Performance Analysis of TCP/AQM with Generalized AIMD under Intermediate Buffer Sizes [2]

⋆

Do Young Eun * and Xinbing Wang [b] [1]

[a] *Department of Electrical and Computer Engineering*
North Carolina State University, Raleigh, NC 27695
Email: dyeun@eos.ncsu.edu

[b] *Department of Electronic Engineering*
Shanghai Jiaotong University, Shanghai, China
Email: xwang8@sjtu.edu.cn

Abstract

For TCP/AQM systems, the issue of buffer sizing has recently received much attention. The classical rule-of-thumb suggests $O(N)$ buffer size to ensure full link utilization when N TCP flows share a bottleneck link of capacity $O(N)$, while recent empirical study shows the buffer of size $O(\sqrt{N})$ is enough to yield high utilization (say, 95%) for large N. However, these results are all limited to the drop-tail scheme and there has been no systematic modeling framework for any buffer sizing between $O(\sqrt{N})$ and $O(N)$. In this paper, we study the limiting behavior of a TCP/AQM system for an intermediate buffer sizing of $O(N^\gamma)$ $(0.5 \leq \gamma < 1)$. We develop a stochastic model in a discrete-time setting to characterize the system dynamics and then show that we can have 100% link utilization and zero packet loss probability for a large number of flows when the buffer size is chosen anywhere between $O(\sqrt{N})$ and $O(N)$. Our model is general enough to cover any queue-based AQM scheme with ECN marking (including the drop-tail) and various generalized AIMD (Additive-Increase-Multiplicative-Decrease) algorithms for each TCP flow. We also provide arguments showing that the discrete-time based modeling can effectively capture all the essential system dynamics under our choice of scaling $(0.5 \leq \gamma < 1)$ for buffer size as well as AQM parameters.

Key words: Congestion Control; Buffer Sizing; Active Queue Management; Discrete-time Stochastic Models; Performance Analysis

⋆ This work was supported in part by NSF CAREER Award CNS-0545893.

[1] The research was conducted when this author was at North Carolina State University.

1 Introduction

TCP is the dominant transport protocol to carry most of Internet data traffic, e.g., email, FTP, P2P. Each TCP source just iterates its own congestion control algorithm to respond to the congestion signal generated by AQM (Active Queue Management) schemes in the network. So, TCP/AQM is essentially a distributed, close-loop congestion control algorithm. On one hand, the distributed nature of TCP makes it easy to implement, because there is no global information needed. On the other hand, the close-loop nature of the system makes it difficult to analyze the overall performance of TCP/AQM under various configuration, compared to the traditional open-loop teletraffic theory.

As an example, the buffer sizing for the Internet router shared by TCP flows has been such a delicate issue. Although extensive research efforts have been made so far [2–5], the problem is still far from being completely understood and solved, mainly because the underlying mechanisms of the close-loop system are quite different from that of usual open-loop queueing theory. Traditionally, the rule-of-thumb for buffer sizing is set to the bandwidth-delay product [2]; for a system with N flows, capacity NC for some constant C, and the round-trip-time (RTT) T, the buffer size should be set to NCT, i.e., $O(N)$.

Recently, it has been shown in [3] that $O(\sqrt{N})$ buffers sizing for drop-tail is enough for high link utilization. While the results in [6] still suggest that the buffer size should be enforced as $O(N)$ under drop-tail for full (100%) link utilization, which is in accordance with the traditional rule-of-thumb in [2]. Although it was empirically shown that the link utilization can be very high, it can be easily seen that the loss probability for drop-tail cannot be arbitrarily close to zero. For instance, it is well-known that $Throughput \approx \frac{1}{RTT}\sqrt{\frac{K}{p}}$ [7,8]. Here, p is the probability that the flow receives 'congestion signal' in the form of either packet-loss or marking. Since throughput is surely bounded above by the nominal link capacity C, we should have $p \geq p'$ for some constant p' unless RTT increases indefinitely. In addition, with the buffer size of $O(N)$, the AQM schemes with ECN marking (say RED, REM [9], PI [10]) in general yield better performance than the simple drop-tail scheme (e.g., zero packet-loss probability for AQMs). Then, it is natural to ask the following three interesting and inter-related questions: First, if we use AQM schemes with marking, are we able to configure the system such that all the necessary congestion signals come from packet marks (and hence packet-loss is kept near

* Corresponding Author: Do Young Eun
Box 7911-EB2, Dept. of Electrical and Computer Engineering
North Carolina State University, Raleigh, NC 27695-7911
Email: dyeun@eos.ncsu.edu
[2] A preliminary version of this paper appears in the proceedings of IEEE IPCCC 2006 [1].

zero) when the buffer size is chosen in between $O(\sqrt{N})$ and $O(N)$? Second, can we achieve 100% link utilization even when the buffer size (and AQM parameters) are chosen far smaller than $O(N)$? Third, under the same choice of buffer size as in the above questions, what is the performance for other types of TCP protocol, i.e., under generalized AIMD or MIMD (Multiplicative-Increase-Multiplicative-Decrease), etc?

To address these questions, in this paper, we study the limiting system behavior as the system size increases under an intermediate buffer size, i.e., the buffer size is chosen as $O(N^\gamma)$ where $0.5 \leq \gamma < 1$. In order to analyze the system, we first propose a stochastic discrete-time model to accurately capture the system dynamics when the buffer size as well as the AQM parameters are chosen according to $O(N^\gamma)$ $(0.5 \leq \gamma < 1)$. Although there already exist some results regarding the stochastic discrete-time model for TCP/RED [11,12], the framework in [11,12] is valid only for the linear scale $O(N)$ for buffer sizing and the RED parameters, while our analysis can be applied to any other scales in general. Based on our model in a discrete-time setting, we show that we can achieve 100% link utilization and zero packet loss (when AQM with packet marking is employed) under the aforementioned scale for the buffer sizing when there are large number of flows. All these analytical results are also supported by simulation results via *ns*-2. Further, we show that the discrete-time modeling (by capturing system 'states' at each RTT instant) becomes accurate and faithful representation of the system under the scale of interest $(O(N^\gamma))$ with $0.5 \leq \gamma < 1$.

The rest of this paper is organized as follows. In Section 2, we describe the scaling used in the paper and present our stochastic discrete-time model for TCP/AQM systems. In Section 3 we analyze the system performance in terms of link utilization and queue-length distribution. We also illustrate that packet loss can be eliminated using AQMs with marking even when the buffer size is chosen far smaller than the traditional rule-of-thumb $(O(N))$. In Section 4, we provide several arguments supporting the use of a discrete-time modeling under our scale (as opposed to a continuous-time modeling). We present simulation results in Section 5 and conclude the paper in Section 6.

2 Model Description

2.1 Scaling the Link

We consider a link with capacity NC shared by N simultaneous connections or flows, where each flow adapts its rate (or window size in TCP) based on whether it receives any packet mark. The rationale behind this scaling is that

as the number of connections increases, the capacity also increases proportionally in order to give each flow approximately the same bandwidth share. Let $Q^N(t)$ be the queue length at time t and $B(N)$ denote the buffer size at that link (or router). See Figure 1 for illustration.

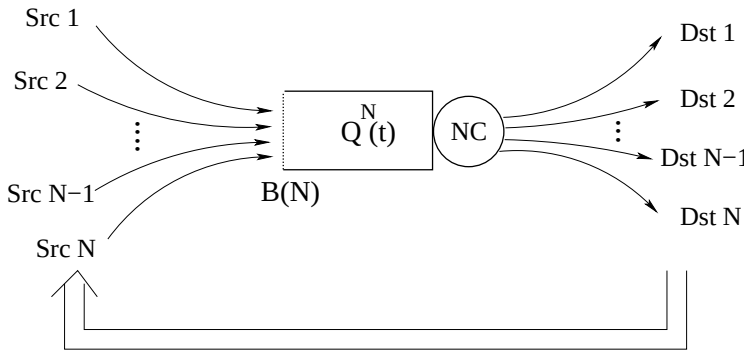


Fig. 1. Simplified link model

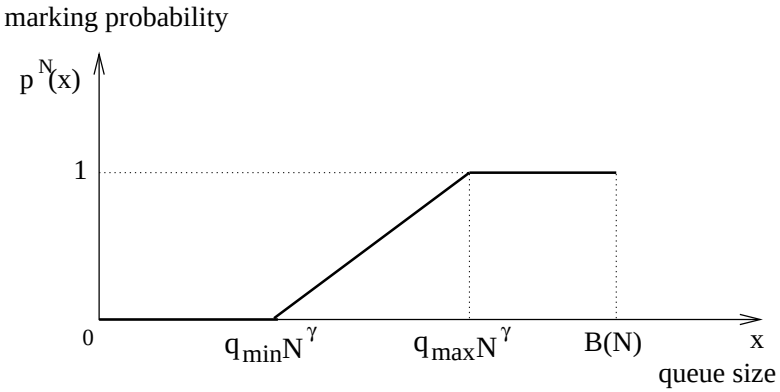


Fig. 2. Example of a marking function (RED type) with $1/2 \leq \gamma < 1$

For AQM schemes, we consider routers using queue-based schemes equipped with ECN capability. One consequence of using ECN is that routers *mark* the packets to notify incipient congestion (based on the current queue-length) and *drop* only if the buffer is physically full and there is no room to accept the incoming packets. When a packet arrives to the queue at time t, this packet will be marked, independently of any other, with probability $p^N(Q^N(t))$, where $Q^N(t)$ denotes the queue size at time t. Hence, for the marking function depicted in Figure 2, if the queue size $Q^N(t)$ is less than $q_{min}N^\gamma$, no packet is marked. If the queue size becomes larger than $q_{max}N^\gamma$, then all the incoming packets are marked (not dropped!). The packet-drop happens only when the queue size goes above the actual buffer size $B(N)$.

Our marking function $p^N(x)$ will satisfy the following:

There exists a continuous, non-decreasing function $p : \mathbb{R}^+ \to [0, 1]$ such that, for all N and x,

$$p^N(N^\gamma x) = p(x), \tag{1}$$

where $1/2 \leq \gamma < 1$. Further, we have $\lim_{x \to 0} p(x) = 0$ and there exists $q_{max} < \infty$ such that

$$p(x) = 1, \quad \text{for all } x \geq q_{max}. \tag{2}$$

The above clearly shows that we use a scale of γ for the marking function. Note that $\gamma = 1$ corresponds to the usual linear scaling as in [11–13,4], while $\gamma = 1/2$ corresponds to the square root scaling for buffer size used in [3].

2.2 How to React to Marks: Generalized AIMD

Let the time be divided into a sequence of slots, each of which has a length of one round-trip-time (RTT). To make our exposition simple, let the length of one time slot be normalized to 1. We assume that all the N flows sharing the common bottleneck link have the same RTT. Let $W_i^N(k)$ be the window size of flow i at the k^{th} time slot (k^{th} RTT), where the superscript N means that there are N flows in the system sharing a link with capacity NC. Let w_{max} be the maximum window size for all flows, i.e., $1 \leq W_i^N(k) \leq w_{max}$ for all i, k, N. Then, given that the current window size is w, each flow increases its window size at the next RTT by $\alpha(w)$ in case of no marked packet and decreases by βw if there is at least one marked packet among w packets. In other words, each window size $W_i^N(k)$ evolves as follows:

$$W_i^N(k+1) = \begin{cases} \left\lfloor W_i^N(k) + \alpha\left(W_i^N(k)\right) \right\rfloor \wedge w_{max}, & \text{if none of } W_i^N(k) \text{ packets is marked,} \\ \left\lfloor W_i^N(k) - \beta W_i^N(k) \right\rfloor \vee 1 & \text{otherwise,} \end{cases} \tag{3}$$

where $x \wedge y := \min\{x, y\}$ and $x \vee y := \max\{x, y\}$, and $\lfloor x \rfloor$ denotes the largest integer smaller than x. Here, $0 < \beta < 1$ is a given constant and the function $\alpha(w)$ is assumed to satisfy the following:

Assumption 1 *The function $\alpha(\cdot)$ is non-decreasing, concave, and satisfy $0 < \alpha(w) < w$ for all $w \in [1, w_{max}]$.*

We can reproduce most of the current TCP algorithms by choosing suitable $\alpha(\cdot)$ and β, For example, setting $\alpha(w) = \alpha$ (constant) gives the well-known AIMD (Additive-Increase-Multiplicative-Decrease). In particular, $\alpha(w) = 1$ and $\beta = 0.5$ correspond to the current TCP-Reno. Another possibility is to choose $\alpha(w) = aw^b$ with $a > 0$ and $0 < b \leq 1$ [14,15]. Note that this case also includes MIMD (Multiplicative-Increase-Multiplicative-Decrease) type of algorithm ($b = 1$) and any other generalized AIMD algorithms considered in the literature.

2.3 Markovian Representation of the System

Given the background on our link model and how each sender changes its window size, in this section, we develop a Markovian representation of the TCP/AQM system.

As before, let $W_i^N(k)$ be the window size of flow i at the k^{th} time slot and $Q^N(k)$ be the corresponding queue-length at the same time slot. We define the system 'state' at time k (k^{th} time slot) by an $(N+1)$-dimensional vector as

$$\vec{X}^N(k) := \left[W_1^N(k), \cdots, W_N^N(k), Q^N(k)\right]. \tag{4}$$

Suppose that $W_i^N(k)$ $(i = 1, 2, \ldots, N)$ and $Q^N(k)$ are given, i.e., $W_i^N(k) = w_i^N(k)$ and $Q^N(k) = q^N(k)$. Then, since each of $w_i^N(k)$ packets will be marked independently of any other with probability $p^N(q^N(k))$, we see that the probability that none of the $w_i^N(k)$ packets is marked becomes

$$p_i^N(k) = \left[1 - p^N(q^N(k))\right]^{w_i^N(k)}, \quad i = 1, 2, \ldots, N. \tag{5}$$

Thus, *given* the window sizes and the queue-length at time k, window sizes for flow i at time $k+1$ will evolve as in (3) with corresponding probability $p_i^N(k)$, and $W_i^N(k+1)$ and $W_j^N(k+1)$ become independent (for $i \neq j$) since each incoming packet is being marked independently.[3] Further, the queue-length at time $k+1$ is given by the following recursion:

$$Q^N(k+1) = \left[Q^N(k) + \sum_{i=1}^{N} W_i^N(k) - NC\right]^+, \tag{6}$$

where $[x]^+ = \max\{x, 0\}$. Hence, $\vec{X}^N(k)$ becomes an $(N+1)$-dimensional homogeneous Markov chain with its evolution described by (6) and (3) with appropriate transition probabilities as in (5).

Before proceeding to our main section, we show that the sum of the window sizes for N flows is always bounded above. In this paper, we provide all the proofs in the Appendix.

Lemma 1 *Let $C' = C + q_{max}$.*[4] *Then, for any k, we have*

$$\frac{1}{N}\sum_{i=1}^{N} W_i^N(k) \leq M,$$

where

$$M := C' + \alpha(C') + \alpha\left(C' + \alpha(C')\right). \tag{7}$$

[3] They are conditionally independent given $\vec{X}^N(k)$.

[4] Here, C should be interpreted as $C \times RTT$ since we normalize RTT to 1.

Proof of Lemma 1: To make the notations simple, let $\bar{w}(k) := \frac{1}{N}\sum_{i=1}^{N} W_i^N(k)$. First, note that we can always find k' such that $\bar{w}(k') < C'$. To see this, suppose otherwise that $\bar{w}(t) \geq C$ for all t. Then, since the network can serve only up to NC packets per time slot, $Q^N(t) > q_{max}N > q_{max}N^\gamma$ for all t and for any $\gamma \in [0.5, 1)$. In other words, all the incoming packets will be marked forever and thus, $\bar{w}(t)$ will keep decreasing and eventually becomes smaller than C', leading to a contradiction.

Now, for such k', observe that

$$\bar{w}(k'+1) - \bar{w}(k') \leq \frac{1}{N}\sum_{i=1}^{N} \alpha\left(W_i^N(k')\right) \leq \alpha(\bar{w}(k')) \leq \alpha(C'), \tag{8}$$

where the second inequality is from the concavity of the function $\alpha(\cdot)$ and Jensen's inequality, and the third one follows since $\alpha(\cdot)$ is non-decreasing and $\bar{w}(k') < C'$. Thus, we have $\bar{w}(k'+1) \leq C' + \alpha(C')$. Without loss of generality, we can assume that $\bar{w}(k'+1) \geq C'$. (If not, simply repeat the argument in (8) until it becomes larger.) Then, regardless of $Q^N(k'+1)$, we see from (6) that $Q^N(k'+2)$ is always larger than $q_{max}N$ (thus larger than $q_{max}N^\gamma$) and so all the incoming packets will be marked during that time slot. This makes all the window sizes at the next time slot keep decreasing until the sum of the window sizes becomes smaller than NC'. However, in the meanwhile, the sum of the window sizes can still increase with some probabilities, but only by up to $N\alpha(C' + \alpha(C'))$ (by repeating the argument in (8)). This completes the proof. ■

Lemma 1 will be invoked to prove one of our main results in Section 3. In the rest of the paper, to avoid any triviality, we will assume that $w_{max} > M$ and $C > 1$, where M is from (7).

3 Main Results

In this section, we provide asymptotic analysis of the system represented by the Markov chain in Section 2.3. First, we show that for fixed N, the chain is ergodic, and thus always converges to its stationary version. Then, we investigate several performance metrics including the link utilization and the queue-length distribution for the stationary chain, as N becomes large.

3.1 Ergodicity of the Markov Chain

Having constructed the Markov chain as in Section 2.3, we then focus on the ergodicity of the chain. First, it is straightforward to see that the chain in (3) and (6) is irreducible and aperiodic. In order to show the required positive recurrence of the chain, consider the following *finite* subset F of the state space

$$F := \{1, 2, \ldots, w_{max}\}^N \times [0, 1, \ldots, q_{max}N^\gamma].$$

In words, the set F contains all the possible window sizes for N flows and all the queue-length fluctuation up to $q_{max}N^\gamma$. Note that for a fixed N, the set F is always finite. Assume that the system is currently at state $j \in F$ and define $\tau_j(F)$ be the return time to set F starting from j. Suppose the queue-length is larger than $q_{max}N^\gamma$ indefinitely. Then, as before, all packets will be marked and all the flows keep decreasing their window sizes forever until all of them are equal to 1, in which case the queue-length has to be smaller than $q_{max}N^\gamma$. Thus, we see that the return time to F starting from any $j \in F$ must be finite, and thus the chain is positive recurrent in view of Lemma 1.1 in [16] (page 168).

As the chain is now ergodic for any fixed N, it always converges to a steady-state in which the distribution of the chain is stationary. Hence, from now on, we assume that the system is in the steady-state and the distribution of $\vec{X}^N(k)$ is *invariant* with k. In the subsequent section, we derive the steady-state system dynamics in terms of the link utilization and the queue-length distribution as N increases.

3.2 Asymptotic Analysis of the System

Since $\vec{X}^N(k)$ is stationary in the steady-state, $Q^N(k)$ is also stationary in k. Let Q^N denote the queue-length random variable in the steady-state. Similarly, whenever there is no ambiguity, we will suppress the time index k from the window size random variables in the steady-state and denote them as W_i^N. We now present our first result below, which will be used later to show that, in the steady-state, the probability that each incoming packet flow i is marked is uniformly bounded away from 0 and 1.

Proposition 1 *For any $N > 0$ and i ($1 \le i \le N$), we have*

$$\frac{\mathbb{E}\{W_i^N\} - 1}{w_{max} + \frac{\alpha(w_{max})}{\beta} - 1} \le \mathbb{E}\left\{\left[1 - p^N(Q^N)\right]^{W_i^N}\right\}. \tag{9}$$

Further, there exists a constant $B \in (0,1)$ (independent of N) such that

$$\mathbb{E}\left\{\left[1-p^N(Q^N)\right]^{w_{max}}\right\} \leq B < 1, \ \forall N > 0. \tag{10}$$

Proof: See Appendix. ■

Proposition 1 tells us that in the stationary regime, the probability of flow i receiving no mark is always bounded above. Later on, we will also show that $\mathbb{E}\{W_i^N\} > 1$ for all i and N, i.e., it is also bounded below. This implies that there is no global synchronization among N flows in the steady-state since there always exists certain fraction of flows decreasing their window sizes at each RTT, thereby making the total arrivals to the queue stationary and the corresponding queue-length fluctuation under control.

From (10), we can show that the steady-state queue-length random variable Q^N is at least on the order of $O(N^\gamma)$ with non-zero probability.

Lemma 2 *There exist constants $q \in (0,\infty)$ and $\delta > 0$ such that, for all $N > 0$,*

$$\mathbb{P}\{Q^N > qN^\gamma\} \geq \delta. \tag{11}$$

Proof of Lemma 2: Since the function $(1-x)^{w_{max}}$ is convex for $x \in [0,1]$, we have from Jensen's inequality and from (10) that

$$\left[1-\mathbb{E}\{p^N(Q^N)\}\right]^{w_{max}} \leq \mathbb{E}\left\{[1-p^N(Q^N)]^{w_{max}}\right\} \leq B < 1.$$

Thus, we obtain

$$0 < A := 1 - B^{1/w_{max}} \leq \mathbb{E}\left\{p^N(Q^N)\right\}, \tag{12}$$

where $0 < A < 1$. Note that, since $p(x)$ is non-decreasing in x and $\lim_{x\to\infty} p(x) = 1$, we have

$$p(x) \leq p(q) + (1-p(q))1_{\{x>q\}},$$

for any $x, q \geq 0$. Thus, from $p^N(N^\gamma x) = p(x)$, we have

$$p^N(N^\gamma x) \leq p(q) + (1-p(q))1_{\{N^\gamma x > N^\gamma q\}}. \tag{13}$$

By choosing $x = Q^N/N^\gamma$ and taking expectation in (13), we have from (12)

$$0 < A \leq \mathbb{E}\{p^N(Q^N)\} \leq p(q) + (1-p(q))\mathbb{P}\{Q^N \geq qN^\gamma\}.$$

Note that, since $\lim_{x\to 0} p(x) = 0$, we can always choose q such that $0 < p(q) < A < 1$. Thus, for such $q \in (0,\infty)$, we have

$$0 < \frac{A-p(q)}{1-p(q)} \leq \mathbb{P}\left\{Q^N \geq qN^\gamma\right\},$$

and (11) follows by setting $\delta = (A - p(q))/(1 - p(q)) > 0$. ■

When the system is in the stationary regime, the total arrival to the queue as well as the queue-length fluctuation becomes stationary in time. In particular, since $Q^N(k)$ has the same distribution as $Q^N(k+1)$, we can rewrite (6) as

$$Q^N(k) \stackrel{d}{=} Q^N(k+1) = \left[Q^N(k) + \sum_{i=1}^{N} W_i^N(k) - NC \right]^+,$$

where $\stackrel{d}{=}$ means equality in distribution. The solution to this distributional equation always exists whenever $\mathbb{E}\{\sum_{i=1}^N W_i^N(k)\} < NC$ and is given by [17]

$$Q^N \stackrel{d}{=} \sup_{t>0} \left[\sum_{k=1}^{t} \sum_{i=1}^{N} W_i^N(k) - NCt \right]. \tag{14}$$

Hence, in some sense, even if the original system is of a closed-loop form, we can view the system dynamics in the steady-state as that of an open-loop queueing system with capacity NC fed by N flows, each of which is characterized by a stationary sequence $W_i^N(k)$ $(k = 1, 2, \ldots)$ whose distribution is yet to be found.

As $W_i^N(k+1)$ for different i evolve independently of each other given the current state at time k, and also from Proposition 1 saying that there is no synchronization among flows, we can assume that in the steady-state, W_i^N $(i = 1, 2, \ldots, N)$ are more or less independent. Further, we can also assume that each arrival process $W_i^N(k)$ is well behaved such that it satisfies the required assumption for the many-sources-asymptotic to hold. [5] Then, we can show the following:

Proposition 2 *For any $\gamma \in [0.5, 1)$ in our TCP/AQM systems, the steady-state link utilization defined by $\rho(N) = \mathbb{E}\{\sum_{i=1}^N W_i^N\}/NC$ approaches to 1 as N increases.*

Proof of Proposition 2: Clearly, in the steady-state, we must have $\rho(N) < 1$, since otherwise the queue-length will grow without bound, thus violating the stationarity of $Q^N(k)$. Suppose that $\liminf_{N\to\infty} \rho(N) = \rho^* < 1$. Choose a subsequence $n_k \uparrow \infty$ such that $\lim_{k\to\infty} \rho(n_k) = \rho^* < 1$. Then, from (14) and since the link utilization along the subsequence is bounded away from 1, we can apply the many-sources-asymptotic upper bound evaluated at zero buffer.

[5] For instance, see [18]. The required assumption is very general and includes almost all the existing traffic models in the literature including long-range dependent processes.

Specifically, we have

$$\limsup_{n_k \to \infty} \frac{1}{n_k} \log \mathbb{P}\{Q^{n_k} > 0\} \leq -I(0),$$

where $I(0) > 0$ for any link utilization $\rho < 1$ [19]. This means that, along the subsequence, $\mathbb{P}\{Q^{n_k} > 0\}$ decreases to zero exponentially fast as n_k grows. But, this contradicts (11) for large N. Therefore, we should have $\rho^* = 1$. ■

We have earlier mentioned that if $\mathbb{E}\{W_i^N\} > 1$, the expectation in (10) is bounded away from 0 and 1. With the help of Proposition 2, we now show this is indeed the case, and further, the probability of queue-length larger than $q'N^\gamma$ is also bounded above.

Lemma 3 *There exist constants $q' \in (0, \infty)$ and $\eta < 1$ such that, for all sufficiently large N,*

$$\mathbb{P}\{Q^N > q'N^\gamma\} \leq \eta < 1. \tag{15}$$

Proof of Lemma 3: Note that, since all the N flows are symmetric, their steady-state window size distributions are identical. In other words, the link utilization simply becomes $\rho(N) = \mathbb{E}\{\sum_{i=1}^N W_i^N\}/NC = \mathbb{E}\{W_i^N\}/C$. So, from Proposition 2 and since $C > 1$, we have $\mathbb{E}\{W_i^N\} > 1$ for all large N. Thus, we have for all sufficiently large N

$$0 < \zeta \leq \frac{\mathbb{E}\{W_i^N\} - 1}{w_{max} + \frac{\alpha(w_{max})}{\beta} - 1},$$

where $\zeta \in (0, 1)$. Thus, from (9), $(1 - p^N(\widehat{Q}_{\overline{W^N}})) \in [0, 1]$, and $1 \leq W_i^N \leq w_{max}$, we have

$$0 < \zeta \leq \mathbb{E}\left\{[1 - p^N(Q^N)]^{W_i^N}\right\} \leq 1 - \mathbb{E}\{p^N(Q^N)\}.$$

Thus, we obtain

$$\mathbb{E}\left\{p^N(Q^N)\right\} \leq 1 - \zeta < 1. \tag{16}$$

Since $p(x)$ is non-decreasing, we clearly have $p(q')1_{\{x \geq q'\}} \leq p(x)$ for all $x, q' \geq 0$. Thus, similarly as in the proof of Lemma 2, we have from (16) that

$$p(q')\mathbb{P}\left\{Q^N \geq q'N^\gamma\right\} \leq \mathbb{E}\left\{p^N(Q^N)\right\} \leq 1 - \zeta < 1.$$

Since $p(x)$ is non-decreasing and continuous, and $\lim_{x \to \infty} p(x) = 1$, we can choose q' such that $1 - \zeta < p(q') < 1$. Hence,

$$\mathbb{P}\left\{Q^N \geq q'N^\gamma\right\} \leq \frac{1 - \zeta}{p(q')} := \eta < 1,$$

and we are done. ■

In order to further characterize the steady-state queue-length fluctuation, we proceed as follows. Suppose that we know the exact distribution of the aggregate arrival process to the queue in the stationary regime, Then, from (14), in principle, we should be able to find the distribution (or at least a very good approximation) of the queue-length random variable Q^N by using well-known results for a queueing system fed by many flows. Examples include many-sources-asymptotic [20,18,21,22], Gaussian traffic modeling [23–25], Poisson limit [26], moderate deviation results [27], or heavy-traffic limits [28,29].

Note that each of the above different approaches holds true only when the system is scaled in some specific way (i.e., the link capacity or the buffer size should be some suitable function of N), and there hardly seems to be any unified limiting result encompassing every possible way of scaling the system. For example, the many-sources-asymptotic requires that both the buffer size and the link utilization increase at the same rate as N increases with the link utilization fixed (less than one), while the Poisson limit is valid especially when the buffer size remains fixed and the heavy-traffic limit becomes useful when the buffer size increases like $O(\sqrt{N})$ with the link utilization approaching to one appropriately.

However, it is by no means feasible to calculate the exact distribution of $\sum_{i=1}^{N} W_i^N(k)$ in (14) as it is a solution of $(N+1)$-dimensional distributional equations in the steady-state. Instead of trying to directly solve the distributional equation, we resort to the Gaussian traffic modeling to serve our purpose. The Gaussian traffic modeling [23–25] turns out to be extremely versatile and accurate over a wide range of network operating points including the moderate deviation scaling [23,27,30]. Also, since N flows are likely to be independent, this approach can further be justified by the Central Limit Theorem as well as also by the recent empirical measurement showing that the marginal distribution of the sum of congestion windows of all flows can be well approximated by a Gaussian distribution [3].

Suppose the amount of arrival $A(0,t)$ to a queue over a time interval $(0,t]$ is a Gaussian process with mean $\mathbb{E}\{A(0,t)\} = \lambda t$ and variance $v(t) = \mathrm{Var}\{A(0,t)\}$. Then, the queue-length distribution is well approximated by [23,24]

$$\mathbb{P}\{Q > x\} \approx \exp\left(-\inf_{t>0} \frac{((\mu-\lambda)t+x)^2}{2v(t)}\right), \tag{17}$$

where μ is the link capacity. In our case, from (14), the link capacity is NC and the mean arrival rate becomes $\mathbb{E}\{A(0,t)\}/t = NC\rho(N)$. Further, we assume that the variance of the sum of N flows' arrival to the queue is approximated by $\sigma^2 N t^{2H}$ where $H \in [0.5, 1)$.[6] Then, the queue-length distribution can be

[6] The exact equality corresponds to the fractional Brownian motion process with parameter H and this includes long-range dependent processes as well.

approximated as [7]

$$\mathbb{P}\{Q^N > O(N^\gamma)\} \approx \exp\left(-\inf_{t>0} \frac{[NC(1-\rho(N))t + O(N^\gamma)]^2}{2\sigma^2 N t^{2H}}\right)$$
$$= \exp\left(-l(H)(1-\rho(N))^{2H} N^{2H-1+\gamma(2-2H)}\right), \tag{18}$$

where $l(H)$ is a function of C, σ^2, H and does not depend on N. Then, since Lemma 2 and 3 guarantees $\mathbb{P}\{Q^N > O(N^\gamma)\} \in (0,1)$ for all sufficiently large N, we must have

$$(1-\rho(N))^{2H} N^{2H-1+\gamma(2-2H)} \in (0,\infty)$$

for all sufficiently large N. This in turns gives, as N increases,

$$\rho(N) \approx 1 - O(N^{-\frac{2H-1+\gamma(2-2H)}{2H}}) \to 1, \tag{19}$$

since $2H - 1 + \gamma(2-2H) > 0$.

Thus far, by relying on a Gaussian traffic modeling approach, we have obtained the convergence rate of $\rho(N)$ to 1 as N grows. From this, we can characterize the queue-length fluctuation in more detail via (18). In particular, consider the probability that the queue-length being even larger than $N^{\gamma+\epsilon}$ for any ϵ. In this case, by substituting (19) back into (18), it follows that $\mathbb{P}\{Q^N > O(N^{\gamma+\epsilon})\}$ decreases to zero as N increases. We have thus shown the following:

Proposition 3 *For any $\gamma \in [0.5, 1)$ in our TCP/AQM systems and for any given $\epsilon > 0$, $\lim_{N\to\infty} Q^N/N^{\gamma+\epsilon} = 0$ in probability.*

Proposition 3 gives us more precise range for the steady-state queue-length fluctuation. In the steady-state, the system behaves in a way that the link utilization is almost full, while the queue-length fluctuation is 'under control', with its fluctuation contained mostly around $O(N^\gamma)$, no more.

3.3 Drop-tail vs. AQM with ECN marks

In the previous section, we assumed an infinite buffer and packet marking is the only sources of 'congestion signal'. Nevertheless, with slight modification, we can still show that our results will continue to hold even under a finite buffer setting. (We only need to modify (6) to reflect the finite buffer.)

[7] Here, $O(N^\gamma)$ means that there exist constants $0 < a < b < \infty$ such that $0 < a < O(N^\gamma)/N^\gamma < b < \infty$ for all large N.

Let $B(N)$ be the buffer size ($B(N) > q_{max}N^{\gamma}$). Obviously, we will have to choose the buffer size $B(N)$ at least on the order of N^{γ}. However, Proposition 3 also asserts that we don't need too large buffer in order to lower the packet-drop ratio. The right choice for 'buffer sizing' involves many performance-related issues and network design consideration and has recently been a key issue in the literature. For instance, it is argued that the buffer size must be large enough (up to the bandwidth-delay product) to keep the packet loss very small and for stability [6,4,13], while $B(N) = O(\sqrt{N})$ is enough to achieve high link utilization [3]. Quite recently, it was also suggested that the buffer size should be chosen much smaller than the current standard [31], as smaller buffer radically decreases queueing delay at the cost of minor degradation of link utilization. All these results are, however, obtained only under the drop-tail policy, and there lacks a systematic and rigorous analysis of the TCP/AQM system under a general queue-based AQM with scaling $O(N^{\gamma})$ $(0.5 \leq \gamma < 1)$.

In Section 3.2, we have shown that the link utilization approaches to 1 as N increases for any choice of N^{γ} for the AQM. In other words, in the limit, the throughput of each flow will approach to the normalized link capacity C. At the same time, since the throughput of a TCP flow can also be approximated by the RTT along its path and the rate of 'packet loss' (or the rate of 'congestion signal' being generated) [7], we can write

$$C \approx \frac{K}{RTT}\frac{1}{\sqrt{p}} \quad \text{for large } N, \tag{20}$$

where K is a constant and p is the probability of receiving any 'congestion signal' during one RTT. Now, let T_p be the two-way propagation delay (constant) and T_q be the queueing delay. We also denote by p_l the probability of packet loss, while by p_m the probability of packet being marked. Then, we have $RTT = T_p + T_q$ and $p \approx p_l + p_m$, and (20) becomes

$$C \approx \frac{K}{T_p + T_q}\frac{1}{\sqrt{p_l + p_m}} \quad \text{for large } N. \tag{21}$$

Clearly, under drop-tail, the only way to generate congestion signal is via packet loss, and we have $p_m = 0$. Further, when the buffer size $B(N)$ is chosen on the order of N^{γ} with $\gamma \in [0.5, 1)$, it follows from Lemmas 2 and 3 (with slight modification) that p_l is also bounded away from 0 and 1. Thus, it cannot be arbitrarily small for any large N, and this is in line with the observation in [4]. Moreover, when the buffer size is chosen far smaller than the bandwidth-delay product, the queueing delay becomes negligible, which makes p_l even larger (see (21)). This also explains why we need large buffer size under drop-tail to decrease the packet loss ratio p_l, and this tradeoff is inevitable.

However, under a queue-based AQM with ECN marks, we can get around such a tradeoff. By choosing $\gamma \in [0.5, 1)$ and the buffer size $B(N) = O(N^{\gamma+\epsilon})$, we can make T_q arbitrarily small and thereby enhance the system stability. The link utilization is still almost 100% from Proposition 2. In addition, we see that a suitably chosen 'extra space' $(B(N) - q_{max}N^{\gamma})$ over which all the packets are marked (not dropped) plays a crucial role in achieving full link utilization, while separating the congestion signal from packet loss in a way that $p_m \in (0, 1)$ and $p_l \approx 0$. As a result, we see that using any queue-based AQM in a link with large capacity and many flows adds much advantage to the system performance over the simple drop-tail policy.

4 Utility of Discrete-Time Models for TCP/AQM

By capturing the system 'state' at each RTT instant and its evolution over the series of RTTs, we have developed a discrete-time Markov chain model for TCP/AQM. This means that all the system dynamics at sub-RTT levels (finer time scale) are ignored and lumped into a single statistic captured at the start (or the end) of each RTT. So, the natural question to ask is: under what situation is this approach valid? We here provide arguments showing that for $\gamma \geq 1/2$, the system dynamics can safely be captured only at certain instants (e.g., at $k \times RTT$, $k = 1, 2, \ldots$), while for $\gamma < 1/2$ (which we do not consider in this paper), all the statistical fluctuation of the system dynamics over finer time scales (such as packet arrival patterns to the queue within each RTT) must be taken into account.

Suppose we keep track of every packet arrival instant to the queue within each RTT. Each flow i will transmit about $C \times RTT$ number of packets onto the network, so within one RTT, there will typically be $O(N)$ number of packets arriving to the queue. Under the current discrete-time model, we only care about the total number of packets to the queue per RTT. However, these packets may arrive in a random fashion; they may arrive infrequently or sometimes back-to-back. When the queue is mostly non-empty, note that all these random fluctuations in packet arrivals per RTT can add up to $O(\sqrt{N})$ additional fluctuation in the queue-length with high probability, provided that packet inter-arrival times are not highly correlated. (This is similar to the Central Limit Theorem regime.) However, since we scale the AQM according to N^{γ} where $\gamma \geq 1/2$, there is enough room in the buffer to absorb such fluctuation $(O(\sqrt{N}))$, and thus the system dynamics is unlikely to be affected by those random arrivals in time. Clearly, if our scale for the AQM is N^{γ} with $\gamma < 1/2$, then any of these typical queue-length fluctuation due to random packet arrivals can cause a large number of packet losses or marks. In such a case, we evidently have to take all the packet-level variations within each RTT into account in order to correctly capture the system dynamics. This case explicitly

comes under scrutiny in our recent paper [32].

In addition to the above sample-path based argument, we can further justify the use of a discrete-time model under our scale via the Gaussian traffic modeling as in Section 3.2 and the notion of the dominant time scale (DTS) [25,33,34]. In the context of teletraffic engineering, the notion of the DTS has been extremely useful in identifying the critical time scale over which the traffic statistics should be captured. Under a Gaussian traffic modeling, the DTS is the time index at which the expression in (17) attains its minimum and coincides the 'critical time scale' in the many-sources-asymptotic via large deviation theory. (See [20,35,25] for details.) Shortly speaking, the DTS is a function of the buffer level, link capacity, and some statistics of the input traffic to the queue. In general, it is increasing as the buffer level or the link utilization increases. In other words, for a larger buffer and higher link utilization, the input traffic should be characterized over a longer time scale to calculate the buffer overflow probability. For example, when there are N flows with capacity NC and the buffer level is kept fixed with constant utilization $\rho < 1$, it turns out that the DTS tends to zero for large N [26]. This means that packet-level dynamics over fine time scales ($o(1)$) should be taken into account, and this was the main idea in [26] and the very reason why Poisson regime kicks in under a fixed buffer.

Consider again the approximation in (17) with $v(t) = \sigma^2 t^{2H}$. Then, it is easy to see that the DTS (minimizer) becomes

$$\hat{t} = \left(\frac{H}{1-H}\right) \frac{x}{\mu - \lambda}.$$

Under our choice of the buffer level $O(N^\gamma)$ and the link utilization $\rho(N)$, this means

$$\hat{t} = \left(\frac{H}{1-H}\right) \frac{O(N^\gamma)}{NC(1-\rho(N))}.$$

Using (19), the above relation becomes

$$\hat{t} = O\left(N^{\frac{2\gamma-1}{2H}}\right). \tag{22}$$

Hence, we see that if $\gamma < 1/2$, then the DTS is decreasing to zero for large N. In other words, statistics over minuscule time scale ($o(1)$) dominate in determining the probability of $\mathbb{P}\{Q^N > O(N^\gamma)\}$. In that case, we should have tracked all the packet-level dynamics over smaller time scale as mentioned earlier. On the other hand, in our current setting with $\gamma \geq 1$, the DTS is getting larger for large N, implying that we can capture the system statistics only at coarser time scales (multiple of RTTs or higher) without much loss of key system dynamics.

5 Simulation Results

We here present simulation results using *ns*-2 [36] under different network configurations to validate the analysis in the previous section. Specifically, we consider two scenarios: (i) persistent flows with homogeneous RTTs; (ii) persistent flows with heterogeneous RTTs. We simulate three AQM schemes with ECN marking: RED, REM [9], and PI [10], compared with Drop-Tail (DT). Our performance metrics are the average queueing delay, jitter (the standard deviation of queueing delay), link utilization and packet loss ratio. To illustrate the benefit of AQM schemes with packet marking, we also consider the ratio between packet loss and the overall congestion signals, i.e., $n_l/(n_l + n_m)$ where n_l is the number of lost packets (due to buffer overflow) and n_m is the number of marked packets.

5.1 System Topology and Configurations

For simulation, we design the system topology displayed in Figure 3, where many persistent TCP flows share a bottleneck router (i.e., Router 1) with link capacity of $NC = 25N$ packets/sec and AQM schemes or DT. Each TCP source sends data to the corresponding sink through the bottleneck link. The local delay from TCP sources to the bottleneck router is 50 ms, the bottleneck link delay is 25 ms and the local delay from Router 2 to the sink nodes is 25 ms. Hence, the total two-way propagation delay is $2 \times (50 + 25 + 25) = 200$ ms.

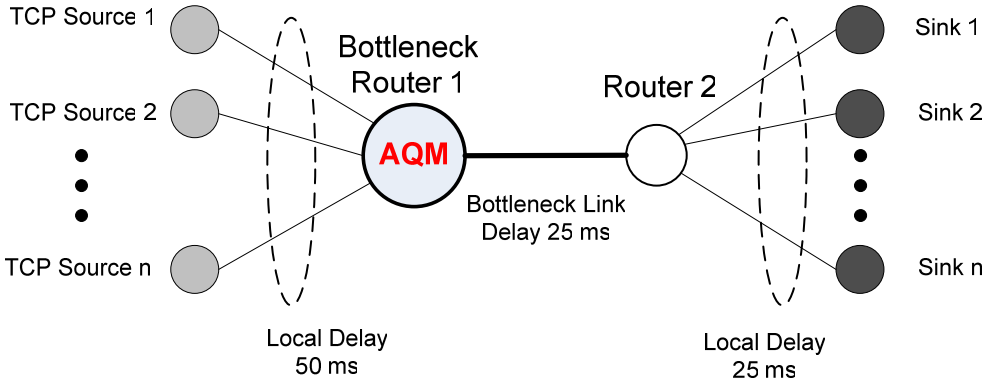
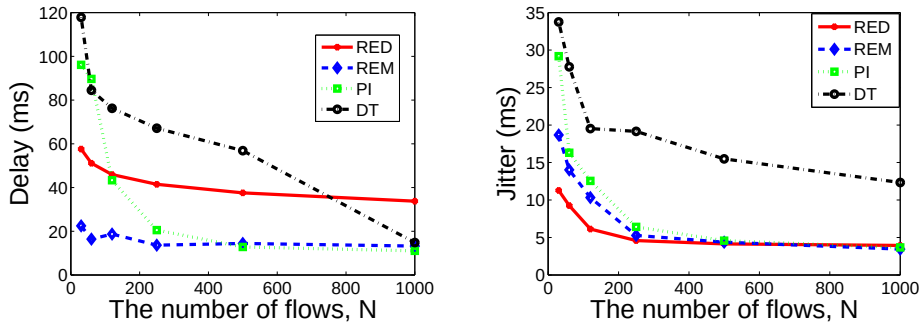


Fig. 3. Simulation topology.

We configure the AQMs and DT in the following way. The buffer size is set to $CT \times N^\gamma$, where C is the average bandwidth share for each flow, T is the RTT, and N is the number of flows. Note that $\gamma = 1$ corresponds to the traditional rule-of-thumb, i.e., bandwidth-delay product, and $\gamma = 0.5$ represents the scheme in [3], i.e., $CT\sqrt{N}$. We configure the system such that C is 25 packets/sec, and the two-way propagation delay is $2T_p = 200$ ms. Note that

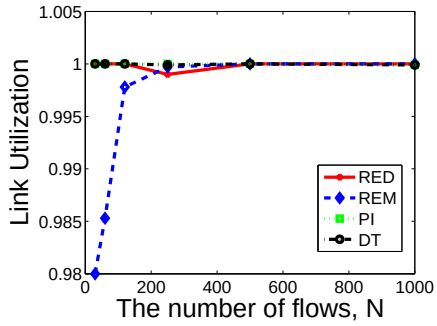
for $\gamma = 0.5$, the queueing delay becomes negligible ($O(\sqrt{N}/NC) = O(1/\sqrt{N})$) and can be ignored, thus we have $CT \approx 2CT_p = 5$ packets. For $\gamma = 1$, the queueing delay will be $O(1)$, so $T > 2T_p$. To account for such a difference, we add 1 packet margin such that the buffer size is $6N^\gamma = (2CT_p + 1)N^\gamma$ for all AQMs and DT. For RED, the minimum and the maximum marking threshold (i.e., min_{th} and max_{th}) are set to N^γ and $5N^\gamma$, respectively, with the maximum marking probability $P_{max} = 0.2$. As to REM and PI, we configure the target queue sizes (pbo_ for REM, q_{ref} for PI) as N^γ, and other parameters of AQMs and DT are set to the default values in *ns*-2.

5.2 Homogeneous RTTs

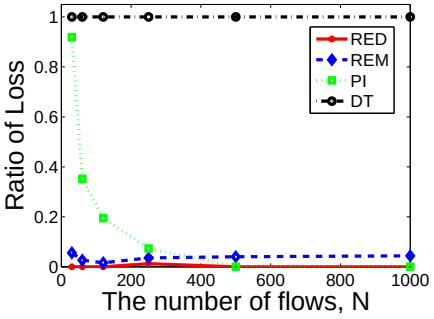


(a) Average queueing delay

(b) Jitter



(c) Link utilization

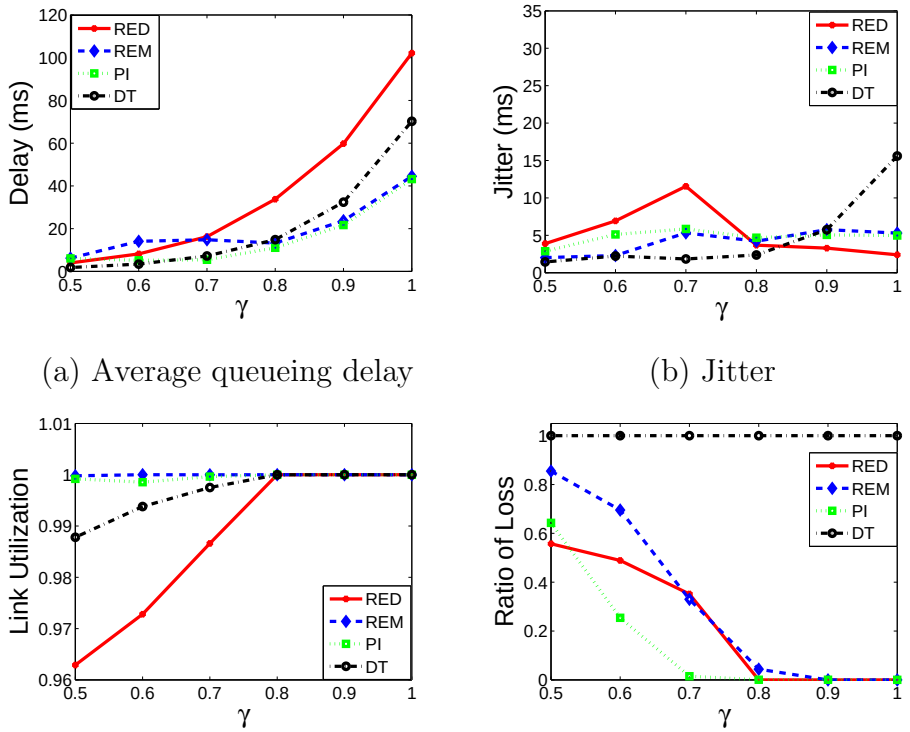


(d) Ratio of packet loss: $\frac{n_l}{n_l+n_m}$

Fig. 4. Homogeneous RTTs: performance metrics with the increase of the number of flows, N for a fixed scaling parameter $\gamma = 0.8$.

Figure 4 shows the performance metrics with the increase of the number of flows, N for a fixed scaling parameter $\gamma = 0.8$. The queueing delay and jitter decrease as N increases for all the schemes (AQMs and DT). This is well expected since the maximum queueing delay is $6N^\gamma/NC = O(N^{\gamma-1})$. When the number of flows is larger than several hundreds, all the schemes considered yield full link utilization. The average packet loss probability is small and less

than 4% in any case. Figure 4 also shows that the congestion signal for DT is purely packet-loss regardless of the number of flows, while for AQM schemes with marking, almost all the congestion signal are from packet marking and there is virtually no packet loss when N is large enough. This clearly indicates added benefit of employing AQMs with packet marking as it results in less retransmission of packets, improving goodput of each flow.



(a) Average queueing delay (b) Jitter

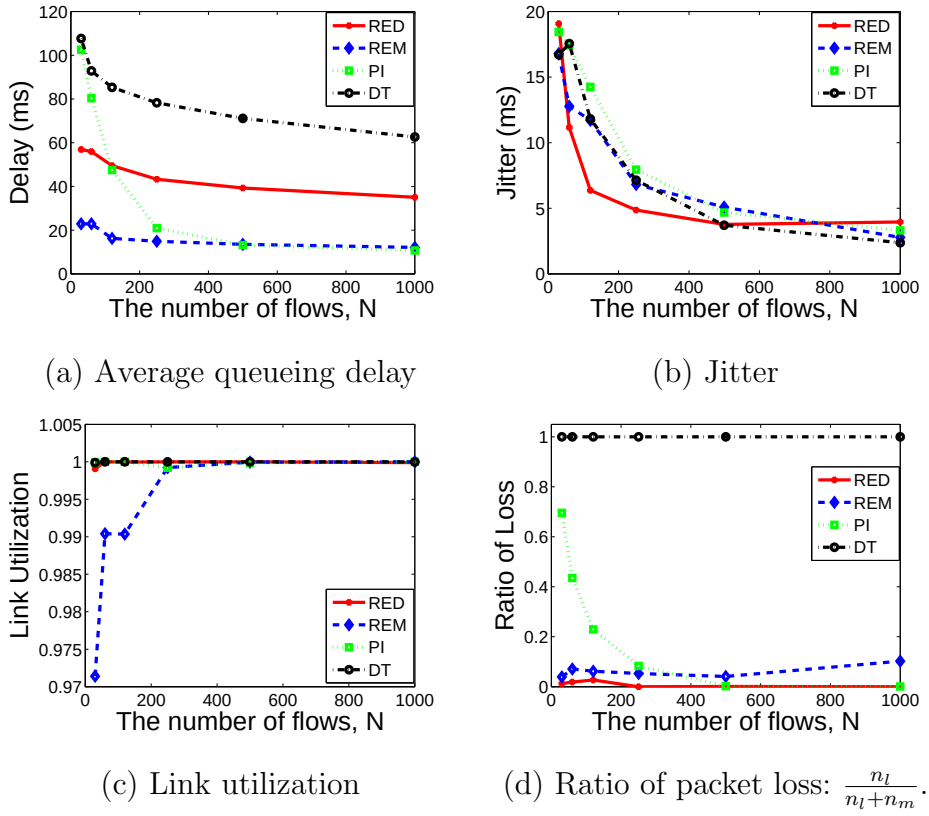
(c) Link utilization (d) Ratio of packet loss: $\frac{n_l}{n_l+n_m}$

Fig. 5. Homogeneous RTTs: performance metrics with the increase of scaling parameter γ for fixed number of flows $N = 1000$.

To illustrate the effect of the scaling parameter γ, we perform the simulation for fixed number of flows $N = 1000$ with various γ in $[0.5, 1]$ and show the results in Figure 5. As expected, the queueing delay increases as γ increases and jitter for AQMs (not for DT) becomes smaller for large γ, since the queue dynamics tend to be stable. As soon as γ becomes larger than 0.8, we have 100% link utilization, and all the congestion signals are from packet marks under all AQMs considered (except DT). If we had more number of flows, say, $N = 10000$, then we would achieve the full link utilization and zero packet loss for smaller γ, resulting in even greater saving of the buffer size. With the help of AQM with packet marking, our result thus in some sense fills the gap between $O(\sqrt{N})$ scheme for the buffer size in [3] for 'high' utilization and much larger buffer size requirement ($O(N)$) as in the traditional rule-of-thumb or as suggested in [6] for 100% utilization.

5.3 *Heterogeneous RTTs*

Next, we repeat the same simulations under heterogeneous RTTs. The two-way propagation delays are randomly drawn from the uniform distribution over $[120, 280]$ ms with mean 200 ms. Specifically, we randomize the local delay from TCP sources to the bottleneck link with the uniform distribution of $[10, 90]$ ms with the same mean of 50 ms as in homogeneous RTT case. As can be seen in Figures 6 and 7, all the performance metrics display the same trend as in the case of homogeneous RTTs.



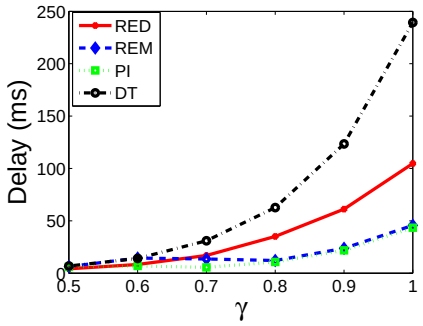
(a) Average queueing delay (b) Jitter

(c) Link utilization (d) Ratio of packet loss: $\frac{n_l}{n_l+n_m}$.

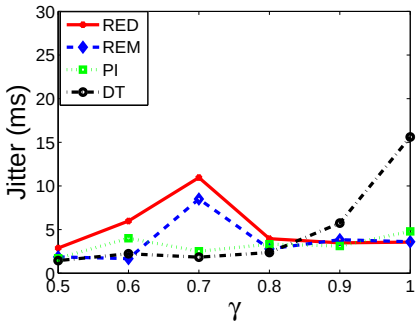
Fig. 6. Heterogeneous RTTs: performance metrics with the increase of the number of flows, N for a fixed scaling parameter $\gamma = 0.8$.

5.4 *Mixture of long-live and short-live flows*

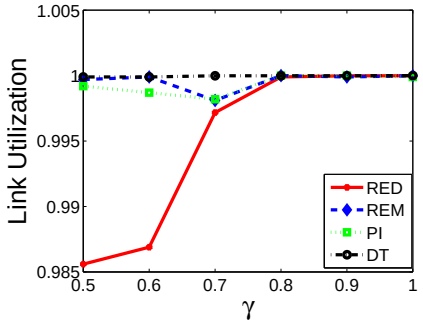
It is well-known that file transfers account for the majority of the traffic over the Internet. Typical file sizes follow Pareto distribution as reported in [37,38], which means the majority of the file transfers has a short life-time and few has a very long life-time. In general, short-lived flows finish transmission within the slow-start phase and never reach the congestion control phase, thus do not react to the congestion signal as the long-lived flows do.



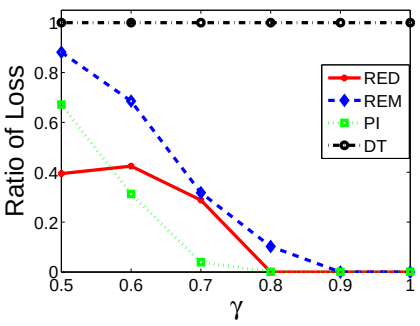
(a) Average queueing delay



(b) Jitter



(c) Link utilization



(d) Ratio of packet loss: $\frac{n_l}{n_l+n_m}$.

Fig. 7. Heterogeneous RTTs: performance metrics with the increase of scaling parameter γ for a fixed number of flows, $N = 1000$.

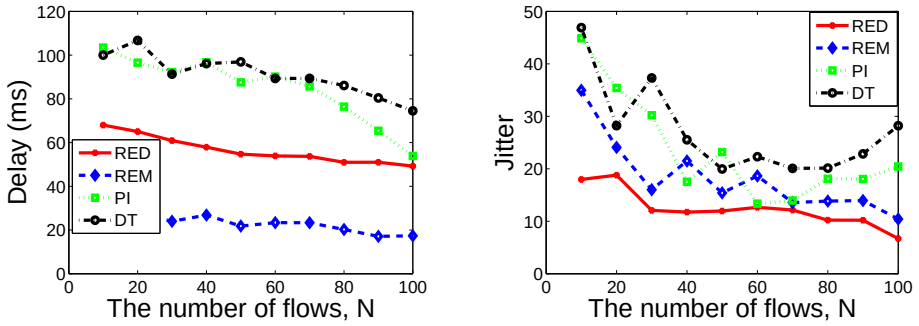
In this section, we simulate TCP/AQM systems with a mixture of long-lived and short-lived flows under heterogeneous RTTs. We generate the traffic as follows. For each persistence TCP flow, we add a number of flows to share the link capacity. The file size for each added flow is characterized by Pareto distribution with shape parameter 1.5 [8] with the minimum file size $x_{min} =$ 1.667 kbyte. So, the distribution of the file size X of each added flow is given by

$$P(X > x) = \left(\frac{x}{x_{min}}\right)^{-1.5}, \quad x \geq x_{min}. \tag{23}$$

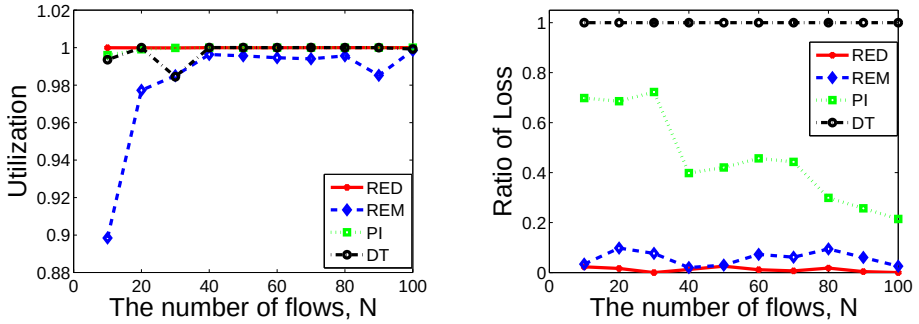
This gives the mean file size of 5 kbyte and a median of 2.645 kbyte, which implies that most of the flows end within 10 packets and nearly half of them end within 6 packets. During the simulation, for each persistent TCP flow, we add about 120 random flows whose file sizes are distributed according to (23) (Most of them are short-lived ones.), and those random flows as a whole consume about 20% of the total bandwidth of the link. We have also examined other situation where the random flows take account for 50% of the total bandwidth, and observed similar results.

We can see that Figures 8 and 9 show the same trend as homogeneous RTTs

[8] The shape parameter is between 1 and 2 [37].



(a) Average queueing delay (b) Jitter



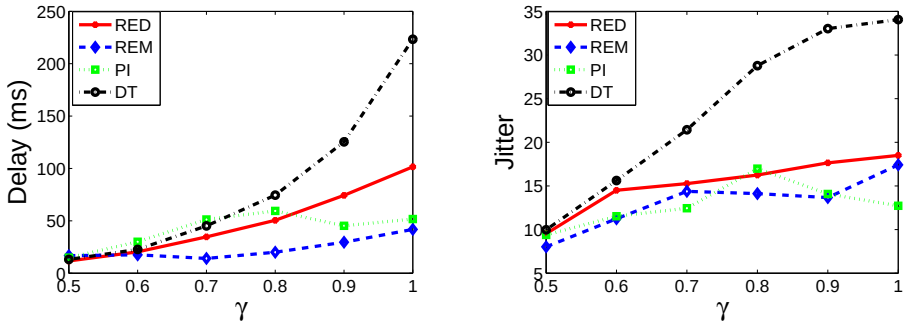
(c) Link utilization (d) Ratio of packet loss: $\frac{n_l}{n_l+n_m}$.

Fig. 8. Mixture of long-live and short-live flows: performance metrics with the increase of the number of flows, N for a fixed scaling parameter $\gamma = 0.8$.

and heterogeneous RTTs cases. Note that for the mixture traffic case, we cannot simulate up 1000 flows as for the other two cases, because the short-life flows consume a great deal of system resource as well.

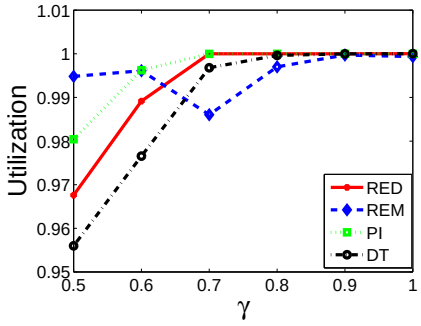
6 Conclusions

Existing results regarding buffer sizing in the Internet routers suggest that the buffer size be $O(\sqrt{N})$ [3] for high link utilization or $O(N)$ [6,2] for 100% utilization where the link has capacity NC serving N flows. These results are based on more or less heuristic arguments and are all limited to the drop-tail scheme. In this paper, we explore the limiting behavior of a TCP/AQM system for intermediate buffer size of $O(N^\gamma)$ $(0.5 \leq \gamma < 1)$ under general queue-based AQM schemes with ECN marking and under generalized AIMD for each TCP flow. We develop a stochastic discrete-time model to characterize the system dynamics and then show that we can have 100% link utilization for a large number of flows when the buffer size requirement is anywhere between $O(\sqrt{N})$ and $O(N)$. In addition, under the same scale, we show that AQM with ECN marking is able to generate the sufficient congestion signal through packet

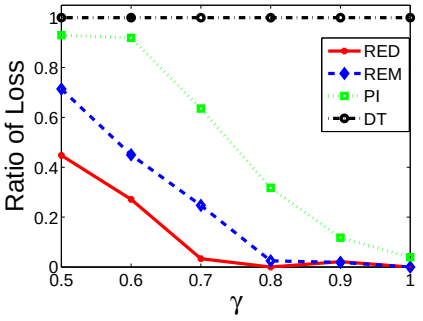


(a) Average queueing delay

(b) Jitter



(c) Link utilization



(d) Ratio of packet loss: $\frac{n_l}{n_l+n_m}$.

Fig. 9. Mixture of long-live and short-live flows: performance metrics with the increase of scaling parameter γ for a fixed number of long-live flows, $N = 100$.

marking and eliminate the packet loss, which would be impossible under the drop-tail. We also illustrate that the choice of scaling (i.e., either $\gamma < 0.5$ or $\gamma \geq 0.5$) for buffer size and AQM parameters determines whether we can use discrete-time model to capture the system dynamics.

References

[1] D. Y. Eun, X. Wang, Performance Modeling of TCP/AQM with Generalized AIMD under Intermediate Buffer Sizes, in: Proceedings of IEEE International Performance Computing and Communications Conference (IPCCC), Phoenix, AZ, 2006.

[2] C. Villamizar, C. Song, High Performance TCP in Ansnet, ACM Computer Communication Review 24 (5) (1994) 45–60.

[3] G. Appenzeller, I. Keslassy, N. McKeown, Sizing Router Buffers, in: Proceedings of ACM SIGCOMM, Portland, OR, 2004.

[4] A. Dhamdhere, H. Jiang, C. Dovrolis, Buffer Sizing for Congested Internet Links, in: Proceedings of IEEE INFOCOM, Miami, FL, 2005.

[5] D. Y. Eun, X. Wang, Stationary Behavior of TCP/AQM with Many Flows under Aggressive Packet Marking, in: IEEE International Conference on Communication, Seoul, Korea, 2005.

[6] J. Sun, M. Zukerman, K. Ko, G. Chen, S. Chan, Effect of Large Buffers on TCP Queueing Behavior, in: Proceedings of IEEE INFOCOM, Hong Kong, 2004.

[7] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, Modeling TCP Throughput: a Simple Model and its Empirical Validation, in: Proceedings of ACM SIGCOMM, 1998.

[8] M. Mathis, J. Semke, J. Mahdavi, The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm, in: Proceedings of ACM SIGCOMM, 1997.

[9] S. Athuraliya, V. H. Li, S. H. Low, Q. Yin, REM: Active Queue Management, IEEE Network 15 (3) (2001) 48–53.

[10] D. T. C. Hollot, V. Misra, W. Gong, Analysis and Design of Controllers for AQM Routers Supporting TCP Flows, IEEE Transactions on Automatic Control 47 (6) (2002) 945–959.

[11] P. Tinnakornsrisuphap, A. M. Makowski, Limit Behavior of ECN/RED Gateways Under a Large Number of TCP Flows, in: Proceedings of IEEE INFOCOM, San Francisco, CA, 2003.

[12] P. Tinnakornsrisuphap, R. J. La, Characterization of Queue Fluctuations in Probabilistic AQM Mechanisms, in: Proceedings of ACM SIGMETRICS, New York, NY, 2004.

[13] S. H. Low, F. Paganini, J. Wang, J. C. Doyle, Linear stability of TCP/RED and a scalable control, Computer Networks 43 (5) (2003) 633–647.

[14] T. Kelly, Scalable TCP: Improving Performance in Highspeed Wide Area Networks, 2002, submitted for publication.

[15] S. Floyd, HighSpeed TCP for Large Congestion Windows, RFC 3649.

[16] P. Brémaud, Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues, Springer-Verlag, 1999.

[17] D. Aldous, A. Bandyopadhyay, A survey of max-type recursive distributional equations, The Annals of Applied Probability 15 (2) (2005) 1047–1110.

[18] N. Likhanov, R. Mazumdar, Cell loss asymptotics for buffers fed with a large number of independent stationary sources, Journal of Applied Probability 36 (1) (1999) 86–96.

[19] D. Y. Eun, N. B. Shroff, Network Decomposition in the Many-Sources Regime, Advances in Applied Probability 36 (3) (2004) 893–918.

[20] C. Courcoubetis, V. A. Siris, G. D. Stamoulis, Application of the many sources asymptotic and effective bandwidths to traffic engineering, Telecommunication Systems 12 (1999) 167–191.

[21] M. Mandjes, J. H. Kim, Large deviations for small buffers: an insensitivity result, Queueing Systems 37 (2001) 349–362.

[22] D. Wischik, Sample path large deviations for queues with many inputs, Annals of Applied Probability 11 (2000) 379–404.

[23] J. Choe, N. B. Shroff, A Central Limit Theorem Based Approach for Analyzing Queue Behavior in High-Speed Networks, IEEE/ACM Transactions on Networking 6 (5) (1998) 659–671.

[24] J. Choe, N. B. Shroff, Use of Supremum Distribution of Gaussian Processes in Queueing Analysis with Long-Range Dependence and Self-Similarity, Stochastic Models 16 (2).

[25] D. Y. Eun, N. B. Shroff, A Measurement-Analytic Approach for QoS Estimation in a Network based on the Dominant Time Scale, IEEE/ACM Transactions on Networking 11 (2) (2003) 222–235.

[26] J. Cao, K. Ramanan, A Poisson Limit for Buffer Overflow Probabilities, in: Proceedings of IEEE INFOCOM, 2002.

[27] D. Wischik, Moderate deviations in queueing theoryDraft.

[28] C. Knessl, J. A. Morrison, Heavy-Traffic Analysis of Data-Handling System with Many Sources, SIAM Journal on Applied Mathematics 51 (1) (1991) 187–213.

[29] W. Whitt, Stochastic-Process Limits, Springer-Verlag, New York, 2002.

[30] A. Ganesh, N. O'Connell, D. Wischik, Big Queues, Springer, Lecture notes in mathematics, volume 1838, 2004.

[31] G. Raina, D. Wischik, Buffer sizes for large multiplexers: TCP queueing theory and instability, in: EuroNGI, Rome, 2005.

[32] D. Y. Eun, X. Wang, A Doubly-Stochastic Approach to the Performance of TCP/AQM in Large-Bandwidth Networks, submitted to *IEEE/ACM Transaction on Networking*, Dec. 2005.

[33] R. J. Gibbens, Y. C. Teh, Critical time and space scales for statistical multiplexing in multiservice networks, in: Proceedings of the 16th International Teletraffic Congress, Edinburgh, UK, 1999, pp. 87–96.

[34] M. Grossglauser, J.-C. Bolot, On the relevance of long-range dependence in network traffic, IEEE/ACM Transactions on Networking (1999) 629–640.

[35] D. Y. Eun, N. B. Shroff, A Measurement-Analytic Framework for QoS Estimation Based on the Dominant Time Scale, in: Proceedings of IEEE INFOCOM, Anchorage, AK, 2001.

[36] The Network Simulator: ns-2, in: http://www.isi.edu/nsnam/ns/, 2004.

[37] M. Crovella, A. Bestravos, Self-similarity in World Wide Web Traffic: Evidence and Possible Cause, in: Proceedings of ACM SIGMETRICS, 1996.

[38] B. Kikdar, S. Kayanaraman, K. S. Vastola, An Integrated Model for the Latency and Steady-State Throughput of TCP Connections, Performance Evaluation 46 (2-3) (2001) 139–154.

Appendix

Proof of Proposition 1: As the system is in steady-state, for any well-defined (measurable) function $g : \mathbb{R} \to \mathbb{R}$, we should have $\mathbb{E}\{g(W_i^N(k+1))\} = \mathbb{E}\{g(W_i^N(k))\}$. Given $\vec{X}^N = \vec{x}^N$, let $f_i \in [0,1]$ denote the probability that none of w_i^N packets from flow i is marked. Specifically, we have

$$f_i\left(\vec{x}^N\right) := \left[1 - p^N(q^N)\right]^{w_i^N}. \tag{24}$$

Then, given that $\vec{X}^N(k) = \vec{x}^N(k)$, we have, for any measurable function g, [9]

$$g\left(W_i^N(k+1)\right) = \begin{cases} g\left(\left(w_i^N(k) + \alpha(w_i^N(k))\right) \wedge w_{max}\right) & \text{w.p. } f_i\left(\vec{x}^N(k)\right), \\ g\left((1-\beta)w_i^N(k) \vee 1\right) & \text{w.p. } 1 - f_i\left(\vec{x}^N(k)\right). \end{cases}$$

Thus,

$$\begin{aligned} \mathbb{E}\left\{g\left(W_i^N(k+1)\right)\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{g\left(W_i^N(k+1)\right) \middle| \vec{X}^N(k)\right\}\right\} \\ &= \mathbb{E}\left\{g\big((W_i^N(k) + \alpha(W_i^N(k))) \wedge w_{max}\big) f_i\big(\vec{X}^N(k)\big)\right\} \\ &\quad + \mathbb{E}\left\{g\left((1-\beta)W_i^N(k) \vee 1\right) \cdot \left(1 - f_i\left(\vec{X}^N(k)\right)\right)\right\}. \end{aligned}$$

From the steady-state assumption, we have $\mathbb{E}\{g(W_i^N(k+1))\} = \mathbb{E}\{g(W_i^N(k))\} = \mathbb{E}\{g(W_i^N)\}$, and we can rewrite the above equation as

$$\begin{aligned} \mathbb{E}\left\{g\left(W_i^N\right)\right\} &= \mathbb{E}\left\{g\left(\left(W_i^N + \alpha(W_i^N)\right) \wedge w_{max}\right) f_i\left(\vec{X}^N\right)\right\} \\ &\quad + \mathbb{E}\left\{g\left((1-\beta)W_i^N \vee 1\right) \cdot \left(1 - f_i\left(\vec{X}^N\right)\right)\right\}, \end{aligned} \tag{25}$$

where the expectation is taken with respect to the stationary distribution of $\vec{X}^N(k)$.

First, we choose $g(x) = x$ in (25). Since $(1-\beta)W_i^N \vee 1 \leq (1-\beta)W_i^N + \beta$ (from $W_i^N \geq 1$) and $0 \leq f_i(\cdot) \leq 1$, we have from (25)

$$\mathbb{E}\left\{W_i^N\right\} \leq \mathbb{E}\left\{\left(W_i^N + \alpha(W_i^N)\right) f_i\left(\vec{X}^N\right)\right\} + \mathbb{E}\left\{((1-\beta)W_i^N + \beta)(1 - f_i\left(\vec{X}^N\right))\right\}.$$

This gives

[9] We drop $\lfloor \cdot \rfloor$ here for notational simplicity and this does not affect our results in the paper.

$$\beta\mathbb{E}\left\{W_i^N - 1\right\} \leq \mathbb{E}\left\{\left(\beta(W_i^N - 1) + \alpha(W_i^N)\right) f_i\left(\vec{X}^N\right)\right\}$$
$$\leq \beta\left(w_{max} + \frac{\alpha(w_{max})}{\beta} - 1\right)\mathbb{E}\left\{f_i\left(\vec{X}^N\right)\right\}.$$

Thus, (9) follows by noting that, from (24),

$$\mathbb{E}\left\{f_i\left(\vec{X}^N\right)\right\} = \mathbb{E}\left\{\left[1 - p^N(Q^N)\right]^{W_i^N}\right\}. \tag{26}$$

To obtain the upper bound, let $g(\cdot) = (\cdot)^2$ in (25). We then choose constants a and b such that

$$a + b = \alpha^2(1) + 2\alpha(1) \qquad \text{and} \qquad aw_{max} + b = 0. \tag{27}$$

This gives $a = -(\alpha^2(1) + 2\alpha(1))/(w_{max} - 1) < 0$ and $b = w_{max}(\alpha^2(1) + \alpha)/(w_{max} - 1) > 0$. Then, for any w with $1 \leq w \leq w_{max}$, we have

$$(w + \alpha(w))^2 \geq w^2 + aw + b. \tag{28}$$

To see this, note that (28) is equivalent to $(2\alpha(w) - a)w + \alpha^2(w) - b \geq 0$. Since $a < 0$, the LHS of this relation is increasing in w, and we have equality at $w = 1$ from (27), thus (28) holds. Similarly, the function $w^2 + aw + b$ is convex in w, so its maximum value over $1 \leq w \leq w_{max}$ occurs only at the boundaries. At $w = w_{max}$, we have $w^2 + aw + b = (w_{max})^2$ from (27), and at $w = 1$, we have $w^2 + aw + b = 1 + a + b = (1 + \alpha(1))^2 < (w_{max})^2$ from (27) and by our choice of w_{max} ($w_{max} > M > C' + \alpha(C') > 1 + \alpha(1)$, see (7)). Thus, for $1 \leq w \leq w_{max}$, we also obtain

$$(w_{max})^2 \geq w^2 + aw + b. \tag{29}$$

Combining (28) and (29) gives

$$\left((w + \alpha(w)) \wedge w_{max}\right)^2 \geq w^2 + aw + b, \ \forall w \in [1, w_{max}].$$

Thus, from (25) with $g(\cdot) = (\cdot)^2$, we get

$$\mathbb{E}\left\{\left(W_i^N\right)^2\right\} \geq \mathbb{E}\left\{\left(\left(W_i^N\right)^2 + aW_i^N + b\right) f_i\left(\vec{X}^N\right)\right\}$$
$$+\mathbb{E}\left\{(1 - \beta)^2 (W_i^N)^2\left(1 - f_i\left(\vec{X}^N\right)\right)\right\}.$$

After rearranging terms, we obtain

$$(2\beta - \beta^2)\mathbb{E}\left\{\left(W_i^N\right)^2\left(1 - f_i\left(\vec{X}^N\right)\right)\right\} \geq \mathbb{E}\left\{\left(aW_i^N + b\right) f_i\left(\vec{X}^N\right)\right\}. \tag{30}$$

Lastly, we define a function $h \in [0, 1]$ to be

$$h(\vec{x}^N) := \left[1 - p^N(q^N)\right]^{w_{max}}$$

Since $1 \leq W_i^N \leq w_{max}$, we have $h\left(\vec{x}^N\right) \leq f_i\left(\vec{x}^N\right)$ for any i (see (24)). Observe

$$
\begin{aligned}
(2\beta-\beta^2)w_{max}^2\mathbb{E}\left\{1-h\left(\vec{X}^N\right)\right\} &\geq (2\beta-\beta^2)w_{max}^2\mathbb{E}\left\{1-f_i\left(\vec{X}^N\right)\right\} \\
&\geq (2\beta-\beta^2)\mathbb{E}\left\{\left(W_i^N\right)^2\left(1-f_i\left(\vec{X}^N\right)\right)\right\} \\
&\geq \mathbb{E}\left\{\left(aW_i^N+b\right)f_i\left(\vec{X}^N\right)\right\} \\
&\geq \mathbb{E}\left\{\left(aW_i^N+b\right)h\left(\vec{X}^N\right)\right\}, \qquad (31)
\end{aligned}
$$

where the third inequality follows from (30) and the last one follows from $aW_i^N+b \geq 0$ for $1 \leq W_i^N \leq w_{max}$. Now, summing (31) over i and dividing by N gives

$$
(2\beta-\beta^2)w_{max}^2\mathbb{E}\left\{1-h\left(\vec{X}^N\right)\right\} \geq \mathbb{E}\left\{\left(a\frac{\sum_{i=1}^N W_i^N}{N}+b\right)h\left(\vec{X}^N\right)\right\}.
$$

Since $\sum_{i=1}^N W_i^N/N < M < w_{max}$ from Lemma 1, and by our choice of a, b in (27), we see that

$$
\begin{aligned}
(2\beta-\beta^2)w_{max}^2\mathbb{E}\left\{1-h\left(\vec{X}^N\right)\right\} &\geq \mathbb{E}\left\{\left(a\frac{\sum_{i=1}^N W_i^N}{N}+b\right)h\left(\vec{X}^N\right)\right\} \\
&\geq (aM+b)\,\mathbb{E}\left\{h\left(\vec{X}^N\right)\right\} \\
&= K\mathbb{E}\left\{h\left(\vec{X}^N\right)\right\}, \qquad (32)
\end{aligned}
$$

where $K := aM+b > 0$ from $M < w_{max}$ and (27). Thus, from (32), we obtain

$$
\mathbb{E}\left\{h\left(\vec{X}^N\right)\right\} \leq \frac{(2\beta-\beta^2)w_{max}^2}{(2\beta-\beta^2)w_{max}+K} := B < 1
$$

for all N. This proves (10) and we are done. ■