Electronics and Communication Engineering, Faculty of Engineering and Technology,
M. S. Ramaiah University of Applied Sciences
470-P Phase IV, Peenya, Bengaluru 560058, India
pvc.research@msruas.ac.in

Abstract *– An exact computational algorithm for the solution of a discrete time multiclass tandem network with a primary class and cross-traffic at each queue is developed. A sequence of truncated Lindley recursions is defined at each queue relating the delays experienced by the first packet from consecutive batches of a class at that queue. Using this sequence of recursions, a convolve-and-sweep algorithm is developed to compute the stationary distributions of the delay and inter-departure processes of each class at a queue, delays experienced by a typical packet from the primary class along its path as well as the mean end-to-end delay of such a packet. The proposed approach is designed to handle the non-renewal arrival processes arising in the network. The algorithmic solution is implemented as an abstract class which permits its easy adaptation to analyze different network configurations and sizes. The delays of a packet at different queues are shown to be associated random variables from which it follows that the variance of total delay is lower bounded by the sum of variances of delays at the queues along the path. The developed algorithm and the proposed lower bound on the variance of total delay are validated against simulation for a tandem network of two queues with three classes under different batch size distributions.*

Keywords*: Delay, Tandem, Communication, Network, Discrete Time, Queueing, Algorithm, Lindley Recursion, End-to-end*

1. INTRODUCTION

Communication networks carry packet traffic through connections between multiple source-destination pairs. Traffic from each connection passes through a network path consisting of a sequence of intermediate nodes and faces contention at each of these nodes from other connections. Delays experienced by the connection along the path are random, and their characterization and estimation are crucial measures of network performance. For this purpose, the connection can be modelled as a tandem network of queues [1-5].

In discrete time queues and their network models, the time axis is divided into equal intervals, termed "slots". In each slot, a batch of packets from each of the sources is generated and enters the queue or network and one or more packets leave the queue or network [3-4].

In a network, the joint distribution of queue lengths and delays as well as the end-to-end delay are useful measures of network performance but have proved to be hard problems to solve [2]. Analogous to the analysis of their continuous time counterparts, decomposition of the stationary joint distribution of the queue lengths at different queues in a discrete time network into a product of marginal distributions is shown for a large class of networks [6]. To employ this for obtaining the delays, one needs to establish arrival theorems which are known for only special models of networks [7]. However, product-form decomposition is employed as a heuristic approximation in general networks. This paper addresses the problem of computing the delays in a discrete tandem network with batch arrival processes and cross-traffic. An exact algorithm for computing the distributions of delays at each queue as well as the mean end-to-delay of a typical packet is developed, along with a lower bound for the variance of the end-to-end delay.

Different models of tandem queueing networks with and without arrivals and with and without departures at the intermediate nodes are studied in the literature. In [8], a tandem network of two queues in continuous

time with Poisson arrivals only at the first queue and identical and independent service times is studied. Joint delays in a continuous time tandem network without intermediate arrivals are analyzed in [9]. In [10], an approximate analysis of the queue lengths and busy periods at each node in a discrete time tandem network with arrivals and departures at intermediate nodes is carried out. The departures are modelled as random packet drops after service. The authors of [11] consider a model identical to the model analyzed in the current paper and propose recursive algorithms for the computation of delay distributions without presenting any results. In the current paper, a sequence of truncated Lindley processes is introduced for expressing the delays in the queues of the tandem network, and a convolve-and-sweep algorithm is devised for computing the delay distributions. This approach has the advantage that the algorithm does not require renewal arrival processes and hence can be applied to all the queues in the tandem network.

Discrete time tandem network with arrivals and departures at intermediate nodes under Furthest-to-Go service discipline is analyzed in [12] and expressions for queue length distributions are obtained. The components of end-to-end delay in a Software Defined Network (SDN) are modelled and experimentally estimated in [5].

Apart from computational approaches, simulation and bounds are also employed in the analysis of tandem networks. Simulation analysis of the performance of high-speed networks is carried out in [13-14]. In [15], the output process of a GI/GI/1 queue is approximated by a renewal process and this is applied to a tandem network of queues without interfering traffic. This approach is applied in [16] to analyze general discrete time networks. Worst case bounds on the end-to-end delay under active queue management scheduling algorithms are derived in [17]. A general network calculus approach to the end-to-end analysis of queueing networks when the inputs are modelled as deterministic or stochastic affine envelop processes is developed in [18-19] and the same has been applied to SDN in [20-21].

Computation of the end-to-end delay in a network is feasible for product-form networks wherein the stationary joint distribution of delays at various queues factors into the product of the marginal distributions. This follows if a stronger assumption that the delays in the queues are independent random variables is made. The product-form decomposition and independence assumption are commonly employed as a heuristic for end-to-end analysis. Lower bounds on the moments of total delay are established by showing that the delays in the individual queues are associated *random variables* [22]. This property is proved in [23] for a tandem network wherein the service times at all but the last server are all a constant. In [23] it is employed to derive an upper bound on the mean total delay. In the current paper, it is established for the tandem network model considered and is employed to obtain a lower bound on the variance of the total delay. The exact mean total delay is computed as the sum of computed marginal distributions.

The rest of the paper is organized as follows: Section 2 describes the model, notation employed and analysis of a first queue in the tandem network as a single discrete queue highlighting the need for careful analysis of the other queues. Section 3 begins by defining a sequence of truncated Lindley processes to recursively relate the delays experienced by the first packet from two consecutive batches of Class 0 at each queue in the tandem network. A computational algorithm is derived from these recursions by adapting the convolve-and-sweep algorithm to the truncated Lindley processes. The delays of a typical packet in the queues are shown to be associated random variables providing a lower bound on the variance of the end-to-end delay. This section also describes the details of implementation and benchmark simulations. The results of computational and their validation via simulation are presented in Section 4. Section 5 concludes the paper with some directions for future work. Appendix 1 presents a proof of Theorem 1 stated in Section 3.

2. MODEL, NOTATION AND PRELIMINARIES

This section presents the tandem discrete time queueing network model, notation employed and approaches to the analysis of a single discrete queue. The model consists of discrete time queues in series, each with arrivals and departures. The tandem network models the path of a packet stream of interest in a general network.

2.1 TANDEM NETWORK MODEL

The discrete time model analyzed in this paper consists of a tandem of one or more discrete time queues and multiple streams of traffic. Each stream can traverse multiple queues. Figure 1 depicts an instance of such a tandem network with two queues Q1 and Q2, and three streams of traffic that are designated as belonging to Classes 0, 1 and 2. Class 0 traffic enters the first queue Q1 and passes through both the queues before leaving the system. Traffic from Class i, i=1,2 enters the ith queue and leaves the system after service at that queue. Classes 1 and 2 can be considered as "cross-traffic" at queues Q1 and Q2, respectively.

Packets from all the classes are assumed to be of the same size and the servers at all the queues have identical service rates. Thus, all the jobs need the same amount of server time at the queues. This constant amount of time is taken as a unit of time and time is discretized into slots of this duration. The system works in discrete time: all the arrivals to a queue arrive at the beginning of the slot and the first job (after ordering for service) will be served at the beginning of the earliest slot that the server is free. The arrivals from each Class

occur as a single batch with a general batch size distribution and it is assumed that the batches are ordered for service according to a stationary policy. One can in principle remove this assumption at the cost of a cumbersome analysis.

The interval $[k,k+1)$ is termed the k^{th} slot. The traffic from each class is assumed to arrive as batches per slot with the batch size following an iid distribution

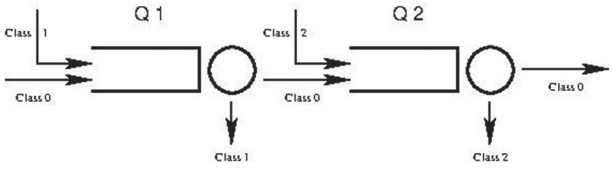

Fig. 1. A Tandem Queueing Network with Cross Traffic

2. 2 NOTATION

The following notation is used throughout the paper:

Qi : The i^{th} Queue in the tandem

$X_{k}^{(i)}$: Number of Class i arrivals in the k^{th} slot

$A_{j,k}^{(i)}$: Arrival slot of the first job of Class i at Qj

$\Delta_{j,k}^{(0)}$: Interval between the arrival of the k^{th} and $(k+1)^{st}$ batch of Class 0 at Qj

$W_{i,k}$: Workload of Qi at the beginning of k^{th} slot

$D_{j,k}^{(0)}$: Delay of the first job of Class the k^{th} batch of Class 0 at Qj

2.3. PRELIMINARIES

The workload $W_{1,k}$, evolves as follows:

$$W_{1,k+1} = \left(W_{1,k} + X_k^{(0)} + X_k^{(1)} - 1\right)^+ \tag{1}$$

where, $x^+=\max\{x,0\}$. Indeed, the total arrivals in the k_{th} slot are $X_k^{(0)}+X_k^{(1)}$ and one packet gets served if there are a non-zero number of packets, leaving $W_{1,k+1}$ packets at the beginning of the $(k+1)^{st}$ slot. The Markov Chain $\{W_{1,k}, k\geq 0\}$ has a unique stationary distribution π under the stability condition $E(X_k^{(0)}+X_k^{(1)})<1$ and can be computed using Ramaswamy Algorithm [25]. By the arrival theorem for geometric arrivals, $D_{j,k}^{(0)}=W_{1,k}\sim\pi$. The delay distribution of a typical Class 0 packet in the k^{th} batch can then be obtained as

$$Y_{j,k}^{(0)} = D_{j,k}^{(0)} + \theta \tag{2}$$

where θ is the rank of the typical packet in its batch.

This analysis does not extend to the rest of the queues in the tandem since the arrival stream from Class 0 packets is non-renewal. Hence, a new approach for developing a computational solution for the tandem networks is needed, as developed in Section 3.

3. COMPUTATIONAL SOLUTION AND ALGORITHM DEVELOPMENT

Eq. (1) is a Lindley Recursion [5, 24] of the form

$$Z_{n+1} = (Z_n + X_n - 1)^+ \tag{3}$$

and the process $\{Z_n\}$ is called a Lindley process. The convolve-and-sweep algorithm for computing the distribution of the Lindley process is given by:

$$P(Z_{n+1} = m) = \begin{cases} P(Z_n = 0)P(X_b \leq 1) \\ + P(Z_n = m)P(X_b = 0), \quad m = 0 \\ \sum_{l=0}^{m+1} P(Z_n = l)P(X_b = m+1-l), \\ m > 0 \end{cases} \tag{4}$$

where it is assumed that Z_0 is independent of $\{X_k\}$.

In this paper, Lindley Recursion and the convolve-and-sweep algorithm will be used to derive a sequence of *truncated* Lindley processes defined by generalizing Eq. (3) (Cf. Eqs. (5)-(7) and (9)-(12)) and build a computational algorithm for obtaining the distributions of delays of each class of packets using the convolve-and-sweep algorithm expressed by Eq. (4).

The advantage of computing the stationary distributions of delays using the convolve-and-sweep algorithm is that no independence assumptions on the arrival processes are required. This is essential for computing the distributions of delays at Q2 and other downstream nodes where the arrival process from Class 0 is not a renewal process. The stationary distributions of delays are computed as limits of the transient distributions.

3.1 TRUNCATED LINDLEY PROCESSES

From the Lindley process $\{W_{1,k}\}$ of Eq. (1), a sequence of truncated Lindley processes is defined:

$$W_{1,k+1}(0) := D_{1,k}^{(0)} + X_k^{(0)} + X_k^{(1)} - 1 \tag{5}$$

$$W_{1,k+1}(l) := \left(W_{1,k+1}(l-1) + X_{A_{1,k}^{(0)}}^{(1)} - 1\right)^+ \tag{6}$$

for l=1, 2,..., $\Delta_{1,k}^{(0)}$-1, using the notation of Sec. 2. Then, the delay of the first packet in the $(k+1)^{st}$ batch of Class 0, $D_{1,k+1}^{(0)}$, is given by:

$$D_{1,k+1}^{(0)} = W_{1,k+1}\left(\Delta_{1,k}^{(0)} - 1\right) \tag{7}$$

Eqs. (5)-(7) can be used to compute the transient distributions of the delays $\{D_{1,k}^{(0)}; k=0,1,\ldots\}$ by extending the convolve-and-sweep algorithm for the standard Lindley Recursion to the sequence of truncated Lindley processes, $\{W_{1,k}(l)\}$. Since $\Delta_{1,k}^{(0)}$ is random, the computation is carried out by conditioning on $\Delta_{1,k}^{(0)}=l$, l=1,2,…, using the distribution of $\Delta_{1,k}^{(0)}$.

3.2 INTER-DEPARTURE DISTRIBUTION FROM Q1

For computing the delay distributions at Q2, the distribution of the inter-arrival time between the first jobs of Class 0 from two successive batches of arrivals at Q1 that have at least one Class 0 job is needed. For this, the following relation is used:

$$\begin{aligned}\Delta_{2,k}^{(0)} &= A_{1,k+1}^{(0)} + D_{1,k+1}^{(0)} - \left(A_{1,k}^{(0)} + D_{1,k}^{(0)}\right) \\ &= D_{1,k+1}^{(0)} - D_{1,k}^{(0)} + \Delta_{1,k}^{(0)}\end{aligned} \tag{8}$$

Eq. (8) follows from the fact that the departure times of the first packet from k^{th} and $(k+1)^{\text{st}}$ Class 0 batches are $A_{1,k}{}^{(0)} + D_{1,k}{}^{(0)}$ and $A_{1,k+1}{}^{(0)} + D_{1,k+1}{}^{(0)}$, respectively, and $\Delta_{j,k}{}^{(0)} = A_{j,k+1}{}^{(0)} - A_{j,k}{}^{(0)}$, j=0,1, by definition.

3.3 DELAY AND INTER-DEPARTURE DISTRIBUTIONS AT Q2

As observed above, the computation of delay distributions at Q1 using Eqs. (5)-(7) can be applied even when the arrival processes are non-renewal as is the case for the Class 0 process at Q2. The corresponding truncated Lindley processes at Q2 are now derived.

To simplify the notation, it is assumed that the Class 0 jobs are arranged to be at the beginning of all the arrivals in a slot at Q2. In Q2, the evolution is given by slight modifications of Eqs. (5)-(7) since now Class 0 packets from its k^{th} batch form a train of $X_{A_{1,k}^{(0)}}^{(1)}$ number arrivals at Q2 starting from slot $A_{2,k}{}^{(0)}$. The workload process evolution is characterized by the Eqs. (9)-(11) below (Cf. Eqs. (5)-(7)):

$$W_{2,k}(0) = D_{2,k}^{(0)} + Z_k^1 + X_{A_{2,k}^{(0)}+X_{A_{1,k}^{(0)}}^{(0)}}^{(1)} - 1 \tag{9}$$

where

$$Z_k^1 = X_{A_{2,k}^{(0)}}^{(1)} + X_{A_{2,k}^{(0)}+1}^{(1)} + \cdots + X_{A_{2,k}^{(0)}+X_{A_{1,k}^{(0)}}^{(0)}-1}^{(1)} \tag{10}$$

and

$$W_{2,k+1}(l) = \left(W_{2,k}(l-1) + X_{A_{2,k}^{(0)}+X_{A_{1,k}^{(0)}}^{(0)}+l}^{(1)} - 1 \right)^+ \tag{11}$$

for l=1,2,…,$\Delta_{2,k}{}^{(0)}$-1. Eq. (10) follows from the fact that, in slots $A_{2,k}^{(0)}, A_{2,k}^{(0)}+1, \ldots, A_{2,k}^{(0)}+X_{A_{1,k}^{(0)}}^{(0)}-1$, there is a single Class 0 packet arriving and being served at Q2. Then, as in Eq. (7), the delay is given by

$$D_{2,k+1}^{(0)} = W_{2,k}\left(\Delta_{2,k}^{(0)} - 1\right) \tag{12}$$

It can be observed that Eq. (10) has a variable number, $X_{A_{1,k}^{(0)}}^{(1)}$ of terms. In addition, $\Delta_{2,k}{}^{(0)}$ is positively correlated with this number, $X_{A_{1,k}^{(0)}}^{(1)}$. Hence, the computation of $D_{2,k+1}{}^{(0)}$ needs to condition on $X_{A_{1,k}^{(0)}}^{(1)}$. With this modification, the computational procedure developed for computing the stationary distribution of delays $D_{2,k+1}{}^{(0)}$ as well as that of the inter–departure times $\Delta_{3,k}{}^{(0)}$, at Q1, carries over to Q2.

3.4 SOLUTION OF TANDEM NETWORK

The computational solutions developed for delays and inter-departure distributions at Q2 can be employed at each of the downstream queues in the tandem network with changes in the input processes. The algorithm for computing the tandem network processes is listed as Algorithm 1.

Algorithm 1

Input: Batch size distributions for Classes with external arrivals, $X_0{}^{(i)}$ and initial queue lengths $W_0{}^{(i)}$

Output: Stationary distributions of delays $D_{j,\infty}{}^{(i)}$ and inter-departure times $\Delta_{j,\infty}{}^{(i)}$ at Qj

Initialization:

a) Set precision value EPSILON (typical: 1e-06)

b) Set the distributions $\{D_{j,0}{}^{(i)}\}$ and $\{\Delta_1{}^{(i)}\}$, for all Classes i to degenerate distribution δ_0

c) Initialize the distribution of batch inter-arrival time $\Delta_{1,\infty}{}^{(0)}$ to Geom(p=$P(X_1{}^{(0)}>0)$)

Iteration:

For each queue Qj, j = 1,2,…, in the tandem do:

For k=1,2,…until convergence do:

Compute the distribution of $D_{j,k+1}{}^{(0)}$ using

Eqs. (5)-(7) for Q1 or Eq. (9)-(12) for Q2, Q3, …

End

Compute the distribution of $\Delta_{j,\infty}{}^{(0)}$ using Eq. (8)

End

The iteration for the computation of the stationary distributions$\{D_{j,\infty}{}^{(0)}\}$ in Algorithm 1 is repeated until a) The CDF of the computed distribution is greater than 1.0-EPSILON, and b) The l_∞ distance between the computed successive distributions is less than EPSILON.

3.5 ASSOCIATION OF DELAYS IN THE TANDEM

The *total delay* in the tandem network is an important performance measure in communication networks. To compute it, the distribution of the *sum* of delays at individual queues in the tandem is needed. Algorithm 1 does not provide this since only the *marginal* distributions of delays are computed.

On the other hand, most network monitoring and control applications employ the moments of the total delay. The mean of the total delay is the sum of mean delays in the individual queues; the latter quantities are readily computed from the obtained marginal distributions of the delays. Regarding the variance of the total delay, the following result is established to provide a computable lower bound:

Theorem 1. The random variables corresponding to the delays experienced by an arbitrary Class 0 job at Q1 and Q2 are associated random variables [24]; i.e., using Eq. (2), $\mathrm{Cov}(f(Y_{1,k}^{(0)}), g(Y_{2,k}^{(0)}) \geq 0$ for all increasing functions f and g.

The theorem is proved in the Appendix.

In particular, the theorem implies $\mathrm{Cov}(Y_{1,k}^{(0)}, Y_{2,k}^{(0)}) \geq 0$. This is used to compute a lower bound on the variance of the total delay $Y_{1,k}^{(0)} + Y_{2,k}^{(0)}$:

$$\mathrm{Var}(Y_{1,k}^{(0)} + Y_{2,k}^{(0)}) \geq \mathrm{Var}(Y_{1,k}^{(0)}) + \mathrm{Var}(Y_{2,k}^{(0)}) \quad (13)$$

3.6 IMPLEMENTATION OF ALGORITHM 1

The computational Algorithm 1 for analysis at each queue in a tandem network is implemented in Python programming language. The iterative computation of the distributions of the delays $\{D_{j,k}^{(0)}\}$ at Q1 using the truncated Lindley recursions given by Eqs. (5)-(7) is implemented abstractly so that it can be employed for both Q1 and downstream queues with suitable parametrization. The abstraction is implemented in an Object-Oriented fashion as a generic Q class with methods tabulated in Table 1.

Q1, Q2, Q3, …, are realized by sub-classing the generic Q class. When instantiated as Q1, the given inter-batch arrival time distribution of Class 0 (resp, Class 1) traffic is returned by the method Delta1. The method Zn_Zn1 implements the Lindley Recursion Eq. (3) for computing the distribution of Z_{n+1} from that of Z_n; it starts with a call to method Z0_Z1 for incorporating a call to the method Delta1. Using Zn_Zn1, the result of l-step convolve-and-sweep $D_{1,l}^{(0)}$ (resp. $D_{1,l}^{(1)}$), starting from $D_{1,0}^{(0)}$ (resp. $D_{1,0}^{(1)}$) and conditioned on the stationary distribution of $\Delta_{1,k}^{(0)}$ (resp. $\Delta_{1,k}^{(1)}$) being l, is carried out by Dn_Dn1. Dn_Dn1 is applied iteratively till convergence to compute the stationary distribution of the delays $D_{1,0}^{(0)}$ (resp. $D_{1,0}^{(1)}$). Finally, the inter-batch *departure* distribution $\Delta_{2,\infty}^{(0)}$ is computed by Delta2 from Eq. (8).

For Q2, this inter-batch arrival distribution of Class 2 is computed by Delta1 and the Algorithm 1 is executed with Class 0 and Class 2 as inputs. This procedure is repeated for the rest of the downstream nodes.

All the methods in Table 1 work with discrete probability distribution functions. To facilitate this, a Python library for definition and manipulation of discrete probability distributions with finite or infinite integer support (spanning positive and negative axis) is developed. The library consists of Dist Class and a set of functions on Dist objects as listed in Table 2.

Specific probability distributions are defined by subclassing Dist. Two of them are used in the results presented in this paper. Geom(10, 0.29825) is the truncated version of the Geometric distribution with parameter 0.29825 with support restricted to {0,1,…,10} and having a mean of 0.425007 and a variance of 0.605592. The mean of about 0.425 for the arrival distributions is chosen so that the average load at a queue with a stream of interest and a cross-traffic stream becomes 0.85 which makes the queue load typical of a communication network node.

Another distribution, termed Prob9by2, with the same support, mean of 0.425002, close to that of Gem(10,0.29825), but with a different variance of 1.31788, is also defined. The distributions are listed in Table 3 and are plotted in Figure 2. Prob9by2 has a heavier tail than Geom(10,0.29825) as can be inferred from Table 3 and Figure 2. The choice of these distributions allows a comparative analysis of the effect of light and heavy-tailed arrival distributions on the network delays.

Table 1. The Q Abstract Class

Method	Purpose
Delta1	For computing the Class 0 batch inter-arrival distribution to Q
Z0_Z1	Basic Convolve-and-Sweep Recursion step for computing the distribution of Z_1 from that of Z_0
Zn Zn1	Basic Convolve-and-Sweep Recursion step for computing the distribution of Z_{n+1} from that of Z_n
D0_D0l	For computing the distribution of l-fold convolution of $D_0^{(0)}$ when $\Delta_{1,k}^{(0)} = l$ using Z0_Z1 and Zn_Zn1 repeatedly
Dn_Dn1	Recursion step for computing the distribution of $D_{n+1}^{(0)}$ from that of $D_n^{(0)}$using Zn Zn1 repeatedly
D_stat	For computing the stationary delay distribution $D_\infty^{(0)}$, by repeatedly applying the Dn_Dn1 iteration
Delta2	For computing the stationary inter-departure distribution from Q

Table 2. Library for Working with Discrete Distributions

Dist Methods	"name", "min_val", "max_val", "prob_mass", "variate", "moment", "var", "std_dev", "display_par_data", "set_name", "set_min_val""set_max_val", "set_prob_dist", "par_data", "set_par_data", "put_mass", "linf_norm", "cdf"
Functions on Dists	"scale_mass", "scale_dist", "plot", "add_dists", "linf_distance_dists", "convolve_dists", "normalize", "truncate_normalize"

Table 3. Two Discrete Probability Distributions

Point	**Mass** (Rounded to 4 decimals for display)	
	Geom(10,0.29825)	**Prob9by2**
0	0.7018	0.7827
1	0.2093	0.1267
2	0.0624	0.0517
3	0.0186	0.0142
4	0.0056	0.0067
5	0.0017	0.0042
6	0.0005	0.0042
7	0.0001	0.0030
8	4.394e-05	0.0022
9	1.310e-05	0.0020
10	5.569e-06	0.0022

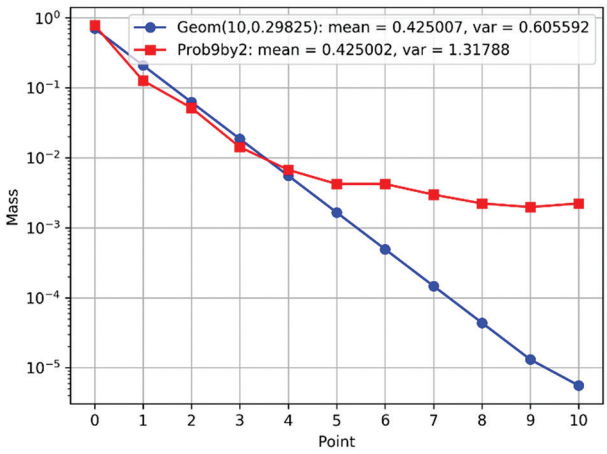

Fig. 2. Two Discrete Probability Distributions

Specific probability distributions are defined by subclassing Dist. Two of them are used in the results presented in this paper. Geom(10, 0.29825) is the truncated version of the Geometric distribution with parameter 0.29825 with support restricted to {0,1,…,10} and having a mean of 0.425007 and a variance of 0.605592. The mean of about 0.425 for the arrival distributions is chosen so that the average load at a queue with a stream of interest and a cross-traffic stream becomes 0.85 which makes the queue load typical of a communication network node.

Another distribution, termed Prob9by2, with the same support, mean of 0.425002, close to that of Gem(10,0.29825), but with a different variance of 1.31788, is also defined. The distributions are listed in Table 3 and are plotted in Figure 2. Prob9by2 has a heavier tail than Geom(10,0.29825) as can be inferred from Table 3 and Figure 2. The choice of these distributions allows a comparative analysis of the effect of light and heavy-tailed arrival distributions on the network delays.

3.7 SIMULATION FOR BENCHMARKING

The results obtained from the computational algorithms are validated against those from simulation. An event-driven simulation program is developed for this purpose. The event-driven program generates columnated output files recording the Queue Lengths and Waiting Times of different classes of jobs at each of the queues as well as the total waiting time in both the queues. The simulations are repeated for multiple runs and empirical quantities are computed using the time and ensemble statistics. Statistical routines for analyzing the data from individual runs as well as aggregated statistics from multiple runs are developed. The statistical quantities estimated include mean, variance and distribution of queue lengths and waiting times.

4. RESULTS AND DISCUSSION

The developed computational algorithm is validated using two instances of the tandem network shown in Figure 1, each with batch sizes given by Geom(10, 0.29825) and Prob9by2 distributions from Table 3 and Figure 2. The results are benchmarked against a simulation-based analysis of the same as outlined in Sec. 3.7.

The computational routines developed (Sec. 3.6) are executed with a precision of 10^{-6}. Stationary distributions of waiting times of an arbitrary Class 0 job at Q1 and Q2 are computed. The computed results are used to obtain a lower bound on the end-to-end delay. For benchmarking, the simulation program is executed with the same arrival distributions and repeated for 25 runs. Results from the numerical computation and simulation are compared and discussed.

4.1 MARGINAL DELAY DISTRIBUTIONS

Figures 3-8 present the results of estimated delay distributions from simulation and computation; Figures 3-5 depict the results for the Geom(10,0.29825) batch size distribution and Figures 6-8 show the results for Prob9by2 batch size distribution.

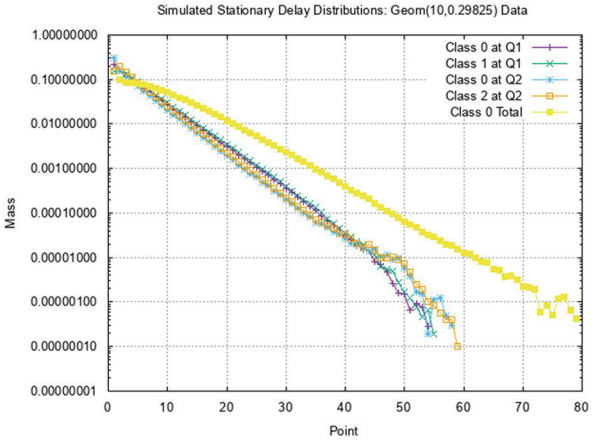

Fig. 3. Simulation Results: Geom(10, 0.29825) Data

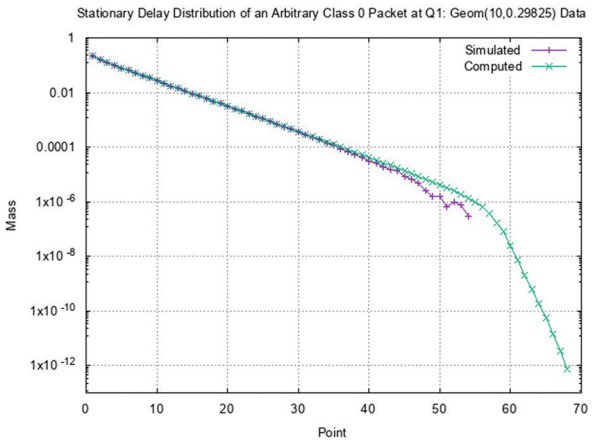

Fig. 4. Stationary Delay Distribution of Class 0 at Q1: Geom(10,0.2985) Batch Sizes

Figures 3 and 6 present the baseline simulation-based estimated distributions of delay experienced by a typical packet from each of the three classes at Q1 and Q2 as well as the total delay of a typical Class

0 packet in the two queue tandem, respectively for Geom(10,0.29825) and Prob9by2 distributions.

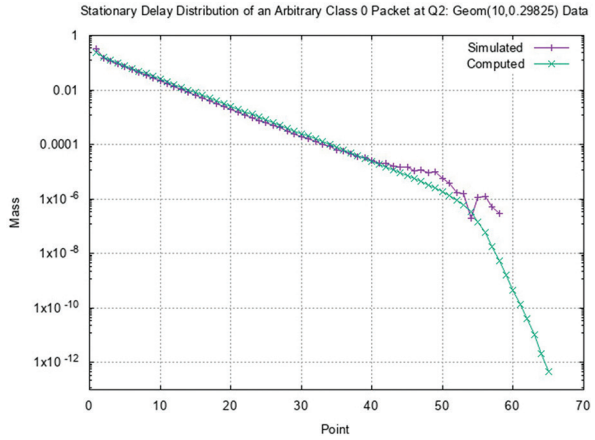

Fig. 5. Stationary Delay Distribution of Class 0 at Q2: Geom(10,0.2985) Batch Sizes

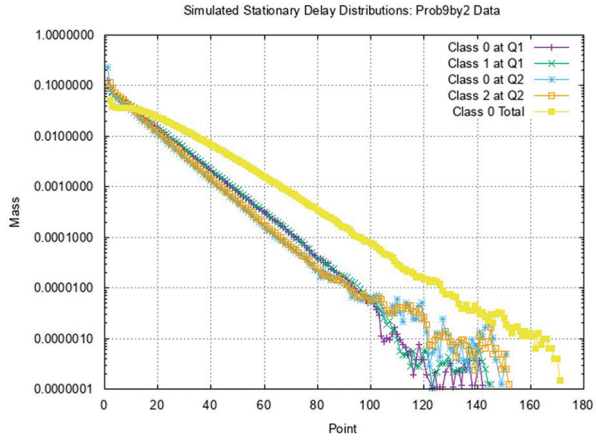

Fig. 6. Simulation Results: Prob9by2) Data

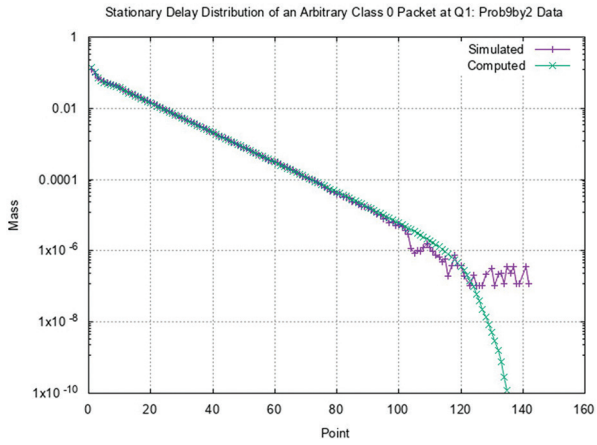

Fig. 7. Stationary Delay Distribution of Class 0 at Q1: Prob9by2 Batch Sizes

Figures 4 and 5 present the comparison of the computed and simulated delay distributions for a typical Class 0 packet in Q1 and Q2, respectively. For values of the delays whose probabilities are reliably estimated by the simulation, there is very good agreement between the simulated and computed quantities. Since large delays are rare events, their estimation via simulation is unreliable, as seen in the plots.

Figures 7 and 8 present the validation of computed delay distributions for a typical Class 0 packet at Q1 and Q2, against simulation for Prob9by2 data. Again, it can be observed that the computed results are validated by their simulation-based estimations.

Since Geom(10,0.29825) has a lighter tail than Prob-9by2 (cf. Figure 2), the delay distributions of the former case are lighter (stochastically less than) compared to those of the latter case, as observed in Figures 4 and 7 as well as in Figures 5 and 8.

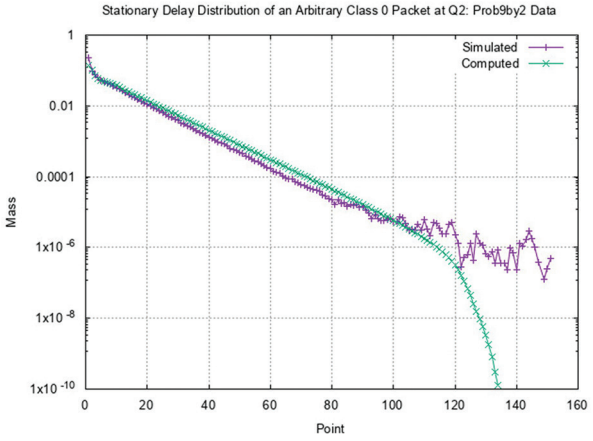

Fig. 8. Stationary Delay Distribution of Class 0 at Q2: Prob9by2 Batch Sizes

4.2 END-TO-END DELAYS

The associativity of the stationary delays at Q1 and Q2 is verified for results of numerical computation as well as simulation for Geom(10,0.29825) and Prob9by2 batch size distributions. The results are summarized in Tables 4 and 5.

In both cases, it is observed that the variance of the stationary total delay (estimated by simulation) is lower bounded by the sum of variances of these stationary delays at individual queues (for both simulation and numerical computation). This provides a validation of the lower bound given by Eq. (13). The variance of delays with Prob9by2 batch arrival distributions is larger than those with Geom(10,0.29825) distributions since the heavier tail of Prob9by2 leads to larger variances.

Table 4. Stationary Delay Variances: Geom(10,0.29825) Data

	Simulated	Computed
Var [Delay at Q 1]	21.0888	21.2862
Var [Delay at Q 2]	16.7995	18.6227
Sum of Variances	**37.8883**	**39.9089**
Var [Total Delay]	**44.7317**	NA

Table 5. Stationary Delay Variances: Prob9by2 Data

	Simulated	Computed
Var [Delay at Q 1]	107.5180	111.9067
Var [Delay at Q 2]	84.4776	108.3647
Sum of Variances	**191.9956**	**220.2714**
Var [Total Delay]	**226.5120**	NA

5. CONCLUSION

A computational approach to the delay distributions in multiclass discrete time tandem networks with general batch size distributions and interfering traffic at each node is developed. The delays of the first packets from consecutive batches of a class are related by using a sequence of truncated Lindley recursions. A convolve-and-sweep algorithm is developed to handle the non-renewal arrival processes and solve these recursions and compute the distributions of the delays and inter-departure distributions.

The delays experienced by a typical packet at different queues are shown to be associated random variables. This is used to compute a lower bound on the variance of the end-to-end delay as the sum of variances of the individual delays experienced at the queues.

An object-oriented implementation of the developed algorithm is carried out facilitating the modelling of different network configurations. A library of routines to compute with discrete probability distributions is also developed to aid the computations.

The developed computational algorithm for the delay distributions and the lower bound on the variance of the total delay are validated using simulation for a tandem network of two queues with cross-traffic at each node under two different batch size distributions.

There are several directions for extending the work presented in this paper. Algorithm 1 can be applied to tandem networks with more than two classes at each node and each stream traversing several nodes. It can be extended to handle more general arrival processes such as Markov modulated arrivals. The solution approach may be extended to compute the joint distributions of the delays in different queues which will allow exact computation of the total end-to-end delays. A non-trivial extension of the developed approach would be to extend it to networks with routing and feedback.

6. APPENDIX

In this Appendix, Theorem 1 is proved using Mathematical Induction.

The delays experienced by an arbitrary Class 0 job at Q1 and Q2 are, respectively, $Y_{j,k}^{(0)} = D_{j,k}^{(0)} + \theta, j=1,2$ (cf. Eq. (2)). Since θ is independent of $D_{j,k}^{(0)}$, it is enough to prove that $D_{j,k}^{(0)}$, $j = 1,2$ are associated random variables.

The idea of the proof from [23] is adapted for showing that $D_{1,k}^{(0)}$ and $D_{2,k}^{(0)}$ are associated.

For any non-negative integers x_0 and x_1,

$$((x_0 - 1)^+ + (x_1 - 1)^+ \geq (x_0 - 2)^+ \quad (14)$$

Applying it iteratively on Eqs. (5)-(6), we obtain

$$D_{1,k+1}^{(0)} \geq \left(D_{1,k}^{(0)} - \Delta_{1,k}^{(0)}\right)^+ \quad (15)$$

from Eq. (7). Similarly, from Eqs. (9)-(12),

$$D_{2,k+1}^{(0)} \geq \left(D_{2,k}^{(0)} + Z_{2,k}^{(0)} - \Delta_{2,k}^{(0)}\right)^+ \quad (16)$$

Inserting Eq. (8) into Eq.(16) gives,

$$D_{2,k+1}^{(0)} \geq \left(D_{2,k}^{(0)} + Z_{2,k}^{(0)} - \Delta_{1,k}^{(0)} + D_{1,k}^{(0)} - D_{1,k+1}^{(0)}\right)^+ \quad (17)$$

Hence, from Eq. (10), it follows that

$$D_{2,k+1}^{(0)} \geq \left(D_{2,k}^{(0)} + X_{A_{1,k}^{(0)}}^{(0)} - \Delta_{1,k}^{(0)} + D_{1,k}^{(0)} - D_{1,k+1}^{(0)}\right)^+ \quad (18)$$

Now, assume $D_{1,k}^{(0)}$ and $D_{2,k}^{(0)}$ are associated random variables. From Eqs. (15) and (18), $D_{1,k+1}^{(0)}$and $D_{2,k+1}^{(0)}$ are increasing functions of the random vector

$$\left(D_{1,k}^{(0)}, D_{2,k}^{(0)}, X_{A_{1,k}^{(0)}}^{(0)}, -\Delta_{1,k}^{(0)}\right) \quad (19)$$

Since $X_{A_{1,k}^{(0)}}^{(1)}$ and $-\Delta_{1,k}^{(0)}$ are independent random variables, they are associated [22]. Also, as a pair, they are independent of pair of random variables $\{D_{1,k}^{(0)}, D_{2,k}^{(0)}\}$ which are assumed to be associated. Hence, the vector in Eq. (19) is a vector of associated random variables [22]. Since $D_{1,k+1}^{(0)}$and $D_{2,k+1}^{(0)}$ are increasing functions of this random vector, they are associated [22]. This proves the induction hypothesis.

For the system starting with zero packets, the same argument as in the above paragraph shows that $D_{1,1}^{(0)}$ and $D_{2,1}^{(0)}$ are associated random variables, establishing the basis case.

Hence, by Mathematical Induction, it follows that $D_{1,k}^{(0)}$ and $D_{2,k}^{(0)}$ are associated random variables, for all $k \geq 0$, completing the proof.

7. REFERENCES

[1] D. Bertsekas, R. Gallager, "Data Networks", 2nd Edition, Prentice-Hall, 1992.

[2] A. Kumar, D. Manjunath, J. Kuri, "Communication Networks: An Analytical Approach", Academic Press, 2004.

[3] H. Bruneel, B. G. Kim, "Discrete–Time Models for Communication Systems Including ATM", Kluwer Academic, 1993.

[4] A. S. Alfa, "Applied Discrete Time Queues", Springer, 2016.

[5] T. Zhang, B. Liu, "Exposing End-to-End Delay in Software-Defined Networking", International Journal of Reconfigurable Computing, Vol. 2019, 2019, p. 7363901.

[6] H. Daduna, "Discrete Time Networks with Product Form Steady States", Queueing Networks: A Fundamental Approach, Springer, 2011, pp. 269-312.

[7] H. Daduna, "Discrete Time Analysis of a State Dependent Tandem with Different Customer Types", Lecture Notes in Computer Science, Vol. 1337, Springer, 1997, pp. 287-296.

[8] O. J. Boxma, "On a Tandem Queueing Model with Identical Service Times at Both the Counters", Advances in Applied Probability, 1979, Parts I, pp. 616-643, Part II, pp. 644-659.

[9] O. P. Vinogradov, "On the Output Stream and the Joint Distribution of Sojourn Times in a Multiphase System with Identical Service", Theory of Probability and Applications, Vol. 40, 1995, pp. 581-588.

[10] S. K. Walley, A. M. Viterbi, "A Tandem of Discrete-Time Queues with Arrivals and Departures at Each Stage", Queueing Systems, Vol. 23, pp. 157-176, 1996.

[11] N. D. Gangadhar, V. Sharma, "Computational Analysis of a Discrete Time Tandem Network of Queues with Intermediate Arrivals and Departures", Proceedings of the Second Canadian Conference on Broadband Research, 1998, pp. 350-359.

[12] M. J. Neely, "Exact Queueing Analysis of Discrete Time Tandem Networks with Arbitrary Arrival Processes", Proceedings of the IEEE International Conference on Communications, Paris, France, 20-24 June 2004, pp. 2221-2225.

[13] W.-C. Lau, S.-Q. Li, "Traffic Distortion and Inter-source Cross-correlation in High Speed Integrated Networks", Computer Networks and ISDN Systems, Vol. 29, 1997, pp. 811-830.

[14] D. Yates, J. Kurose, D. Towsley, M. G. Hluchj, "On Per Session End to End Delay Distributions and the Call Admission Problem for Real Time Applications with QoS Requirements", Proceedings of ACM SIGCOMM, 1993, pp. 2-12.

[15] W. Whitt, "Approximations for Departure Processes and Queues in Series", Naval Research Logistics Quarterly, Vol. 31, 1984, pp. 499-521.

[16] G. Hasselinger, E. S. Rieger. "Analysis of Open Discrete Time Queueing Networks: A Refined Decomposition Approach", Journal of the Operations Research Society, Vol. 47, 1996, pp. 640-653.

[17] S. J. Golestani, "Network Delay Analysis of a Class of Fair Queueing Algorithms", IEEE Journal on Selected Areas in Communications, Vol. 13, 1995, pp. 1057-1070.

[18] R. L. Cruz, "Quality of Service Guarantees in Virtual Circuit Switched Networks", IEEE Journal of Selected Areas in Communications, Vol. 13, 1995, pp. 1048-1056.

[19] C. S. Chang, "Queue length and Delay of Deterministic and Stochastic Queueing Networks", IEEE Transactions on Automatic Control, Vol. 39, 1994, pp. 913-931.

[20] S. Azodolmolky, R. Nejabati, M. Pazouki, P. Wieder, R. Yahyapour, D. Simeonidou, "An Analytical Model for Software Defined Networking: A Network Calculus-based Approach", Proceedings of IEEE Global Communications Conference, Atlanta, GA, USA, 9-13 December 2013, pp. 1397-1402.

[21] C. Lin, C. Wu, M. Huang, Z. Wen, Q. Zheng, "Performance Evaluation for SDN Deployment: An Approach based on Stochastic Network Calculus. China Communications, Vol. 13 (Supplement 1), 2016, pp. 98-106.

[22] J. D. Esary, F. Proschan, D. W. Walkup, "Association of Random Variables with Applications", Annals of Mathematical Statistics, Vol. 38, 1967, pp. 1466-1474.

[23] S.-C. Niu, "Bounds for The Expected Delays in Some Tandem Queues", Journal of Applied Probability, Vol. 17, 1980, pp. 831-838.

[24] M. Zukerman, "Introduction to Queueing Theory and Stochastic Teletraffic Models", arXiv:1307.2968, 2021

[25] V. Ramaswami, "A Stable Recursion for the Steady State Vector in Markov Chains of M/G/1 Type", Stochastic Models, Vol. 4, 1990, pp. 151-161.