The Notion of End-to-End Capacity and its Application to the estimation of End-to-End Network Delays

Han S. Kim* and Ness B. Shroff**
*Samsung Electronics
Suwon, Korea
hs365.kim@samsung.com
**School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
shroff@ecn.purdue.edu
http://yara.ecn.purdue.edu/∼shroff

Abstract

In this paper we develop a new notion called the *end-to-end capacity* in terms of input and output processes of an end-to-end path. The end-to-end capacity is defined for a path of interest, and we show that the end-to-end path can be represented by a single-node model with the end-to-end capacity in the sense that the single-node model is equivalent to the original path in terms of the queue-length and departure traffic. This allows us to estimate the end-to-end delay distribution based on endpoint measurements. We also investigate the applicability of this approach to admission control. This is the first attempt to estimate the end-to-end delay distribution itself.

1 Introduction

The queue-length distribution and loss probability for a single node have been extensively studied [1, 2, 3, 4, 5, 6]. However, while much work has focused on single-node analysis, extensions to the end-to-end case have not been very successful. If we simply apply single-node analysis to each node on the end-to-end path to stay within end-to-end quality of service (QoS) guarantees, it could result in a highly inefficient utilization of network resources, and also cause scalability problems. For example, suppose that the QoS requirement for each flow is to maintain the probability of exceeding the end-to-end delay threshold D (also called the delay violation probability) to be less than ϵ. Further, assume that we have a tool for estimating this delay violation probability only for single-node systems. A simple way to guarantee the end-to-end QoS is to estimate the delay violation probability p_D for delay threshold $D' = D/n$ (where n is the number of nodes on the path) at each node on the path and maintain p_D less than ϵ. This simple approach may result in an unnecessarily small end-to-end delay violation probability, hence a lower utilization and a waste of network resources. Moreover, it could also result in scalability problems because the per-flow delay violation probability would need to be managed even at core nodes within the network that usually serve a very large number of flows.

There has been an attempt to reduce the inefficient usage of network resources by optimally setting the QoS level at each node depending on traffic models and the type of QoS metric [7] being used. It has been investigated via a simulation study that the convolution of delay distributions of all the nodes on the path is quite close to the actual end-to-end delay distribution [8]. However, such approaches are still not scalable. Moreover, in the latter approach, in order to estimate the end-to-end delay violation probability at just one threshold D, the delay distribution of each node needs to be calculated for the entire range of the convolution.

In this paper we develop a new notion called the "end-to-end capacity" and demonstrate its applicability to the estimation of the end-to-end delay distribution[1] based on endpoint measurements. The underlying idea is the following. We define the end-to-end capacity as the maximum capacity that can be allocated to a path for a given traffic and set of connections. Then, a single-node model with this end-to-end capacity is equivalent to the original end-to-end path in the sense that they have the same end-to-end queue length,[2] and hence, they also have the same departure and end-to-end delay. On the other hand, it is well known for a single-node model that when the capacity is constant, say c, the delay violation probability is equal to the tail probability scaled by c. Let W and Q represent the steady state versions of the queueing delay and the queue length, respectively. Then, $\mathbb{P}\{W > x\} = \mathbb{P}\{Q > cx\}$. We find a similar relationship in the case of non-constant capacity, $\mathbb{P}\{W > x\} \approx \mathbb{P}\{Q > \bar{c}x\}$, where $\bar{c}$ is the mean of the capacity. In particular, we have shown that for any $\delta > 0$, $\mathbb{P}\{Q > (\bar{c} + \delta)x\} \leq \mathbb{P}\{W > x\} \leq \mathbb{P}\{Q > (\bar{c} - \delta)x\}$ for all sufficiently large x. Based on these results, we first estimate the queue-length distribution in the single-queue model by endpoint measurements in order to avoid the scalability problem, and then obtain the end-to-end delay distribution.

In fact, the idea was initiated in our previous work [9]. However, we have found that the method proposed in [9] to measure the statistics of the end-to-end capacity during *busy periods* could result in inaccuracies because busy-period statistics may not accurately represent the entire statistics, especially for long-range dependent traffic. In this paper, we propose a new method that is more tolerant of traffic that exhibit long-range dependence.

The main contributions of this paper are:

(i) We introduce a notion of the end-to-end capacity and demonstrate its applicability to estimating the end-to-end delay distribution based on endpoint measurements. Our approach is the first attempt to estimate the end-to-end delay distribution itself.

(ii) We show that the end-to-end path can be represented by a single-node with the end-to-end capacity in

[1] Throughout the paper we consider only the delay due to queueing unless stated otherwise. The other terms contributing to the end-to-end delay are typically constant factors such as the transmission time and the propagation delay, and are ignored.

[2] We refer the queue length as the amount of workload.

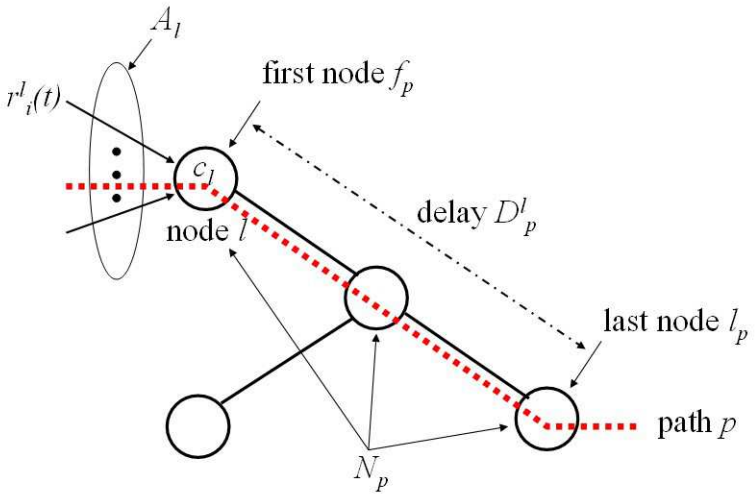

Figure 1: Symbol definitions

the sense that they are identical in terms of the end-to-end queue length.

(*iii*) We apply the above results to admission control providing end-to-end QoS.

This paper is organized as follows. In Section 2, we derive a single node representation and its equivalence to the end-to-end path of interest. We also show certain properties of the end-to-end capacity. In Section 3, we describe how to apply the notion of the end-to-end capacity in estimating the end-to-end delay distribution. By proposing the idea of measuring the end-to-end capacity, we can approximately compute the end-to-end delay distribution. In Section 4, we provide simulation results showing the accuracy of the approximation and demonstrate its applicability to admission control. In Section 5, we further discuss our approach and compare it with related works. We conclude in Section 6. All proofs are provided in the Appendix.

2 System Model

We consider a discrete time system serving flows within a class using a FIFO discipline. A path is defined as a set of links and nodes connecting the source to the destination.

2.1 Definitions

We provide the following notation that will be used throughout the text. The corresponding symbols can be found in Fig. 1.

- N_p := set of nodes belonging to path p
- f_p := first node (ingress node) of path p

- $l_p :=$ last node (egress node) of path p
- $A_l :=$ set of flows on node l
- $B_p := \bigcap_{l \in N_p} A_l =$ set of flows traversing path p
- $c_l :=$ capacity of node l
- $D_p^l :=$ constant delay between node l and the last node of path p, excluding the queueing delay
- $r_i^l(t) :=$ rate (or the input workload) of flow i entering node l at time t
- $d_i^l(t) :=$ rate (or the output workload) of flow i departing node l at time t
  (If flow i moves from node 1 to node 2 with delay D, $r_i^2(t) = d_i^1(t - D)$.)
- $a_l(t) := c_l - \sum_{i \in A_l} d_i^l(t) =$ unused capacity of node l at time t
- $c_p(t) := \sum_{i \in B_p} d_i^{l_p}(t) + \min_{l \in N_p} a_l(t - D_l) =$ end-to-end capacity of path p (defined as the maximum capacity that can be allocated to the path for given flows and connections)
- $q_p(t) := \sum_{k=1}^{t} \sum_{i \in B_p} r_i^{f_p}(k - D_{f_p}) - \sum_{k=1}^{t} \sum_{i \in B_p} d_i^{l_p}(t) =$ end-to-end queue length of path p (the summation of the workload, belonging B_p, at each node on path p)
- $w_p(t) := \min\{s : \sum_{k=1}^{t} \sum_{i \in B_p} r_i^{f_p}(k - D_{f_p}) - \sum_{k=1}^{t+s} \sum_{i \in B_p} d_i^{l_p}(k) \le 0\} =$ end-to-end (queueing) delay[3] of path p

2.2 Equivalent Single-Queue Representation

We now develop a single-queue representation of an end-to-end path. For simplicity, let us first consider a path consisting of two nodes, as shown in Fig. 2. In this case, f_p and l_p are node 1 and node 2, respectively. Let $r_1(t)$ represent the aggregate traffic rate generated by the flows entering node 1 and traversing path p, i.e. $r_1(t) = \sum_{i \in B_p} r_i^{f_p}(t)$. Similarly, $r_2(t) = \sum_{i \in B_p} r_i^{l_p}(t)$. Let $r_1'(t)$ and $r_2'(t)$ represent the cross traffic, i.e., $r_1'(t) = \sum_{i \in A_{f_p} \cap B_p^c} r_i^{f_p}(t)$ and $r_2'(t) = \sum_{i \in A_{l_p} \cap B_p^c} r_i^{l_p}(t)$, and $d_1'(t)$ and $d_2'(t)$ represent the corresponding departures. Assume that the departure at the first node arrives at the second node D time slots later. By replacing $c_i(t) = C_i - d_i'(t), i = 1, 2$, we can re-draw the path as if the cross traffic were a part of the capacity (Fig. 3). We will compare this (Fig. 3) with the single-node model (Fig. 4) where $c_p(t) = d_2(t) + \min\{a_1(t - D), a_2(t)\}$.

Proposition 2.1 *Assume that* $q_1(0) = q_2(D) = q_p(D) = 0$. *Then,* $q_1(t - D) + q_2(t) = q_p(t)$, *and* $d_2(t) = d_p(t)$, $\forall t \ge D$.

[3] $w_p(t)$ defined here is the delay seen by the last packet arriving at the first node at time slot t.

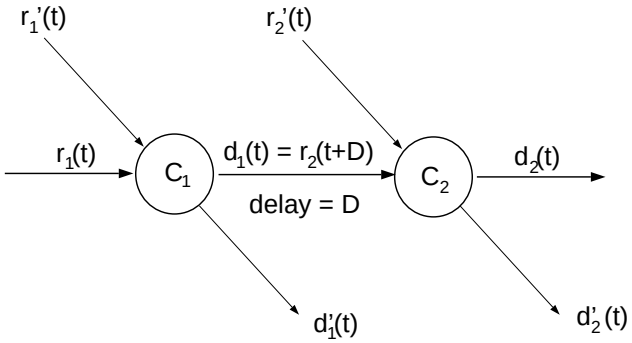

Figure 2: Original path

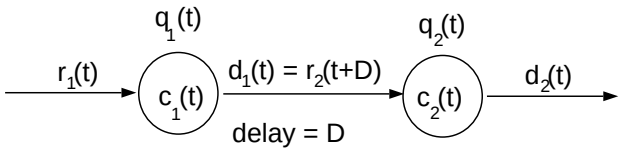

Figure 3: Two-node model

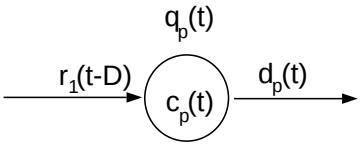

Figure 4: Single-node model

Proposition 2.1 tells us that once both queues become empty, which is ensured as long as the system is stable, the queue length and the departure processes in the single-node model are thereafter identical to those on the original path.

Thus far, we have considered a path with two nodes. It can easily be extended to a path of more than two nodes as follows. Find $c_p(t)$ for the last two nodes, and replace them with a single node. Then, repeat this for the new last two nodes, etc. We can also extend this to multi-class cases simply by treating other classes as cross traffic.

Note that $c_p(t) = \min\{c_1(t-D) + q_2(t-1), c_2(t)\}$. Hence, it is possible that $c_p(t) > c_1(t-D)$ for some instant t. Nonetheless, we are able to show the following result.

Proposition 2.2 *Assume that all flows are stationary and ergodic,*[4] *and that* $\mathbb{E}\{r_i(t) + r_i'(t)\} < C_i$ *for stability. Then,*

$$\mathbb{E}\{c_p(t)\} \leq \min_{i=1,2} \mathbb{E}\{c_i(t)\}.$$

[4] A process is said to be stationary when the expectations of the process are time-invariant, and ergodic when the expectations are same as its time averages. If the above statement is true up to 2nd order, it is said to be stationary and ergodic in the wide sense.

Since $\mathbb{E}\{c_i(t)\} = \mathbb{E}\{C_i - d'_i(t)\} = C_i - \mathbb{E}\{r'_i(t)\} \geq \mathbb{E}\{c_p(t)\}$, we also have $\mathbb{E}\{r_1(t-D) - c_p(t)\} \geq \mathbb{E}\{r_i(t) + r'_i(t) - C_i\}$, the following conjecture is likely to be true.

Conjecture A:

$$\mathbb{P}\{r_1(t-D) > c_p(t)\} \geq \max_{i=1,2} \mathbb{P}\{r_i(t) + r'_i(t) > C_i\}. \quad (1)$$

Conjecture A has a practically important meaning. We can infer from the conjecture that, for example, if an admission decision is made such that the overflow probability of the single-node model is less than ϵ, the overflow probability at each node on the path is also less that ϵ (this is demonstrated later in the paper via numerical studies).

3 Estimation of the End-to-End Delay Distribution

An important application of $c_p(t)$ is that it can be used to estimate the end-to-end delay distribution by endpoint measurements. We first assume that the steady state versions of $Q_p(t)$ and $W_p(t)$ exist, and denote them by Q_p and W_p, respectively. The existence of the steady state distribution has been shown by Loynes [10] if the input process is stationary and ergodic and the capacity is strictly larger than the mean input rate. We believe (but have not proven) that if all the input flows are stationary and ergodic, so is $c_p(t)$. However, the focus of Section 3 is not to provide a thorough justification of the estimation but to demonstrate a possible application of $c_p(t)$. Hence, in this paper, we simply assume that the steady state versions exist.

We have empirically found that the end-to-end queue-length distribution scaled by $\bar{c}_p := \mathbb{E}\{c_p(t)\}$ closely matches the end-to-end delay distribution. Fig. 5 shows an example. The path consists of two nodes as in Fig. 3, $r_1(t) = 100$ Markov-Modulated Fluid (MMF) processes,[5] $c_1(t) =$ Gaussian process with mean 45 (or 53) and $\mathrm{Cov}(t) = 10 \times 0.9^t$, and $c_2(t) =$ Gaussian process with mean 47 (or 53) and $\mathrm{Cov}(t) = 10 \times 0.8^t$ (the resulting $\bar{c}_p$ is 42 (or 51)). This figure is obtained assuming perfect knowledge of $c_p(t)$. In practice, however, $c_p(t)$ is not known without obtaining information from all nodes on the path, and should be estimated by endpoint measurements. From Fig. 5, we can see that $\mathbb{P}\{Q_p > x\} \approx \mathbb{P}\{W_p > x/\bar{c}_p\}$ where Q_p and W_p represent the steady state versions of the end-to-end queue length and the end-to-end delay, respectively. Hence, we can approximate the end-to-end delay distribution by means of the end-to-end queue-length distribution and the end-to-end capacity. The reason we are first dealing with the queue-length distribution (which is later scaled to approximate the end-to-end delay distribution) rather than directly focusing on the delay distribution is that the end-to-end queue length at time t can be represented by the summation of the queue length at each node at

[5] MMF processes are widely used to model voice sources [11, 12]. Throughout the paper, all MMF sources will have the transition matrix of [0.99666, 0.00334; 0.005, 0.995] and the rate vector of [0, 0.17]. These values are chosen to represent a voice source for a 45 Mbps link with 2 msec time slot and 53 byte ATM cell.

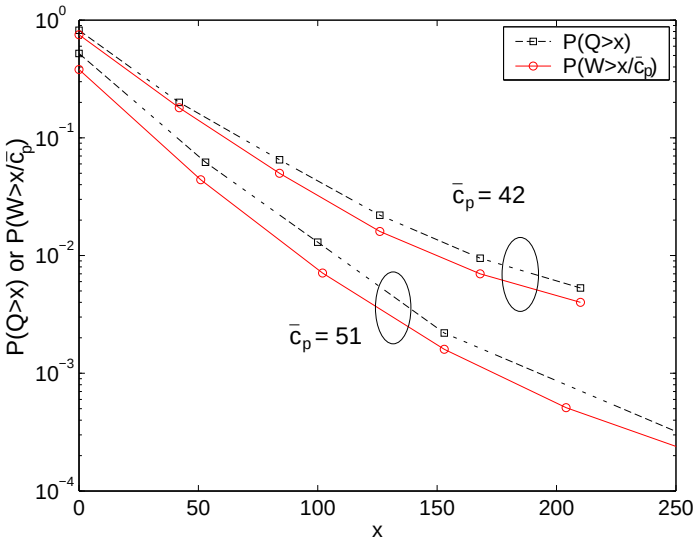

Figure 5: Comparison of $\mathbb{P}\{Q_p > x\}$ and $\mathbb{P}\{W_p > x/\bar{c}_p\}$.

time t. However, the end-to-end delay seen by the last packet of time slot t at the first node is not the simple sum of the delays seen by the last packet of time slot t at each node because last packets at different nodes will be different.

To support the approximation, $\mathbb{P}\{W_p > x\} \approx \mathbb{P}\{Q_p > \bar{c}_p x\}$, we investigate the relationship between $\mathbb{P}\{W_p > x\}$ and $\mathbb{P}\{Q_p > \bar{c}_p x\}$. For long-range dependent (LRD) input traffic,[6] which results in $\mathbb{P}\{Q_p > x\}$ to decay slower than an exponential, i.e., $\lim_{x\to\infty} \frac{e^{-\alpha x}}{\mathbb{P}\{Q_p > x\}} = 0$ for any $\alpha > 0$, we have the following result.[7]

Proposition 3.1 *Assume that* $\{c_p(t)\}_{t\geq 1}$ *are independent of* $Q_p(0)$ *and* $\frac{1}{\sqrt{x}} \sum_{t=1}^{x} c_p(t)$ *converges in distribution to a Gaussian, and that for any* $\alpha > 0$ $\lim_{x\to\infty} \frac{e^{-\alpha x}}{\mathbb{P}\{Q_p > x\}} = 0$. *Then, for any* $\delta > 0$, *there exist an* x_0 *such that*

$$\mathbb{P}\{Q_p > (\bar{c}_p + \delta)x\} \leq \mathbb{P}\{W_p > x\} \leq \mathbb{P}\{Q_p > (\bar{c}_p - \delta)x\}, \qquad \forall x \geq x_0. \tag{2}$$

What the proposition means is that the delay distribution is captured by queue-length distributions scaled by either $(\bar{c}_p + \delta)$ or $(\bar{c}_p - \delta)$. Since δ can be arbitrarily small, $\mathbb{P}\{W_p > x\}$ and $\mathbb{P}\{Q_p > \bar{c}_p x\}$ are likely to be similar.

3.1 MVA Approximation for Delay

As illustrated in Fig. 5, once we have the mean end-to-end capacity and an estimate of the end-to-end queue-length distribution, we can approximate the end-to-end delay distribution by scaling the end-to-end queue-length distribution. So we first estimate the tail of the end-to-end queue-length distribution.

[6] It is well known that a queue serving long-range dependent input processes has a heavy-tailed queue-length distribution. Since the end-to-end queue length is a sum of queue lengths at each node along the path, any queue serving long-range dependent traffic results in heavy-tailed end-to-end queue-length distribution.

[7] We believe that this proposition is also true for short-range dependent traffic, and thus Eq. (2) may hold more generally. However, we have been able to prove it only for the LRD case.

We treat the path p as a virtual single node with input $\sum_{i\in B_p} r_i^{f_p}(k-D_p^{f_p})$ and capacity $c_p(t)$. For simplicity, we rewrite the input as $r_1(t-D)$ to represent the aggregate input $\sum_{i\in B_p} r_i^{f_p}(k-D_p^{f_p})$. We then estimate the tail probability of the virtual single node queue-length distribution by applying an existing single-node technique. It has been found that the Maximum Variance Asymptotic (MVA) approach (first named in [4]) provides an accurate estimate of the tail probability for a large class of input processes including long-range dependent processes. The MVA method was originally developed for the case when a large number of network applications are multiplexed so that the aggregate traffic can be characterized by a Gaussian process. It has been numerically shown that the MVA method also works well for non-Gaussian cases, even when the number of multiplexed sources is moderate (e.g., tens of sources) [4, 6]. Hence, although the net input $r_1(t-D)-c_p(t)$ may not be accurately modeled as Gaussian, we can still apply the MVA method to approximately estimate the tail probability by

$$\mathbb{P}\{Q_p > x\} \approx e^{-m_x/2} \tag{3}$$

where

$$m_x := \min_{t\geq 1} \frac{(x+(\bar{c}_p-\bar{r}_1)t)^2}{\mathrm{Var}\{X(1,t)\}}, \tag{4}$$

and $X(1,t)=\sum_{k=1}^{t}[r_1(k-D)-c_p(k)]$. Note that $X(1,t)$ is defined as the accumulation process. It is introduced here for analysis because the queue length is a result of all the past inputs as well as the current input.

An important question is how to obtain $\bar{c}_p$ and $\mathrm{Var}\{X(1,t)\}$. This will be explained in the following subsection.

Then, as Fig. 5 suggests, we approximate the end-to-end delay by

$$\mathbb{P}\{W_p > x\} \approx e^{-m_{\bar{c}_p x}/2}, \tag{5}$$

and we call this *MVA approximation for delay*, or shortly *MVA-delay*.

3.2 Measuring the Moments of $X(1,t)$

The MVA method requires the first two moments of $X(1,t)$. We describe here how the egress node can learn about $r_1(t)$ and $c_p(t)$. We assume that the ingress node inserts time-stamps to record the arrival time of packets [13, 14]. The egress node can infer $r_1(t)$ from this time-stamp.

A naive way to deliver the information about $c_p(t)$ to the egress node is to update the unused-capacity field in the packet header at each node. However, this results in scalability issues at core nodes. Fortunately, in our previous work, it has been shown that core nodes serving large numbers of flows with a large capacity compared to edge nodes can be ignored from the point of view of end-to-end analysis [15, 16] (See also Fig. 8 in the next

section). Thanks to this result, it is possible for the egress node to calculate $c_p(t)$ at the cost of additional functionality in the ingress node only.

When the ingress node f_p inserts a time-stamp for each arriving packet to record its arrival time, it could insert one more value to denote the amount of unused capacity in the previous time slot. Then, for a departing packet with time-stamp t, the egress node l_p can determine $c_p(t+D_p^{f_p}-1)$ by comparing its own unused capacity, $a_{l_p}(t+D_p^{f_p}-1)$, with the value recorded in the packet, $a_{f_p}(t-1)$. This comparison only needs to be done for one packet per path per time slot. If no packet of path p with time-stamp t is found at the egress node, we can simply set $c_p(t+D_p^{f_p}-1) = a_{l_p}(t+D_p^{f_p}-1)$. Then, we can calculate $X(1,t) = \sum_{k=1}^{t}[r_1(k-D) - c_p(k)]$. Since the unused capacity $a_l(t)$ is the same for all paths on node l, the implementational complexity is not that high.

3.3 Direct Measurement of $\mathbb{P}\{W_p > x\}$

Before we investigate the accuracy of our approach by numerical experiments, we pose the following question: "Can we not directly measure the delay distribution, especially when the estimation is after all based on *measurements*"? The problem in directly measuring the delay distribution is that it may require a very long time to measure a small value.

For example, measuring a probability of 10^{-6} typically requires more than 10^7 samples. If the link speed is 45Mbps and the packet size is 53Bytes, 10^7 packet time is about 100sec which is too long. However, measuring moments of flows, which is required in the MVA method, can be done over a much shorter duration and usually with higher reliability. Table 1 compares the duration of the simulation runs required to keep the 90% confidence interval to be within an order of magnitude on the logarithmic scale. In this experiment, 200 MMFs and an AR Gaussian input with mean 30 and covariance $\mathrm{Cov}(t) = 10 \times 0.9^t$ are multiplexed in a queue with capacity 118. The first row of Table 1 tells us that direct measurement requires 2×10^6 simulation cycles to measure a delay probability of 10^{-5} while measuring moments requires only 2×10^4 simulation cycles. When the target value becomes 10 times smaller (changed from 10^{-5} to 10^{-6}), direct measurement requires a 50 times longer duration (changed from 2×10^6 to 1×10^8) while measuring moments for the MVA method requires only a 4 times longer duration (changed from 2×10^4 to 8×10^4).

Table 1: Comparison of the required simulation cycles

target value	direct measuring	moments measuring
10^{-5}	2×10^6 cycles	2×10^4 cycles
10^{-6}	1×10^8 cycles	8×10^4 cycles

4 Numerical Experiments

In this section we investigate how the MVA approximation for delay performs and how it can be applied for admission control for providing end-to-end QoS. In the numerical experiments conducted, the propagation delay between nodes is set to 0. Here, we use fractional Gaussian noise (fGn),[8] which is a classical example of a self-similar process. There are several methods to generate self-similar traffic [17, 18, 19], and we use the random midpoint replacement algorithm [19]. When the value of the input process is not an integer, as many packets as the integer part of the value are generated and the remainder is passed to the next time slot. The duration of simulation is set to be over a million cycles to obtain reasonable reliability.

4.1 Approximation of the End-to-End Delay

In the first experiment, we use a five node network with 4 paths (Fig. 6). The input processes of each path are

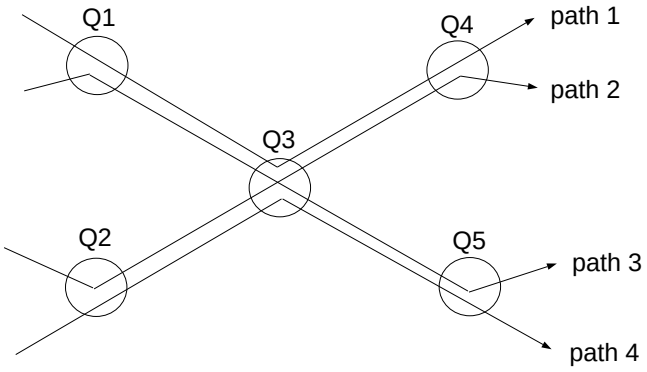

Figure 6: Five node network

either MMF processes (with the same parameters as before)[9] or fGn as follows:

200 MMFs and fGn(150,50,0.8) for path 1,

100 MMFs and fGn(150,50,0.8) for path 2,

100 MMFs and fGn(150,50,0.6) for path 3,

200 MMFs and fGn(150,50,0.8) for path 4.

We consider three scenarios (A), (B), and (C). The capacities for each scenario is given by:

(A) $C_1 = C_4 = 410, C_2 = C_5 = 420, C_3 = 840$;

[8] Fractional Gaussian noise is the increment process of fractional Brownian Motion (fBM) process and has autocovariance function is given by

$$C_\lambda(l) = \frac{\sigma^2}{2}(|l-1|^{2H} + |l+1|^{2H} - 2|l|^{2H}),$$

where $H \in [0.5, 1)$ is the Hurst parameter. A fGn with mean $= \bar{\lambda}$, variance $= \sigma^2$, and Hurst value $= H$ will be denoted as fGn$(\bar{\lambda}, \sigma^2, H)$.

[9] The MVA approach has been found to be more accurate when input processes are either only short range dependent processes such as MMF and AR processes or only one kind of fGn [4, 6]. So, we provide results when both short and long range dependent processes are mixed.

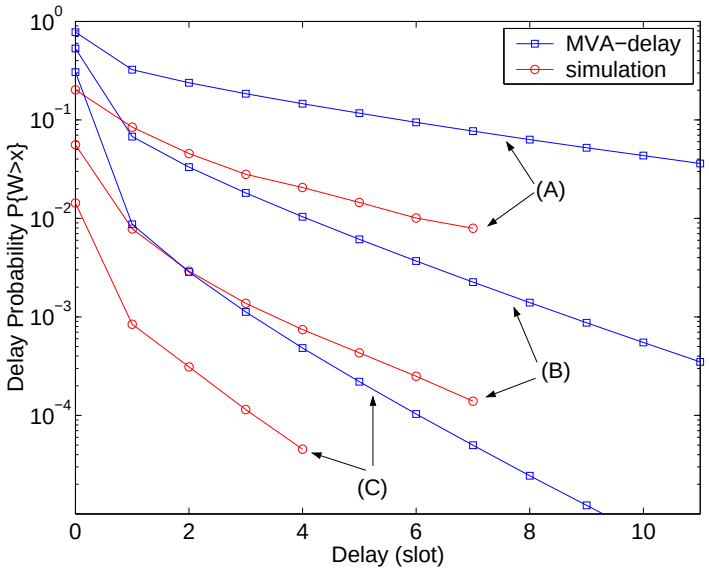

Figure 7: Approximation of the end-to-end delay for path 4.

(B) $C_1 = C_4 = 415, C_2 = C_5 = 425, C_3 = 845$;

(C) $C_1 = C_4 = 420, C_2 = C_5 = 430, C_3 = 850$.

These values are chosen so that both short and the long range dependent traffic are mixed and long range dependent traffic is dominant. In particular, the Hurst parameter of 0.8 is chosen to be similar to the Hurst parameter of the MPEG video trace, which will be used in the next experiment. It should be noted that Q3 with large capacity where all the flows are multiplexed is a core node that does not alter the unused-capacity field of the packet header. The result is shown in Fig. 7 for three cases, (A), (B) and (C). From Fig. 7 we can see that the approximation bounds the actual delay within an order of magnitude.

In the second experiment, we investigate a more practical scenario where the network is serving voice and video traffic. The same five node network with 4 paths (Fig. 6) is used and each path carries the following traffic types:

800 voice and 14 video sources for path 1,

1000 voice and 9 video sources for path 2,

700 voice and 14 video sources for path 3,

500 voice and 19 video sources for path 4.

The capacities are: $C_1 = C_2 = C_4 = C_5 = 210$pkt/slot, $C_3 = 450$pkt/slot, which are chosen for a 45Mbps link with 2ms time slot and 53byte packet.[10] We use the MMF model to represent a voice source and actual MPEG

[10]If we are interested in rare events of delay violation with large threshold, the impact of the slot size is negligible. We choose 2ms slot size here, but the result with a 10ms slot size is almost the same.

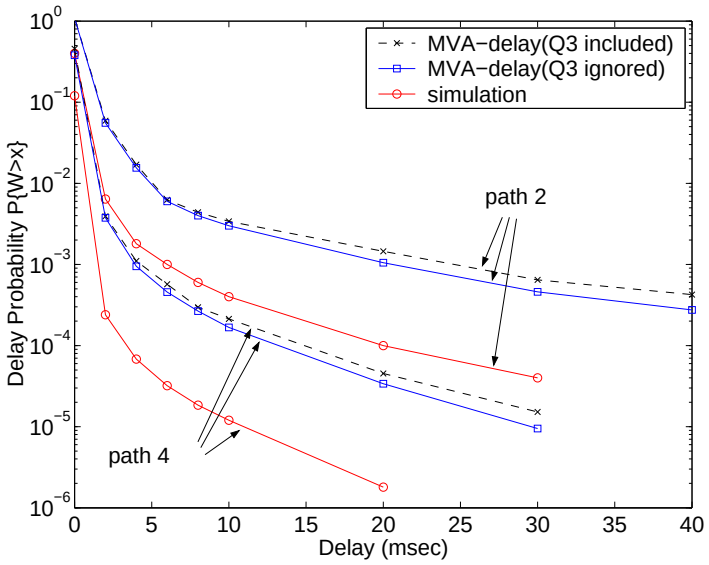

Figure 8: Approximation of the end-to-end delay for path 2 and 4.

video trace for the video sources. The video trace comes from an MPEG-encoded action movie (007 series) which has been found to exhibit long-range dependence [20]. Fig. 8 shows that the approximation still captures the actual delay. Note that there are two types of estimation curves and that the difference is quite small. The solid lines are for the case where the core node (Q3 in Fig. 6) does not change the unused-capacity field in the packet header, and the dashed lines are for the case where the core node also updates the unused-capacity field in the packet header. This figure implies that the core node can be ignored as explained before, and expected from [15, 16].

In the next experiment, we increase the network size with 9 nodes and 7 paths where path 4 consists of 5 nodes (Fig. 9). The traffic configuration is similar to the previous experiment. Paths 1 through 4 have the same traffic as in Fig. 6. Paths 5, 6, and 7 have the same traffic with paths 1, 2, and 3, respectively. The capacities are: $C_1 = C_2 = C_4 = C_6 = C_8 = C_9 = 210$pkt/slot, $C_3 = C_5 = C_7 = 450$pkt/slot, where Q3, Q5, and Q7 are core nodes that do not touch the unused-capacity field of the packet header. Fig. 10 shows the MVA-delay curve for path 4. From Fig. 10 we can see that if core nodes have a relatively large capacity, the estimation error does not increase much. Although the error is one or two orders of magnitude, it should be noted that the MVA-delay captures the shape of the actual delay curve. This means that, as will be demonstrated in Section 4.2, such an approximation can be used for admission control and the achievable utilization will be conservative, but still quite close to the maximum utilization for a given QoS.

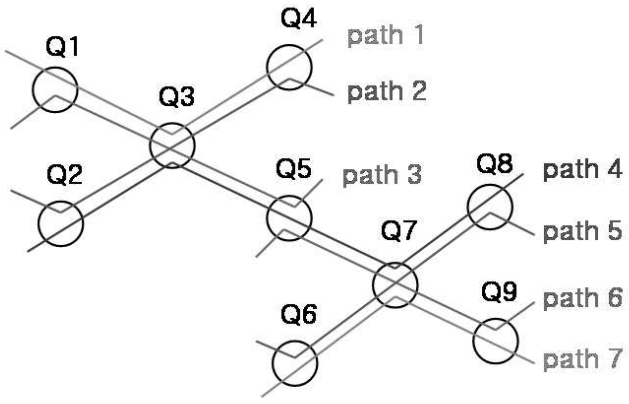

Figure 9: Nine node network

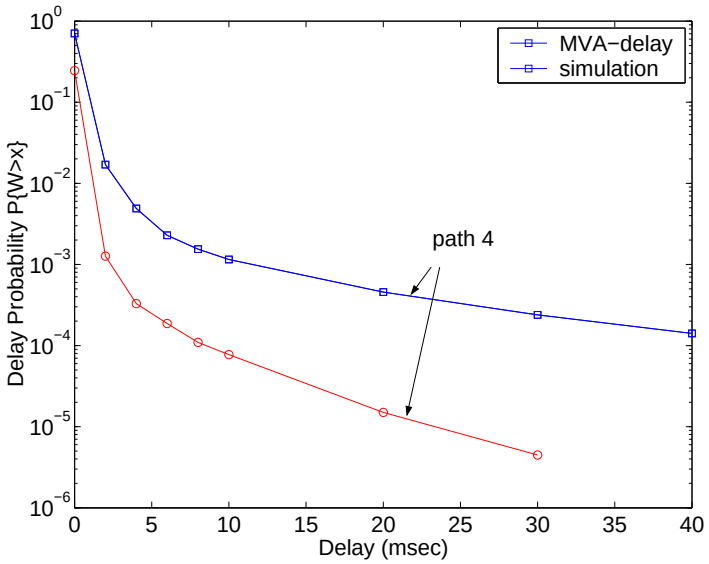

Figure 10: Approximation of the end-to-end delay for path 4.

4.2 Application to Admission Control

Many admission control algorithms are based on single-node analysis [21, 22, 23] and the admission decision is made by estimating the QoS at a node, for example, the overflow probability that the aggregate flow rate is greater than the capacity of the node. In order to provide a sort of end-to-end QoS, this type of test needs to be performed at all nodes on the path including core routers that serve a large number of flows, thus causing the scalability problem. We can apply the concept of the end-to-end capacity and the end-to-end delay distribution to admission control without the scalability problem because admission control is done only by the edge nodes.

One way of implementing admission control is based on the end-to-end overflow probability. Once a new flow request for a path arrives, the edge node on that path will estimate the end-to-end overflow probability, which is defined as the probability that the aggregate input to the path is greater than the end-to-end capacity. In this experiment we use a Gaussian approximation to estimate the overflow probability. Let μ and σ^2 be the mean and variance of a new flow, $\hat{c}_p$ and $\hat{\sigma}_c^2$ be the measured mean and variance of the end-to-end capacity, $\hat{\mu}_r$ and $\hat{\sigma}_r^2$ be the measured mean and variance of the existing aggregate flow on the path, and ϵ be the target QoS. Then, a new flow is admitted if

$$Q\left(\frac{\hat{c}_p - \hat{\mu}_r - \mu}{\sqrt{\hat{\sigma}_c^2 + \hat{\sigma}_r^2 + \sigma^2}}\right) < \epsilon, \tag{6}$$

where $Q(\cdot)$ is the complementary cdf of a standard Gaussian random variable, $N(0,1)$. Note that this admission control assumes that existing flows are stationary. The effect of flow dynamics has been taken into account in [22] but only for a system with zero buffers. In this experiment, we fix path 2,3, and 4 with 1400 voice flows, and perform admission control for path 1 in the same five node network in Fig. 6. Table 2 compares the number of flows admitted by the proposed algorithm with the maximum obtained by simulation. From Table 2, we can see that the number of admitted flows by our algorithm is conservative but close to the maximum and that the target QoS is met.

Table 2: Admission control by the end-to-end overflow probability

target QoS	# by our algo.	max #	actual QoS at Q_1, Q_3, Q_4
10^{-3}	1452	1466	$8.6 \times 10^{-4}, 1.2 \times 10^{-4}, 8.3 \times 10^{-4}$
10^{-4}	1240	1249	$2.6 \times 10^{-5}, 4.2 \times 10^{-6}, 1.9 \times 10^{-5}$

Another way of implementing admission control is based on the end-to-end delay violation probability. Let μ and $v(t)$ be the mean and the variance function[11] of a new flow, $\hat{c}_p$ and $\hat{v}_c(t)$ be the measured mean and variance function of the end-to-end capacity, $\hat{\mu}_r$ and $\hat{v}_r(t)$ be the measured mean and variance function of the

[11] The variance function $v(t)$ is defined as the variance of the accumulated input during $[1, t]$.

existing aggregate flow on the path, and ϵ be the target QoS. Then, a new flow is admitted if

$$\sup_{t\geq 1} \frac{\hat{v}_c(t) + \hat{v}_r(t) + v(t)}{[x + (\hat{c}_p - \hat{\mu}_r - \mu)t]^2} > -2\log\epsilon. \tag{7}$$

In this experiment, we fix path 2,3, and 4 with 35 video flows, and perform admission control for path 1 in the same five node network in Fig. 6. We set D to 20, i.e., $40ms$. From the result in Table 3, we can see that the number of admitted flows by our algorithm is again conservative but close to the maximum, and the target QoS is met.

Table 3: Admission control by the end-to-end delay violation probability

target QoS	# by our algo.	max #	actual QoS
10^{-5}	32	33	3.4×10^{-7}
10^{-6}	29	31	1.6×10^{-8}

5 Related Work

Although our approach is motivated by empirical observations, it is nevertheless valuable because the MVA approximation for delay is the first attempt to estimate the end-to-end delay distribution itself. Existing works on delay have focused on a deterministic end-to-end delay bound [24], or a statistical per-node delay bound [25, 26]. In [24], the maximum (or worst-case) end-to-end delay is calculated for regulated traffic. In [25, 26], an upper bound on the delay distribution at a single node is obtained when the amount of input is statistically bounded by a traffic envelope.

The admission control algorithm in [13] is based on the end-to-end delay violation probability. Based on this algorithm, a new flow with peak-rate envelope $r(t)$ is admissible with delay bound x and confidence level $\Phi(\alpha)$ if

$$t\bar{R}(t) + tr(t) - \bar{S}(t+x) + \alpha\sqrt{t^2\sigma^2(t) + \psi^2(t+x)} < 0 \tag{8}$$

for all interval lengths $0 \leq t \leq T$, and $\lim_{t\to\infty} \bar{R}(t) + r(t) \leq \lim_{t\to\infty} \frac{\bar{S}(t)}{t}$. Here, the existing aggregate flow on the path has a maximum arrival envelope with mean $\bar{R}(t)$ and variance $\sigma^2(t)$, and the end-to-end path has a minimum service envelope with mean $\bar{S}(t)$ and variance $\psi^2(t)$, T is the length of the measurement window, $\Phi(\alpha) = \exp(-\exp(-\frac{\alpha-\lambda}{\delta}))$, $\delta = \sqrt{\frac{6}{\pi^2}(t^2\sigma^2 + \psi^2(t+x))}$, and $\lambda = t\bar{R}(t) + tr(t) - \bar{S}(t+x) - 0.57772\delta$. Hence, we can infer from the result of [13] that $\mathbb{P}\{W_p > x\} \leq 1 - \min_{0\leq t\leq T} \Phi(\alpha_x)$ where α_x is the minimum value such that (8) is satisfied for given x. This is an upper bound on the end-to-end delay distribution. Since the bound is obtained by the *maximum* arrival envelope and the *minimum* service envelope, it could be quite loose in terms

of predicting the actual delay probability. The performance of this algorithm also depends on the value of T. In [21], the impact of T has been investigated when only the arrival envelope is used. Considering both arrival envelope and service envelope, what we have found is that the performance for different values of T can be quite different, and that either very small or very large values of T may cause a significant error. It is expected that a very large T will result in significant underutilization because the envelopes become deterministic as T goes to ∞ so that the delay bound provided by the test (8) will be the worst-case delay.

Based on (4), it appears at first that $\mathrm{Var}\{X(1,t)\}$ needs to be evaluated for the entire range of t due to the *min* operation over $\{t \geq 1\}$. However, it has been shown that the value of t (or the *dominant time scale*) at which $\frac{(x+(\bar{c}_p-\bar{r}_1)t)^2}{\mathrm{Var}\{X(1,t)\}}$ takes its minimum can be determined by measuring $\mathrm{Var}\{X(1,t)\}$ for values of t only up to a bound on the dominant time scale [27]. This makes the MVA approach amenable for on-line measurements.

6 Conclusion

In this paper, we have proposed the notion of the end-to-end capacity and demonstrated its applicability to the estimation of the end-to-end delay distribution. We have shown that the end-to-end path can be represented by a single-node with a certain end-to-end capacity. These two systems are identical in terms of the end-to-end queue length. Thus, they also have the same departure process and the same end-to-end delay distribution. Further, we have empirically found that $\mathbb{P}\{W_p > x\} \approx \mathbb{P}\{Q_p > \bar{c}_p x\}$. We have shown that for LRD traffic, for any $\delta > 0$, $\mathbb{P}\{Q_p > (\bar{c}_p + \delta)x\} \leq \mathbb{P}\{W_p > x\} \leq \mathbb{P}\{Q_p > (\bar{c}_p - \delta)x\}$ for all sufficiently large x.

Based on these results, we have proposed an estimation technique for the end-to-end delay distribution (called the *MVA approximation for delay*). In particular, we estimate the delay distribution for the single-node model that is equal to the end-to-end delay distribution for the original path. We obtain the estimation of the delay distribution for the single-node model by estimating the queue-length distribution first and then scaling it by the mean end-to-end capacity. Since the estimation is done by using only endpoint measurements, the scheme is scalable. We have also validated our estimation by numerical experiments with various traffic types including long-range dependent traffic. Unlike existing work on the delay that has focused on the maximum (or worst-case) end-to-end delay [24] or a bound on the per-session delay distribution at a single node [25, 26], our approach is the first attempt to estimate the end-to-end delay distribution itself. The error caused due to such an approach is mainly because we are not measuring the end-to-end capacity but calculating the queue-length distribution. A possible avenue for future work would be to improve the approximation by removing the assumption that $X(1,t)$ is Gaussian (at the expense of some measurement complexity). Another avenue of

future work is extending the work to incorporate multi-class traffic.

References

[1] R. G. Addie and M. Zukerman, "An Approximation for Performance Evaluation of Stationary Single Server Queues," *IEEE Transactions on Communications*, vol. 42, no. 12, pp. 3150–3160, Dec. 1994.

[2] N. G. Duffield and Neil O'Connell, "Large Deviations and Overflow Probabilities for the General Single Server Queue, with Application," *Proc. Cambridge Philos. Soc.*, vol. 118, pp. 363–374, 1995.

[3] P. W. Glynn and W. Whitt, "Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single-Server Queue," *Journal of Applied Probability*, pp. 131–155, 1994.

[4] J. Choe and N. B. Shroff, "A Central Limit Theorem Based Approach for Analyzing Queue Behavior in High-Speed Networks," *IEEE/ACM Transactions on Networking*, vol. 6, no. 5, pp. 659–671, Oct. 1998.

[5] N. Likhanov and R. R. Mazumdar, "Cell-Loss Asymptotics in Buffers fed with a Large Number of Independent stationary sources," in *Proceedings of IEEE INFOCOM*, San Francisco, CA, 1998.

[6] H. S. Kim and N. B. Shroff, "Loss Probability Calculations and Asymptotic Analysis for Finite Buffer Multiplexers," *IEEE/ACM Transactions on Networking*, vol. 9, no. 6, pp. 755–768, Dec. 2001.

[7] R. Nagarajan, J. Kurose, and D. Towsley, "Local Allocation of End-to-End Quality-of-Service in High-Speed Networks," *IFIP Transactions C-Communication Systems*, vol. 15, pp. 99–118, 1993.

[8] D. Yates, J. Kurose, and D. Towsley, "On Per-Session End-to-End Delay and the Call Admission Problem for Real-Time Applications with QOS Requirements," *Journal of Highspeed Networks*, vol. 3, no. 4, pp. 429–458, 1994.

[9] H. S. Kim and N. B. Shroff, "An Approximation of the End-to-End Delay Distribution," in *Proceedings of 11th International Workshop on Quality of Service*, Monterey, CA, 2003, pp. 59–75.

[10] R. M. Loynes, "The Stability of a Queue with Non-independent Inter-arrival and Service Times," *Proc. Cambridge Philos. Soc.*, vol. 58, pp. 497–520, 1962.

[11] J. N. Daigle and J. D. Langford, "Models for Analysis of Packet Voice Communication Systems," *IEEE Journal on Selected Areas in Communications*, vol. 4, pp. 847–855, Sep. 1986.

[12] K. Sriram and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexer for Voice and Data," *IEEE Journal on Selected Areas in Communications*, vol. 4, pp. 833–846, Sep. 1986.

[13] C. Cetinkaya, V. Kanodia, and E. Knightly, "Scalable Services via Egress Admission Control," *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 69–81, 2001.

[14] P. Yuan, J. Schlembach, A. Skoe, and E. Knightly, "Design and Implementation of Scalable Admission Control," *Computer Networks Journal:Special Issue on Quality of Service in IP Networks*, vol. 37, no. 5, pp. 507–518, Nov. 2001.

[15] D. Eun, H. S. Kim, and N. B. Shroff, "End-to-End Traffic Analysis in Large Networked Systems," in *Proceedings of Allerton Conference*, Monticello, IL, 2001.

[16] D. Eun and N. B. Shroff, "Simplification of Network Analysis in Large-Bandwidth Systems," in *Proceedings of IEEE INFOCOM*, San Francisco, CA, April 2003.

[17] V. Paxson, "Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic," *ACM SIGCOMM Computer Communication Review*, , no. 5, pp. 5–18, Oct. 1997.

[18] T. Taralp, M. Devetsikiotis, and I. Lambadaris, "Efficient fractional Gaussian noise generation using the spatial renewal process," in *Proceedings of the IEEE International Conference on Communications*, Atlanta, GA, 1998, pp. 1456–1460.

[19] W.-C. Lau, A. Erramilli, J.L. Wang, and W. Willinger, "Self-similar Traffic Generation: The random Midpoint Displacement Algorithm and Its Properties," in *Proceedings of the IEEE International Conference on Communications*, Seattle, WA, 1995, pp. 466–472.

[20] E. W. Knightly, "Second Moment Resource Allocation in Multi-Service Networks," in *Proceedings of ACM SIGMETRICS*, Seattle, WA, 1997, pp. 181–191.

[21] J. Qiu and E. Knightly, "Measurement-Based Admission Control with Aggregate Traffic Envelopes," *IEEE/ACM Transactions on Networking*, vol. 9, no. 2, pp. 199–210, April 2001.

[22] M. Grossglauser and D. Tse, "A Time-Scale Decomposition Approach to Measurement-Based Admission Control," in *Proceedings of IEEE INFOCOM*, New York, NY, 1999.

[23] G. Bianchi, A. Capone, and C. Petrioli, "Throughput Analysis of End-to-End Masurement-based Admission Control in IP," in *Proceedings of IEEE INFOCOM*, Tel Aviv, Israel, 2000.

[24] R. L. Cruz, "A Calculus for Network Delay, Part II : Network Analysis," *IEEE Transactions on Information Theory*, vol. 37, pp. 132–142, Jan. 1991.

[25] J. Kurose, "On Computing Per-Session Performance Bounds in High-Speed Multi-Hop Computer Networks," in *Proceedings of ACM SIGMETRICS*, June 1992, pp. 128–139.

[26] H. Zhang and E. W. Knightly, "Providing End-to-End Statistical Performance Guarantee with Bounding Interval Dependent Stochastic Models," in *Proceedings of ACM SIGMETRICS*, 1994, pp. 211–220.

[27] D. Eun and N. B. Shroff, "A Measurement-Analytic Approach for QoS Estimation in a Network based on the Dominant Time Scale," *IEEE/ACM Transactions on Networking*, March 2003, to appear.

[28] W. Feller, *An Introduction to Probability Theory and its Applications I*, John Wiley & Son, New York, 1968.

Appendix

Proof of Proposition 2.1: We will prove by mathematical induction.

When $t = D$, $q_1(0) + q_2(D) = 0 = q_p(D)$. Suppose $q_1(t-D-1) + q_2(t-1) = q_p(t-1)$ for $t \geq D+1$.

$$\begin{aligned}
q_1(t-D) + q_2(t) &= (q_1(t-D-1) + r_1(t-D) - c_1(t-D))^+ + (q_2(t-1) + r_2(t) - c_2(t))^+ \\
&= (q_1(t-D-1) + r_1(t-D) - d_1(t-D)) + (q_2(t-1) + d_1(t-D) - d_2(t)) \\
&= q_1(t-D-1) + q_2(t-1) + r_1(t-D) - d_2(t) \\
&= q_p(t-1) + r_1(t-D) - d_2(t) \qquad (9) \\
q_p(t) &= (q_p(t-1) + r_1(t-D) - c_p(t))^+ \\
&= (q_p(t-1) + r_1(t-D) - d_2(t) - \min\{a_1(t-D), a_2(t)\})^+ \qquad (10)
\end{aligned}$$

We will show that (10) is equal to (9).

Case 1) $a_1(t-D) = 0$ or $a_2(t) = 0$:

$$\begin{aligned}
q_p(t) &= (q_p(t-1) + r_1(t-D) - d_2(t) - \min\{a_1(t-D), a_2(t)\})^+ \\
&= (q_p(t-1) + r_1(t-D) - d_2(t))^+ \\
&= q_p(t-1) + r_1(t-D) - d_2(t) \\
&= q_1(t-D) + q_2(t) \qquad (\Leftarrow \text{from } (9))
\end{aligned}$$

Case 2) $a_1(t-D) > 0$ and $a_2(t) > 0$: Note that $q_1(t-D) = q_2(t) = 0$ in this case.

$$\begin{aligned} d_1(t-D) &= q_1(t-D-1) + r_1(t-D), \\ d_2(t) &= q_2(t-1) + r_2(t) \;=\; q_2(t-1) + d_1(t-D) \\ &= q_2(t-1) + q_1(t-D-1) + r_1(t-D) \\ &= q_p(t-1) + r_1(t-D). \end{aligned}$$

Thus,

$$\begin{aligned} q_p(t) &= (q_p(t-1) + r_1(t-D) - d_2(t) - \min\{a_1(t-D), a_2(t)\})^+ \\ &= (d_2(t) - d_2(t) - \min\{a_1(t-D), a_2(t)\})^+ \\ &= (-\min\{a_1(t-D), a_2(t)\})^+ \\ &= 0 \;=\; q_1(t-D) + q_2(t). \end{aligned}$$

So we have (10) $\equiv$ (9), from which it follows that $d_2(t) = d_p(t)$.

$$\begin{aligned} d_2(t) &= r_2(t) + q_2(t-1) - q_2(t) \\ &= d_1(t-D) + q_2(t-1) - q_2(t) \\ &= r_1(t-D) + q_1(t-D-1) - q_1(t-D) + q_2(t-1) - q_2(t) \\ &= r_1(t-D) + [q_1(t-D-1) + q_2(t-1)] - [q_1(t-D) + q_2(t)] \\ &= r_1(t-D) + q_p(t-1) - q_p(t) \\ &= d_p(t). \end{aligned}$$

■

Proof of Proposition 2.2:

Note that $c_p(t) = \min\{c_1(t-D) + q_2(t-1), c_2(t)\}$. Hence, $\mathbb{E}\{c_p(t)\} \le \mathbb{E}\{c_2(t)\}$.

For a given sample path, let I be an interval from the time when $q_2(t)$ becomes positive to the time when $q_2(t)$ becomes zero. Because of stability, there will be infinitely many intervals, and index them as $I_k, k = 1, 2, \cdots$. Let t_k be the last moment of I_k. Note that $q_2(t) > 0$ and $c_p(t) = c_2(t)$ for all $t \in I_k - \{t_k\}$, and $q_2(t_k) = 0$. Then, for all $t \notin \bigcup_k I_k, q_2(t-1) = 0$, and hence, $c_p(t) = \min\{c_1(t-D), c_2(t)\} \le c_1(t-D)$. Now, to show that

$\sum_{t\in I_k} c_p(t) \le \sum_{t\in I_k} c_1(t-D)$ will complete the proof.

$$\begin{aligned} q_2(t_k-1) &= \sum_{t\in I_k-\{t_k\}} (r_2(t)-c_2(t)) = \sum_{t\in I_k-\{t_k\}} (d_1(t-D)-c_1(t)) \\ &\le \sum_{t\in I_k-\{t_k\}} (c_1(t-D)-c_1(t)) = \sum_{t\in I_k-\{t_k\}} (c_1(t-D)-c_p(t)). \end{aligned}$$

Thus,

$$\sum_{t\in I_k-\{t_k\}} c_1(t-D) \ge \sum_{t\in I_k-\{t_k\}} c_p(t) + q_2(t_k-1).$$

$$\begin{aligned} \sum_{t\in I_k} c_p(t) &= \sum_{t\in I_k-\{t_k\}} c_p(t) + c_p(t_k) \\ &= \sum_{t\in I_k-\{t_k\}} c_p(t) + r_2(t_k) + q_2(t_k-1) + \min\{a_1(t-D), a_2(t)\} \\ &\le \sum_{t\in I_k-\{t_k\}} c_1(t-D) + r_2(t_k) + \min\{a_1(t-D), a_2(t)\} \\ &\le \sum_{t\in I_k-\{t_k\}} c_1(t-D) + c_1(t_k-D) \;=\; \sum_{t\in I_k} c_1(t-D). \end{aligned}$$

■

Proof of Proposition 3.1:

Since we are interested in the asymptotics, assume that x is integer for simplicity. Let σ^2 be the variance of $c_p(t)$, $F_Q()$ be the distribution function of Q_p, and $F_Z()$ be the distribution function of $Z := \frac{\sum_{t=1}^{x} c_p(t) - \bar{c}_p x}{\sigma\sqrt{x}}$.

Note that $\{W_p > x\} = \{\sum_{t=1}^{x} c_p(t) < Q_p(0)\}$. Since $\{c_p(t)\}_{t\ge 1}$ are independent of $Q_p(0)$,

$$\begin{aligned} \mathbb{P}\{W_p > x\} &= \int_{\{Q_p(0)>0\}} \mathbb{P}\left\{ \sum_{t=1}^{x} c_p(t) < Q_p(0) \,\middle|\, Q_p(0) \right\} dF_Q \\ &= \int_0^\infty \mathbb{P}\left\{ \sum_{t=1}^{x} c_p(t) < q \right\} dF_Q(q) \\ &= \int_0^\infty \mathbb{P}\left\{ \frac{\sum_{t=1}^{x} c_p(t) - \bar{c}x}{\sigma\sqrt{x}} < \frac{q-\bar{c}x}{\sigma\sqrt{x}} \right\} dF_Q(q) \\ &= \int_0^\infty F_Z\left(\frac{q-\bar{c}x}{\sigma\sqrt{x}} \right) dF_Q(q). \end{aligned}$$

First we prove the left inequality: $\mathbb{P}\{Q_p > (\bar{c}_p+\delta)x\} \le \mathbb{P}\{W_p > x\}$.

$$\begin{aligned} \int_0^\infty F_Z\left(\frac{q-\bar{c}x}{\sigma\sqrt{x}} \right) dF_Q(q) &\ge \int_{(\bar{c}+\delta)x}^\infty F_Z\left(\frac{q-\bar{c}x}{\sigma\sqrt{x}} \right) dF_Q(q) \\ &\ge \int_{(\bar{c}+\delta)x}^\infty F_Z\left(\frac{\delta x}{\sigma\sqrt{x}} \right) dF_Q(q) \\ &= F_Z\left(\frac{\delta}{\sigma}\sqrt{x} \right) \mathbb{P}\{Q > (\bar{c}+\delta)x\}. \end{aligned}$$

Since Z converges (in distribution) to a standard Gaussian random variable as x goes to ∞, $F_Z(\frac{\delta}{\sigma}\sqrt{x})$ can be arbitrarily close to 1 for sufficiently large x, say, larger than $1-\epsilon$. Thus, $F_Z(\frac{\delta}{\sigma}\sqrt{x})\mathbb{P}\{Q > (\bar{c}+\delta)x\} \geq (1-\epsilon)\mathbb{P}\{Q > (\bar{c}+\delta)x\}$ for all sufficiently large x, and we have the left inequality.

We next prove the right inequality: $\mathbb{P}\{W_p > x\} \leq \mathbb{P}\{Q_p > (\bar{c}_p - \delta)x\}$.

$$\begin{aligned}
\int_0^\infty F_Z\left(\frac{q-\bar{c}x}{\sigma\sqrt{x}}\right) dF_Q(q) &= \int_0^{(\bar{c}-\delta)x} F_Z\left(\frac{q-\bar{c}x}{\sigma\sqrt{x}}\right) dF_Q(q) + \int_{(\bar{c}-\delta)x}^\infty F_Z\left(\frac{q-\bar{c}x}{\sigma\sqrt{x}}\right) dF_Q(q) \\
&\leq \int_0^{(\bar{c}-\delta)x} F_Z\left(\frac{-\delta x}{\sigma\sqrt{x}}\right) dF_Q(q) + \int_{(\bar{c}-\delta)x}^\infty F_Z\left(\frac{q-\bar{c}x}{\sigma\sqrt{x}}\right) dF_Q(q) \\
&\leq F_Z\left(\frac{-\delta}{\sigma}\sqrt{x}\right) + \int_{(\bar{c}-\delta)x}^\infty F_Z\left(\frac{q-\bar{c}x}{\sigma\sqrt{x}}\right) dF_Q(q) \\
&\leq F_Z\left(\frac{-\delta}{\sigma}\sqrt{x}\right) + \int_{(\bar{c}-\delta)x}^\infty dF_Q(q) \\
&= F_Z\left(\frac{-\delta}{\sigma}\sqrt{x}\right) + \mathbb{P}\{Q > (\bar{c}-\delta)x\}.
\end{aligned}$$

Since Z converges to a standard Gaussian, and since $\int_{-\infty}^{-x} e^{-y^2/2}dy \sim \frac{1}{x}e^{-x^2/2}$ for large x [28], $F_Z\left(\frac{-\delta}{\sigma}\sqrt{x}\right)$ can be as small as $\frac{K_1}{\sqrt{x}}e^{-K_2 x}$ for some $K_1 > 0$ and $K_2 > 0$. Thus, for large x

$$F_Z\left(\frac{-\delta}{\sigma}\sqrt{x}\right) + \mathbb{P}\{Q > (\bar{c}-\delta)x\} \leq \frac{K_1}{\sqrt{x}}e^{-K_2 x} + \mathbb{P}\{Q_p > (\bar{c}-\delta)x\}.$$

Since $\lim_{x\to\infty} \frac{e^{-\alpha x}}{\mathbb{P}\{Q_p > x\}} = 0$ for any $\alpha > 0$ from the assumption, we have for any $\epsilon > 0$ that $\frac{K_1}{\sqrt{x}}e^{-K_2 x} \leq \epsilon\mathbb{P}\{Q_p > (\bar{c}-\delta)x\}$ for all large x. Hence,

$$\frac{K_1}{\sqrt{x}}e^{-K_2 x} + \mathbb{P}\{Q_p > (\bar{c}-\delta)x\} \leq (1+\epsilon)\mathbb{P}\{Q_p > (\bar{c}-\delta)x\}$$

for all sufficiently large x, and we have the right inequality.

■