# An Approximation of the End-to-End Delay Distribution

Han S. Kim and Ness B. Shroff

School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
{hkim,shroff}@ecn.purdue.edu
http://yara.ecn.purdue.edu/~newsgrp

**Abstract.** In this paper we propose an approximation for the end-to-end (queueing) delay distribution based on endpoint measurements. We develop a notion of the end-to-end capacity which is defined for a path of interest. 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 the departure. Our study is motivated by the case where the end-to-end delay distribution can be approximated by an appropriately scaled end-to-end queue-length distribution. We investigate the accuracy of our approximation and demonstrate its application to admission control providing end-to-end QoS.

## 1 Introduction

There are a plethora of papers that have analyzed the queue-length distribution and loss probability for a single node [1] [2] [3] [4] [5] [6]. However, if we simply apply the single-node analysis to each node on the end-to-end path for the end-to-end QoS guarantee, it could result in an inefficient utilization of network resources, and also cause a scalability problem. For example, suppose that the QoS requirement per each flow is to maintain the end-to-end delay violation probability with threshold $D$ less than $\epsilon$, and that we have a tool for estimating the 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 with threshold $D' = D/n$ (where $n$ is the number of nodes on the path) at each node on the path and maintain it less than $\epsilon$. This may result in an unnecessarily small end-to-end delay violation probability and, hence an unnecessary waste of network resources. Moreover, it will also cause a scalability problem because per-flow delay violation probability would need to be managed even at the core nodes inside the network that serve a very large number of flows.

There has been an attempt to reduce the unnecessary waste of network resources by optimally setting the QoS level at each node depending on traffic models and the type of the QoS metric [7]. It has been investigated via a simulation study that the convolution of delay distributions of all the nodes on the

K. Jeffay, I. Stoica, and K. Wehrle (Eds.): IWQoS 2003, LNCS 2707, pp. 59–75, 2003.

path is quite close to the actual end-to-end delay distribution [8]. But such approaches still have the scalability problem. Moreover, in the latter approach, to estimate the end-to-end delay violation probability at one point of threshold $D$, the delay distribution of each node need to be calculated for the entire range of the convolution.

In this paper we propose an approximation of the end-to-end (queueing) 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 the path for the given traffic and connections. Then, a single-node model with the end-to-end capacity is equivalent to the original path in the sense that they have the same end-to-end queue length, and hence, they also have the same departure and end-to-end delay. On the other hand, it is well known that when the capacity is constant, say $c$, the delay violation probability, $\mathbb{P}\{W > x\}$, is equal to the tail probability scaled by $c$, $\mathbb{P}\{Q > cx\}$, for integer $x$ in a discrete-time FIFO queue, i.e., $\mathbb{P}\{W > x\} = \mathbb{P}\{Q > cx\}$. 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 $x$. Precise definitions of these quantities will be provided later. 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.

The main contributions of this paper are:
(*i*) We propose an approximation for 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 the sense that they are identical in terms of the end-to-end queue length.
(*iii*) We apply the above results to admission control.

This paper is organized as follows. In Section 2 we define the end-to-end capacity and the end-to-end queue-length and motivate our study. In Section 3, we provide estimation methods for the end-to-end capacity and the end-to-end queue-length distribution, and then, based on these estimates, we approximately compute the end-to-end delay distribution. We provide simulation results showing how the approximation works and demonstrate its applicability to admission control in Section 4. In Section 5 we validate our approach by showing certain properties of the end-to-end capacity. In Section 6 we discuss our approach comparing with related works. We conclude in Section 7. All proofs are provided in the Appendix.

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

## 2 System Model

We consider a discrete time system. Define a path as a set of links and nodes connecting the source to the destination.

### 2.1 Definitions

- $N_p :=$ 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 number of packets) of flow $i$ entering node $l$ at time $t$
- $d_i^l(t) :=$ rate 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 number of packets, belonging $B_p$, at each node on path $p$)
- $w_p(t) := \min\{s : \sum_{k=1}^{t} \sum_{i \in B_p} r_i^{f_p}(k - D_{f_p}) - \sum_{k=1}^{t+s} \sum_{i \in B_p} d_i^{l_p}(k) \leq 0\} =$ end-to-end (queueing) delay[2] of path $p$

### 2.2 Motivation

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. 1 shows an example. The path consists of two nodes as in Fig. 5, $r_1(t) = 100$ on-off sources[3], $c_1(t) =$ 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 the perfect knowledge of $c_p(t)$. In practice, however, $c_p(t)$ is not known without information from all nodes on the path, and should be estimated by endpoint measurements. From Fig. 1, 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,

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

[3] The same traffic parameters in [6] are used.

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 handling 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 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.

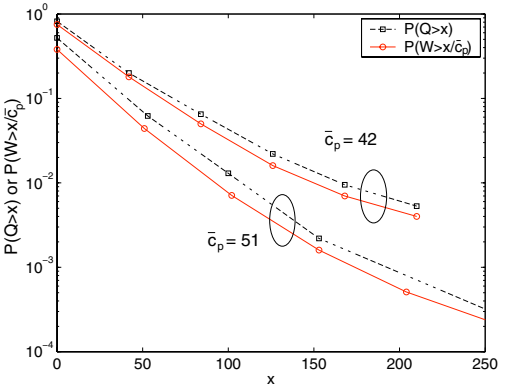


**Fig. 1.** Comparison of $\mathbb{P}\{Q_p > x\}$ and $\mathbb{P}\{W_p > x/\bar{c}_p\}$.

## 3 Estimation of the End-to-End Delay Distribution

In this section, we propose an approximation of the end-to-end delay distribution. As illustrated in Fig. 1, 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.

### 3.1 MVA Approximation for Delay

We first estimate the tail of the end-to-end queue-length distribution. 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 Asymptotics (MVA)

approach (first named in [4]) provides an accurate estimate of the tail probability. Although the net input $r_1(t-D) - c_p(t)$ may not be modeled as Gaussian, which is assumed in the MVA approach, it have been investigated that the MVA method also works well for non-Gaussian cases including a case where a small number of flows are multiplexed [4] [6]. Hence, we estimate the tail probability by

$$\mathbb{P}\{Q_p > x\} \approx e^{-m_x/2} \tag{1}$$

where

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

and $X(1,t) = \sum_{k=1}^{t} [r_1(k-D) - c_p(k)]$. An important question is how to obtain $\bar{c}_p$ and $\mathrm{Var}\{X(1,t)\}$. They are estimated from the measurement during the *busy period*. This will be explained in the following subsection.

Then, as Fig. 1 suggests, we approximate the end-to-end delay by

$$\mathbb{P}\{W_p > x\} \approx e^{-m_{\bar{c}_p x}/2}, \tag{3}$$

and we call this *MVA approximation for delay*.

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

The MVA method requires the first two moments of $X(1,t)$. We assume that $c_p(t)$ and $r_1(t)$ are independent, and hence, we estimate their moments separately and add their variance to get $\mathrm{Var}\{X(1,t)\}$.[4] We assume that the ingress node inserts timestamps to record the arrival time of packets [10] [11]. Then, there will be no problem in measuring the moments of $r_1(t)$.

We now explain how to measure the moments of $c_p(t)$. If a minimally backlogging probe packet which makes $\min_{l \in N_p} a_l(t - D_p^l) = 0$ were inserted to the path, the departure would be exactly $c_p(t)$. But determination of such input itself would be problematic because it also requires the knowledge of all flows at each node. Inspired by the idea in [10],[5] we define the busy period as an interval in which the end-to-end queue length $q_p(t)$ is greater than 0. The determination of the busy period can be done by comparing the accumulated input and departure, assuming that $D_p^{f_p}$ is known since it is fixed for a given path. Since the departure is equal to $c_p(t)$ during the busy period, the moments of $c_p(t)$ can be estimated by measuring the departure during the busy period.

One may wonder how well the moments measured during the busy period represent the actual moments. We have found that the moments are a little underestimated compared to their actual values but the errors tend to compensate

[4] Though $c_p(t)$ and $r_1(t)$ are somewhat correlated, $\mathrm{Var}\{X(1,t)\}$ is quite close to the sum of $\mathrm{Var}\{\sum c_p(t)\}$ and $\mathrm{Var}\{\sum r_1(t)\}$. We have also found empirically that as $r_1(t)$ becomes smaller compared to the capacity, their statistics become independent.

[5] In [10], the service envelope is measured during a backlogging interval in which there is at least one more packet arrival between the departure time and the arrival time of each packet.

each other because the first moment is the numerator and the second moment is the denominator in (1). We have also found that the estimation during busy period becomes more accurate as $c_i(t)$ has relatively smaller variance (compared to the mean) and less correlation, which is expected as the system size increases.

## 4 Numerical Experiments

In this section we investigate how the MVA approximation for delay performs and how it can be applied for admission control.

### 4.1 Approximation of the End-to-End Delay

In this experiment, we use a five node network with 4 paths (Fig. 2). Each path is carrying voice and video traffic: 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: $C_1 = C_2 = C_4 =$

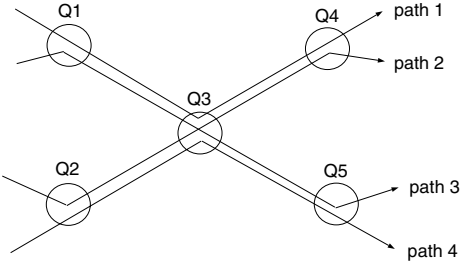


**Fig. 2.** Five node network

$C_5 = 210$pkt/slot, $C_3 = 420$pkt/slot, which are chosen for a 45Mbps link with 2ms time slot and 53byte packet.[6] We use the on-off model as a voice source[7] and the real MPEG trace (007 movie) as a video source. The propagation delay is set to 0. From Fig. 3 we can see that the approximation bounds the actual delay within an order of magnitude. 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 quite close to the maximum utilization for given QoS.

### 4.2 Application to Admission Control

Many admission control algorithms are based on single-node analysis [12] [13] [14] and the admission decision is made by estimating the QoS at a node, for

[6] 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 10ms slot size is almost same.

[7] The same traffic parameters in [6] are used.

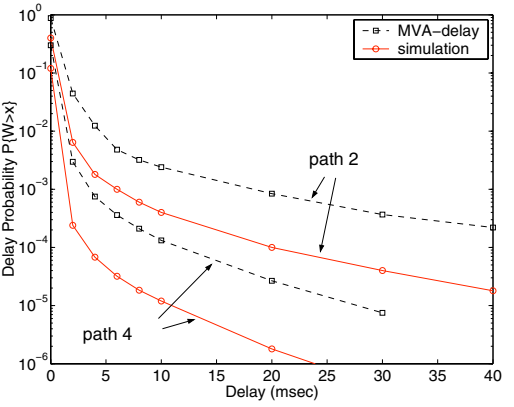


**Fig. 3.** Approximation of the end-to-end delay for path 2 and 4.

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 need to be performed at all nodes on the path including core routers that serve a large number of flows, thus causing a 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 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 Gaussian approximation for the estimation of the overflow probability. Let $\mu$ and $\sigma^2$ 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 $\epsilon$ be the target QoS. Then, a new flow is admitted if

$$Q\left(\frac{\hat{c}_p - \hat{\mu}_r - \mu}{\sqrt{\hat{\sigma}_c^2 + \hat{\sigma}_r^2 + \sigma^2}}\right) < \epsilon \tag{4}$$

where $Q()$ is the complementary cdf of a standard Gaussian random variable, $N(0,1)$. In this experiment, we fix path 2,3, and 4 with 1400 voice flows, and do the admission control for path 1 in the same five node network in Fig. 2. Table 1 compares the number of flows admitted by the proposed algorithm with the maximum obtained by simulation. From Table 1, we can see that the number of admitted flows by our algorithm is conservative but close to the maximum and the target QoS is met.

**Table 1.** 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 $\mu$ and $v(t)$ be the mean and the variance function[8] 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 existing aggregate flow on the path, and $\epsilon$ be the target QoS. Then, a new flow is admitted if

$$\sup_{t \geq 1} \frac{\hat{v}_c(t) + \hat{v}_r(t) + v(t)}{[x + (\hat{c}_p - \hat{\mu}_r - \mu)t]^2} > -2 \log \epsilon. \tag{5}$$

In this experiment, we fix path 2,3, and 4 with 35 video flows, and do the admission control for path 1 in the same five node network in Fig. 2. We set $D$ to 20, i.e., $40ms$. From the result in Table 2, 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 2.** Admission control by the end-to-end delay violation probability

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

## 5 Properties of the End-to-End Capacity

We now provide some theoretical properties on the end-to-end capacity that further support our methodology. More specifically, we show that the single-node model with the end-to-end capacity has the identical queue length with the original path, and that the end-to-end capacity plays a role like a lower bound to the capacities of the nodes on the path in some sense.

### 5.1 Equivalent Single-Queue Representation

For simplicity, start with a path of two nodes as shown in Fig. 4. $r'_1(t)$ and $r'_2(t)$ are cross traffic, and $d'_1(t)$ and $d'_2(t)$ are the corresponding departures. Assume

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

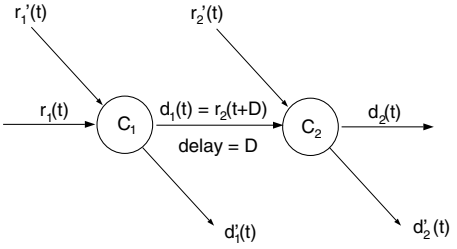


**Fig. 4.** Original path

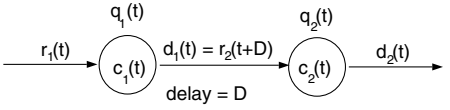


**Fig. 5.** Two-node model

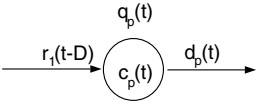


**Fig. 6.** Single-node model

that the departure of 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. 5). We will compare this(Fig. 5) with the single-node model(Fig. 6) where $c_p(t) = d_2(t) + \min\{a_1(t - D), a_2(t)\}$.

**Proposition 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* $d_2(t) = d_p(t)$, $\forall t \geq D$.

Proposition 1 tells us that once both queues become empty, which is ensured as long as the system is stable, the single-node model is identical with the original path thereafter.

We have considered so far a path with two nodes. It can be easily extended to a path of more than two nodes. 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. We can also extend this to multi-class cases simply by treating other classes as cross traffic.

### 5.2 Properties of $c_p(t)$

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 instants. However, we have the following result.

**Proposition 2.** *Assume that all flows are stationary and ergodic, and that* $\mathbb{E}\{r_i(t) + r'_i(t)\} < C_i$ *for stability. Then,*

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

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 $\mathbb{E}\{r_1(t-D) - c_p(t)\} \geq \mathbb{E}\{r_i(t) + r'_i(t) - C_i\}$, and we expect that the following conjecture is true.

**Conjecture A:**

$$\mathbb{P}\{r_1(t-D) > c_p(t)\} \geq \max_{i=1,2} \mathbb{P}\{r_i(t) + r'_i(t) > C_i\}. \tag{6}$$

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 $\epsilon$, the overflow probability at each node on the path is also less that $\epsilon$ (it has been demonstrated in numerical experiments).

### 5.3 The Relationship between $\mathbb{P}\{W_p > x\}$ and $\mathbb{P}\{Q_p > \bar{c}_p x\}$

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 $\mathbb{P}\{Q_p > \bar{c}_p x\}$. Though we believe that they are asymptotically similar (which is obvious when $c_p$ is constant), we have shown so far for a particular case only. For a heavy-tailed case where $\mathbb{P}\{Q_p > x\}$ is decaying slower than 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.

**Proposition 3.** *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*

$$\mathbb{P}\{Q_p > (\bar{c}_p + \delta)x\} \leq \mathbb{P}\{W_p > x\} \leq \mathbb{P}\{Q_p > (\bar{c}_p - \delta)x\}, \qquad \forall x \geq x_0. \tag{7}$$

## 6 Discussion

### 6.1 Related Work

Our approach is motivated by empirical observations. Nevertheless, it is 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 [15], or a statistical per-node delay bound [16] [17]. In [15], the maximum (or worst-case) end-to-end delay is calculated for regulated traffic. In [16] and [17], an upper bound to 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 [10] 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 $\lim_{t\to\infty} \bar{R}(t) + r(t) \leq \lim_{t\to\infty} \frac{\bar{S}(t)}{t}$, where 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 [10] that $\mathbb{P}\{W_p > x\} \leq 1 - \min_{0\leq t\leq T} \Phi(\alpha_x)$ where $\alpha_x$ 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 [12], 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 $\infty$ so that the delay bound provided by the test (8) will be the worst-case delay.

Based on (2), it seems that $\mathrm{Var}\{X(1,t)\}$ needs to be evaluated for the entire range of $t$ due to the *min* operation over $\{t \geq 1\}$. Fortunately, 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 [19]. This makes the MVA approach amenable for on-line measurements.

It has been shown that core nodes serving large flows with large capacity compared to edge nodes can be ignored for the end-to-end analysis [18] [20]. Thanks to this, it is possible to improve the accuracy in measuring moments of $c_p(t)$ at the cost of additional functionality in the ingress node. When the ingress node $f_p$ inserts a timestamp for each arriving packet to record its arrival time, it could insert one more to denote the amount of unused capacity at the previous time slot. Then, for a departing packet with timestamp $t$, 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, $a_{f_p}(t-1)$, and it does not have to check the busy period. This comparison only needs to be done at least for one packet per path per time slot. If no packet of path $p$ with time stamp $t$ is found at the egress node, simply set $c_p(t + D_p^{f_p} - 1) = a_{l_p}(t + D_p^{f_p} - 1)$. Since the unused capacity $a_l(t)$ is the same for all paths on node $l$, the implementation complexity is not that high.

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

One may ask why not directly measure the delay (or queue-length) distribution when the estimation is based on *measurements* after all. The problem in directly measuring the delay distribution is that it may require too long time to measure a small value. For example, measuring a value of $10^{-6}$ 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 in shorter duration with more reliability. Table 3 shows the result of an experiment comparing the length of time to measure(or estimate) the delay probability within the 90% confidence interval less than an order of magnitude. In this experiment, 200 on-off sources 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. When the target value becomes 10 times smaller (changed from $10^{-5}$ to $10^{-6}$), direct measuring requires 50 times longer duration while moments measuring for the MVA method requires only 4 times longer duration.

**Table 3.** Comparison of the required simulation cycles

| target value | direct measuring | moments measuring |
|---|---|---|
| $10^{-5}$ | $2 \times 10^6$ cycles | $2 \times 10^4$ cycles |
| $10^{-6}$ | $1 \times 10^8$ cycles | $8 \times 10^4$ cycles |

## 7 Conclusion

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 and the same end-to-end delay. Further, we have empirically found that $\mathbb{P}\{W_p > x\} \approx \mathbb{P}\{Q_p > \bar{c}_p x\}$. We have also shown for a particular case that 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 (*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 endpoint measurements only, the scheme is scalable. We have also validated our estimation by numerical experiments. Unlike existing works on the delay that have focused on the maximum (or worst-case) end-to-end delay [15] or a bound on the per-session delay distribution at a single node

[16] [17], our approach is the first attempts to estimate the end-to-end delay distribution itself.

## References


1. Addie, R. G. and Zukerman, M.: An Approximation for Performance Evaluation of Stationary Single Server Queues. *IEEE Transactions on Communications* **42,** (1994) 3150–3160.
2. Duffield, N. G. and O'Connell, N.: Large Deviations and Overflow Probabilities for the General Single Server Queue, with Application. *Proc. Cambridge Philos. Soc.* **118,** (1995) 363–374.
3. Glynn, P. W. and Whitt, W.: Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single-Server Queue. *Journal of Applied Probability* (1994) 131–155.
4. Choe, J. and Shroff, N. B.: A Central Limit Theorem Based Approach for Analyzing Queue Behavior in High-Speed Networks. *IEEE/ACM Transactions on Networking* **6,** (1998) 659–671.
5. Likhanov, N. and Mazumdar, R. R.: Cell-Loss Asymptotics in Buffers fed with a Large Number of Independent stationary sources. In *Proceedings of IEEE INFOCOM.* San Francisco, CA (1998).
6. Kim, H. S. and Shroff, N. B.: Loss Probability Calculations and Asymptotic Analysis for Finite Buffer Multiplexers. *IEEE/ACM Transactions on Networking* **9,** (2001) 755–768.
7. Nagarajan, R., Kurose, J. and Towsley, D.: Local Allocation of End-to-End Quality-of-Service in High-Speed Networks. *IFIP Transactions C-Communication Systems* **15,** (1993) 99–118.
8. Yates, D., Kurose, J. and Towsley, D.: On Per-Session End-to-End Delay and the Call Admission Problem for Real-Time Applications with QOS Requirements. *Journal of Highspeed Networks* **3,** (1994) 429–458.
9. Kim, H. S.: Queueing Analysis of Network Multiplexers: Loss Ratio and End-to-End Delay Distribution. *PhD thesis.* School of Electrical and Computer Engineering Purdue University, West Lafayette, IN (2003).
10. Cetinkaya, C., Kanodia, V. and Knightly, E.: Scalable Services via Egress Admission Control. *IEEE Transactions on Multimedia* **3,** (2001) 69–81.
11. Yuan, P., Schlembach, J., Skoe, A. and Knightly, E.: Design and Implementation of Scalable Admission Control. *Computer Networks Journal:Special Issue on Quality of Service in IP Networks* **37,** (2001) 507–518.
12. Qiu, J. and Knightly, E.: Measurement-Based Admission Control with Aggregate Traffic Envelopes. *IEEE/ACM Transactions on Networking* **9,** (2001) 199–210.
13. Grossglauser, M. and Tse, D.: A Time-Scale Decomposition Approach to Measurement-Based Admission Control. In *Proceedings of IEEE INFOCOM.* New York, NY (1999).
14. Bianchi, G., Capone, A. and Petrioli, C.: Throughput Analysis of End-to-End Masurement-based Admission Control in IP. In *Proceedings of IEEE INFOCOM.* Tel Aviv, Israel (2000).
15. Cruz, R. L.: A Calculus for Network Delay, Part II : Network Analysis. *IEEE Transactions on Information Theory* **37,** (1991) 132–142.
16. Kurose, J.: On Computing Per-Session Performance Bounds in High-Speed Multi-Hop Computer Networks. In *Proceedings of ACM SIGMETRICS.* (1992) 128–139.

17. Zhang, H. and Knightly, E. W.: Providing End-to-End Statistical Performance Guarantee with Bounding Interval Dependent Stochastic Models. In *Proceedings of ACM SIGMETRICS*. (1994) 211–220.
18. Eun, D., Kim, H. S. and Shroff, N. B.: End-to-End Traffic Analysis in Large Networked Systems. In *Proceedings of Allerton Conference*. Monticello, IL (2001).
19. Eun, D. and Shroff, N. B.: A Measurement-Analytic Approach for QoS Estimation in a Network based on the Dominant Time Scale. *IEEE/ACM Transactions on Networking* (2003) to appear.
20. Eun, D. and Shroff, N. B.: Simplification of Network Analysis in Large-Bandwidth Systems. In *Proceedings of IEEE INFOCOM*. San Francisco, CA (2003).
21. Feller, W.: An Introduction to Probability Theory and its Applications I. John Wiley & Son, New York (1968).

# Appendix

**Proof of Proposition 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 $t \geq D+1$.

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

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) ≡ (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:**

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:**

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

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

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

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

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

Since $Z$ converges (in distribution) to a standard Gaussian random variable as $x$ goes to $\infty$, $F_Z(\frac{\delta}{\sigma}\sqrt{x})$ 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: $\mathbb{P}\{W_p > x\} \leq \mathbb{P}\{Q_p > (\bar{c}_p - \delta)x\}$.

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

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

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 $x$ [21], $F_Z\left(\frac{-\delta}{\sigma}\sqrt{x}\right)$ 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.

■