Packet Delay in Optical Circuit-Switched Networks

Zvi Rosberg, Andrew Zalesky, and Moshe Zukerman, *Senior Member, IEEE*

***Abstract*— A framework is provided for evaluation of packet delay distribution in an optical circuit-switched network. The framework is based on a fluid traffic model, packet queueing at edge routers, and circuit-switched transmission between edge routers. Packets are assigned to buffers according to their destination, delay constraint, physical route and wavelength. At every decision epoch, a subset of buffers is allocated to end-to-end circuits for transmission, where circuit holding times are based on limited and exhaustive circuit allocation policies. To ensure computational tractability, the framework approximates the evolution of each buffer independently. 'Slack variables' are introduced to decouple amongst buffers in a way that the evolution of each buffer remains consistent with all other buffers in the network. The delay distribution is derived for a single buffer and an approximation is given for a network of buffers. The approximation entails finding a fixed point for the functional relation between the 'slack variables' and a specific circuit allocation policy. An analysis of a specific policy, in which circuits are probabilistically allocated based on buffer size, is given as an illustrative example. The framework is shown to be in good agreement with a discrete event simulation model.**

***Index Terms*— Circuit switching, packet delay, WDM network, fixed point approximation.**

I. Introduction

The advancement of optical technology in recent years [1] positions the *All-Optical Network (AON)* as a viable option for core backbone networks. An AON consists of core routers interconnected by fiber links carrying hundreds of wavelength channels, referred to as the core network. Edge routers are located at the periphery of the core network and given the task of assembling and disassembling many data streams arriving from or destined to users connected to the core network via access networks. For such a task, edge routers possess buffering capabilities, and from the viewpoint of the core network, may be considered as source and destination nodes.

An AON transmits data streams by way of all-optical lightpaths established using *wavelength division multiplexing* (WDM). Data remains in the optical domain throughout transmission from source to destination. However, signaling and switching functions may occur in the electronic domain. The primary advantage of an AON is that data streams do not undergo optical-electrical-optical (OEO) conversion, which increases end-to-end latency.

Next generation dense-WDM (DWDM) fiber technology is likely to offer a single fiber containing hundreds of wavelength channels, each modulated at 10 Gb/s. Hence, links with a total capacity of tens of Tb/s may be attached to a single core router requiring switching capacity not available by present day electronics. Consequently, bufferless core routers are envisaged, where all data queueing is shifted to the edge routers. An AON architecture motivates the framework considered herein.

Any network deployment is a tradeoff between cost and performance implying some amount of packet loss and/or queueing delay, depending on the edge router architecture. Since substantial data buffering in core routers is not a valid option in an AON, route reservation procedures should not mandate buffering at core routers. Two such route reservation procedures are as follows.

The first is a one-way reservation procedure known as tell-and-go [15], in which a reservation request is sent before the data is transmitted. Then, without waiting for an acknowledgment, the data is transmitted after some predefined offset time. Two tell-and-go procedures have gained the most attention. One is just-in-time (JIT) [14] and the other is just-enough-time (JET) [10]. In both reservation procedures, packets with a common destination are aggregated in the edge routers into large transmission units called bursts, each of which is transmitted separately. This approach is known as *optical burst switching (OBS)*.

The other reservation procedure is classical two-way reservation [2], [3], [16], in which data transmission does not commence until the edge router receives acknowledgement of all resource reservations.

With one-way reservation, transmitted data can be blocked by any core router along its route. Thus, a major performance measure is blocking probability, which has been derived for various reservation procedures [12]. With two-way reservation, blocking at core routers is averted by delaying data transmission until the edge router receives acknowledgement of all resource reservations, and thus, the main performance measure is queueing delay at the edge routers.

The focus of this paper is to provide a framework for evaluation of packet delay distribution in an optical circuit-switched network. The framework allows for delay differentiation as well as routing and wavelength assignment (RWA) algorithms. As explained in Section II, optical circuits may have a more complex structure than circuits in the classical models [7]. Thus, they will be referred to as *optical circuit-switched (OCS)* networks.

OCS is typically used as an umbrella term to encompass many network architectures based on two-way reservation. This paper focuses on an OCS architecture that operates as follows. Packets are enqueued in logical buffers located at the periphery of the network and it is the delay distribution of a packet's queuing time that we are interested in estimating. Time is divided into discrete (circuit) periods. At the boundary of each period a central controller determines whether or not a

Z. Rosberg is with the Department of Communication Systems Engineering, Ben Gurion University, Beer-Sheva, 84105, Israel; rosberg@bgumail.bgu.ac.il

A. Zalesky and M. Zukerman are with the Centre for Ultra-Broadband Information Networks (CUBIN), Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, VIC 3010, Australia; {a.zalesky,m.zukerman}@ee.unimelb.edu.au

CUBIN is an affiliated program of National ICT Australia.

buffer is to be allocated a circuit during the next period based on the number of packets enqueued in that buffer as well as the number of packets enqueued in all other buffers. Circuit periods (i.e., holding times) can either be based on limited or exhaustive circuit allocation policies.

Performance of circuit-switched networks has been studied only with respect to blocking probabilities [7], [8], [11]. The study in [7] is concerned with routing data or voice in a classical circuit-switched network and the studies in [8] and [11] are concerned with RWA in optical circuit-switched networks. In these studies, blocking probabilities have been derived using the reduced-load fixed point approximation based on solving Erlang's formula, under the assumption blocking events occur independently on each link.

It is worth noting, maximizing the carried traffic in a circuit-switched network with an arbitrary topology and an arbitrary RWA algorithm given a static traffic demand can be formulated as an integer linear program [5] [11]. Although the integer linear program is in NP, its solution can be carried out off line and then used in a lookup table whose entries represent different traffic demands. This supports a network model, where an RWA algorithm is regarded as a black box.

The rest of this paper is organized as follows. In Section II, we formulate the general problem and define the model. Sections III and IV are devoted to the delay evaluation framework. In particular, Section III considers a single buffer, while Section IV uses the single buffer case as a foundation to evaluate delay distribution for a network of buffers. The framework is illustrated by an example of an RWA algorithm in Section V, and model extensions are explained in Section VII. Some important practical considerations are discussed in Section VI. In Section VIII, we present numerical data validating our illustrative example and in Section IX, we draw our conclusions.

II. Model Formulation

Our objective is to develop a framework for evaluation of packet delay distribution in OCS networks. To this end, we consider J data streams, each associated with a source-destination pair of edge routers, delay constraint, a route and wavelength assignment sequence from the source to the destination, and other external classifications. Data packets from stream $j, 1 \leq j \leq J$, that cannot be transmitted immediately are queued in logical buffer j at its corresponding source edge router.

A *circuit* in our framework is a unidirectional lightpath connecting a pair of source-destination edge routers capable of transmitting C b/s uninterruptedly for a period of T seconds. A circuit is set up by selecting a unidirectional route between the source-destination pair and allocating a dedicated sequence of wavelengths and switching resources along the selected route as dictated by the given RWA algorithm. The wavelength sequence must be aligned with the wavelength conversion rules along the route.

Circuits are allocated to the logical buffers using a policy R based on the queue lengths at all logical buffers. A strict requirement of a circuit allocation policy is that any allocated set of circuits can serve their associated buffers concurrently and continuously. That is, their lightpaths are disjoint. When a circuit is allocated to a logical buffer, it is drained at a maximum rate of C b/s. An allocated circuit that is not reselected T seconds after its allocation is torn down.

One means of providing delay differentiation is to assign buffers with more stringent delay requirements to a greater number of allocated set of circuits, which results in more frequent service allocations. Other means are policy dependent, as explained in Section V, where a threshold randomized policy is presented.

The assumption that circuits are selected in a synchronized manner and at fixed time intervals does not impose limitations on our framework. Neither is the assumption that a circuit period must be of a fixed length. In Section VII, we explain how to extend our analysis to variable circuit lengths and asynchronous allocations.

Circuit setup begins by evaluating all queue lengths and then a circuit allocation policy R is called to compute the set of circuits, which can be allocated concurrently. If the edge routers are time synchronized, the overall procedure can be implemented centrally or distributively.

Assume that bits arrive at each logical buffer j according to a continuous fluid stream with an integral constant bit rate of A_j b/s. Considering the expected Tb/s nature of multiplexed input streams, such a fluid approximation is a natural traffic model. Modeling data transmission as a continuous fluid stream is also tangible due to the nature of an optical circuit, in which an arriving bit can be served on-the-fly without waiting for its encapsulating data packet.

Without loss of generality, we normalize all rates by dividing them by their largest common integral denominator, say B. Henceforth, we refer to a unit of B bits as 'B-bit' and let K and $\{A_j\ , 1 \leq j \leq J\}$ be the normalized lightpath transmission and arrival rates (in 'B-bits per circuit period'), respectively. We further assume that every A_j is an integral fraction of K. That is, there are integers $\{m_j^0\}$ such that

$$K = m_j^0 A_j, \quad \forall\ j. \tag{1}$$

Also, without loss of generality, we assume that $T = 1$, and transmission and arrival rates are specified in circuit periods. To summarize, time units are specified in circuit periods and data units in B-bits (normalized bits).

Let n denote a circuit switching decision epoch, $X_j(n)$ denote the queue length (in 'B-bits') in logical buffer j at epoch n, $1 \leq j \leq J$, and $\boldsymbol{X}(n) = \big(X_1(n), X_2(n), \ldots, X_J(n)\big)$ denote the system state at epoch n, $n = 0, 1, 2, \ldots$.

Given a circuit allocation policy R, let $\boldsymbol{\delta}^R(\boldsymbol{x}) = \big(\delta_1^R(\boldsymbol{x}), \delta_2^R(\boldsymbol{x}), \ldots, \delta_J^R(\boldsymbol{x})\big)$ be a binary vector indicating which of the logical buffers are allocated circuits at state $\boldsymbol{X}(n) = \boldsymbol{x} \stackrel{\text{def}}{=} (x_1, x_2, \ldots, x_J)$. That is, $\delta_j^R(\boldsymbol{x})$ is 1 or 0 depending on whether or not R allocates a circuit to logical buffer j at state $\boldsymbol{x}$, respectively.

The process $\big(\boldsymbol{X}(n),\ n \geq 0\big)$ is a Markov chain and each of its components $X_j(n)$ evolves according to

$$X_j(n+1) = \big[X_j(n) + A_j - \delta_j^R(\boldsymbol{x})K\big]^+, \tag{2}$$

where $[y]^+ = \max\{0, y\}$.

Let $\mathcal{S}_j(i)$ be the set of system states, where logical queue j comprises of i B-bits and let $\alpha_j(i,n)$ be the probability that algorithm R allocates a circuit to buffer j at epoch n given that $\boldsymbol{X}(n) \in \mathcal{S}_j(i)$. That is,

$$\mathcal{S}_j(i) = \{\boldsymbol{x} \ ; \ \boldsymbol{X}(n) = \boldsymbol{x}, X_j(n) = i\} \tag{3}$$

$$\alpha_j(i,n) = P\left(\delta_j^R(\boldsymbol{X}(n)) = 1 \mid \boldsymbol{X}(n) \in \mathcal{S}_j(i)\right). \tag{4}$$

The marginal process $\big(X_j(n),\ n \geq 0\big)$ is not Markovian. Nevertheless, its evolution in time can be expressed in the probability space of the Markov chain $\big(\boldsymbol{X}(n),\ n \geq 0\big)$ as follows. By (2), given $X_j(n) = i$, we have

$$X_j(n+1) = \begin{cases} \big[i + A_j - K\big]^+, & \text{w.p. } \alpha_j(i,n); \\ \\ i + A_j \ , & \text{w.p. } 1 - \alpha_j(i,n), \end{cases} \tag{5}$$

for every $1 \leq j \leq J$, where 'w.p.' stands for 'with probability'.

The Markov chain $\big(\boldsymbol{X}(n),\ n \geq 0\big)$ may or may not be periodic, depending on the allocation policy R. For instance, if R allocates circuits based on a deterministic set function of the queue length vector $\boldsymbol{x}$ (i.e., a deterministic stationary policy), then the resulting Markov chain is periodic. For these policies, periodicity follows from the deterministic fluid arrival processes and the fact that only a finite number of states can be visited by the Markov chain under appropriate positive recurrent conditions. The performance of circuit allocation policies under which the Markov chain is periodic have been exactly analyzed elsewhere [13] and not considered herein.

If the Markov chain $\big(\boldsymbol{X}(n),\ n \geq 0\big)$ is aperiodic and positive recurrent (i.e., has a stationary state distribution function), the probabilities $\{\alpha_j(i,n)\}$ under stationary conditions exist and are independent of n, but do depend on the entire system state. Thus, under stationary conditions, (5) translates into

$$X_j(n+1) = \begin{cases} \big[i + A_j - K\big]^+, & \text{w.p. } \alpha_j(i); \\ \\ i + A_j \ , & \text{w.p. } 1 - \alpha_j(i), \end{cases} \tag{6}$$

given $X_j(n) = i$.

According to (6), it may be suggested that the stationary distribution of a Markov chain evolving according to (6) with probabilities $\{\alpha_j(i)\}$ can approximate the multidimensional Markov chain $\big(\boldsymbol{X}(n),\ n \geq 0\big)$. The probabilities $\{\alpha_j(i)\}$ may be regarded as 'slack variables'.

The concept underpinning our approximation is as follows. For every logical buffer j, consider a one dimensional Markov chain evolving according to (6) and independently of the other buffers. In the original multidimensional process, the J sets of allocation probabilities $\{\alpha_j(i)\}$, $1 \leq j \leq J$, are clearly inter-dependent. Therefore, the J sets of allocation probabilities must be resolved in a way that consistency is maintained across all sets. The consistency conditions give rise to a set of fixed-point equations, each of which describes one of the one-dimensional Markov chains, assuming they evolve independently.

In the next section, we derive the queue length and the packet delay distributions in a generic single buffer evolving according to (6).

III. A Single Logical Queue

A. Definition and ergodicity

For notational clarity, we omit the logical buffer index j in this section and denote a generic one-dimensional queueing system by $\big(X(n),\ n \geq 0\big)$. Assuming independent evolution of the marginal processes of $\big(\boldsymbol{X}(n),\ n \geq 0\big)$, (1) and (6) imply that given $X(n) = i$,

$$X(n+1) = \begin{cases} \big[i + A - m^0 A\big]^+, & \text{w.p. } \alpha(i); \\ \\ i + A \ , & \text{w.p. } 1 - \alpha(i), \end{cases} \tag{7}$$

where A and $K = m^0 A$ are the arrival and transmission rates, respectively.

The upper event in (7) represents an allocated circuit period and the lower event represents an unallocated period. Observe that after every unallocated circuit period, the queue length increases by A and after every allocated circuit period, the queue length decreases by $\min\{i, (m^0 - 1)A\}$, where i is the queue length at the beginning of the circuit period. Consequently, $X(n)$ assumes only integral multiples of A. That is, its state space is $\{iA \ ; \ i = 0, 1, 2, \ldots\}$. Without loss of generality, we relabel the process states and denote them by the set of non-negative integers, with the convention that $X(n) = i$ denotes iA B-bits reside in the queue. With relabelling, (7) becomes

$$X(n+1) = \begin{cases} \big[i + 1 - m^0\big]^+, & \text{w.p. } \ \alpha(i); \\ \\ i + 1 \ , & \text{w.p. } \ 1 - \alpha(i). \end{cases} \tag{8}$$

Since the transmission rate for $X(n) \geq (m^0 - 1)$ is K, it is reasonable to approximate $\alpha(i) = \overline{\alpha}$ for $i \geq m^0 - 1$. This approximation is motivated by the fact the transmission rate is always $K = Am^0$ B-bits if $(m^0 - 1)A$ B-bits or more reside in a buffer. We further have $0 \leq \alpha(i) \leq 1$.

Given that we consider only policies R under which the multidimensional Markov chain $\big(\boldsymbol{X}(n),\ n \geq 0\big)$ is aperiodic, we may restrict attention to aperiodic one-dimensional Markov chains $\big(X(n),\ n \geq 0\big)$. Since there is a positive probability to return to state zero from any other state it can be shown that the Markov chain is irreducible and aperiodic. A necessary and sufficient condition for ergodicity is

$$\overline{\alpha}\ m^0 > 1. \tag{9}$$

Indeed, assuming (9) holds, the expected drift in one transition is

$$E[X(n+1) - X(n)/X(n) = i] = 1 - \overline{\alpha} m^0 < 0,$$

for $i \geq m^0 - 1$. Thus, by the Foster-Lyapunov drift criterion [4], the Markov chain is ergodic.

B. Queue length probability generation function

The probability generation function (pgf) under stationary conditions, $G(z) = \lim_{n\to\infty} E\big[z^{X(n)}\big]$, $|z| \leq 1$, is derived in

Appendix A and is given by

$$G(z)=\frac{\sum_{i=0}^{m^0-2}\big[(\alpha(i)z^{m^0-1}-\overline{\alpha}z^i)+(\overline{\alpha}-\alpha(i))z^{m^0+i}\big]p(i)}{z^{m^0-1}-(1-\overline{\alpha})z^{m^0}-\overline{\alpha}}, \tag{10}$$

where $p(i)$ is the stationary probability of having iA B-bits in the buffer.

The pgf in (10) is expressed by a function of the m^0-1 boundary probabilities $p(i), i=0,1,\ldots,m^0-2$, that are yet to be determined. Standard application of Rouche's Theorem and the analyticity of $G(z)$ in the unit disk $|z|\le 1$ yield these boundary probabilities (see [6, pp. 121-124]).

Specifically, as we prove in Appendix B, the denominator of $G(z)$ has m^0-1 distinct zeros within and onto the unit disk $|z|\le 1$. To find the boundary probabilities $p(i), i=0,1,\ldots,m^0-2$, we exploit the analyticity of $G(z)$ in the unit disk $|z|<1$. Namely, the numerator of $G(z)$ must be zero for every zero of its denominator within the unit disk. One zero of the denominator is clearly 1 for which all the coefficients of $p(i)$ in the numerator are zero and therefore useless. All other m^0-2 zeros, denoted by z_m, $m=1,2,\ldots,m^0-2$, are within the unit disk and define the following m^0-2 linear equations:

$$\sum_{i=0}^{m^0-2}\big[(\alpha(i)z_m^{m^0-1}-\overline{\alpha}z_m^i)+(\overline{\alpha}-\alpha(i))z_m^{m^0+i}\big]p(i)=0, \tag{11}$$

$$m=1,2,\ldots,m^0-2.$$

Another equation is obtained from the normalization condition $G(1)=1$. Applying L'hopital's rule to (10), we have

$$\sum_{i=0}^{m^0-2}\big[m^0\overline{\alpha}-(1+i)\alpha(i)\big]p(i)=\overline{\alpha}m^0-1. \tag{12}$$

Equations (11)–(12) form a set of m^0-1 independent linear equations whose solution determine $p(i), i=0,1,\ldots,m^0-2$. The independence is verified by checking the positivity of the corresponding determinant as in [6, pp. 121-124].

Once the boundary probabilities are determined, $G(z)$ is completely specified. The stationary probabilities, $\{p(i)\}$, are given by $p(i)i!=\frac{d^{(i)}G(z)}{dz}|_{z=0}$ and the expected queue length under stationary conditions, $E(X)$, is given by $E(X)=\frac{dG(z)}{dz}|_{z=1}$. Higher moments are derived by taking higher derivatives at $z=1$.

It is well known that moment and probability derivations from $G(z)$ are very tedious. In the next subsections, we apply simpler methods to derive $E(X)$ and $p(i)$ for $i\ge m^0-1$.

C. Expected queue length

First, we derive the expected queue length at a circuit period boundary under stationary conditions, $E(X)$, and then the long-run time-average queue length, $E(\tilde{X})$.

A simple method to derive $E(X)$ is to express the one-step evolution of $X^2(n+1)$ (similar to (8)) and then equate between the expected values of both sides. This method yields the expression in (13), which is presented on the page that follows. The expected number of B-bits at a circuit boundary is therefore $A\cdot E(X)$.

To find the time average queue length we note that the queue length evolution between two consecutive circuit period boundaries, $\{X(t),\ 0\le t\le 1\}$, is as follows. Given $X(n)=i$,

$$X(t)=\begin{cases}\big[i+t-m^0t\big]^+, & \text{w.p. } \alpha(i);\\ i+t\ , & \text{w.p. } 1-\alpha(i),\end{cases} \tag{14}$$

By the mean ergodic Theorem, $E(\tilde{X})=\int_{t=0}^{1-}E\big(X(t)\big)dt$. Note that for $i\ge m^0-1$, we have $i+t-m^0t\ge 0$ for every $0\le t\le 1$; and for $i<m^0-1$, we have $i+t-m^0t\ge 0$ for $0\le t\le i/(m^0-1)$. Integrating yields

$$\begin{aligned}E(\tilde{X}) &= E(X)+\tfrac{1}{2}-\tfrac{\overline{\alpha}m^0}{2}\left(1-\sum_{i=0}^{m^0-2}p(i)\right)\\ &\quad+\tfrac{1}{2}\sum_{i=0}^{m^0-2}p(i)\alpha(i)\left(\tfrac{i^2}{m^0-1}-2i-1\right).\end{aligned} \tag{15}$$

The time-average number of B-bits is therefore $A\cdot E(\tilde{X})$.

D. Queue length distribution

In subsection III-B, we derived the probabilities $p(i), i=0,1,\ldots,m^0-2$. In this subsection, we derive a simple recursion for $p(i)$, $i\ge m^0-1$.

From (8), the balance equations are given by

$$p(0)=\sum_{i=0}^{m^0-1}p(i)\alpha(i), \tag{16}$$

and

$$p(i)=p(i-1)\big(1-\alpha(i-1)\big)+p(i+m^0-1)\overline{\alpha}, \tag{17}$$

for $i\ge 1$.

Given $\{p(i)\ ;\ 0\le i\le m^0-2\}$, by (16),

$$p(m^0-1)=\frac{p(0)-\sum_{i=0}^{m^0-2}p(i)\alpha(i)}{\overline{\alpha}}; \tag{18}$$

and by (17),

$$p(m^0+i)=\frac{p(i+1)-p(i)\big(1-\alpha(i)\big)}{\overline{\alpha}},\quad i\ge 0. \tag{19}$$

E. Delay distribution

In non-fluid models, where packet arrivals and departures occur at particular time instances, packet delay is a well defined notion. In a fluid traffic model, however, a packet can be served while it is still arriving. Thus, the time interval during which a packet arrives could overlap with its transmission interval and multiple notions of packet delay can be defined. Regardless of the notion of delay defined, a packet scheduling rule is required and we assume a FIFO regime.

Consider a notion of delay defined as the time elapsed from the arrival instance of the first bit of a packet to the departure instance of the last bit of a packet. Such a notion of delay

$$E(X) = \frac{\overline{\alpha}\big(m^0-1\big)^2 + (1-\overline{\alpha}) + \sum_{i=0}^{m^0-2} p(i)\Big[\overline{\alpha}m^0\big(2(i+1)-m^0\big) - \alpha(i)\big(i+1\big)^2\Big]}{2(\overline{\alpha}m^0-1)}. \tag{13}$$

must be defined in terms of packet length and is considered below to derive the delay distribution for a special case.

An alternative notion of delay, referred to as B-bit delay, is the time elapsed from the arrival to the departure instant of a B-bit. The B-bit notion of delay is not defined in terms of packet length, however, it does reflect packet delay in the following sense. At a B-bit arrival instant, the portion of the packet preceding the B-bit is either enqueued or has undergone transmission; at a B-bit departure instant, the portion of the packet preceding the B-bit has undergone transmission. Thus, B-bit delay reflects the delay of an arbitrary packet prefix.

The expected B-bit delay under stationary conditions is derived from $E(\tilde{X})$ by Little's Theorem. Since $A\cdot E(\tilde{X})$ is the expected queue length in B-bits in the buffer at an arbitrary instant, and the B-bit arrival rate is A, the expected B-bit delay (queueing time) is $E(\tilde{X})$ given in (15).

We now return to the former notion of delay defined as the time elapsed from the arrival instance of the first bit of a packet to the departure instance of the last bit of a packet and we assume each packet comprises of L B-bits. We further assume that during each circuit period there is an integral number M of packet arrivals, i.e., $A = M \cdot L$, and all packets are served according to the FIFO regime.

Let D be the packet delay, measured in circuit periods, defined as the time elapsed from the arrival instance of the first bit of a packet to the departure instance of the last bit of a packet. We now derive the packet delay distribution for a special symmetric case. For definiteness, assume that the packet arrival process begins at the boundary of a circuit period. There are M packets arriving during every circuit period, each having a different delay. Let D_m, $1 \le m \le M$, be the delay of a packet whose arrival begins $(m-1)/M$ circuit periods after a circuit boundary. The delay of an arbitrary packet is given by

$$D = \frac{1}{M}\sum_{m=1}^{M} D_m. \tag{20}$$

The difficulty in deriving packet delay distribution is attributable to the fact that $\{\alpha(i)\}$ is different, for every i. Therefore, the time between two consecutive circuit allocations is not identically distributed. To simplify the derivation, we consider the special symmetric case, in which $\alpha(i) = \alpha$, for every i. The derivation of delay distribution for the special symmetric case may serve as a guide to deriving the delay distribution for the general case. We derive the delay distribution by way of a computational procedure rather than a closed form expression. The procedure produces the delay distribution histogram. The details of the derivation are deferred until Appendix C.

IV. A Network of Edge Routers

Deriving the exact stationary distribution for the multidimensional Markov chain determined by an arbitrary circuit allocation policy R is intractable. To ensure computational tractability, consider approximating the evolution of each buffer independently. To decouple amongst buffers in a way that the evolution of each buffer remains consistent with all other buffers in the network, the stationary circuit allocation probabilities, $\{\alpha_j(i)\}$, must be chosen in agreement with the policy R.

For any given R, let $\mathcal{S}_j^R(i)$ be the subset of $\mathcal{S}_j(i)$, defined in (3), where $\delta_j^R(\boldsymbol{x}) = 1$. Namely, the set of states, where buffer j comprises i B-bits and is allocated a circuit.

By the independence assumption and (4), the following *R-consistency equations* must hold:

$$\begin{aligned}\alpha_j(i) &= P\Big(\delta_j^R\big(\boldsymbol{X}(n)\big) = 1 \ / \ X_j(n) = i\Big) \\ &= \textstyle\sum_{\boldsymbol{x}\in\mathcal{S}_j^R(i)} \prod_{m\neq j} p_m(x_m), \quad \forall\ (j,i),\end{aligned} \tag{21}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_J)$ and $p_m(x_m) = P\big(X_m(n) = x_m\big)$, $1 \le m \le J$, are the stationary marginal probabilities.

If (21) does hold, we say that *the independent Markov chains* $\big\{X_j(n),\ 1 \le j \le J,\ n \ge 0\big\}$ *are consistent with policy* R.

For every logical buffer j, let $\boldsymbol{\alpha}_j = \{\alpha_j(i)\ ;\ i \ge 0\}$ and $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_j\ ;\ 1 \le j \le J\}$. A set $\boldsymbol{\alpha}$ is a *consistent set of allocation probabilities* if it satisfies (21).

Since the stationary probabilities $\{p_j(i)\}$ depend on $\boldsymbol{\alpha}_j$ we use the notation $p_j(\boldsymbol{\alpha}_j, i)$ rather than $p_j(i)$.

To find the consistent set of allocation probabilities, define the transformations:

$$T_j^i(\boldsymbol{\alpha}) = \sum_{\boldsymbol{x}\in\mathcal{S}_j^R(i)} \prod_{m\neq j} p_m(\boldsymbol{\alpha}_j, x_m), \qquad 1 \le j \le J, \quad i \ge 0. \tag{22}$$

The *R-consistency equations* (21) are satisfied if and only if there is an $\boldsymbol{\alpha}^*$ such that

$$T_j^i\big(\boldsymbol{\alpha}^*\big) = \alpha_j^*(i), \quad \forall\ (i,j). \tag{23}$$

Observe that each transformation set $\big\{T_j^i\big(\boldsymbol{\alpha}\big)\big\}$ is a continuous mapping from the compact set $[0,1]^{|J|\cdot(K-A)}$ to itself and therefore it has a fixed point by the Brouwer fixed-point theorem [9].

To find the consistent set of allocation probabilities $\boldsymbol{\alpha}^*$, we invoke the following successive substitution algorithm with some initial set $\boldsymbol{\alpha}^{(0)}$:

$$\alpha_j^{(n+1)}(i) = T_j^i\big(\boldsymbol{\alpha}^{(n)}\big), \quad \forall\ (i,j) \text{ and } \ n \ge 0. \tag{24}$$

Once a consistent set $\boldsymbol{\alpha}^*$ is found, the delay distribution using policy R is computed for every logical buffer as given in Section III-E.

As demonstrated in the example presented in Sections V and VIII, the successive substitution algorithm is not guaranteed to converge to the consistent set of allocation probabilities $\boldsymbol{\alpha}^*$, furthermore, there is no guarantee that the transformation

$T_j^i(\boldsymbol{\alpha})$ admits a unique set of consistent allocation probabilities.

However, as demonstrated by all test instances considered, the successive substitution algorithm does indeed converge to a set of consistent allocation probabilities, which are in good agreement with a simulation model used to verify the approximation. The successive substitution algorithm usually requires only a few iterations to converge within a sufficiently small error criterion. The delay evaluation framework can therefore accurately approximate the expected B-bit delay in a fraction of the computational time required by the simulation model.

V. A Circuit Allocation Policy Example

Let $\mathcal{J}$ be the set of all logical buffers. A building block to define general circuit allocation policies is a *maximal transmission (MT)* set. An MT set is a subset $\mathcal{J}_i$ of $\mathcal{J}$, satisfying: (i) all buffers in $\mathcal{J}_i$ can be allocated a circuit concurrently without resulting in data loss; (ii) there is no superset of $\mathcal{J}_i$ that satisfies (i). Allocating circuits to a set of buffers that does not define an MT set is suboptimal.

The set of all MT sets, denoted by $\mathcal{J}^* = \{\mathcal{J}_i \subseteq \{1, 2, \ldots, J\}\ ;\ 1 \leq i \leq N\}$, can be mapped to a realizable network consisting of a topology and routing policy. Restrictions are not imposed to avoid overlapping MT sets. In particular, a buffer j may reside in more than one MT set.

A general circuit allocation policy is one that selects a single MT set at every circuit period based on some measurable information about all buffers. Any deterministic stationary policy allocating an MT set as a function of all queue lengths defines a weighted time division multiplexing (TDM) policy and results in a periodic Markov chain. The performance of these policies have been analyzed elsewhere [13] and not considered herein.

Here, we demonstrate the delay evaluation framework for the following threshold randomized policy implemented with the aid of a common pseudo random number generator. Each MT set $\mathcal{J}_i$, is associated with a triplet (t_i, ξ_i^1, ξ_i^2), where t_i is a threshold value and $\xi_i^2 > \xi_i^1$ are positive weights.

An MT set constellation is a binary vector $\boldsymbol{b} = (b_1, b_2, \ldots, b_N)$, where $b_i = 0$, if and only if $\sum_{j\in\mathcal{J}_i} X_j(n) \leq t_i$. Let $\xi_i(b_i) = \xi_i^1$, if $b_i = 0$; and $\xi_i(b_i) = \xi_i^2$, if $b_i = 1$

The policy is defined as follows: *For every given MT set constellation* $\boldsymbol{b}$*, MT set* $\mathcal{J}_i$ *is selected with probability* $\xi_i(b_i)/\sum_{l=1}^N \xi_l(b_l)$.

A distributed implementation requires to pass around the constellation vector and to use the same pseudo random generator in all buffers. The latter guarantees that exactly one MT set is chosen for each constellation.

For every j, let $\mathcal{N}_j$ be the set of all MT sets not containing buffer j; N_j be its cardinal number; Y_j be the number of MT sets not containing j, where each one of them has a total buffer size less than or equal to its corresponding threshold; and $Z_j(i)$ be the number of MT sets containing j, where each one of them has a total buffer size less than or equal to its corresponding threshold, given $X_j(n) = i$.

Given the current $\boldsymbol{\alpha}$ and the events $Y_j = y$ and $Z_j(i) = z$,

$$\alpha_j(i, y, z) \stackrel{\text{def}}{=} T_j^i\big(\boldsymbol{\alpha}, y, z\big) = \frac{z\xi_1 + (N - N_j - z)\xi_2}{(z + y)\xi_1 + (N - z - y)\xi_2}.$$

To uncondition the events $Y_j = y$ and $Z_j(i) = z$, given the current $\boldsymbol{\alpha}$ and the buffer independent assumption, we invoke the Central Limit Theorem and use the following Gaussian approximation to compute $P\big(Y_j = y\big)$ and $P\big(Z_j(i) = z\big)$.

Since $Y_j = \sum_{l\in\mathcal{N}_j}(1 - b_l)$, it can be approximated, for a large value of N_j, by a Gaussian random variable with mean $\sum_{l\in\mathcal{N}_j} p_l$ and variance $\sum_{l\in\mathcal{N}_j}(1 - p_l)p_l$, where

$$p_l = P\left(\sum_{k\in\mathcal{J}_l} X_k(n) \leq t_l\right).$$

Similarly for $Z_j(i)$, since $Z_j(i) = \sum_{l\notin\mathcal{N}_j}(1-b_l)$, it can also be approximated, for a large value of $N - N_j$, by a Gaussian random variable with mean $\sum_{l\notin\mathcal{N}_j} q_l$ and variance $\sum_{l\notin\mathcal{N}_j}(1-q_l)q_l$, where

$$q_l = P\left(\sum_{k\in\mathcal{J}_l\setminus\{j\}} X_k(n) \leq t_l - i\right).$$

When an MT set contains a large number of buffers, the probabilities $P\Big(\sum_{j\in\mathcal{J}_i} X_j(n) \leq t_i\Big)$ can also be approximated by a Gaussian distribution. The required first two moments are computed from the stationary distributions of the individual buffers.

Finally,

$$\alpha_j(i) = \int_y \int_z \alpha_j(i, y, z) dF_{Y_j}(y) dF_{Z_j(i)}(z),$$

where $F_{Y_j}(y)$ and $F_{Z_j(i)}(z)$ are the respective Gaussian random variables. The integral is numerically evaluated using a 'continuity correction' to account for the fact that Y_j and $Z_j(i)$ are discrete random variables.

Observe that the structure of the threshold randomized policies facilitates delay differentiation. Firstly, buffers with different delay requirements can be differentiated by assigning them into different MT sets. Secondly, the thresholds and weights of each MT set $\mathcal{J}_i$, (t_i, ξ_i^1, ξ_i^2), are calibrated so as to provide higher allocation priorities to MT sets with more stringent delay constraints. This is indeed possible, since by lowering the threshold t_i and/or increasing the weights (ξ_i^1, ξ_i^2), the allocation priority of MT set $\mathcal{J}_i$ is increased.

VI. Practical Considerations

At this stage, it may be of benefit to the reader to make clear some important practical considerations. A pressing question is how should the length of a circuit period, denoted by T, be chosen in practice? To minimize the expected B-bit queuing delay, T should be chosen as small as possible. In fact, as long as the set of allocation probabilities ensure ergodicity, the expected B-bit queuing delay can be made arbitrarily small by choosing T arbitrarily small. This is an artifact of modeling the packet arrival process as being deterministic.

However, in practice several considerations impose constraints on the choice of T. In particular, it is essential that

T must exceed the time required to reconfigure a logical topology, which encompasses the time required to rearrange the switching fabric of an optical cross-connect and the time required for control signaling to propagate. Other considerations that may each impose a lower bound on T include:

- the processing capability of the circuit allocation decision maker may be overwhelmed for small enough T since a circuit allocation decision must be made so often;
- control signaling may consume exorbitant amounts of capacity for small enough T; and
- fast oscillating power fluctuations may appear at the input of an optical amplifier for small enough T since the logical topology undergoes such frequent reconfiguration.

Although it is hard to assign an exact numerical lower bound for T, it is clear from the above considerations that such a lower bound must exist for a practical implementation.

Another consideration that needs to be drawn to the attention of the reader is that for a stochastic packet arrival process, the circuit allocation decision maker must make a decision based on a slightly outdated record of the number of packets enqueued in each buffer, which is regarded as the buffer state. In particular, the state conveyed to the decision maker is outdated at the time a circuit allocation decision is made because the state of each buffer continues to evolve in the time it takes for the state to propagate to the decision maker and for the decision maker to process the updated state information.

It is common practice to resolve the uncertainty in the buffer state information by replacing it in the decision function with estimators based on the best available information. Note that for a deterministic packet arrival process, there is no uncertainty in the buffer state information maintained by the decision maker since the decision maker itself can exactly infer the state of each buffer based on the past decisions it made. However, if the arrival process diverts from a deterministic process, the rate used in this model is set to the long-run average rate. In such cases, the predicted performance of this model would be optimistic. Note that for wide-bandwidth networks such as optical networks, the multiplexing level is extremely high resulting in an almost deterministic arrival process. Nevertheless, the performance with 'on-off' arrival processes is a subject for future work.

Finally, its worth noting that although propagation delay is not explicitly accounted for within the framework, it is nothing more than a deterministic additive constant. Indeed it is possible that for small enough T queuing delay may be considered negligible relative to propagation delay. However, it is reasonable to suggest that the considerations listed above will require T to be set such that queueing delay will certainly not be negligible. In fact, the framework can be used to determine the range of T for which propagation delay overshadows queueing delay and vice versa.

VII. Adaptive Circuit Allocation and Implementation Aspects

To ensure computational tractability, the delay evaluation framework approximates the evolution of each buffer independently. Thus, allowing asynchronous circuit allocations of fixed durations does not invalidate the analysis derived in Sections III and IV. The circuit allocation policy, however, needs to be dynamic. That is, upon a circuit period completion, the policy must be capable of allocating one or more new circuits given that a set of circuits have been assigned.

An implementation of a circuit setting black box requires queue length monitoring and messaging to feed the allocation policy. Whether implemented centrally or distributively, a latency between the time stamps of the monitored queue lengths and the circuit setup time will always occur. Thus, a queue length prediction problem rises. In a system where our fluid traffic model applies, the predication problem is trivial since input and transmission rates are determined from the allocated circuits (which are known). Since the rates are fixed, the queue length at any moment in the future is known in advance.

A useful extension to the framework in Section II is to allow policies where the circuit allocation period may depend on the queue length. Specifically, for every queue length $X(n) = i$, a circuit is allocated with probability $\alpha(i)$ and the allocated circuit period is of length $t(i)$, which is specified in circuit periods. With probability $(1 - \alpha(i))$, the allocation attempt fails and another attempt is made after b circuit periods. (Here we adopt the same notations and definitions as in previous sections.)

An interesting case is an exhaustive policy obtained from the function $t(i) = i/(K - A)$. Here, the allocated duration is selected to exactly clear the B-bits in the queue and those that will arrive during the allocation time. Note that with this policy, if a current allocation attempt is successful, then the queue length drops to zero at the next allocation attempt. Thus, to prevent artificial steps of length zero, we fix $\alpha(0) = 0$. Moreover, since an unsuccessful allocation attempt is followed by another attempt after b circuit periods, letting $\alpha(i)$ be state dependent is redundant. Therefore, we confine ourselves to the case where $\alpha(i) = \alpha$ for $i > 0$.

To derive an expression for packet delay, the Markov chain with states given by the embedded points at which circuit allocation attempts are made is considered.

With the exhaustive policy above, the single queue length, $\big(X(n),\ n \geq 0\big)$, evolves as follows. Given $X(n) = i > 0$,

$$X(n+1) = \begin{cases} 0, & \text{w.p. } \alpha; \\ i + bA\ , & \text{w.p. } 1 - \alpha. \end{cases} \tag{25}$$

Given $X(n) = 0$,

$$X(n+1) = bA. \tag{26}$$

The expected drift in the process state in one transition is $E[X(n+1) - X(n)/X(n) = i] = (1-\alpha)bA - \alpha i$ (for $i > 0$), which is strictly negative if $i > (1-\alpha)bA/\alpha$. Thus, by the Foster-Lyapunov drift criterion [4], the Markov chain is positive recurrent.

Under stationary conditions, the derivation of the pgf is simple and yields

$$G(z) = \frac{\alpha\big[1 - p(0)(1 - z^{bA})\big]}{1 - (1-\alpha)z^{bA}}, \qquad |z| \leq 1. \tag{27}$$

To find $p(0)$, notice that the queue length drops to zero only after a successful allocation attempt, after which the queue length rises to bA in the following step. From that step forward, independent allocation attempts are made every b circuit periods, each with a probability of α of succeeding. Thus, the expected return time to state zero is $1 + 1/\alpha$. By definition, we have

$$p(0) = \frac{1}{1 + 1/\alpha} = \frac{\alpha}{1 + \alpha}.$$

The expected queue length at the embedded points is given by the derivative of $G(z)$ evaluated at $z = 1$. Simple calculation yields:

$$E(X) = \frac{\alpha b A}{1 + \alpha}. \tag{28}$$

Under stationary conditions, the time average queue length, $E\big(\tilde{X}\big)$, can be derived based on the following observation. For every given state $X(n) = i > 0$, with probability α, the queue length decreases to zero at rate $1/(K - A)$; with probability $1-\alpha$, it increases to $i+bA$ at rate A. For state $X(n) = 0$, the queue length increases to bA at rate A. Considering a simple triangle and rectangular area calculation yields

$$\begin{aligned} E\big(\tilde{X}\big) \quad &= \tfrac{\alpha^2 b^2 A}{2(1+\alpha)} + \tfrac{\alpha}{2(K-A)} E\big(X^2\big) \\ &\quad + \big(E(X) + \tfrac{bA}{2}\big)(1-\alpha)b. \end{aligned} \tag{29}$$

The time average queue length, $E\big(\tilde{X}\big)$ in (29) is expressed in terms of $E(X)$ and $E\big(X^2\big)$, where the former is given by (28). The second moment, $E\big(X^2\big)$, can be derived either from the 2^{nd} derivative of $G(Z)$, or by representing the one step evolution of $X^2(n+1)$ in a similar manner as in (25)-(26) and equating the expected values in both sides. This latter procedure is less tedious and provides the equation:

$$E\big(X^2\big) = \frac{(\alpha b A)^2}{1+\alpha} + (1-\alpha)\left[E\big(X^2\big) + (bA)^2 + 2bAE(X)\right]. \tag{30}$$

Replacing $E(X)$ in (30) with the right hand side of (28) yields the following closed form expression:

$$E\big(X^2\big) = \frac{\big[1 + 2\alpha(1-\alpha)\big](bA)^2}{\alpha(1+\alpha)}. \tag{31}$$

By Little's Lemma, the time average queueing delay of an arbitrary B-bit is $E(\tilde{D}) = E(\tilde{X})/A$. The packet delay distribution can be obtained in a similar manner as in Section III-E.

VIII. Numerical Examples

A diverse collection of symmetric, asymmetric and randomly generated networks are defined to serve as test instances for the delay evaluation framework. A discrete event simulation model is used to quantify the error introduced by approximating the evolution of each buffer independently.

For the purpose of numerical evaluation, test instances are specified by a collection of MT sets. Each test instance, or collection of MT sets, can be mapped to a realizable network consisting of a topology and routing policy. However, the network itself is irrelevant, only the collection of MT sets is of concern.

All test instances entail 100 buffers and 400 MT sets; that is, $J = 100$ and $N = 400$. Test instances are distinguished by the following two attributes:

(i) the cardinality of each MT set; and
(ii) the number of MT sets resided in by each buffer.

To reduce the number of free parameters: $m_j^0 = m^0$, $1 \le j \le 100$; $\xi_i^2/\xi_i^1 = 10$, $1 \le i \le 400$; and $t_i = |\mathcal{J}_i|$, $1 \le i \le 400$. In words: the proportionality between the arrival bit rate and the service bit rate is the same for all buffers, the ratio between the upper and lower weights is 10 for all MT sets, and the threshold of an MT set is chosen as its cardinality.

Test instances are classified as symmetric (S), asymmetric (A) and random (R). In a symmetric test instance, the cardinality of all MT sets is equal and all buffers reside in an equal number of MT sets. An asymmetric test instance allows the cardinality of each MT set and the number of MT sets resided in by each buffer to vary in a strictly *deterministic* manner. Finally, a randomly generated test instance is such that the cardinality of each MT set and the number of MT sets resided in by each buffer varies according to a statistical distribution. For all test instances, MT sets are necessarily unique. An integer programming approach is used to ensure the feasibility of all symmetric and asymmetric test instances, in which not all combinations of attributes (i) and (ii) are feasible.

Test instances are chosen to reflect the full range of accuracies that may be expected with the delay evaluation framework and are defined in the following.

(S1) Each of the 100 buffers resides in n, $n = 160, 200, 240$, of the 400 MT sets. Thus, the cardinality of each MT is given by $100n/400$; that is, a cardinality of 40, 50 and 60, respectively.

(A1) Each of the 100 buffers resides in 240 MT sets. The 400 MT sets are evenly divided such that 200 are of cardinality 40, and 200 are of cardinality 80. Thus, each buffer resides in 80 MT sets of cardinality 40, and 160 MT sets of cardinality 80.

(A2) A variation of (A1). Each of the 100 buffers resides in 180 MT sets. The 400 MT sets are evenly divided such that 100 are of cardinality 30, 100 are of cardinality 40, 100 are of cardinality 50 and 100 are of cardinality 60.

(A3) Of the 100 buffers, 50 reside in 240 MT sets, and 50 reside 160 MT sets, referred to as *Class 1* and *Class 2* buffers, respectively. Thus, the cardinality of each MT set is 50 and the composition of each MT set is such that 30 of the 50 buffers reside in 240 MT sets and 20 of the 50 buffers reside in 160 MT sets.

(R1) Each of the 100 buffers resides in a random number of MT sets according to the discrete uniform distribution on the interval $[160, 240]$. Thus, the expected MT set cardinality is 50. Buffers are randomly allocated to MT sets and it is ensured each MT set is unique.

(R2) Of the 100 buffers, 50 reside in a random number of MT sets according to the discrete uniform distribution on the interval $[230, 240]$ and 50 according to the discrete uniform distribution on the interval $[160, 170]$, referred

to as *Class 1* and *Class 2* buffers, respectively. Thus, the expected MT set cardinality is 50.

For each test instance, the expected B-bit delay, which quantifies the expected queueing time of an arbitrary B-bit, $E(D)$, is computed both, by the delay evaluation framework in (15), and by the simulation model. The results are plotted as a function of m^0, $m^0 = 3, 4, 5, 6, 7$. Recall that m^0 is the ratio of the service bit rate to the arrival bit rate. The expected B-bit delay is expressed in units of circuit periods. That is, unity B-bit delay corresponds to the length of a circuit period, T. Plots generated by the simulation model are shown within 95% confidence intervals. For random test instances, the expected B-bit delay is quantified as an average across 3 independent trials. Plots are shown in Figs 1-5.

All test instances demonstrate that the expected delay generated by the evaluation framework and simulation model are in good agreement, particularly for a high load, which is represented by $m^0 = 3$.

For larger values of m^0, the quality of the error margin varies and the analytical frameworks always provide an upper bound. Specifically, an error margin of less than 1% is attained for $m^0 = 3$. The maximum error margins for all test instances are given in Table I. Observe that test instances, in which all buffers do not reside in the same number of MT sets, such as test instances (A3) and (R2), give rise to the greatest error margin.

Five approximations contribute to the error margin, they are as follows:

(i) approximating the evolution of each buffer independently;
(ii) approximating the probability $P(Y_j = y)$, for each buffer j, with a normal distribution;
(iii) approximating the probability $P(Z_j(i) = z)$, for each buffer j and buffer size i, with a normal distribution;
(iv) approximating the probability $P(\sum_{k\in\mathcal{J}_l} X_k(n) \leq t)$, for each MT set $\mathcal{J}_l$, with a normal distribution; and,
(v) approximating $\alpha(i) = \overline{\alpha}$ for $i \geq m^0 - 1$.

Secondary approximations, such as assuming integral arrival and service bit rates, are implemented in the simulation model, and thus do not contribute to the error margin.

By normal approximation, it is meant the Central Limit Theorem is invoked to approximate the distribution of a sum of independent random variables. The normal approximation is accurate if the number of MT sets is sufficiently large and the number of buffers residing in each MT set is sufficiently close to half the total number of MT sets. The accuracy of the normal approximation is compromised if the number of MT sets is small, in which case the probabilities $P(\sum_{k\in\mathcal{J}_l} X_k(n) \leq t)$ will be poorly approximated, or if the number of buffers residing in each MT set is either small or almost equal to the total number of MT sets, in which case the probabilities $P(Z_j(i) = z)$ and $P(Y_j = y)$ are poorly approximated, respectively. The normal approximation may be avoided in such instances where the number of buffers residing in each MT set is small, by computing the appropriate probabilities exactly by summing over all possible permutations.

Approximating $\alpha(i) = \overline{\alpha}$, for $i \geq m^0 - 1$ introduces error, if the threshold $t \geq m^0 - 1$. For example, if $t \geq m^0 - 1$, $P(Z_j(m^0 - 1) = z) \neq P(Z_j(t + 1) = z) = 0$, however, $\alpha(m^0 - 1) = \alpha(t + 1) = \overline{\alpha}$, since $\alpha(i) = \overline{\alpha}$, for $i \geq m^0 - 1$. Therefore, if $t \geq m^0 - 1$, $\overline{\alpha}$ is approximated such that $\overline{\alpha} = \sum_{i=m^0-1}^{K} \overline{p}(i)\alpha(i)$, where $\overline{p}(i) = \frac{p(i)}{\sum_{i=m^0-1}^{K} p(i)}$ and $K \gg m^0 - 1$ represents a numerical truncation point.

To quantify the error introduced by approximating $\overline{\alpha}$ as such, three symmetric test instances are defined, in which the normal approximations are avoided by considering only 12 buffers residing in MT sets of cardinality c, $c = 5, 6, 7$. The expected B-bit delay, $E(D)$, is plotted as a function of the threshold t, $t = 0, 1, \ldots, 7$, for $m^0 = 4$ in Fig. 6. Observe the increased error margin for $t \geq m^0 - 1 = 3$. For $t \leq m^0 - 2 = 2$, the error margin is less than one percent and is completely attributable to approximating the evolution of each buffer independently.

As shown in Figs. 1-5, the expected B-bit delay is monotonic in the proportionality between the arrival and service bit rate, m^0. For most test instances, the expected B-bit delay is less than one circuit period for $m^0 = 6, 7$, indicating a bit is transmitted in its arriving circuit period with high probability. The expected B-bit delay is not plotted for $m^0 = 1, 2, 3$, because the underlying Markov chain is not ergodic for some test instances given $m^0 \leq 3$.

The computational time required by the framework to generate an estimate of B-bit queuing delay for a test instance never exceeds one minute. In contrast, the simulation demands several days of computation time to generate an equivalent estimate within acceptable confidence intervals. This is one of the key advantages the delay evaluation framework has to offer.

Test Instance	Max Error Margin
(S1) n = 160	0.36%
(S1) n = 200	6.6%
(S1) n = 240	1.6%
(A1)	13.9%
(A2)	6.4%
(A3) Class 1	24.5%
(A3) Class 2	24.1%
(R1)	5.8%
(R2) Class 1	2.2%
(R2) Class 2	9.3%

TABLE I
MAXIMUM ERROR MARGIN

IX. CONCLUSIONS

A framework was provided for evaluation of packet delay distribution in an optical circuit-switched network. The framework was based on a fluid packet arrival and service rate model, in which packets are assigned to a buffer of an edge router, based on delay constraint and destination, and enqueued.

Two types of circuit allocation policies were integrated into the framework. First, circuit holding times were of fixed duration and allocated at the boundary of fixed time frames ('limited'), and second, circuit holding times were adaptive to

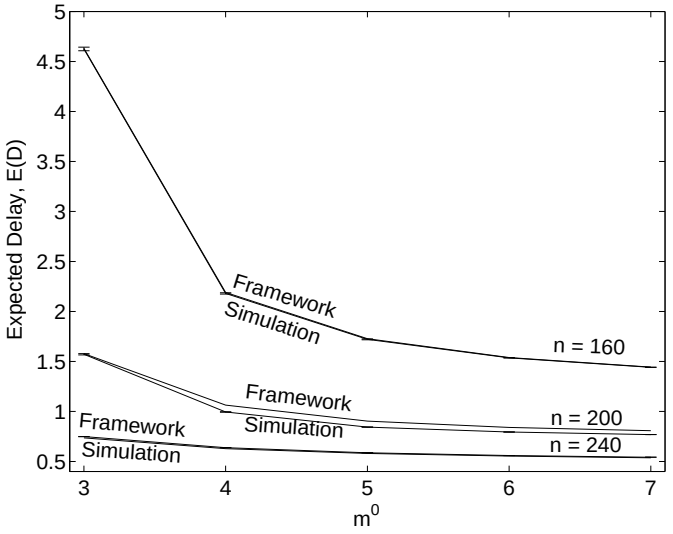

Fig. 1. Test instance (S1). Expected B-bit delay in units of T as a function of proportionality between arrival bit rate and service bit rate.

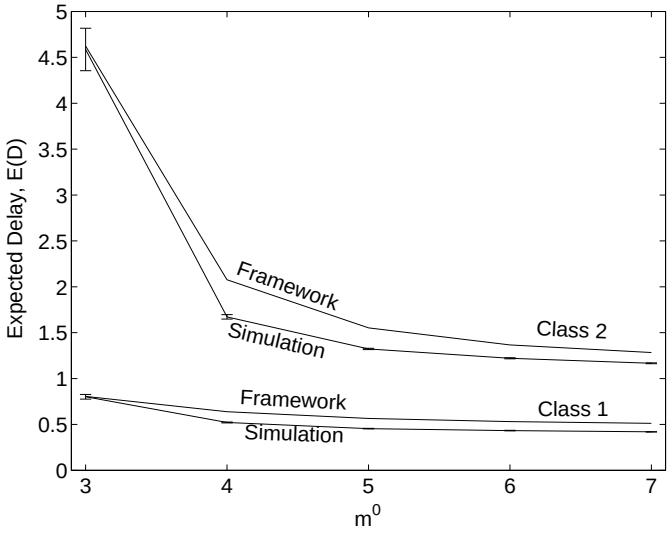

Fig. 3. Test instance (A3). Expected B-bit delay in units of T as a function of proportionality between arrival bit rate and service bit rate.

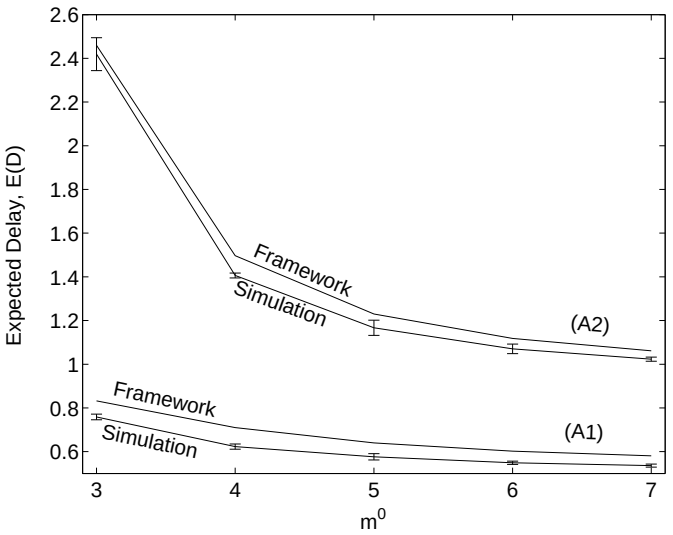

Fig. 2. Test instance (A1) and (A2). Expected B-bit delay in units of T as a function of proportionality between arrival bit rate and service bit rate.

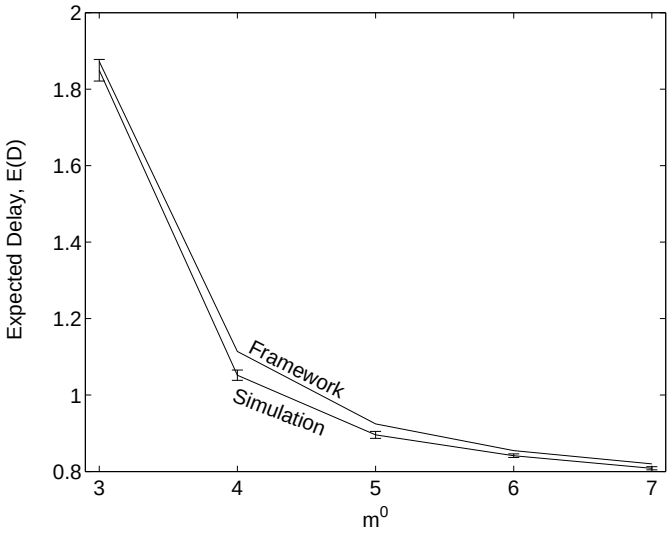

Fig. 4. Test instance (R1). Expected B-bit delay in units of T as a function of proportionality between arrival bit rate and service bit rate.

buffer size such that the holding time was sufficient to empty a buffer ('exhaustive').

The framework approximates the evolution of each queue length independently. 'Slack variables' were introduced to decouple amongst buffers in a way that the evolution of each queue length remains consistent with all other queue lengths in the network. The exact delay distribution was derived for a single buffer and an approximation was given for a network of buffers. The approximation entailed finding a fixed point for the functional relation between the 'slack variables' and a specific allocation policy.

An analysis of a circuit allocation policy, in which circuits are probabilistically allocated based on queue lengths, was given as an illustrative example. The framework was shown to be in good agreement with a discrete event simulation model.

APPENDIX

A. Derivation of $G(z)$

Using (8), the state relabelling, the definition of $G(z)$ and the fact that $\alpha(i) = \overline{\alpha}$ for $i \geq m^0 - 1$, $G(z)$ can be separated into the following two summations:

$$G(z) = \sum_{i=0}^{m^0-2} [\alpha(i) z^{(i+1-m^0)^+} + (1-\alpha(i)) z^{i+1}] p(i)$$

$$+ \sum_{i=m^0-1}^{\infty} [\overline{\alpha} z^{(i+1-m^0)^+} + (1-\overline{\alpha}) z^{i+1}] p(i).$$

For the first summation, $i+1-m^0 < 0$, and for the second

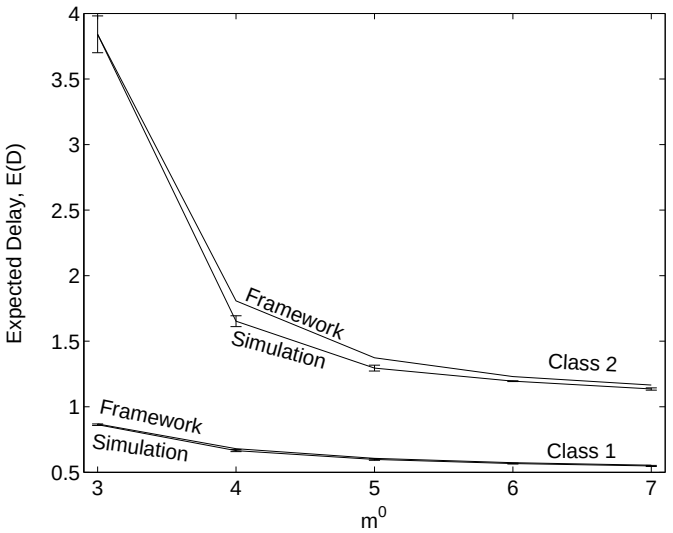

Fig. 5. Test instance (R2). Expected B-bit delay in units of T as a function of proportionality between arrival bit rate and service bit rate.

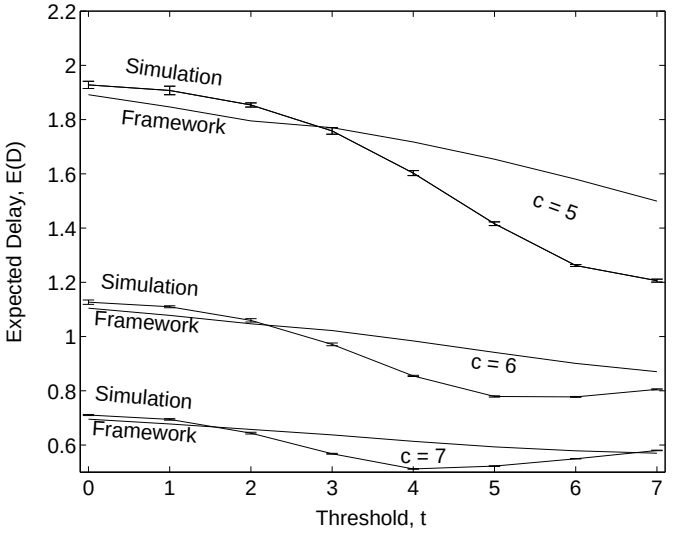

Fig. 6. Effect of varying threshold. Expected B-bit delay in units of T as a function of threshold, t. Observe the increased error margin for $t \geq m^0 - 1 = 3$.

summation, $i + 1 - m^0 \geq 0$, thus

$$G(z) = \sum_{i=0}^{m^0-2} [\alpha(i) + (1-\alpha(i))z^{i+1}]p(i) + \sum_{i=m^0-1}^{\infty} [\overline{\alpha} z^{i+1-m^0} + (1-\overline{\alpha})z^{i+1}]p(i).$$

Multiplying by z^{m^0-1} and rearranging the second summation yields

$$G(z) = \sum_{i=0}^{m^0-2} [\alpha(i) + (1-\alpha(i))z^{i+1}]p(i) + (\overline{\alpha} z^{1-m^0} + z - \overline{\alpha} z) \sum_{i=m^0-1}^{\infty} p(i) z^i.$$

Since $\sum_{i=m^0-1}^{\infty} p(i)z^i = G(z) - \sum_{i=0}^{m^0-2} p(i)z^i$, the second summation can be written in terms of $G(z)$ giving the following implicit equation for $G(z)$,

$$G(z) = \sum_{i=0}^{m^0-2} [\alpha(i) + (1-\alpha(i))z^{i+1}]p(i) + (\overline{\alpha} z^{1-m^0} + z - \overline{\alpha} z)\left(G(z) - \sum_{i=0}^{m^0-2} p(i) z^i \right).$$

Elementary rearrangements give

$$G(z) = \frac{\sum_{i=0}^{m^0-2} [\alpha(i) z^{m^0-1} - \overline{\alpha} z^i + (\overline{\alpha} - \alpha(i)) z^{m^0+i}] p(i)}{z^{m^0-1} - \overline{\alpha} - z^{m^0} + \overline{\alpha} z^{m^0}}.$$

B. Properties of $G(z)$

First we show that the denominator of $G(z)$ has $K - A$ distinct zeros. Represent the denominator of $G(z)$, $h(z)$, as a sum of the two functions $f(z) = z^{K-A}$ and $g(z) = -\big(\overline{\alpha} + (1-\overline{\alpha})z^K\big)$.

Clearly, $f(z)$ has a single zero of order $K - A$ at 0. Furthermore, for every z on the unit contour $|z| = 1$,

$$|g(z)| \leq |f(z)| \tag{32}$$

and the derivatives of $f(z)$ and $g(z)$ satisfy $\frac{df(Z)}{d|z|} = K - A$ and $\frac{dg(Z)}{d|z|} = (1-\overline{\alpha})K$, respectively.

From the ergodicity condition (9), $\frac{df(Z)}{d|z|} > \frac{dg(Z)}{d|z|}$ on the contour $|z| = 1$. Combined with (32), it follows that $|g(z)| < |f(z)|$ for every z on any contour $|z| = 1 + \delta$, where $\delta > 0$. Invoking Rouche's Theorem, $f(z)$ and $f(z) + g(z)$ have the same number of zeros within every contour $|z| = 1 + \delta$, where $\delta > 0$. That is, within and onto the unit disk $|z| = 1$. Since $f(z)$ has $K - A$ zeros, so does the denominator of $G(z)$, $h(z) = f(z) + g(z)$.

Next we show that all zeros must be distinct (i.e., of order one). Suppose in contradiction that they are not distinct. Then the derivative of $h(z)$ at any multiplicative must vanish. However, the derivative of $h(z)$, $h'(z)$, is given by: positive in $|z| \leq 1$:

$$\begin{aligned} |h'(z)| &= |(K-A)z^{K-A-1} - (1-\overline{\alpha})Kz^{K-1}| \\ &\geq |(K-A)z^{K-A-1}| - |(1-\overline{\alpha})Kz^{K-1}| \\ &= (K-A)|z|^{K-A-1} - (1-\overline{\alpha})K|z|^{K-1}. \end{aligned}$$

It is easily verified that the ergodicity condition (9) is equivalent to $|h'(z)| > 0$, for every z in $|z| \leq 1$. Thus, all zeros are distinct.

C. Derivation of Delay Distribution for $\alpha(i) = \alpha$

Under stationary conditions, assume a circuit period begins at time 0. That is, $P(X(0) = i) = p(i)$, where the set $\{p(i)\}$ is derived in Subsections III-B and III-D. The duration of each packet arrival is $1/M$ circuit periods. The 1^{st} packet starts its arrival at time 0 and every subsequent packet m, $2 \leq m \leq M$, starts its arrival upon the arrival completion of

packet $m-1$. (Note that the $\{D_m\}$ are statistically dependent.) First, we derive the distribution of D_1 and then we express the remaining $M-1$ distributions recursively.

By assuming $\alpha(i)=\alpha$, the number of circuit periods between two consecutive circuit allocations, S, is geometrically distributed with a success probability of α. The pgf of S is given by

$$G_S(z)=\frac{z\alpha}{1-z(1-\alpha)},\quad |z|\le 1. \tag{33}$$

Let τ_j, $j\ge 1$, be the number of circuit periods between the $j-1$ and the j^{th} circuit allocation, using the convention that allocation 0 is done at time 0. The random variables $\{\tau_j\ ;\ j\ge 1\}$ are independent and geometrically distributed taking values $1,2,3,\ldots$. Note that τ_j includes the j^{th} allocated circuit period used for transmission. From (33), the pgf of the summation $\tau^{(n)}=\sum_{j=1}^{n}\tau_j$ is given by $[G_S(z)]^n$.

It is now shown that an integral number of packets reside within a buffer at every circuit period boundary. Let $b(k)$ be the number of packets transmitted during an allocated circuit period, given that there are k packets at the beginning of the circuit period. If $k\ge(m^0-1)M$, the queue at the buffer is drained at rate m^0M packets per period, and therefore $b(k)=m^0M$. If $k<(m^0-1)M$, the buffer queue is drained at rate m^0M during the first period fraction of $k/(m^0-1)M$, and at rate M during the rest of the period, implying $b(k)=k+M$. Thus,

$$b(k)=\begin{cases} m^0\cdot M, & \text{if}\quad k\ge(m^0-1)M;\\ \\ k+M, & \text{otherwise.}\end{cases} \tag{34}$$

Consequently, at every circuit period boundary, there is an integral number of packets whose distribution is given by

$$q(k)\stackrel{\text{def}}{=}p(kL),\quad k\ge 0. \tag{35}$$

The number of circuits period needed to transmit k packets at rate m^0M is $n(k)\stackrel{\text{def}}{=}\lceil k/m^0M\rceil$. All, but possibly the last circuit period, are fully used to transmit the k packets. The utilization of the last circuit period is given by $1-frac(k)$, where $frac(k)\stackrel{\text{def}}{=}n(k)-k/m^0M$.

Let $d_m^1(k)$ ($d_m^2(k)$) be the delay of the m^{th} arriving packet given that there are k packets at time 0 and the first circuit period is allocated (not allocated), where $1\le m\le M$.

Suppose that k packets are present at time 0. If the first circuit period is not allocated, the k present packets and the M first arrivals are all transmitted at rate $1/m^0M$. Thus, for $m=1$,

$$\begin{aligned} d_1^2(k) &= 1+\tau^{n(k)}+\tfrac{1}{m^0M}\\ &\quad+\mathcal{I}\{k\in\mathcal{Z}_0\}(\tau-1)-\mathcal{I}\{k\notin\mathcal{Z}_0\}frac(k)\\ &=\tau^{n(k)}+\tfrac{1}{m^0M}+\mathcal{I}\{k\in\mathcal{Z}_0\}\tau\\ &\quad+\mathcal{I}\{k\notin\mathcal{Z}_0\}(1-frac(k)),\end{aligned} \tag{36}$$

where $\mathcal{Z}_0$ is the set of all positive integer multiples of m^0M; $\mathcal{I}\{E\}$ is the set indicator function; and τ is an independent geometric random variable with success probability α.

For $m\ge 2$,

$$\begin{aligned} d_m^2(k) &= d_{m-1}^2(k)-\tfrac{1}{M}+\tfrac{1}{m^0M}\\ &\quad+\mathcal{I}\{k+m-1\in\mathcal{Z}_0\}(\tau-1).\end{aligned} \tag{37}$$

If the first circuit period is allocated, then (34) implies that the m^{th} arriving packet is served in the first circuit period if and only if $k+m\le m^0M$. Moreover, packet m completes its transmission when it completes its arrival if and only if the queue length drops to zero no later than m/M. That is, if and only if $k/(m^0-1)M\le m/M$, which is equivalent to $k\le(m^0-1)m$. Therefore, implying the following.

For $m\ge 1$ and $0\le k\le(m^0-1)m$,

$$d_m^1(k)=\frac{1}{M}. \tag{38}$$

For $m\ge 1$ and $(m^0-1)m<k\le m^0M-m$, the k present packets and the first m arrivals are all served at rate m^0M. Since the m^{th} arrival starts at time $(m-1)/M$ and the $k+m$ packets complete their transmission at time $(m+k)/m^0M$, we have

$$d_m^1(k)=\frac{m+k}{m^0M}-\frac{m-1}{M}. \tag{39}$$

From (38) and (39) it follows that for $k\le m^0M-m$,

$$d_m^1(k)=\max\left\{\frac{1}{M},\frac{m+k}{m^0M}-\frac{m-1}{M}\right\}. \tag{40}$$

For $k\ge m^0M-m+1$, the m^{th} packet is not transmitted during the first circuit period. Similar to the derivations of (36)–(37), we have for $m=1$

$$\begin{aligned} d_1^1(k) &= \tau^{n(k-m^0M)}+\tfrac{1}{m^0M}+\mathcal{I}\{k\in\mathcal{Z}_0\}\tau\\ &\quad+\mathcal{I}\{k\notin\mathcal{Z}_0\}(1-frac(k)).\end{aligned} \tag{41}$$

For $m\ge 2$,

$$\begin{aligned} d_m^1(k) &= d_{m-1}^1(k)-\tfrac{1}{M}+\tfrac{1}{m^0M}\\ &\quad+\mathcal{I}\{k+m-1\in\mathcal{Z}_0\}(\tau-1).\end{aligned} \tag{42}$$

Let $d^i(k)=\frac{1}{M}\sum_{m=1}^{M}d_m^i(k)$, $i=1,2$. From (20),

$$D=\begin{cases} d^1(k), & \text{w.p.}\ \ \alpha q(k);\\ \\ d^2(k), & \text{w.p.}\ \ (1-\alpha)q(k).\end{cases} \tag{43}$$

By (36) and (37), the distribution of the random variable $d^2(k)$ is expressed by

$$\begin{aligned} d^2(k) &= \tau^{n(k)}+\tfrac{M+1}{2m^0M}-\tfrac{M-1}{2M}+\mathcal{I}\{k\in\mathcal{Z}_0\}\tau\\ &\quad+\mathcal{I}\{k\notin\mathcal{Z}_0\}(1-frac(k))\\ &\quad+\tfrac{\tau-1}{M}\sum_{m=1}^{M-1}(M-m)\mathcal{I}\{k+m\in\mathcal{Z}_0\}.\end{aligned} \tag{44}$$

Similar to the above derivation but using (40)–(42) results in the following expression for $d^1(k)$.

For $k \geq m^0 M$,

$$\begin{aligned} d^1(k) \quad &= \tau^{n(k-m^0M)} + \tfrac{M+1}{2m^0M} - \tfrac{M-1}{2M} \\ &+\mathcal{I}\{k \in \mathcal{Z}_0\}\tau + \mathcal{I}\{k \notin \mathcal{Z}_0\}(1 - frac(k)) \\ &+\tfrac{\tau-1}{M}\sum_{m=1}^{M-1}(M-m)\mathcal{I}\{k+m \in \mathcal{Z}_0\}. \end{aligned} \tag{45}$$

For $k < m^0 M$,

$$\begin{aligned} d^1(k) \quad &= \tfrac{1}{M}\sum_{m=1}^{\min\{M, m^0M-k\}} \max\left\{\tfrac{1}{M}, \tfrac{m+k}{m^0M} - \tfrac{m-1}{M}\right\} \\ &+\mathcal{I}\{k > (m^0-1)M\}\tfrac{S(k)}{M}, \end{aligned} \tag{46}$$

where $S(k)$ is the total delay contribution of the packets transmitted during the second circuit holding time, derived as follows.

Note that for this case we have $(m^0-1)M < k < m^0M$, and packets $m = 1, \ldots, m^0M - k$ are transmitted during the first circuit holding time and packets $m = m^0M - k + 1, \ldots, M$ are transmitted during the second circuit holding time.

Let $d_m(k)$ be the delay of packet m, $m = m^0M - k + 1, \ldots, M$, given k packets at time 0. Thus,

$$d_{m^0M-k+1}(k) = \tau + \frac{1}{m^0M} - \frac{m^0M-k+1}{M}, \tag{47}$$

and for $2 \leq j \leq k - (m^0-1)M$,

$$d_{m^0M-k+j}(k) = d_{m^0M-k+j-1}(k) + \frac{1}{m^0M} - \frac{1}{M}. \tag{48}$$

From (47)–(48), $S(k) = \sum_{m=m^0M-k+1}^{M} d_m(k)$ and can be computed by recursion.

Note that both random variables, $d^1(k)$ and $d^2(k)$, are linear combinations of independent geometric distributions. Therefore, the conditional histogram of D, given k packets at a circuit boundary, can be computed from (44)–(48). The unconditional histogram of D is then derived using the distribution $\{q(k)\}$ as derived in (18), (19) and (35).

REFERENCES

[1] D. J. Blumenthal, J. E. Bowers, L. Rau, H. F. Chou, S. Rangarajan, W. Wang and H. N. Poulsen, "Optical signal processing for optical packet switching networks," *IEEE Optical Communication*, Feb. 2003, pp. S23-S29.

[2] M. Duser and P. Bayvel, "Analysis of a dynamically wavelength routed optical burst switched network architecture," *IEEE J. Lightwave Technology*, vol. 20, no. 4, April 2002, pp. 574–585.

[3] M. Duser and P. Bayvel, "Performance of a dynamically wavelength routed optical burst switched network," *IEEE Photonic Technology Letters*, vol. 14, no. 2, Feb. 2002, pp. 239–241.

[4] F. G. Foster, "On the stochastic matrices associated with certain queueing processes," *Annals of Mathematical Statistics*, vol. 24, no. 3, 1953, pp. 355-360.

[5] M. Gondran and M. Minoux, *Graphs and Algorithms*, John Wiley and Sons, 1986.

[6] J. F. Hayes, *Modelling and Analysis of Computer Communication Networks*, Plenum, 1984.

[7] F. P. Kelly, "Blocking probabilities in large circuit-switched networks," *Advances in Applied Probability*, vol. 18, 1986, pp. 473-505.

[8] A. Mokhtar and M. Azizoglu, "Adaptive wavelength routing in all-optical networks," *IEEE/ACM Trans. Networking*, vol. 6, no. 6, April 1998, pp. 197-206.

[9] J. R. Munkres, *Elements of Algebraic Topology*, Addison-Wesley, 1984.

[10] C. Qiao, "Labeled optical burst switching for IP-over-WDM integration," *IEEE Communications Mag.*, Sept. 2000, pp. 104-114.

[11] R. Ramaswami and K. N. Sivarajan, "Routing and wavelength assignment in all-optical networks," *IEEE/ACM Trans. Networking*, vol. 3, Oct. 1995, pp. 489-500.

[12] Z. Rosberg, H. L. Vu, M. Zukerman, and J. White, "Performance analyses of optical burst switching networks," *IEEE J. Selected Areas in Communications*, vol. 21, no. 7, Sept. 2003.

[13] Z. Rosberg, "Circuit allocation in optical networks with average packet delay cost criterion," Oct. 30, 2003. (Submitted for publication.)

[14] J. Y. Wei, J. L. Pastor, R. S. Ramamurthy and Y. Tsai, "Just-in-time optical burst switching for multi-wavelength networks," in *Proc. the 5th Int. Conf. on Broadband Communications*, 1999, pp. 339-352.

[15] I. Widjaja, "Performance analysis of burst admission-control Protocols," *IEE Proceedings – Communications*, vol. 142, no. 1, Feb. 1995, pp. 7-14.

[16] A. Zalesky, E. W. M. Wong, M. Zukerman, H. L. Vu and R. S. Tucker, "Performance analysis of an OBS edge router," *IEEE Photonic Technology Letters*, vol. 16, no. 2, pp. 695-697, Feb. 2004.