The Effects of Active Queue Management and Explicit Congestion Notification on Web Performance

Long Le Jay Aikat Kevin Jeffay F. Donelson Smith
Department of Computer Science
University of North Carolina at Chapel Hill
http://www.cs.unc.edu/Research/dirt

ABSTRACT

We present an empirical study of the effects of active queue management (AQM) and explicit congestion notification (ECN) on the distribution of response times experienced by a population of users browsing the Web. Three prominent AQM schemes are considered: the Proportional Integral (PI) controller, the Random Exponential Marking (REM) controller, and Adaptive Random Early Detection (ARED). The effects of these AQM schemes were studied with and without ECN. Our primary measure of performance is the end-to-end response time for HTTP request-response exchanges. For this measure, our major results are:

- If ECN is not supported, ARED operating in byte-mode was the best performing AQM scheme, providing better response time performance than drop-tail FIFO queuing at offered loads above 90% of link capacity. However, ARED operating in packet-mode (with or without ECN) was the worst performing scheme, performing worse than drop-tail FIFO queuing.
- ECN support is beneficial to PI and REM. With ECN, PI and REM were the best performing overall schemes, providing significant response time improvement over ARED operating in byte-mode. In the case of REM, the benefit of ECN was dramatic. Without ECN, response time performance with REM was worse than drop-tail FIFO queuing at all loads considered.
- ECN was not beneficial to ARED. Under current ECN implementation guidelines, ECN had no effect on ARED performance. However, ARED performance with ECN improved significantly after reversing a guideline that was intended to police unresponsive flows. Nonetheless, overall, the best ARED performance was achieved without ECN.
- Whether or not the improvement in response times with AQM is significant (when compared to drop-tail FIFO), depends heavily on the range of round-trip times (RTTs) experienced by flows. As the variation in flows' RTT increases, the impact of AQM and ECN on response-time performance is reduced.

We conclude that AQM can improve application and network performance for Web or Web-like workloads. In particular, it appears likely that with AQM and ECN, provider links may be operated at near saturation levels without significant degradation in user-perceived performance.

1 INTRODUCTION AND MOTIVATION

The random early detection (RED) algorithm, first described over ten years ago [8], inspired a new focus for congestion control research on the area of *active queue management* (AQM). AQM is a router-based form of congestion control wherein routers notify end-systems of incipient congestion. The common goal of all AQM designs is to keep the average queue size in routers small. This has a number of desirable effects including (1) providing queue space to absorb bursts of packet arrivals, (2) avoiding lock-out and bias effects from a few flows dominating queue space, and (3) providing lower delays for interactive applications such as Web browsing [4].

All AQM designs function by detecting impending queue buildup and notifying sources before the queue in a router overflows. The various designs proposed for AQM differ in the mechanisms used to detect congestion and in the type of control mechanisms used to achieve a stable operating point for the queue size. Another dimension that has a significant impact on performance is how the congestion signal is delivered to the sender. In today's Internet where the dominant transport protocol is TCP (which reacts to segment loss as an indicator of congestion), the signal is usually delivered implicitly by dropping packets at the router when the AQM algorithm detects queue buildup. An IETF proposed standard adds an explicit signalling mechanism, called *explicit congestion notification* (ECN) [15], by allocating bits in the IP and TCP headers for this purpose. With ECN a router can signal congestion to an end-system by "marking" a packet (setting a bit in the header).

In this work we report the results of an empirical evaluation of three prominent examples of AQM designs. These are the Proportional Integral (PI) controller [10], the Random Exponential Marking (REM) controller [3] and a contemporary redesign of the classic RED controller, Adaptive RED [7] (here called ARED). While these designs differ in many respects, each is an attempt to realize a control mechanism that achieves a stable operating point for the size of the router queue. Thus a user of each of these mechanisms can determine a desired operating point for the control mechanism by simply specifying a desired mean queue size. Choosing the desired queue size may represent a tradeoff between link utilization and queuing delay — a short queue reduces latency at the router but setting the target queue size too small may reduce link utilization by limiting the router's ability to buffer short bursts of arriving packets.

Our goal in this study was first and foremost to compare the performance of control theoretic AQM algorithms (PI and REM) with the more traditional randomized dropping found in RED. For performance metrics we chose both user-

To appear: *IEEE/ACM Transactions on Networking*, 2005.

centric measures of performance such as response times for the request-response exchanges that comprise Web browsing, as well as more traditional metrics such as achievable link utilization and loss rates. The distribution of response times that would be experienced by a population of Web users is used to assess the user-perceived performance of the AQM schemes and is our primary metric for assessing overall AQM performance. Of particular interest was the implication of ECN on performance. ECN requires changes to end-system protocol stacks and hence it is important to quantify the performance gain to be had at the expense of a more complex protocol stack and migration issues for the end-system.

Our experimental platform was a laboratory testbed consisting of a large collection of computers arranged to emulate a peering point between two ISPs operated at 100 Mbps (see Figure 1). We emulated the Web browsing behaviour of tens of thousands of users whose traffic transits the link connecting the ISPs and investigated the performance of each AQM scheme in the border-routers connecting the ISPs. Each scheme was investigated both with and without ECN support across a variety of AQM parameter settings that represented a range of target router-queue lengths. For each target queue length we varied the offered load on the physical link connecting the ISPs to determine how (or if) AQM performance was affected by load.

Our primary results were that AQM and ECN can provide significant benefit to application and network performance, however, (1) this benefit occurs only at very high levels of network load and (2) the degree of benefit provided by AQM is influenced by the round-trip times experienced by HTTP connections.

Concerning network load, it was previously shown that AQM (with or without ECN) only improved response time performance (compared to drop-tail FIFO queuing) at offered loads above 80% of link capacity [11]. For offered loads greater than or equal to 90% of link capacity, the control theoretic designs PI and REM give the best performance but only when deployed with ECN-capable end-systems and routers. However, in these environments the improvement in performance can be substantial. Response times for HTTP request-response exchanges approximate those achieved on an uncongested network at the cost of slightly lower achievable link utilization (compared to drop-tail FIFO queue management). If ECN support is not present in the network, then ARED operated in byte-mode, gives the best performance. Moreover, for offered loads of 90% of link capacity, ARED byte-mode performance without ECN approximated that of PI and REM with ECN. At higher loads (98% of link capacity), ARED improved response time performance compared to drop-tail FIFO, however, the improvement was small compared to the more substantial improvements realized by PI and REM with ECN.

An additional aspect of our study was the effect of round-trip times (RTTs) on response-time performance. Response time is a function of round-trip time which in turn is a function of transmission, propagation, and queuing delays. AQM affects only the queuing delay component of RTT and hence the impact of AQM on response time depends on the magnitude of the queuing delay's contribution to total RTT. Experiments were run with two distributions of RTTs: a uniform distribution of RTTs (used in [5, 11]), and a more variable, empirical distribution of RTTs (from data reported in [1]). The results for AQM experiments performed with uniformly distributed RTTs are those recited above. When the empirical RTT distribution was used, the same relative conclusions hold, however, the magnitude of the performance improvements achieved with AQM and ECN were less dramatic.

In total, our results suggest that with the appropriate choice and configuration of AQM, providers may be able to operate links dominated by Web traffic at load levels as high as 90% of link capacity without significant degradation in application or network performance. Thus unlike a similar earlier study [5] which was negative on the use of a specific form of AQM (RED), we view the present results as a significant indicator that the stated goals of AQM can be realized in practice.

Our results also demonstrate some shortcomings in the design of AQM algorithms. Specifically we show that ARED performance is critically a function of whether the router's queue length is measured in units of bytes or packets. We also show that the current guidelines for forwarding ECN-marked packets are counter-productive. When ARED measures queue length in packets it consistently resulted in response time performance that was worse than that achieved with simple drop-tail FIFO queuing. Moreover, unlike PI and REM whose performance was significantly improved by the addition of ECN, ARED performance in "packet-mode" was unaffected by ECN. However, by reversing an implementation guideline for ECN, specifically by allowing ECN-marked packets to be forwarded and not dropped when the average queue length is in the "gentle region," ARED performance with ECN was substantially improved (resulting in better performance than drop-tail). However, overall, the best ARED performance was always obtained when queue length was measured in bytes rather than packets.

While the results of this study are intriguing, the study was nonetheless limited. The design space of AQM schemes is large with each algorithm typically characterized by a number of independent parameters. We limited our consideration of AQM algorithms to a comparison between two classes of algorithms: those based on control theoretic principles and those based on the original randomized dropping paradigm of RED. Moreover, we studied a link carrying only Web-like traffic. More realistic mixes of HTTP and

other TCP traffic as well as traffic from UDP-based applications need to be examined. However, unfortunately, at present, good source-level models of general TCP and UDP traffic suitable for synthetic traffic generation do not exist.

The following section reviews the salient design principles of current AQM schemes and reviews the major algorithms that have been proposed. Section 3 presents our experimental methodology and discusses the generation of synthetic Web traffic. Section 4 presents our results for AQM with packet drops and Section 5 presents our results for AQM with ECN. Section 6 presents additional experiments that show the sensitivity of performance results to round-trip times. The results are discussed in Section 7. We conclude in Section 8 with a summary of our major results.

2 BACKGROUND AND RELATED WORK

The original RED design uses a weighted-average queue size as a measure of congestion. When this weighted average is smaller than a minimum threshold (min_{th}), no packets are marked or dropped. When the average queue length is between the minimum threshold and the maximum threshold (max_{th}), the probability of marking or dropping packets varies linearly between 0 and a maximum drop probability (max_p, typically 0.10). If the average queue length exceeds max_{th}, all packets are marked or dropped. (The actual size of the queue must be greater than max_{th} to absorb transient bursts of packet arrivals.) A modification to the original design introduced a "gentle mode" in which the mark or drop probability increases linearly between max_p and 1 as the average queue length varies between max_{th} and $2 \times max_{th}$. This fixed a problem in the original RED design caused by the non-linearity in drop probability (increasing from max_p to 1.0 immediately when max_{th} is reached).

An alleged weakness of RED is that it does not take into consideration the number of flows sharing a bottleneck link [6]. Given TCP's congestion control mechanism, a packet mark or drop reduces the offered load by a factor of $(1 - 0.5n^{-1})$ where n is the number of flows sharing the bottleneck link. Thus, RED is not effective in controlling the queue length when n is large. On the other hand, RED can be too aggressive and can cause under-utilization of the link when n is small. Feng *et al.* concluded that RED needs to be tuned for the dynamic characteristics of the aggregate traffic on a given link [6]. They proposed a self-configuring algorithm for RED by adjusting max_p every time the average queue length falls out of the target range between min_{th} and max_{th}. When the average queue length is smaller than min_{th}, max_p is decreased multiplicatively to reduce RED's aggressiveness in marking or dropping packets; when the queue length is larger than max_{th}, max_p is increased multiplicatively. Floyd *et al.* improved upon this original adaptive RED proposal by replacing the MIMD (multiplicative increase multiplicative decrease) approach with an AIMD (additive increase multiplicative decrease) approach [7]. They also provided guidelines for choosing min_{th}, max_{th}, and the weight for computing a target average queue length. The RED version that we implemented and studied in our work (referred to herein as "ARED") includes both the adaptive and gentle refinements to the original design. It is based on the description in [7].

Misra *et al.* applied control theory to develop a model for TCP and AQM dynamics and used this model to analyze RED [14]. They asserted two limitations in the original RED design: (1) RED is either unstable or has slow responses to changes in network traffic, and (2) RED's use of a weighted-average queue length to detect congestion and its use of loss probability as a feedback signal to the senders were flawed. Because of this, in overload situations, flows can suffer both high delay and a high packet loss rate. Hollot *et al.* simplified the TCP/AQM model to a linear system and designed a Proportional Integrator (PI) controller that regulates the queue length to a target value called the "queue reference," q_{ref} [10]. The PI controller uses instantaneous samples of the queue length taken at a constant sampling frequency as its input. The drop probability is computed as

$$p(kT) = a \times (q(kT) - q_{ref}) - b \times (q((k-1)T) - q_{ref}) + p((k-1)T)$$

where $p(kT)$ is the drop probability at the k^{th} sampling interval, $q(kT)$ is the queue length sample, and T is the sampling period. A close examination of this equation shows that the drop probability increases in sampling intervals when the queue length is higher than its target value. Furthermore, the drop probability also increases if the queue has grown since the last sample (reflecting an increase in network traffic). Conversely, the drop probability in a PI controller is reduced when the queue length is lower than its target value or the queue length has decreased since its last sample. The sampling interval and the coefficients in the equation depend on the link capacity, the expected number of active flows using the link, and the maximum RTT among those flows.

Athuraliya *et al.* proposed the Random Exponential Marking (REM) AQM scheme [3]. REM periodically updates a congestion measure called "price" that reflects any mismatch between packet arrival and departure rates at the link (*i.e.*, the difference between the demand and the service rate) and any queue size mismatch (*i.e.*, the difference between the actual queue length and its target value). The price measure p at time t is computed by:

$$p(t) = \max(0, p(t-1) + \gamma \times (\alpha \times (q(t) - q_{ref}) + x(t) - c))$$

where c is the link capacity (in packet departures per unit time), $q(t)$ is the queue length, and $x(t)$ is the packet arrival rate, all determined at time t. As with ARED and PI, the control target is only expressed by the queue size.

The mark/drop probability in REM at time t is $1 - \phi^{-p(t)}$, where $\phi > 1$ is a constant. In overload situations, the congestion price increases due to the rate mismatch and the queue mismatch. Thus, more packets are dropped or marked to signal TCP senders to reduce their transmission rate. When congestion abates, the congestion price is reduced because the mismatches are now negative. This causes REM to drop or mark fewer packets and allows the senders to potentially increase their transmission rate. It is easy to see that a positive rate mismatch over a time interval will cause the queue size to increase. Conversely, a negative rate mismatch over a time interval will cause the queue to drain. Thus, REM is similar to PI because the rate mismatch can be detected by comparing the instantaneous queue length with its previous sampled value. Furthermore, when the drop or mark probability is small, the exponential function can be approximated by a linear function [2].

An additional aspect of each AQM scheme is whether the algorithm measures the length of the router's queue (and specifies target queue length, thresholds, *etc.*) in units of bytes or packets. When measuring queue length in bytes, the AQM algorithms bias the initial drop probability p by the size of the arriving packet according to the following formula:

$$p_b = p \frac{arriving\ packet\ size}{average\ packet\ size}$$

Thus all other factors being equal, AQM algorithms operated in "byte-mode" assign lower drop probabilities to small packets (*e.g.*, SYNs, FINs, pure ACKs, *etc.*) than to large packets. For PI and REM it is recommended that queue length be measured in bytes while for ARED the recommendation is to measure queue length in packets. However, to better compare ARED to PI and REM we will evaluate ARED performance in both byte- and packet-mode.

3 EXPERIMENTAL METHODOLOGY

For our experiments we constructed a laboratory network that emulates the interconnection between two Internet service provider (ISP) networks. Specifically, we emulate one peering link that carries Web traffic between sources and destinations on both sides of the peering link and where the traffic carried between the two ISP networks is evenly balanced in both directions.

The laboratory network used to emulate this configuration is shown in Figure 1. All systems shown in this figure are Intel-based machines running FreeBSD 4.5. At each edge of this network are a set of machines that run instances of a Web request generator (described below) each of which emulates the browsing behavior of thousands of human users. Also at each edge of the network is another set of machines that run instances of a Web response generator (also described below) that creates the traffic flowing in response

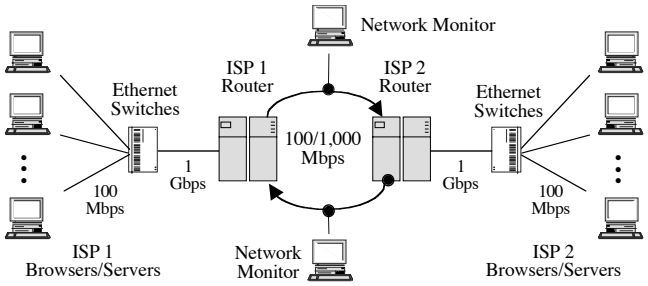

Figure 1: Experimental network setup.

to the browsing requests. A total of 44 traffic-generating machines are in the testbed. In the remainder of this paper we refer to the machines running the Web request generator simply as the "browser machines" (or "browsers") and the machines running the Web response generator as the "server machines" (or "servers").

At the core of this network are two router machines running the ALTQ extensions to FreeBSD. ALTQ extends IP-output queuing at the network interfaces to include alternative queue-management disciplines [13]. The ALTQ infrastructure was used to implement PI, REM, and ARED. The routers are interconnected via three point-to-point Ethernet segments (two 100 Mbps Fast Ethernet segments and one fiber Gigabit Ethernet segment) as illustrated in Figure 1. The gigabit interconnection is used to perform experiments in an uncongested environment while the 100 Mbps connections are used to perform experiments in a congested environment. When conducting experiments on the uncongested network, static routes are configured on the routers so that all traffic uses the full-duplex Gigabit Ethernet segment. When we need to create a bottleneck between the two routers, the static routes are reconfigured so that all traffic flowing in one direction uses one 100 Mbps Ethernet segment and all traffic flowing in the opposite direction uses the other 100 Mbps Ethernet segment. These configurations allow us to emulate the full-duplex behavior of the typical wide-area network link.

Another important factor in emulating this network is the effect of end-to-end latency. We use a locally modified version of the *dummynet* [12] component of FreeBSD to configure out-bound packet delays on browser machines to emulate different round-trip times on *each* TCP connection (giving *per-flow* delays). This is accomplished by extending the *dummynet* mechanisms for regulating per-flow bandwidth to include a mode for adding a randomly chosen minimum delay to all packets from each flow. The same minimum delay is applied to all packets in a given flow (identified by IP addressing 5-tuple). The minimum delay in milliseconds assigned to each flow is randomly sampled from an RTT distribution that is provided for each experiment. Two RTT distributions are used. The first is a discrete uniform distribution. For the experiments reported in Sections 4 and 5, a uniform distribution of minimum RTTs on

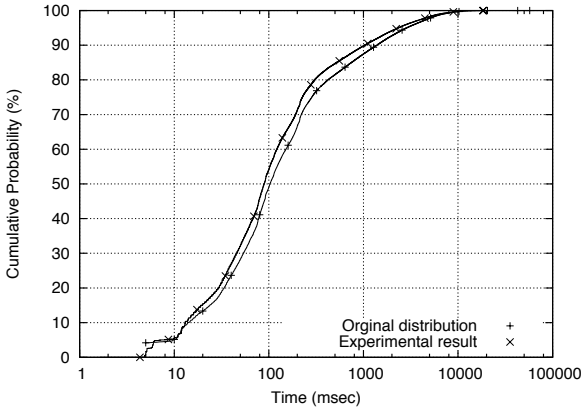

Figure 2: CDF of the generalized minimum RTT distribution, measured versus experimentally reproduced values.

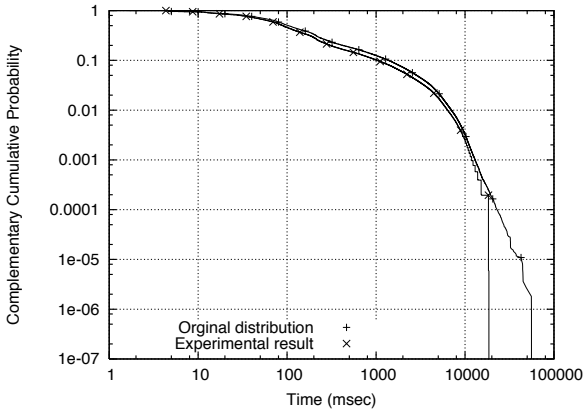

Figure 3: CCDF of generalized minimum RTT distribution measured versus experimentally reproduced values.

the range [10, 150] (a mean of 80 milliseconds) was used. The minimum and maximum values for this distribution were chosen using the method described in [5] to approximate a typical range of Internet round-trip times within the continental U.S. and the uniform distribution ensures a large variance in the values selected over this range.

The second minimum RTT distribution is a more general distribution that comes from a recent measurement study of the RTTs experienced by the TCP connections transiting a university campus-to-Internet gateway [1]. Figures 2-3 show the cumulative distribution function (CDF) and complementary CDF (CCDF) of the general RTT distribution. (Note that the uniform distribution of minimum RTTs used in Sections 4-5 is a good approximation for the body of the more general distribution (*e.g.*, the 5th to 80th percentile).) Figure 2 shows both the general distribution used as an input to the traffic generation program and the range of minimum RTTs actually achieved in our experiments. The general RTT distribution is used for the experiments reported in Section 6.

In all experiments the actual round-trip times experienced by the TCP senders (servers) will be the combination of the flow's minimum RTT (*dummynet* delay) plus the queuing delays at the routers. (End systems are configured to ensure no resource constraints were present, hence delays there are insignificant.) A TCP window size of 16K bytes was used on all the end systems because widely used OS platforms, *e.g.*, most versions of Windows, typically have default windows this small or smaller.

3.1 Web-Like Traffic Generation

The traffic that drives our experiments is based on a recent large-scale analysis of Web traffic [16]. The resulting model is an application-level description of the critical elements that characterize how HTTP/1.0 and HTTP/1.1 protocols are used in practice. It is based on empirical data and is intended for use in generating synthetic Web workloads. An important property of the model is that it reflects the use of persistent HTTP connections as implemented in many contemporary browsers and servers. Further, the analysis presented in [16] distinguishes between Web objects that are "top-level" (typically an HTML file) and those that are embedded objects (*e.g.*, an image file). At the time these data were gathered, approximately 15% of all TCP connections carrying HTTP protocols were effectively persistent (were used to request two or more objects) but more than 50% of all objects (40% of bytes) were transferred over these persistent connections.

The model is expressed as empirical distributions describing the elements necessary to generate synthetic HTTP workloads. The elements of the model that have the most pronounced effects on generated traffic are summarized in Table 1. Most of the behavioral elements of Web browsing are emulated in the client-side request-generating program (the "browser"). Its primary parameter is the number of emulated browsing users (typically several hundred to a few thousand). For each user to be emulated, the program implements a simple state machine that represents the user's state as either "thinking" or requesting a Web page. If requesting a Web page, a request is made to the server-side portion of the program (executing on a remote machine) for the pri-

Table 1: Elements of the HTTP traffic model.

Element	Description
Request size	HTTP request length in bytes
Response size	HTTP reply length in bytes (top-level & embedded)
Page size	Number of embedded (file) references per page
Think time	Time between retrieval of two successive pages
Persistent connection use	Number of requests per persistent connection
Servers per page	Number of unique servers used for all objects in a page
Consecutive page retrievals	Number of consecutive top-level pages requested from a given server

mary page. Then requests for each embedded reference are sent to some number of servers (the number of servers and number of embedded references are drawn as random samples from the appropriate distributions). The browser also determines the appropriate usage of persistent and non-persistent connections; 15% of all new connections are randomly selected to be persistent. Another random selection from the distribution of requests per persistent connection is used to determine how many requests will use each persistent connection. One other parameter of the program is the number of parallel TCP connections allowed on behalf of each browsing user to make embedded requests within a page. This parameter is used to mimic the parallel connections used in Netscape (typically 4) and Internet Explorer (typically 2).

For each request, a message of random size sampled from the request size distribution is sent over the network to an instance of the server program. This message specifies the number of bytes the server is to return as a response (a random sample from the distribution of response sizes depending on whether it is a top-level or embedded request). The server transmits this number of bytes back to the browser. For each request/response exchange, the browser logs its response time. Response time is defined as the elapsed time between either the time of the socket *connect()* operation (for a non-persistent connection) or the initial request (on a persistent connection) or the socket *write()* operation (for subsequent requests on a persistent connection) and the time the last byte of the response is returned. Note that this response time is for each object of a page, not the total time to load all objects of a page.

When all the request/response exchanges for a page have been completed, the emulated browsing user enters the thinking state and makes no more requests for a random period of time sampled from the think-time distribution. The number of page requests the user makes in succession to a given server machine is sampled from the distribution of consecutive page requests. When that number of page requests has been completed, the next server to handle the next top-level request is selected randomly and uniformly from the set of active servers. The number of emulated users is constant throughout the execution of each experiment.

3.2 Experiment Calibrations

Offered load for our experiments is defined as the network traffic resulting from emulating the browsing behavior of a fixed-size population of Web users. It is expressed as the long-term average throughput (bits/second) on an uncongested link that would be generated by that user population. There are three critical elements of our experimental procedures that had to be calibrated before performing experiments:

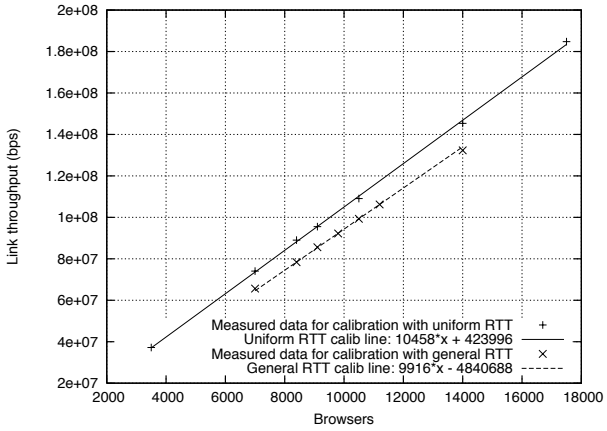

Figure 4: Link throughput *v*. number of emulated browsing users for the uniform and general minimum round-trip time distribution.

1. Ensuring that no element on the end-to-end path represented a primary bottleneck other than the links connecting the two routers when they are limited to 100 Mbps,

2. The offered load on the network can be predictably controlled using the number of emulated users as a parameter to the traffic generators, and

3. Ensuring that the resulting packet arrival time-series (*e.g.*, packet counts per millisecond) is long-range dependent as expected because the distribution of response sizes is a heavy-tailed distribution [16].

To perform these calibrations, we first configured the network connecting the routers to eliminate congestion by running at 1 Gbps. All calibration experiments were run with drop-tail queues with length equal to 2,400 packets (the reasons for this choice are discussed in Section 4). We ran one instance of the browser program on each of the browser machines and one instance of the server program on all the server machines. Each browser was configured to emulate the same number of active users and the total active users varied from 7,000 to 35,000.

Two sets of calibration experiments were performed: one with the uniform minimum RTT distribution, and one with the more general minimum RTT distribution. Figure 4 shows the aggregate traffic on one direction of the 1 Gbps link as a function of the number of emulated users for both RTT distributions. The load in the opposite direction was measured to be essentially the same and is not plotted in this figure. The offered load expressed as link throughput is a linear function of the number of emulated users indicating there are no fundamental resource limitations in the system and generated loads can easily exceed the capacity of a 100 Mbps link.

For each of our minimum RTT distributions, these data can be used to determine the number of emulated users that would generate a specific offered load in the absence of a bottleneck link. This capability is used in subsequent ex-

periments to control the offered loads on the network. For example, if we want to generate an offered load equal to the capacity of a 100 Mbps link, we would need to emulate a user population in ISP1 and a user population in ISP2 (see Figure 1), such that the aggregate requests flowing from the population of emulated users in ISP1 to servers in ISP2, plus the aggregate responses flowing from servers in ISP1 to the population of emulated users in ISP2, equals 100 Mbps on average. The analogous situation would also have to hold for the traffic flowing from ISP2 to ISP1. To generate an offered load of 100 Mbps, Figure 4 is used to determine that with uniformly distributed minimum RTTs, approximately 9,520 users must be emulated on each side of the 1 Gbps link (*i.e.*, 9,520 users in ISP1 and 9,520 users in ISP2 for a total of 19,040 emulated users). Note that as expected, more users must be emulated to realize a given target load with the more general minimum RTT distribution. To generate an offered load of 100 Mbps with the more general RTT distribution, approximately 10,570 users must be emulated in ISP1 and ISP2. Note further that for offered loads approaching saturation of the 100 Mbps link, the actual link utilization will, in general, be less than the intended offered load. This is because as response times become longer, users have to wait longer before they can generate new requests and hence generate fewer requests per unit time.

A motivation for using Web-like traffic in our experiments was the assumption that properly generated traffic would exhibit demands on the laboratory network consistent with those found in empirical studies of real networks, specifically, a long-range dependent (LRD) packet arrival process. The empirical data used to generate our Web traffic showed heavy-tailed distributions for both user "think" times and response sizes [16]. For example, while the median response size generated in experiments is approximately 1,000 bytes, responses as large as 10^9 bytes are also generated. We analytically verified that the number of packets and bytes arriving to the router interfaces on the 1 Gbps link indeed constituted an LRD arrival process [11]. Thus, although our study considers only web traffic, the dynamics of the arrival process seen at router queues is indicative of arrival processes observed on real networks.

3.3 Experimental Procedures

Each experiment was run using a fixed population of emulated users chosen, as described above, to place a nominal offered load on an unconstrained network. Each browser program emulated an equal number of users. The offered loads used in experiments were chosen to represent user populations that could consume 90% or 98% of the capacity of the 100 Mbps link connecting the two router machines (*i.e.*, consume 90 or 98 Mbps, respectively). In [11] we demonstrated that at offered loads up to 80% of link capacity, the distribution of response times achieved with AQM was virtually identical to that achieved with conventional drop-tail FIFO queuing. Because these distributions were also quite similar to the response-time distribution on the uncongested network, we concluded that AQM offered no advantage over drop-tail at or below 80% load. For this reason we begin our study here at 90% load. ([11] also reports the results of additional experiments, identical to those performed here, for offered loads of 105% of link capacity.) It is important to emphasize again that terms like "98% load" are used as a shorthand notation for "a population of Web users that would generate a long-term average load of 98 Mbps on a 1 Gbps link."

Each experiment was run for 120 minutes to ensure very large samples (over 10,000,000 request/response exchanges in each experiment) but data were collected only during a 90-minute interval to eliminate startup effects at the beginning and termination synchronization anomalies at the end. Each experiment for a given AQM scheme was repeated three times with a different set of random number seeds for each repetition. To facilitate comparisons among different AQM schemes, experiments for different schemes were run with the same sets of initial seeds for each random number generator.

The key indicator of performance we use in reporting our results is the end-to-end response time for each HTTP request/response exchange. We report these as plots of the cumulative distributions of response times up to 2 seconds. In these plots we show the combined results from three independent repetitions for each experiment. We also show the results obtained on an uncongested 1 Gbps link to provide a baseline for comparison. On all plots, the "uncongested network" line represents the best possible response time distribution. We also report the fraction of IP datagrams dropped at the link queues, the link utilization on the bottleneck link, and the number of request/response exchanges completed in the experiment. The values we report for these measures are means over the three repetitions of an experiment.

4 AQM EXPERIMENTS WITH PACKET DROPS

For PI and REM, target queue lengths of 24 and 240 packets were evaluated. These values were chosen to represent two operating points: one that potentially yields minimum latency (24) and one that potentially provides high link utilization (240). The values used for the coefficients in the control equations above are those recommended in [2, 10] and confirmed by the algorithm designers. For ARED the same two target queue lengths were evaluated. The calculations for all the ARED parameter settings follow the guidelines given in [7] for achieving the desired target delay (queue size). For all three algorithms we set the maximum queue size to a number of packets sufficient to ensure tail drops do

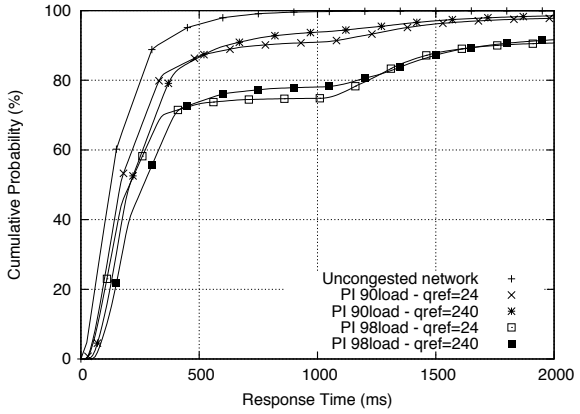

Figure 5: Response time distribution for PI with packet drops.

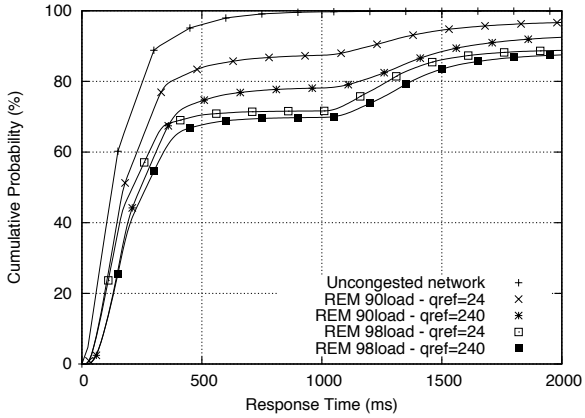

Figure 6: Response time distribution for REM with packet drops.

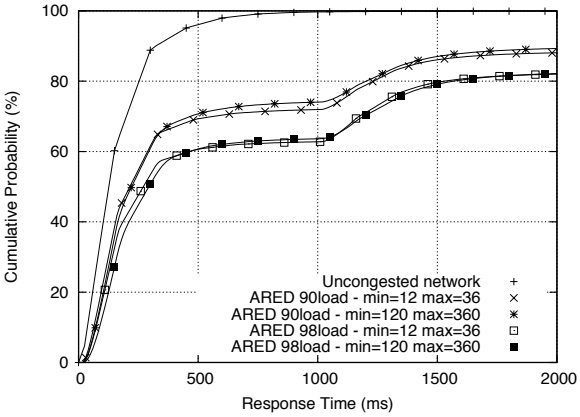

Figure 7: Response time distribution for ARED in packet-mode with drops.

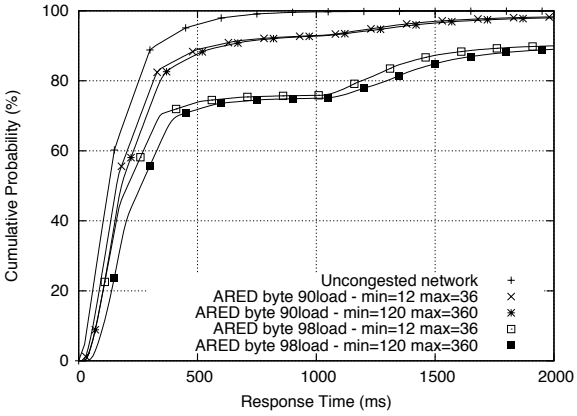

Figure 8: Response time distribution for ARED in byte-mode with drops.

not occur. All experiments in this section use the uniform minimum RTT distribution.

4.1 Results for PI with Packet Drops

Figure 5 gives the results for PI at target queue lengths of 24 and 240 packets, and offered loads of 90% and 98%. At 90% load, a target queue size of 24 results in lower response times for all but the largest 10% of request/response exchanges, those requiring more than approximately 500 milliseconds to complete. For these largest exchanges, the longer target size of 240 is slightly better. At 98% load, the tradeoff between optimizing the response time of "shorter" exchanges, those requiring less than approximately 400 milliseconds to complete in this case, versus "longer" exchanges, those requiring more than 400 milliseconds to complete, is more clear. At 98% load, a target queue size of 24 packets results in lower response times for only the shortest 70% of request/response exchanges. At both loads, both target queue lengths result in equivalent performance for the very largest exchanges (those requiring more than 2 seconds to complete). Overall, we conclude PI provides the best response time performance when used with a target queue reference of 24 packets. Table 2 summarizes the loss rates, link utilization, and number of completed requests for the PI experiments.

Note that in Figure 5 we see a feature that is found in all our results at high loads where a significant number of packets are dropped (see Table 2). The flat area in the curves between approximately 500 milliseconds and 1 second shows the impact of RTO granularity in TCP — request/response exchanges that experience a timeout take at least 1 second to complete on FreeBSD.

4.2 Results for REM with Packet Drops

Figure 6 gives the results for REM at target queue lengths of 24 and 240 packets, and offered loads of 90% and 98%. At 90% load, a queue reference of 24 performs significantly better than a target queue of 240. At 98% load, a queue reference of 24 continues to perform slightly better than 240. Overall, like PI, REM provides the best response time per-

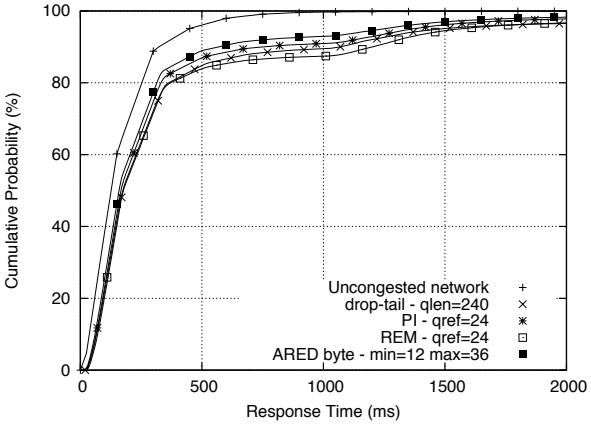

Figure 9: Comparison of all schemes at 90% load.

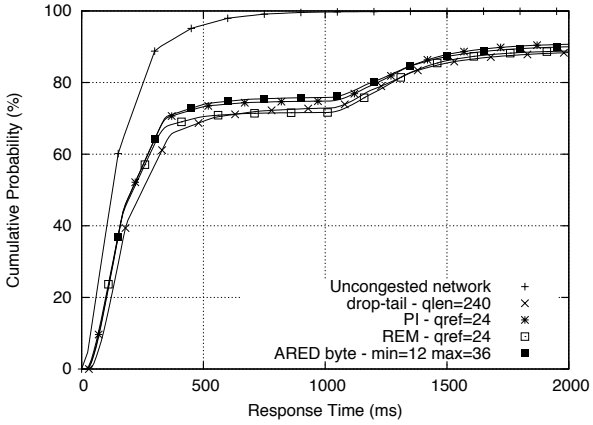

Figure 10: Comparison of all schemes at 98% load.

formance when used with a target queue reference of 24 packets.

4.3 Results for ARED with Packet Drops

ARED experiments were performed in both packet-mode and byte-mode (*i.e.*, with ARED computing the average queue length in terms of either packets or bytes). Previous results for ARED operating in packet-mode with packet drops were negative. ARED was shown to increase response time for HTTP transfers when compared to drop-tail FIFO queuing at all load levels considered [11]. These results are confirmed here. For ARED operating in packet-mode, Figure 7 shows a significant shift in the response time distribution compared to PI and REM for both target queue lengths and both load levels. However, as shown in Figure 8, ARED operating in byte-mode provides significantly better response times. Interestingly, as shown in Table 3, at 98% load, ARED in byte-mode results in a (slightly) higher loss rate than in packet-mode, however, more responses complete (are delivered) during the experiment and a higher network utilization is observed. Similar to PI and REM, the best performance is obtained with queue thresholds corresponding to a target queue length of 24 (th_{min} = 12, th_{max} = 36).

4.4 Comparing all Schemes with Packet Drops

We use the results from a conventional drop-tail queue of size equal to either 24 or 240 packets as a baseline for evaluating the performance of the AQM designs. In addition, we also attempted to find a queue size for drop-tail that would represent a "best practice" choice. Guidelines (or "rules of thumb") for determining the "best" allocations of queue size have been widely debated in various venues including the IRTF *end2end-interest* mailing list. One guideline that appears to have attracted a rough consensus is to provide buffering approximately equal to 2-4 times the bandwidth-delay product of the link. Bandwidth in this expression is that of the link and the delay is the mean round-

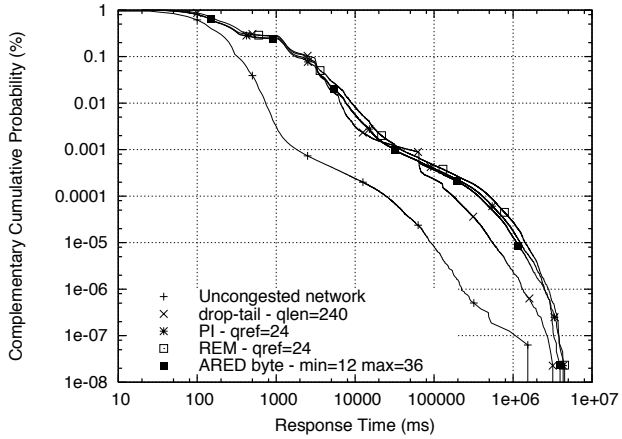

Figure 11: Response time CCDF of all schemes with packet drops, 98% load.

trip time for all connections sharing the link — a value that is, in general, difficult to determine. Other mailing list contributors have recently tended to favor buffering equivalent to 100 milliseconds at the link's transmission speed. In our experimental environment where the link bandwidth is 100 Mbps and mean frame size is a little over 500 bytes, 100 milliseconds of buffering implies a queue length of approximately 2,400 packets.

In [11] we evaluated the response-time performance of a drop-tail queue with length equal to 24, 240 and 2,400 packets for offered loads of 80%, 90%, and 98%. Here, we use a drop-tail queue of 240 packets as a baseline for comparing with AQM mechanisms because it corresponds to one of the targets selected for AQM and provides reasonable performance for drop-tail even though it provides only about 10 milliseconds of buffering at 100 Mbps.

Figures 9 and 10 compare the response time performance of PI, REM, and ARED under the best settings for each algorithm at offered loads of 90% and 98%. To calibrate these curves, the response time performance under drop-tail on

Table 2: Loss, completed requests, and link utilizations for PI and REM.

	Offered Load	Loss ratio (%)		Completed requests (millions)		Link utilization/ throughput (Mbps)	
		No ECN	ECN	No ECN	ECN	No ECN	ECN
1 Gbps network	90%	0		15.0		91.3	
	98%	0		16.2		98.2	
drop-tail $q = 240$	90%	1.9		14.7		90.0	
	98%	5.8		15.1		91.9	
PI $q_{ref} = 24$	90%	1.1	0.2	14.5	14.7	88.1	88.1
	98%	4.1	1.7	14.9	14.9	89.4	89.5
PI $q_{ref} = 240$	90%	0.4	0.04	14.6	14.7	88.3	88.2
	98%	3.7	1.5	15.0	15.1	90.0	90.4
REM $q_{ref} = 24$	90%	1.6	0.1	14.3	14.6	86.4	88.2
	98%	4.9	1.7	14.6	14.9	87.5	89.5
REM $q_{ref} = 240$	90%	3.2	0.1	13.7	14.7	83.3	88.5
	98%	5.4	1.6	14.4	15.0	86.2	90.4

the congested 100 Mbps network and the uncongested 1 Gbps network is also shown. The uncongested network curve represents the best possible response time distribution and provides a basis for an absolute comparison of AQM schemes. The drop-tail curve on the 100 Mbps network (the curve labeled "drop-tail" on all plots) represents the baseline performance that ideally all AQM schemes should beat. Thus in evaluating a AQM algorithm, its performance will be considered acceptable in the absolute if the response time CDF is better (above) drop-tail's. In comparing results for two AQM schemes, we claim that the response time performance is better for one of them if its CDF is clearly above the other's (closer to that of the uncongested network) in some substantial range of response times, and comparable in the remaining range.

Comparing AQM schemes at 90% load, ARED operating in byte-mode is the best performing algorithm, providing better response times for virtually all request/response exchanges. PI, REM, and drop-tail provide equivalent performance for approximately the 40% of exchanges that can be completed in approximately 125 milliseconds or less. For the remainder of the distribution out to 2 seconds, PI outperforms REM and drop-tail while REM either underperforms or performs the same as drop-tail.

At 98% load, PI, REM, and ARED in byte-mode, result in nearly identical performance for the approximately 65% of request/response exchanges that can be completed in 300 milliseconds or less. In addition, all three schemes outperform drop-tail. For the remaining 35% of exchanges, ARED and PI provide similar or slightly better response times than drop-tail while REM provides similar or slightly worse response times. However, overall, no AQM scheme can offset the performance degradation at this extreme load.

Tables 2 and 3 show that at 90% and 98% offered loads, drop-tail with a queue of 240 packets gives slightly better link utilization than any of the AQM schemes. It also completes slightly more request-response exchanges than the other schemes. However, drop-tail has higher loss ratios than the other schemes. ARED in byte-mode has slightly better loss ratios than PI and REM at all loads. ARED and PI complete more requests, and have better link utilization than REM at all loads.

Figures 9 and 10 show that at least 90% of all request/response exchanges complete in under 2 seconds for the best AQM parameter settings at 98% load. Figure 11 shows the remainder of the distribution for this load level. The conclusions drawn from Figures 9 and 10 also hold for exchanges that experience response times up to approximately 50 seconds (~99.95% of all request/ response exchanges). The remaining exchanges perform best under drop-tail. For the 0.05% of request/response exchanges in the tail of the distribution, ARED in byte-mode outperforms PI and REM.

Table 3: Loss, completed requests, and link utilizations for ARED.

	Offered Load	Loss ratio (%)			Completed requests (millions)			Link utilization/ throughput (Mbps)		
		No ECN	ECN	Gentle	No ECN	ECN	Gentle	No ECN	ECN	Gentle
ARED $min = 12$ $max = 36$	90%	0.9	0.7	0.7	13.8	13.8	14.4	85.2	84.7	87.2
	98%	2.1	2.1	1.8	13.9	14.0	14.4	86.2	86.0	88.0
ARED byte $min = 12$ $max = 36$	90%	0.8	1.1	1.1	14.6	14.5	14.6	88.0	87.8	87.5
	98%	3.6	4.0	3.1	14.8	14.6	14.6	89.4	88.0	88.0
ARED $min = 120$ $max = 360$	90%	1.1	1.2	1.0	13.9	13.9	14.6	84.9	85.0	88.4
	98%	3.3	3.9	3.1	14.0	13.9	14.6	86.1	85.9	88.7
ARED byte $min = 120$ $max = 360$	90%	0.9	1.8	1.0	14.6	14.2	14.2	87.6	85.7	86.0
	98%	4.2	4.5	3.9	14.6	14.4	14.4	87.8	86.4	87.1

The major conclusion from the experiments with packet drops is that AQM, specifically, PI and ARED in byte-mode, can improve response times of Web request/response exchanges when compared to drop-tail FIFO queue management. This improvement comes at the cost of a very slight decrease in link utilization.

5 AQM EXPERIMENTS WITH ECN

AQM schemes drop packets as an indirect means of signaling congestion to end-systems. The explicit congestion notification (ECN) packet-marking scheme was developed as a means of explicitly signaling congestion to end-systems [15]. To signal congestion a router can "mark" a packet by setting a specified bit in the

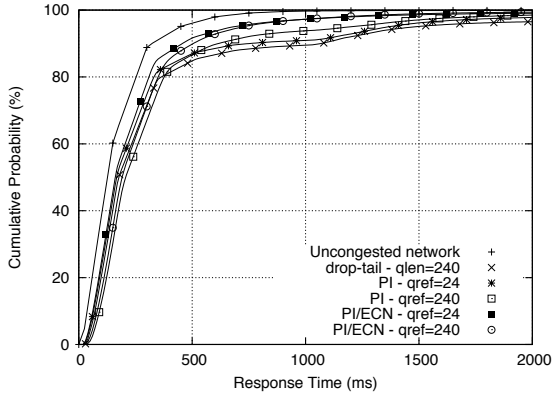

Figure 12: Response time distribution for PI with and without ECN, 90% load.

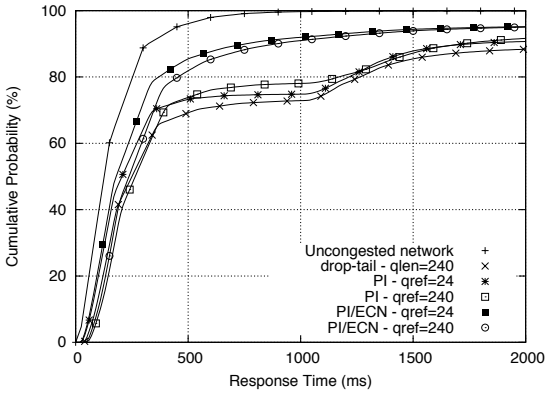

Figure 13: Response time distribution for PI with and without ECN, 98% load.

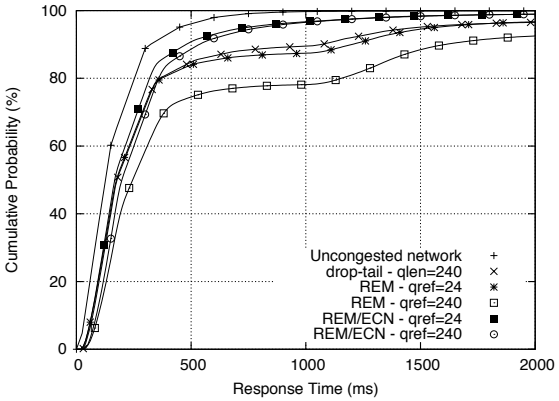

Figure 14: Response time distribution for REM with and without ECN, 90% load.

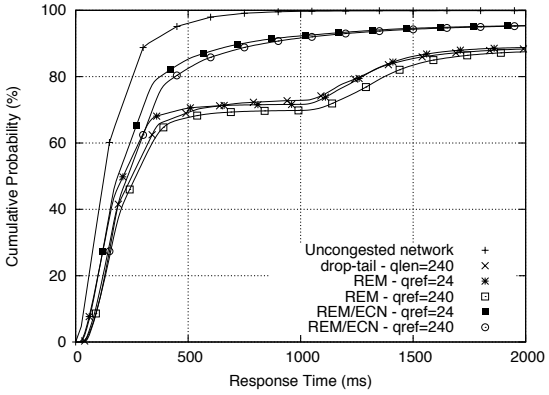

Figure 15: Response time distribution for REM with and without ECN, 98% load.

TCP/IP header of the packet. This marking is not modified by subsequent routers. Upon receipt of a marked packet, a TCP receiver will mark the TCP header of its next outbound packet (typically an ACK) destined for the sender of the original marked packet. Upon receipt of this marked packet, the original sender will react as if a single packet had been lost within a send window. In addition, the sender will mark its next outbound packet (with a different marking) to confirm that it has reacted to the congestion.

We repeated each of the above experiments with PI, REM, and ARED using packet marking and ECN instead of packet drops for offered loads of 90% and 98%. The uniform distribution of minimum RTTs is again used throughout.

5.1 Results for PI and REM with ECN

Figures 12-15 show the results for PI and REM with ECN. At 90% load, PI with ECN performs best with a target queue length of 24 packets. However, with a target queue length of 240, there is little change in performance. At 98% load, ECN significantly improves performance for PI at both target queue lengths.

REM shows significant improvement in performance with ECN at both loads. Whereas without ECN, PI and drop-tail outperformed REM at both 90% and 98% load, with ECN, REM outperforms drop-tail and gives performance similar to PI.

Table 2 again presents the link utilization, loss ratios, and the number of completed requests for each ECN experiment. PI with ECN clearly seems to have better loss ratios, although there is little difference in link utilization and number of requests completed. REM's improvement when ECN is used derives from lowered loss ratios, increases in link utilization, and increases in number of completed requests.

5.2 Results for ARED with ECN

Figures 16-19 show the results for ARED with ECN. Contrary to the PI and REM results, for ARED in both packet-mode and byte-mode, ECN has very little effect on response times. In particular, at all tested target queue lengths, ARED packet-mode performance with ECN is worse than drop-tail at all loads. In byte-mode, only ARED with ECN and queue thresholds of (12, 36) outperforms drop-tail. However, even in this case, performance is slightly worse than ARED byte-

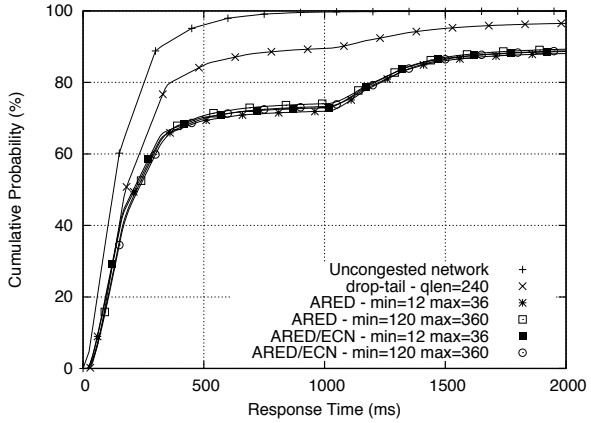

Figure 16: ARED packet-mode response time distribution with/without ECN, 90% load.

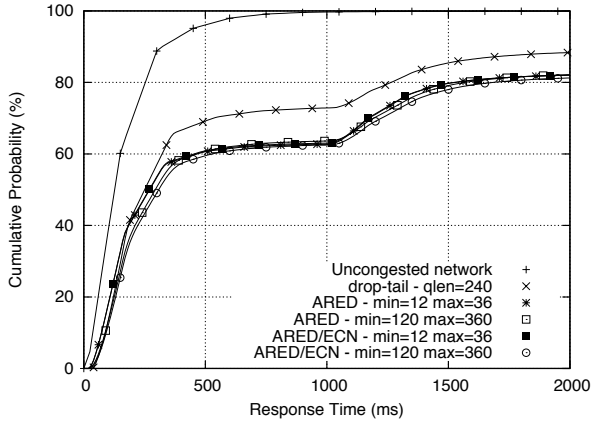

Figure 17: ARED packet-mode response time distribution with/without ECN, 98% load.

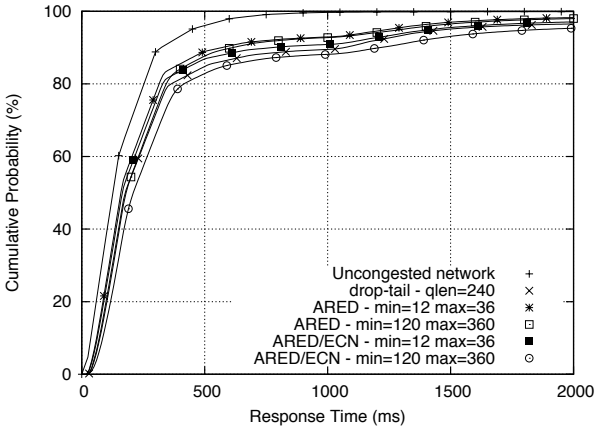

Figure 18: ARED byte-mode response time distribution with/without ECN, 90% load.

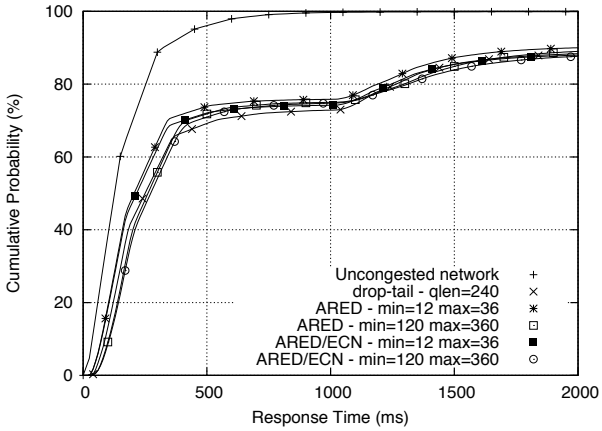

Figure 19: ARED byte-mode response time distribution with/without ECN, 98% load.

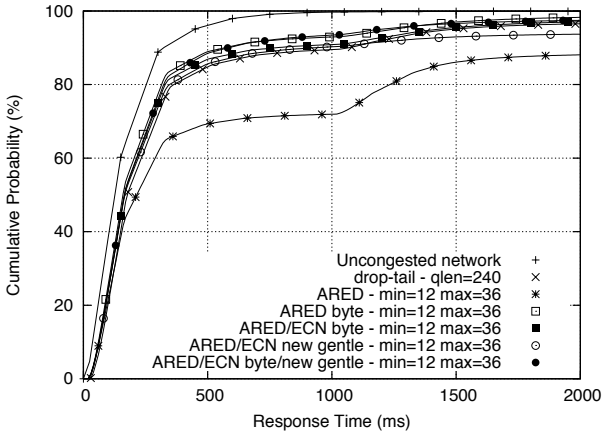

Figure 20: ARED response time comparison with/without ECN forwarding in the gentle region, 90% load.

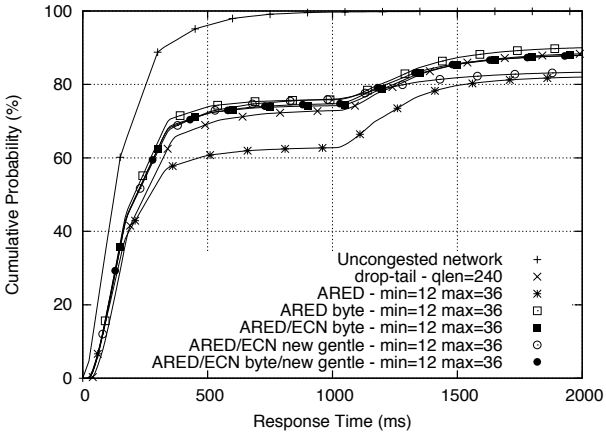

Figure 21: ARED response time comparison with/without ECN forwarding in the gentle region, 98% load.

mode without ECN with the same thresholds. Moreover, as shown in Table 3, in almost all the ARED experiments, the loss rate is higher with ECN than without ECN.

Additional analysis of these experiments indicates that the performance anomalies observed with ECN are due to a subtle aspect of ARED's design. In ARED's "gentle region," when the average queue size is between max_{th} and 2 x

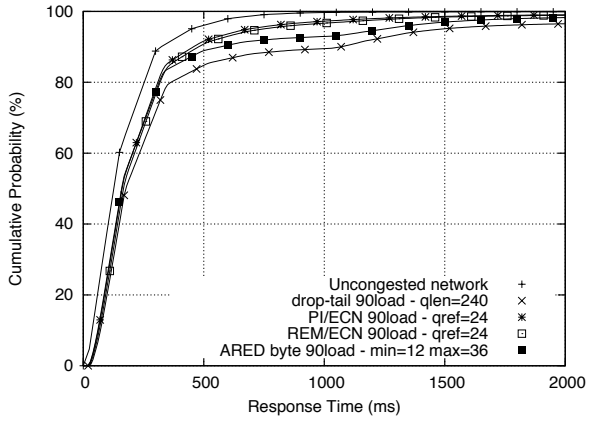

Figure 22: Comparison of the best of all schemes, 90% load.

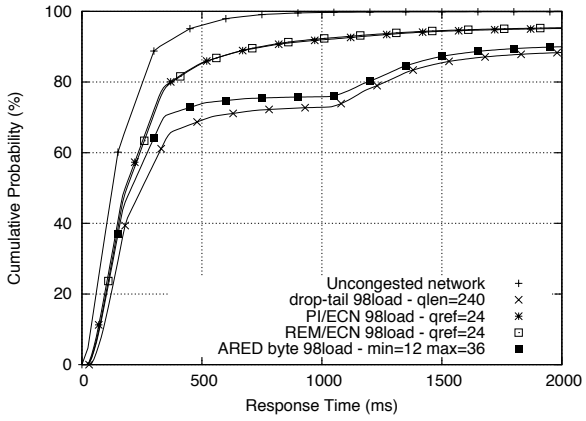

Figure 23: Comparison of the best of all schemes, 98% load.

max_{th}, ARED drops packets even if the packets carry ECN-markings. This is keeping with ECN guidelines that state packets should be dropped when the AQM scheme's max_{th} queue length threshold is exceeded. The stated motivation for this rule is to more effectively deal with potential non-responsive flows that are ignoring congestion indications and thereby increasing the average queue length [15]. We believe this rule to be counter-productive in environments such as ours where there are no non-responsive flows.

To test this hypothesis we allow ARED to forward all packets with ECN-markings in the gentle region. Figures 20-21 compare the performance of ARED with ECN in both packet-mode and byte-mode with and without our "new gentle" ECN forwarding behavior.[1] With the new gentle ECN behavior, performance in packet-mode at both load levels is substantially improved, outperforming drop-tail for the vast majority of request/response exchanges.

The results are less dramatic for ARED in byte-mode. At 90% load, new gentle ECN forwarding in byte-mode improves performance over original gentle ECN forwarding in byte-mode. However, overall, new gentle ECN forwarding in byte-mode does not improve performance over original ARED in byte-mode without ECN. Moreover, at 98% load, new gentle ECN forwarding in byte-mode neither improves response time performance over original gentle ECN forwarding in byte-mode, nor gives better performance then original ARED in byte-mode without ECN.

In summary, ECN provides no benefit to ARED in byte-mode. However, with ECN forwarding in the gentle region, ECN significantly ameliorates the otherwise poor performance of ARED operating in packet-mode. Nonetheless, overall, we conclude that the best ARED response time performance is achieved in byte-mode without ECN.

With respect to loss rate and link utilization, Table 3 shows that ARED in packet-mode (queue thresholds (12, 36)) with new gentle ECN forwarding has loss rates lower than ARED with or without (original) ECN, and comparable to PI and REM with ECN. ARED in byte-mode without ECN (queue thresholds (12, 36)) experiences a comparable loss rate at 90% load but a higher loss rate at 98% load. Nonetheless, ARED in byte-mode without ECN results in slightly more completed requests per experiments and higher link utilization.

5.3 Comparisons of PI, REM, & ARED

Recall that at 80% load, previous work showed no AQM scheme provides better response time performance than a simple drop-tail queue. This result was not changed by the addition of ECN [11]. Here we compare the performance obtained for PI, REM, and ARED with best parameter settings for loads of 90% and 98%. Figures 22-23 show these results.

At 90% load, both PI and REM perform best with ECN while ARED performs best in byte-mode without ECN. All provide response time performance that is close to that on an uncongested link for the shortest 85% of request/response exchanges. For the remaining 15% of exchanges, PI and REM perform somewhat better than ARED. In addition, all the AQM schemes perform better than drop-tail for well over 95% of all exchanges.

At 98% load there is noticeable response time degradation with both PI and REM, however, the results are far superior to those obtained with drop-tail and ARED. Further, both PI and REM with ECN have substantially lower packet loss rates than drop-tail and link utilizations that are only modestly lower. For the best performing ARED, byte-mode

[1] For clarity, Figures 20-21 show only the results for ARED with thresholds of (12, 36). Experiments were performed with the new gentle ECN forwarding behavior at thresholds of (120, 360) and the results were similar to those shown here.

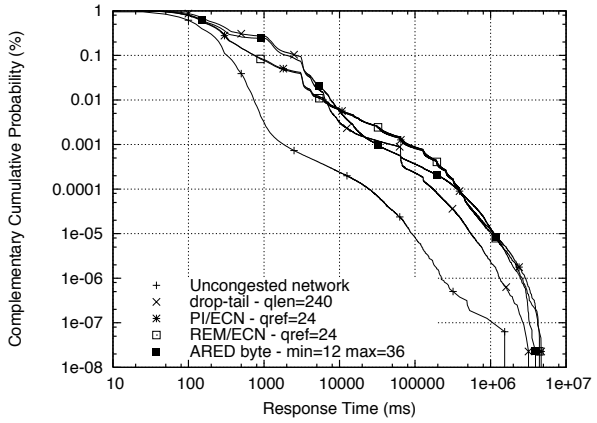

Figure 24: CCDF of the best of all schemes, 98% load.

without ECN, response time performance at 98% load is somewhat better than drop-tail but significantly worse than PI and REM (except for the shortest 45% of request/response exchanges where performance is comparable).

Figure 24 shows the tails of the response time distribution at 98% load. For the best AQM settings, drop-tail again eventually provides better response time performance, however, the crossover point occurs earlier than in the non-ECN case, at approximately 5 seconds. The 1% of request/response exchanges experiencing response times longer than 5 seconds complete sooner under drop-tail. ARED performance in byte-mode again eventually beats PI and REM for a handful of exchanges.

The major conclusion from the experiments with ECN, is that with the addition of ECN support in routers and end-systems, the control theoretic AQM designs PI and REM, can provide significantly improved response time performance over drop-tail FIFO queuing. This is especially true at loads approaching link saturation. However, as was the case with packet drops, these response time improvements come at the cost of slightly decreased link utilizations.

6 THE EFFECTS OF ROUND-TRIP TIME ON AQM PERFORMANCE

To study the sensitivity of response time to round trip time (RTT), we reran several experiments applying a more general distribution of minimum RTTs to our method of source-level generation of Web traffic (see Section 3). We repeated the experiments of Sections 4 and 5 to test the effects of AQM with and without ECN. As described in Section 3, the use of the general minimum RTT distribution required a recalibration of the network. Experiments were still performed with offered loads of 90% and 98% of the capacity of the bottleneck 100 Mbps link, however, different (larger) populations of emulated users were required to realize these loads (see Figure 4).

Figures 25-28 show the major results for the settings of algorithm parameters that previously resulted in the best performance. Without ECN, at 90% load, PI, REM, and ARED byte-mode provide response time performance indistinguishable from drop-tail and surprisingly close to the performance achieved on the uncongested network. ARED packet-mode significantly underperforms drop-tail and all other algorithms. At 98% load, overall performance decreases and slightly more differentiation is visible between PI, REM, ARED byte-mode, and drop-tail. However, again, all give near identical performance and ARED packet-mode still gives poor performance.

With ECN, at 90% load, all queue management paradigms give identical performance that is nearly the best possible performance. At 98% load, PI, REM, ARED byte-mode with new gentle ECN forwarding, and drop-tail provide identical performance for the first 50% of request/response exchanges (those completing in approximately 250 milliseconds or less). For the remainder of the distribution out to 2 seconds, PI and REM perform best and ARED byte-mode with new gentle forwarding performs better than drop-tail. ARED packet-mode with new gentle forwarding very slightly underperforms drop-tail initially and then approximates drop-tail performance.

Table 4 gives the summary statistics for the experiments with the generalized minimum RTT distribution. Note that as expected, loss-rates decrease with the addition of ECN.

Overall we conclude that with the general round-trip time distribution, AQM adds no value without ECN. Only the control theoretic AQM schemes can improve performance, but only when used with ECN and only at extreme network loads (loads approaching network saturation). A possible explanation for these results is that the characteristics of the arrival process at router queues under the general RTT distribution are such that AQM has less opportunity to effect response time (*e.g.*, the arrival process is less bursty). This conjecture is supported by the fact simple drop-tail queuing performs surprisingly well in this environment.

7 DISCUSSION

Our experiments have demonstrated several interesting differences in the performance of Web traffic under the different operating modes of AQM schemes as well as interesting differences between control theoretic and pure randomized-dropping AQM. Our most striking result is the improvement in ARED performance in byte-mode over packet-mode. ARED in packet-mode (the recommended mode of operation for ARED) consistently gave worse response time performance than drop-tail and all other AQM schemes. However, if ECN was not used, ARED operating in byte-mode resulted in the best performance at 90% load and, along with PI, resulted in the best performance at 98% load. We conjecture that the positive effects of byte-mode are primar-

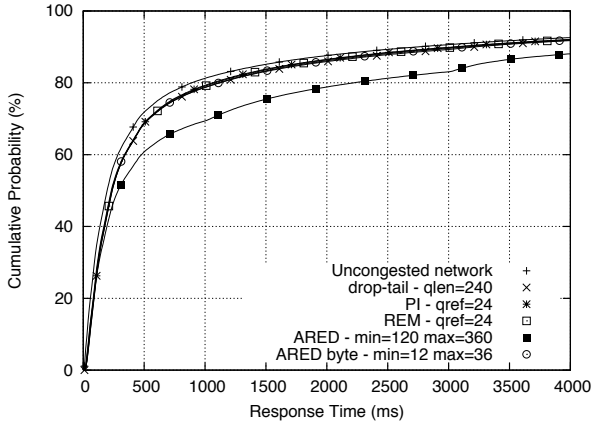

Figure 25: Response time distribution with measured RTT distribution without ECN, 90% load.

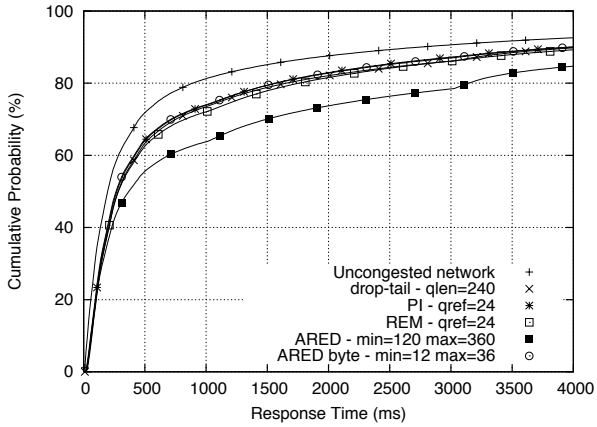

Figure 26: Response time distribution with measured RTT distribution without ECN, 98% load.

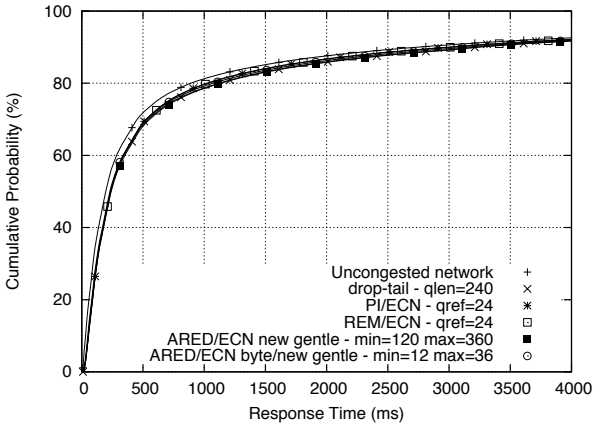

Figure 27: Response time distribution with measured RTT distribution with ECN, 90% load.

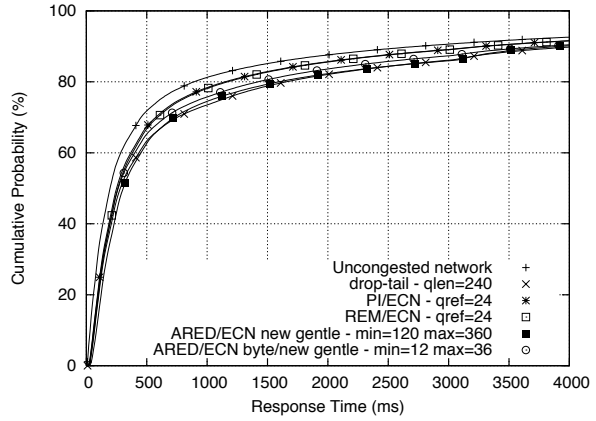

Figure 28: Response time distribution with measured RTT distribution with ECN, 98% load.

ily due to its lowering of the drop probability for small data segments, SYNs, FINs, and pure ACKs.

A second striking result is that once ARED is operating in byte-mode, the addition of ECN provides little benefit. This is sharp contrast to PI and REM which both provide better response times with ECN. ECN similarly had little effect on ARED performance in packet-mode.

In addition to the ARED byte-mode results, the performance of the new gentle forwarding behavior suggests that the design decision to drop ECN-marked packets in ARED's gentle region deserves reconsideration. Although we did not evaluate the effectiveness of ARED (or any scheme) in controlling unresponsive flows, such control cannot come at the expense of decreasing the performance of responsive flows (such as the ones in our experiments).

Regarding the differences in the performance of Web traffic under control theoretic and pure random-dropping AQM, for offered loads up to 90%, comparable good performance is possible under all schemes. Of note is the fact that ECN is

Table 4: Summary statistics for all queue management schemes with the generalized minimum RTT distribution.

	Offered Load	Loss ratio (%)		Completed requests (millions)		Link utilization/ throughput (Mbps)	
		No ECN	ECN	No ECN	ECN	No ECN	ECN
Uncongested network	90%	0		14.7		89.7	
	98%	0		16.0		97.8	
drop-tail $q = 240$	90%	0.3		14.4		86.9	
	98%	1.5		15.0		89.9	
PI $q_{ref} = 24$	90%	0.1	0.02	14.5	14.6	87.3	87.4
	98%	1.0	0.2	15.0	15.1	88.6	88.8
REM $q_{ref} = 24$	90%	0.2	0.02	14.5	14.6	87.1	87.4
	98%	1.4	0.2	14.7	15.1	87.1	89.1
ARED-byte $min = 12$ $max = 36$	90%	0.2	0.07	14.5	14.4	86.7	86.7
	98%	1.0	0.8	14.9	15.0	89.0	88.9
ARED $min = 120$ $max = 360$	90%	0.2	0.05	13.5	14.4	84.4	86.5
	98%	2.1	1.3	13.7	15.0	85.9	89.7

required for the best performance with PI and REM while ECN is not required for the best performance with ARED. However, at 98% load the control theoretic schemes significantly outperform ARED. It remains an open question to see if ECN can be effectively combined with an ARED design to bridge this performance gap.

Considering only control theoretic AQM, an interesting result is that performance varied substantially between PI and REM with packet dropping and this performance gap was closed through the addition of ECN. A preliminary analysis of REM's behavior suggests that ECN is not so much improving REM's behavior as it is ameliorating a fundamental design problem. Without ECN, REM consistently causes flows to experience multiple drops within a source's congestion window, forcing flows more frequently to recover the loss through TCP's timeout mechanism rather than its fast recovery mechanism. When ECN is used, REM simply marks packets and hence even if multiple packets from a flow are marked within a window the timeout will be avoided. Thus ECN appears to improve REM's performance by mitigating the effects of its otherwise poor (compared to PI) marking/dropping decisions.

Finally, the experiments with the general minimum RTT distribution show that AQM performance is clearly sensitive to round-trip time. Further experimentation is required to understand this result. In particular, we need to understand how longer RTTs effect measures of traffic such as the burstiness of the packet-arrival process at the router in our experiments.

Our study of AQM performance concerned only its effect on Web traffic. Ideally we would like to study the effect of AQM on more general models of TCP traffic, however, at present, good source-level models of general TCP traffic suitable for synthetic traffic generation do not exist. Remedying this problem is the subject of our future work [9].

8 CONCLUSIONS

From the results reported above we draw the following conclusions. These conclusions are based on a premise that user-perceived response times are the primary yardstick of performance and that link utilization and packet loss rates are important but secondary measures.

To begin, it is useful to recall one of the primary conclusions from our initial AQM study [11]:

> *For offered loads up to 80% of bottleneck link capacity, no AQM scheme provides better response time performance than simple drop-tail FIFO queue management. Further, the response times achieved on a 100Mbps link are not substantially different from the response times on a 1 Gbps link with the same number of active users that generate this load. This result is not changed by combining any of the AQM schemes with ECN.*

Thus for Web or Web-like traffic, any benefit AQM can provide to application and network performance is limited to occurring only at very high loads. For loads of 90% and 98% of the bottleneck link's capacity, we conclude:

- ARED in byte-mode significantly outperforms ARED in packet-mode. Moreover, ARED in packet-mode, the current recommended mode of ARED usage, was the worst performing AQM design while ARED in byte-mode was the best performing AQM design when ECN is not used. When ECN is not used, ARED in byte-mode outperformed both PI and REM and provided a modest response time improvement over drop-tail.

- ECN does not improve the performance of ARED in either byte- or packet-mode and in cases actually degrades performance. However, allowing ARED to forward ECN marked packets when the weighted average queue length is in the "gentle region" significantly improves the performance of ARED in packet-mode. This improvement, however, results in absolute performance that is still lower than that achieved by ARED in byte-mode without ECN.

- With ECN, both PI and REM provide significant response time improvement at offered loads at or above 90% of link capacity. In particular, at a load of 90%, PI and REM with ECN provide performance on a 100 Mbps link competitive with that achieved with a 1 Gbps link with the same number of active users. While PI and REM with ECN are the best overall performers, it is noteworthy that at 90% load, ARED in byte-mode without ECN matches PI and REM's performance with ECN for the shortest 85% of all request/response exchanges.

- Without ECN, REM and ARED in packet-mode underperform drop-tail (*i.e.*, degrade application and network performance).

Overall we conclude that AQM can improve application and network performance for Web or Web-like workloads. If arbitrarily high loads on a network are possible then the control theoretic designs PI and REM give the best performance but only when deployed with ECN-capable end-systems and routers. In this case the performance improvement at high loads may be substantial. Whether or not the improvement in response times with AQM is significant (when compared to drop-tail FIFO), depends heavily on the range of round-trip times (RTTs) experienced by flows. As the variation in flows' RTT increases, the impact of AQM and ECN on response-time performance is reduced. If network saturation is not a concern then ARED in byte-mode, without ECN, gives the best performance. Combined, these results suggest that with the appropriate choice of AQM, providers may be able to operate links dominated by Web traffic at load levels as high as 90% of link capacity without

significant degradation in application or network performance.

9 ACKNOWLEDGEMENTS

We are indebted to Sanjeewa Athuraliya, Sally Floyd, Steven Low, Vishal Misra, and Don Towsley, for their assistance in performing the experiments described herein. We also thank the numerous reviewers of this paper, especially Sally Floyd, for their constructive comments.

This work was supported in parts by the National Science Foundation (grants ANI 03-23648, EIA 03-03590, CCR 02-08924, and ITR 00-82870), Cisco Systems Inc., and the IBM Corporation.

10 REFERENCES

[1] J. Aikat, J. Kaur, D. Smith, K. Jeffay, *Variability in TCP Roundtrip Times*, Proc. 2003 ACM Internet Measurement Conference, Miami Beach, FL, October 2003, pp. 279-284.

[2] S. Athuraliya, A Note on Parameter Values of REM with Reno-like Algorithms, *http://netlab.caltech.edu*, March 2002.

[3] S. Athuraliya, V. H. Li, S.H. Low, Qinghe Yin, *REM: Active Queue Management*, IEEE Network, Vol. 15, No. 3, May 2001, pp. 48-53.

[4] B. Braden, *et al*, *Recommendations on Queue Management and Congestion Avoidance in the Internet*, RFC 2309, April, 1998.

[5] M. Christiansen, K. Jeffay, D. Ott, and F.D. Smith, *Tuning RED for Web Traffic*, Proc., ACM SIGCOMM 2000, Sept. 2000, pp. 139-150.

[6] W. Feng, D. Kandlur, D. Saha, K. Shin, *A Self-Configuring RED Gateway*, Proc., INFOCOM '99, March 1999, pp. 1320-1328.

[7] S. Floyd, R. Gummadi, S. Shenker, *Adaptive RED: An Algorithm for Increasing the Robustness of RED's Active Queue Management*, *http://www.icir.org/floyd/papers/ adaptiveRed.pdf*, August 1, 2001.

[8] S. Floyd, and V. Jacobson, *Random Early Detection Gateways for Congestion Avoidance,* IEEE/ACM Transactions on Networking, Vol. 1 No. 4, August 1993, p. 397-413.

[9] F. Hernández-Campos, F.D. Smith, K. Jeffay, *Generating Realistic TCP Workloads*, Proc., Computer Measurement Group's 2004 Intl. Conference, Las Vegas, NV, December 2004.

[10] C.V. Hollot, V. Misra, W.-B. Gong, D. Towsley, *On Designing Improved Controllers for AQM Routers Supporting TCP Flows*, Proc., IEEE INFOCOM 2001, April 2001, pp. 1726-1734

[11] L. Le, J. Aikat, K. Jeffay, F.D. Smith, *The Effects of Active Queue Management on Web Performance*, ACM SIGCOMM 2003, August 2003, pp. 265-276.

[12] L. Rizzo, *Dummynet: A simple approach to the evaluation of network protocols*, ACM CCR, Vol. 27, No. 1, January 1997, pp. 31-41.

[13] C. Kenjiro, *A Framework for Alternate Queueing: Towards Traffic Management by PC-UNIX Based Routers,* Proc., USENIX 1998 Annual Technical Conf., New Orleans LA, June 1998, pp. 247-258.

[14] V. Misra, W.-B. Gong,, D. Towsley, *Fluid-based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED*, Proc., ACM SIGCOMM 2000, pp. 151-160.

[15] K. Ramakrishnan, S. Floyd, D. Black, *The Addition of Explicit Congestion Notification (ECN) to IP*, RFC 3168, September 2001.

[16] F.D. Smith, F. Hernandez Campos, K. Jeffay, D. Ott, *What TCP/IP Protocol Headers Can Tell Us About the Web*, Proc. ACM SIGMETRICS 2001, June 2001, pp. 245-256.

[17] W. Willinger, M.S. Taqqu, R. Sherman, D. Wilson, *Self-similarity through high variability: statistical analysis of ethernet LAN traffic at the source level*, IEEE/ACM Transactions on Networking, Vol. 5, No. 1, February 1997, pp. 71-86.