Web Servers Under Overload: How Scheduling Can Help

BIANCA SCHROEDER and MOR HARCHOL-BALTER
Carnegie Mellon University

This article provides a detailed implementation study on the behavior of web serves that serve static requests where the load fluctuates over time (transient overload). Various external factors are considered, including WAN delays and losses and different client behavior models. We find that performance can be dramatically improved via a kernel-level modification to the web server to change the scheduling policy at the server from the standard FAIR (processor-sharing) scheduling to SRPT (shortest-remaining-processing-time) scheduling. We find that SRPT scheduling induces no penalties. In particular, throughput is not sacrificed and requests for long files experience only negligibly higher response times under SRPT than they did under the original FAIR scheduling.

Categories and Subject Descriptors: C.4 [**Computer Systems Organization**]: Performance of Systems—*Reliability, availability, and serviceability*; C.5.5 [**Computer System Implementation**]: Servers; D.4.1 [**Operating Systems**]: Process Management—*Scheduling*; D.4.8 [**Operating Systems**]: Performance—*Queueing theory*

General Terms: Performance, Reliability, Algorithms, Design, Experimentation, Measurement

Additional Key Words and Phrases: Web server, overload, starvation, scheduling, SRPT, unfairness

1. INTRODUCTION

Most well-managed web servers perform well most of the time. Occasionally, however, every popular web server experiences transient *overload*. Overload is defined as the point when the demand on at least one of the web server's resources exceeds the capacity of that resource. While a well designed web server should not be persistently overloaded, transient periods of overload are often inevitable, since the traffic increase at the server that leads to the transient period of overload is difficult to predict. As an example, the amount of traffic received by a Web site might rise because of an unexpected increase of the site's popularity, e.g., after being featured on national television or in a major newspaper. Another situation that could lead to transient periods of overload

This work was supported by NSF Career Grant CCR-0133077, by NSF ITR Grant 99-167 ANI-0081396, and by Spinnaker Networks via Pittsburgh Digital Greenhouse Grant 01-1.
Author's address: B. Schroeder, Computer Science Dept., Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213–3891; email: biancas@andrew.cmu.edu.

is under-provisioning for sales-boosting holidays. Although an online retailer knows to expect more Web site hits during those days, it is difficult to predict exactly the associated increase in traffic volume.

An overloaded Web server typically displays signs of its affliction within a few seconds. Work enters the Web server at a greater rate than the Web server can complete it. This causes the number of connections at the Web server to build up. Very quickly, the Web server reaches the limit on the number of connections that it can handle. From the client perspective, the request for a connection will either never be accepted or will get through only after several trials. Even when the client's request for a connection does get accepted, the time to service that request may be very long because the request has to *timeshare* with all the other requests at the server.

The solution most commonly suggested to avoid overload is admission control [Cherkasova and Phaal 1998; Iyer et al. 2000; Voigt et al. 2001; Welsh and Culler 2003; Elnikety et al. 2004]: the Web server or a front-end monitors the load and the capacity of the server and, if necessary, rejects incoming requests, in order to ensure satisfactory service to those requests being accepted to the server. The decision of when and which request to reject is based on sophisticated algorithms, for example, leveraging application-level knowledge in the admission decision, applying techniques from control theory, or by combining admission control with QoS. Nevertheless, in the end it comes down to denying some customers service.

In this article we suggest a different solution to server overload that provides stable server performance without dropping requests. Our solution is based on connection scheduling. Traditional Web servers fairly proportion their resources among all requests (referred to as ***FAIR*** scheduling in the remainder of the article). By contrast we propose *unfair* scheduling: instead of giving a fair share to each request, we give priority to requests for small files, in accordance with an approximation of the well-known scheduling algorithm, preemptive, shortest-remaining-processing-time (***SRPT***) first.

SRPT scheduling is not new. The SRPT algorithm makes sense in any context where the service requirement of jobs is known a priori. In such a context, the SRPT algorithm can be thought of as a greedy strategy to minimize the number of jobs in the system by always working on the job that is closest to completion. SRPT is provably the optimal scheduling policy for minimizing mean response times [Schrage and Miller 1966]. SRPT has been proposed for Web systems that serve static content before [Harchol-Balter et al. 2003; Rawat and Kshemkayani 2003], where it was observed that the service demand of a request is proportional to the size of the file requested (hence known a priori). However, SRPT has never been proposed in the context of overload, and has typically only been evaluated under simulation [Gong and Williamson 2003; Murta and Corlassoli 2003] or using application-level scheduling, which does not always provide sufficient scheduling control (see Crovella et al. [1999]).

Using SRPT under lower-load situations is vindicated because of analytical work proving that SRPT does not hurt requests for long files, despite favoring requests for short files [Wierman and Harchol-Balter 2003; Bansal and Harchol-Balter 2001a].

What is new in this work is that we propose using SRPT as a solution for coping with transient overload. Overload is a regime where SRPT has never been evaluated (neither analytically nor via simulation). There is a strong belief that long requests will starve under SRPT [Bender et al. 1998; Stallings 2001, p. 410; Silberschatz et al. 2002, p. 162].

This article makes two contributions. First, we provide a detailed performance study of a Web server under overload, showing just how bad overload can be. We experiment with both *persistent overload* (the request rate exceeds the server capacity during the entire experiment) and *transient overload* (alternate periods of overload and low load where the overall mean load is below 1). Load will be defined formally in the next section. We find that within a very short period, even low amounts of overload cause the server to experience instability, dropping requests from its SYN queue, while the client experiences very high response times (response time is defined as the time from when the client submits the request until receiving the last byte of the request). We evaluate a full range of complex environmental conditions, summarized in Table II, including: the effect of WAN delay and loss, the effect of user aborts, the effect of persistent connections, the effect of SYN cookies, the effect of the RTO TCP timer, the effect of the packet length, and the effect of the SYN queue and ACK queue length limits.

Second, the paper proposes SRPT-like scheduling as a means to combat overload. We show that contrary to intuition, SRPT scheduling at the Web server under transient overload does not unduly harm requests for large files as compared with traditional FAIR scheduling used in Web servers. Furthermore, SRPT scheduling significantly improves mean response times overall by up to an order of magnitude. Lastly, even the mean time, until the client receives the first byte, is improved by nearly an order of magnitude under SRPT compared to FAIR. Our implementation results are corroborated via theoretical approximations which match the trends we observe. Our evaluation of SRPT involves a kernel-level implementation whereby we control the order in which the socket buffers at the server are drained into the network.

The outline of this article is as follows: In Section 2, we describe the experimental setup, including the workload, the bottleneck resource, and how overload is generated. Section 3 explains our implementation of SRPT. Section 4 studies exactly what happens in a traditional (FAIR) Web server under persistent overload and contrasts that with the performance of the modified (SRPT) Web server. Section 5 compares the performance of the FAIR server and the SRPT server under transient overload. Section 6 analyzes where exactly SRPT's performance gains come from. Section 7 discusses relevant previous work, and Section 8 concludes.

2. EXPERIMENTAL SETUP

2.1 Trace-Based Workload

In this article we focus on servicing *static* requests, of the form "Get me a file." While Web sites are increasingly generating dynamic content, studies

from 1997–2001 [Krishnamurthy and Rexford 2001; Manley and Seltzer 1997; Feldmann] suggest that the request stream at Web sites is still dominated by static requests. Even logs as recent as 2004 from proxy servers suggest that 67–73% of requests are for static content [Home 2004].

Serving static requests quickly is the focus of many companies for example, Akamai Technologies, and much ongoing research. In the context of overload, static requests are especially important since popular Web sites often convert their dynamic content into static content when the Web site is overloaded [LeFebvre 2002].

The workload in our experiments is based on a one-day trace from the 1998 Soccer World Cup[1] obtained from the Internet Traffic Archive. The trace contains 4.5 million mostly static HTTP requests. In our experiments, our clients generate requests according to an arrival process based on the interarrival times in the trace, as explained in Section 2.3. The trace is also used to specify the request size in bytes. Results for experiments with other logs are given in the Appendix.

Some statistics about our trace workload follow. The mean file size requested is 5K bytes. The minimum size file requested is a 41-byte file. The maximum size file requested is a 2.02MB file. The distribution of the file sizes requested has been analyzed in earlier work [Arlitt and Jin 2000] and found to be *heavy-tailed*: while the body of the distribution can be reasonably well modeled by a log-normal distribution, the tail is best fit by a Pareto distribution with an α-parameter of less than 1.5. We find that the largest $< 3\%$ of the requests make up $> 50\%$ of the total load (in terms of total number of bytes), exhibiting a strong heavy-tailed property. 50% of files have sizes less than 1K bytes, and 90% of files have sizes less than 9.3K bytes. Figure 1 shows the full complementary cumulative distribution of requested file sizes.

2.2 Load and the Bottleneck Resource

To understand the performance of a Web server under overload, we need to understand which resource in a Web server experiences overload first, that is, which is the bottleneck resource. The three contenders are the CPU; the disk to memory bandwidth and the server's limited fraction of its ISP's bandwidth. On a site consisting primarily of *static content*, a common performance bottleneck is the limited bandwidth that the server has bought from its ISP [Microsoft TechNet Insights 2001; Cockcroft 1996; Maggs 2001]. Even a fairly modest server can completely saturate a T3 connection or a 100Mbps Fast Ethernet connection. Also, buying more bandwidth is typically relatively more costly than upgrading memory or CPU. In fact, most Web sites buy sufficient memory so that all their files fit within memory (keeping disk utilization low) [Maggs 2001]. For static workloads, CPU load is typically not an issue.

[1]The original trace contains three months of data. We choose a period of this trace that creates a load close to 1 in our experimental setup, to minimize the scaling necessary to achieve the different load levels we experiment with (see Section 2.3 on scaling). The period of the trace we choose exhibits the same statistical characteristics as the entire trace.

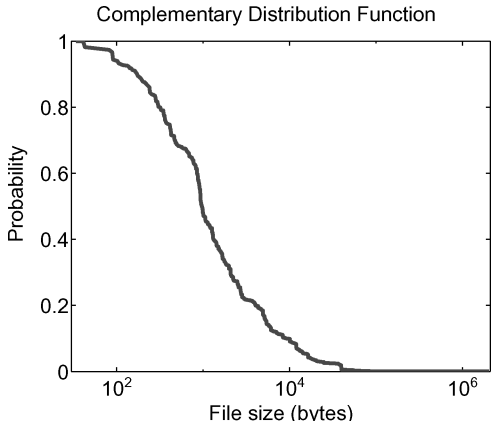


Fig. 1. Complementary cumulative distribution function, $\bar{F}(x)$, for the trace-based workload. $\bar{F}(x) = \Pr\{\text{size of the file requested } > x\}$.

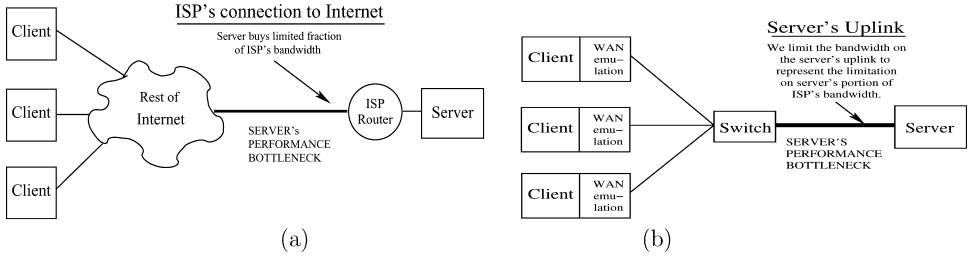


Fig. 2. (a) Server's bottleneck is the limited fraction of bandwidth that it has purchased from its ISP; (b) Our implementation setup models this bottleneck by limiting the server's uplink bandwidth.

We model the limited bandwidth that the server has purchased from its ISP by placing a limitation on the server's uplink, as shown in Figure 2. In all our experiments the bandwidth on the server's uplink is the bottleneck resource. *System load* is therefore defined in terms of the load on the server's uplink. For example, if the Web server has a 100Mbps uplink, and the average amount of data requested by the clients is 80Mbps, then the system load ρ is 0.8. Although in this article we assume that the bottleneck resource is the limited bandwidth that the server has purchased from its ISP, the main ideas can also be adapted for alternative bottleneck resources.

2.3 Defining Persistent and Transient Overload

In our experiments we consider two types of overload, *persistent overload* and *transient overload*.

Persistent overload is used to describe a situation where the server is run under a fixed load $\rho > 1$ during the whole experiment. The motivation behind experiments with persistent overload is mainly to gain insight into what happens under overload. The overloaded state is unlikely to persist for too long in practice due to system upgrades. Nevertheless, due to the burstiness of Web

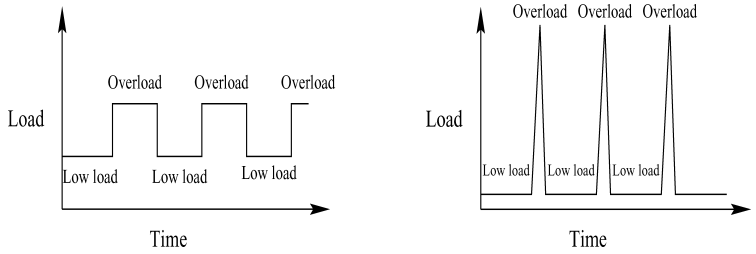


Fig. 3. Two different types of transient overload, alternating (left) and intermittent (right). Experimentally, the load will never look as constant as above, since arrival times and request sizes come from a trace.

traffic, even in the case of regular upgrades, a popular Web server is still likely to experience transient periods of overload.

We consider two different types of transient overload In the first type, called *alternating overload*, the load alternates between overload and low load, where the length of the overload period is equal to the length of the low load period (see Figure 3 (left)). In the second, called *intermittent overload*, the load is almost always low, but there are occasional spikes of overload that are evenly spaced (see Figure 3 (right)). In all cases, the overall mean system load is less than 1.

Throughout, since the bandwidth on the uplink is the bottleneck resource, we define *load* to be the ratio of the bandwidth requested and the maximum bandwidth available on the uplink. To obtain a particular load, we scale the interarrival times in the trace as follows: We first measure the system load (i.e., the bandwidth utilization) in an experiment, using the original (unscaled) interarrival times from the trace. In order to achieve a system load that is x times higher than the original load, we divide all interarrival times by x. Scaling the interarrival times (while keeping the sequence of requested Web objects constant) models a general increase in traffic at the Web server, for example, due to sudden popularity or due to holiday shopping. Note that this type of overload is different from the case where a sudden rise in the popularity of one single object at a site causes overload. In this article we are addressing only overload conditions that are caused by an increase in the arrival rate at the server, while the distribution of the documents requested remains the same (or similar) to that before overload.

We run all experiments in this article for several different alternating and intermittent workloads, which are defined in Table I. All our results are based on 30 minute experiments, although we show only shorter fragments in the figures for better readability.

2.4 Why Generating Overload is Difficult

Experimenting with persistent overload is inherently more difficult than running experiments where the load is high but remains below 1. The main issue in experimenting with overload is that running the server under overload is very taxing on client machines. While both the client machines and the server must allocate resources (such as TCP control blocks or file descriptors) for all accepted requests, the client machines must additionally allocate

Table I. Definition of Trace-Based Workloads

Workload	Type	Duration low (seconds)	Duration overload load (seconds)	Avg. low load	Avg. overload	Avg. load
W1	Alternating	25	25	$\rho = 0.2$	$\rho = 1.2$	0.7
W2	Alternating	10	10	$\rho = 0.2$	$\rho = 1.2$	0.7
W3	Alternating	50	50	$\rho = 0.2$	$\rho = 1.6$	0.9
W4	Alternating	25	25	$\rho = 0.4$	$\rho = 1.4$	0.9
W5	Alternating	10	10	$\rho = 0.4$	$\rho = 1.4$	0.9
W6	Alternating	40	40	$\rho = 0.1$	$\rho = 1.7$	0.9
W7	Intermittent	63	6	$\rho = 0.4$	$\rho = 5$	0.8
W8	Intermittent	63	10	$\rho = 0.45$	$\rho = 3$	0.8
W9	Intermittent	20	3	$\rho = 0.735$	$\rho = 2$	0.9
W10	Intermittent	13.3	2	$\rho = 0.735$	$\rho = 2$	0.9

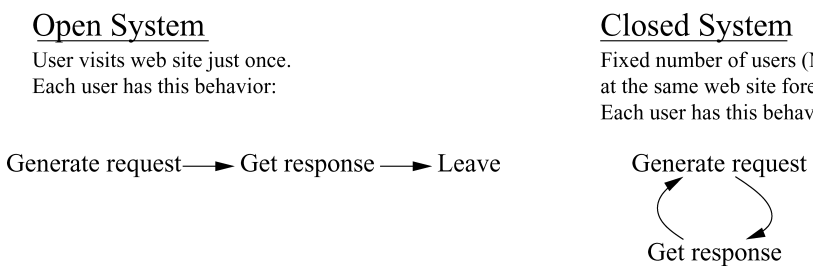


Fig. 4. Two models for how the requests to a Web server are generated. In creating overload, one must use an open system model.

resources for all connections that the server has repeatedly refused (due to a full SYN-queue).

Requests to a Web server may be generated either using an open system, a closed system, or some hybrid combination of the two (see e.g., Harchol-Balter et al. [2003]). To obtain overload, an open or partly-open system is necessary [Banga and Druschel 1999]. Existing Web workload generators include `Surge` [Barford and Crovella 1998] and `Sclient` [Banga and Druschel 1999]. `Surge` mimics a closed system with several users where each user makes a request, and after receiving the response, waits for a certain think time before it makes the next request. Note that in a closed system it is not possible to create overload since a new request will be made only if another request finishes (see Figure 4 (right)). By contrast, an open system allows one to create any amount of overload by simply generating a rate of requests where the sum of the sizes of the files requested exceeds the server uplink bandwidth.

Since none of the existing Web-workload generators support all the features we want to experiment with (open system, persistent connections, user abort, and reload, etc.), we choose to implement our own trace-based Web-workload generator based on the `libwww` library [W3C]. `Libwww` is a client-side Web API based on *select()*. One obstacle in using `libwww` to build a Web workload generator is that it does not support multiple parallel connections *to the same host*. We modify `libwww` in the following way to perform our experiments with persistent connections: Whenever our application passes a URL to our modified `libwww`,

it first checks whether there is a socket open to this destination that is (a) idle and (b) that has not reached the limit on the maximum number of times that we want to reuse a connection. If it doesn't find such a socket, it establishes a new connection. We validate our workload generator by running experiments involving only the features supported by `Sclient`, and we find that `Sclient` and our new workload generator yield the same results. We also verify that the workload generator never becomes the bottleneck in the experiments by checking that all requests are actually made at the desired times.

2.5 Machine Configuration

Our experimental setup involves six machines connected by a 10/100 Ethernet switch. Each machine has an Intel Pentium III 700MHz processor, and a 256MB RAM, and runs Linux 2.2.16. One of the machines is designated as the server and runs Apache 1.3.14. The other five machines act as Web clients and generate HTTP 1.1 requests, as described in Section 2.1. The switch and the network cards of the six machines are forced into 10Mbps mode to make it easier to create overload at the bottleneck device.[2]

On the server machine, we increased the size of the SYN and the ACK-queue to 512, as is common practice for high performance web servers. We also increased the upper limit on the number of Apache processes from 150 to 350.[3]

Furthermore, we have instrumented the kernel of the server to provide detailed information on the internal state of the server. This includes the length of the SYN- and ACK-queues, the number of packets dropped inside the kernel, and the number of incoming SYNs that are dropped. We also log the number of active sockets at the server, which includes all TCP connections that have resources in the form of buffers allocated to them, except for those in the ACK-queue. Essentially, this means sockets being serviced by an Apache process, and sockets in the FIN-WAIT state.

Following we provide a brief tutorial of the processing of requests and sources of delays within a Linux-based Web server.

A connection begins with a client sending a SYN. When the server receives the SYN, it allocates an entry in the SYN-queue and sends back a SYN-ACK. After the client ACKs the server's SYN-ACK, the server moves the connection record to its ACK-queue, also known as the listen queue. The connection waits in the ACK-queue until an Apache process becomes idle and processes it. Each Apache process can handle, at most, one request at a time. When an Apache process finishes handling a connection, the connection sits in the FIN-WAIT state until an ACK and, FIN are received from the client.

There are standard limits on the length of the SYN-queue (128), the length of the ACK-queue (128), and the number of Apache processes (150). These limits are often increased for high-performance Web servers.

---

[2]Experiments were also performed under 100Mbps mode, but not in overload, since that puts too much strain on the client machines (see Section 2.4).

[3]Our FAIR server performed best with the above values: SYNQ = ACKQ = 512, and #Apache Processes = 350. The SRPT server is not affected at all by these limits, since SYNQ occupancy is always very low under SRPT.

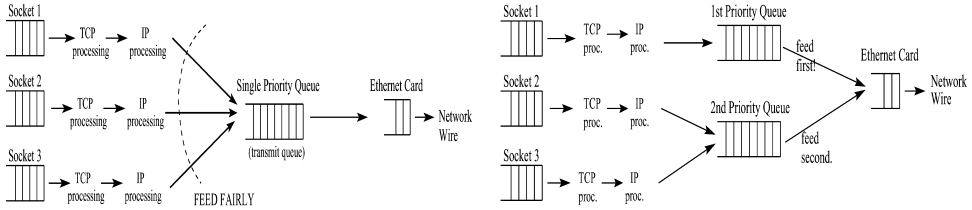


Fig. 5. Data flow in standard Linux (left) and in Linux with priority queueing (right).

The limits above impose many sources of delays. If the server receives a SYN while the SYN-queue is full, it discards the SYN, thus forcing the client to wait for a timeout and then retransmit the SYN. Similarly, if the server receives the ACK for a SYN-ACK while the ACK-queue is full, the ACK is dropped and must be retransmitted. The timeouts are long (typically 3 seconds), since at this time the client doesn't have an estimate for the retransmission timer (RTO) for its connection to the server. Lastly, if no Apache process is available to handle a connection, the connection must wait in the ACK-queue.

3. IMPLEMENTING AN SRPT WEB SERVER

In this section, we describe our implementation of SRPT in an Apache Web server running on Linux.

3.1 Achieving Priority Qeueing in linux

Figure 5(a) shows the data flow in standard Linux. We will refer to the unmodified Apache server running on this (unmodified) standard Linux as FAIR scheduling throughout.

There is a socket buffer corresponding to each connection. Data streaming into each socket buffer is then processed by the TCP and the IP. Throughout this processing, the packet stream corresponding to each connection is kept separate from every other connection. Finally, there is a single[4] priority queue into which all streams feed. Importantly, all streams get equal turns draining into the priority queue, subject to TCP windowing constraints. This single priority queue can get as long as 100 packets. Packets leaving this queue drain into a short Ethernet card queue and out to the network.

To implement SRPT, we need more priority levels. We build the Linux kernel with support for the user/kernel Netlink Socket, QOS and Fair Queueing, and the Prio Pseudoscheduler. These kernel options allow us to switch the device queue from the default 3-band queue to a 16-band priority queue, through the `tc` [Almesberger] user, space tool.

Figure 5(b) shows the flow of data in Linux after our kernel modifications. Again, the data is passed from the socket buffers through TCP and IP processing, during which the packet streams corresponding to each connection are kept separate. There are 16 priority queues (Figure 5 shows only two). All the

[4]The queue actually consists of three priority queues (also, called bands). By default, however, all packets are queued to the same band.

connections of priority i feed fairly into the ith priority queue. The priority queues then feed, in a prioritized fashion, into the Ethernet Card queue. The important point here is that priority queue i is only allowed to drain if priority queues 0 through $i - 1$ are all empty. Note that, since scheduling occurs after TCP/IP processing, it does not interfere with the TCP or the IP.

3.2 Modifications to Apache

In order to approximate SRPT using the priority queues, for each request we first have to initialize its socket to a priority corresponding to the requested file's size. The idea is to have those sockets corresponding to smaller files drain into the higher-priority queues. We later need to update the priority, in agreement with the remaining size of the file. Both are done via the `setsockopt()`, system call within the Web server code.

The only remaining problem is that SRPT assumes infinite precision in ranking the remaining processing requirements of the requests. In practice, we are limited to 16 priority bands.

It turns out that the way in which request sizes are partitioned among these priority levels is somewhat important with respect to the server performance. We have experimentally derived some good guidelines that apply to heavy-tailed Web workloads. Denoting the cutoffs by $x_1 < x_2 < \ldots < x_n$,

—the lowest size cutoff x_1 should be such that about 50% of requests have a size smaller than x_1. Intuitively, the smallest 50% of requests comprise so little total load in a heavy-tailed distribution that there's no point in separating them further.

—the highest cutoff, x_n, needs to be low enough so that the largest (approximately) 0.5%–1% of the requests have a size $> x_n$. This is necessary to prevent the largest requests from starving.

—the middle cutoffs are far less important. A logarithmic spacing works well.

In the experiments throughout this article, we use only five priority classes (cutoffs: 1KB, 2KB, 5KB, and 50KB) to approximate SRPT. Using more classes improves performance only slightly. While in our experiments we have hard-coded these size cutoffs in the Apache source code, they could be made accessible to the system administrator via a handle in the Apache configuration file. The system administrator or an automated mechanism could optimize the cutoffs if the file size distribution at a Web site changes.

A potential problem with our approach is the overhead of the `setsockopt` system call used to modify priorities. However, this overhead is mitigated by the low system-call overhead in Linux and the limited number of system calls: since there are only five priority classes the priority changes (at most) four times, and only for the very largest of the file requests.

4. EXPERIMENTAL RESULTS – PERSISTENT OVERLOAD

In this section, we study exactly what happens in a standard (FAIR) Web server during persistent overload and contrast that with the performance of our modified (SRPT) server. We run the Web server under a persistent overload of 1.2,

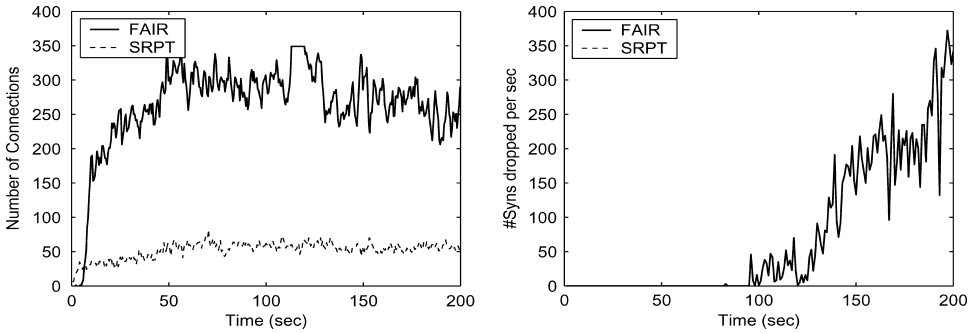


Fig. 6. Results for a persistent overload of 1.2 shown from the perspective of the server; (left): buildup in connections at the server; (right): number of incoming SYN packets dropped by the server.

that is, the average amount of data requested by the clients per second exceeds the bandwidth on the uplink by a factor of 1.2. We analyze our observations from two different angle: the server's view and the client's view.

We start with the server's view. One indication of a server is the buildup in the number of connections at the server, shown in Figure 6 (left). In FAIR, the number of connections grows rapidly, until (after around 50 seconds) the number of connections reaches 350—the maximum number of Apache processes. At this time all the Apache processes are busy, and consequently, the SYN and the ACK queues fill up. After about 70 seconds, the first incoming SYNs are dropped and the rate of SYN drops increases steadily and rapidly from that point on as shown in Figure 6 (right). By contrast, in the SRPT server the number of connections grows at a much slower rate. The reason is that the SRPT server queues up only requests for large files. Observe that for an overload of 1.2 and an uplink capacity of 10Mbps, each second the server is unable to complete 2Mbit worth of requests, on average. It turns out that largest requests compromising load between 1 and 1.2 have a mean size of about 1Mbyte. Thus, the SRPT server accumulates only one-quarter of one request per second. After 200 seconds, it thus makes sense that the number of accumulated requests under SRPT is only 50, as shown in Figure 6 (left). In fact, our experiments show that the SRPT server does not start dropping requests until approximately the half-hour mark.

Another server-oriented metric that we consider is *byte throughput*. We find that the byte throughput is the same under FAIR and SRPT. Despite the differences in the way FAIR and SRPT schedule requests, they both manage to keep the link fully utilized.

Next we describe the clients' experience in terms of *mean response time* and the *mean time until the first byte of a request is received*. In computing these metrics for persistent overload, we consider only those requests that finish before the final request arrives (subsequently, load will drop). Figure 7 (left) shows that, for the FAIR server, response times grow rapidly over time. After only 40 seconds (long before the SYN-queue fills up), the mean response time is 5 seconds, already intolerable. By contrast, under SRPT the response times

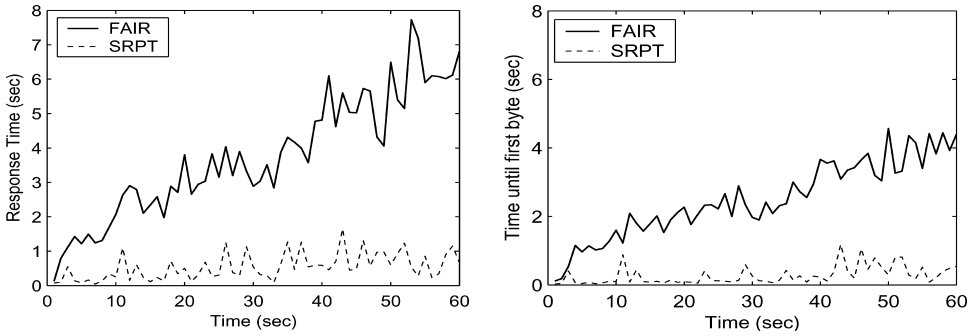


Fig. 7. Results for a persistent overload of 1.2 shown from the perspective of the client, (left): mean response time; (right): time until first byte is received.

are significantly lower and hardly grow over time. Figure 7 (right) shows that the mean time until the first byte is received follows a trend very similar to that of the mean response time. We will therefore in the remainder of this article use only the response time metric.

To understand the difference in response times between SRPT and FAIR under persistent overload, we need to examine the effect of those requests that don't complete on the requests that do complete (and factor that complete into the mean response time). Under FAIR, many of the requests that don't complete still steal a fair share of bandwidth from the other requests. Hence, they cause the response times of even requests for short files to increase. By contrast, under SRPT all of the requests for the largest files (those that increase load from 1.0 to 1.2) do not receive any bandwidth under persistent overload; requests for small files are completely isolated from those for large files. Thus, response times of the completing requests do not increase over time, even after tens of minutes.

To summarize the observations discussed, we see that, after less than 100 seconds of very modest overload, the FAIR server starts to drop incoming requests and the response times reach values that are not tolerable to users. The SRPT server significantly extends the time until SYNs are dropped and improves the client experience notably. We emphasize that the above experiments assume very modest overload (as in the high load portion of workload W1 of Table I). Some of the other workloads in Table I have more severe high loads. For example, under a persistent high load of 1.4, we find that SYNs are dropped after only 18 seconds of persistent overload under FAIR, whereas SRPT avoids SYN drops for nearly half-an-hour of persistent overload. Response time growth is also much more dramatic under a persistent overload of 1.4.

5. EXPERIMENTAL RESULTS—TRANSIENT OVERLOAD

In this section, we evaluate the performance of the standard (FAIR) Web server and our modified (SRPT) server for transient overload. To level the playing field between SRPT and FAIR, we purposely choose to start by evaluating in detail the performance for the transient workload W1 from Table I (see Section 5.1). Workload W1 has the property that the duration and intensity of overload are modest (high load is only 1.2 for a duration of only 25 seconds) so that the SYN

Table II. Columns 1 and 2 list the various factors. Column 3 specifies one value for each factor. This value corresponds to Figure 9 (left, middle) and to Figure 10 (left, middle). Column 4 provides a range of values for each factor. The range is evaluated in Figure 9 (right) and in Figure 10 (right)

Setup	Factor	Specific case shown in the left and middle columns of Figure 9 and Figure 10.	Range of values studied. Shown in rightmost column of Figure 9 and Figure 10.
(A)	Baseline Case	RTT = 0, Loss = 0%, No Persistent Conn., RTO = 3 sec, packet length = 1500, No SYN cookies, SYNQ = ACKQ = 512, #ApacheProcesses = 350	
(B)	WAN Delays	baseline + 100 ms RTT	RTT = 0–150 ms
(C)	Loss	baseline + 5% loss	Loss = 0–15%
(D)	WAN delay & loss	baseline + 100 ms RTT + 5% loss	RTT = 0–150 ms, Loss = 0–15%
(E)	Persistent Connections	baseline + 5 req. per conn.	0–10 requests/conn.
(F)	Initial RTO value	baseline + RTO 0.5 sec	RTO = 0.5sec–3sec
(G)	SYN Cookies	baseline (SYN Cookies OFF)	SYN cookies = ON/OFF
(H)	User Abort/Reload	baseline + user aborts: User aborts after 10 sec and retries up to 3 times	Abort after 3-15 sec with up to 2, 4, 6, or 8 retries
(I)	Packet length	baseline + 536 bytes packet length	Packet length = 536-1500 bytes
(J)	Realistic Scenario	RTT = 100 ms, Loss = 5%, 5 req. per conn., RTO = 3 sec, pkt. len. = 1500, No SYN cookies, SYNQ = ACKQ = 512, #ApacheProcs = 350, User aborts after 7 sec and retries up to 3 times	

queue does not fill up under FAIR. Thus, FAIR does not suffer the expensive timeouts under workload W1 (due to a SYN drop) which appear under some of the other workloads. Our study considers all factors described in Table II. In Section 5.2, we consider the other transient workloads in Table I.

5.1 Results for Workload W1

(*A*) *The simple baseline case*. In this section, we study the simple baseline case described in Table II, row (A). We first consider how the health of the server is affected during the low load and overload periods. We observe, as in the case of persistent overload, that the number of connections grows at a much faster rate under FAIR than under SRPT. While under SRPT the number of connections never exceeds 50, it frequently exceeds 200 under FAIR. However, neither server reaches the maximum SYN-queue capacity (since the overload period is short), and therefore, no SYNs are dropped.

Figure 8(a) and (b) shows the response times over time of all jobs (averaged over 1 second intervals) under FAIR and SRPT. These plots show just 200 seconds out of a 30-minute experiment. The overload periods are shaded light gray and the low-load periods are shaded dark gray. Observe that under FAIR the

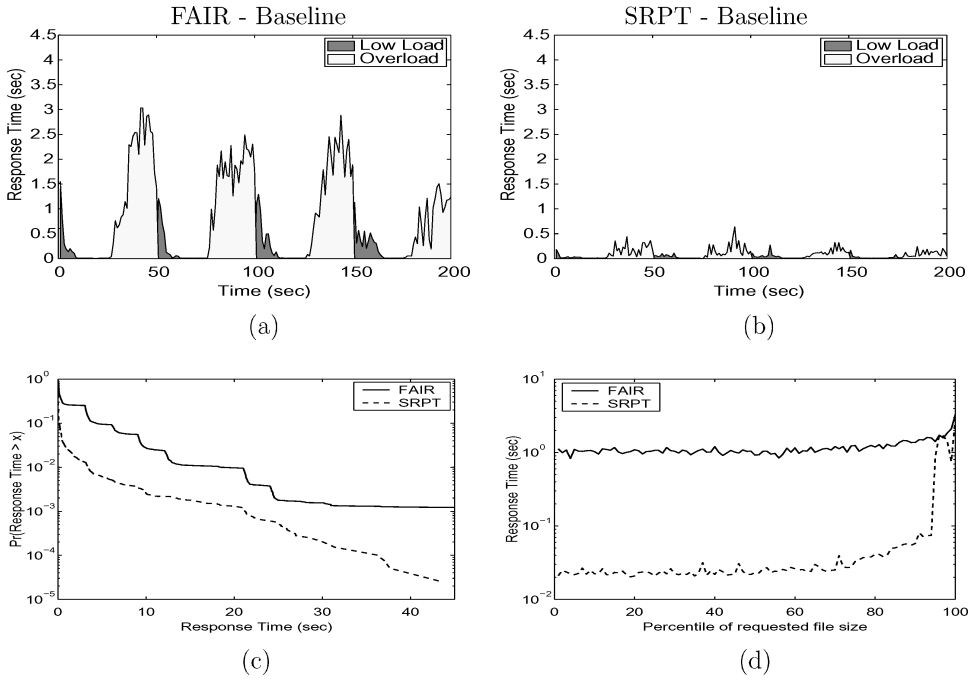


Fig. 8. Detailed performance under alternating workload W1 in the baseline case (setup (A)) of Table II: (a) mean response time under FAIR; (b) mean response time under SRPT; (c) the complementary cumulative distribution of response times under FAIR and SRPT; (d) response times as a function of the request size, showing that requests for large files do not suffer.

mean response times go up to more than 3 seconds during the overload period,[5] while under SRPT they hardly ever exceed 0.5 seconds.

Figure 8(c) shows the complementary cumulative distribution of the response times. Note that there is an order of magnitude of separation between the curve for FAIR and SRPT. The mean response time taken over the entire length of the experiment is 1.1 seconds under FAIR, as compared to only 138 milliseconds under SRPT. Furthermore, the variability in response times measured by the squared coefficient of variation is 6.39 under FAIR, compared to 1.08 under SRPT. Another interesting observation is that the FAIR curve has bumps at regular intervals. These bumps are due to TCP's exponential backoff in the case of packet loss during connection setup. Given that we have a virtually loss-free LAN setup and the SYN-queue never fills up, one may wonder where these packets are dropped. Our measurements inside the kernel show that the timeouts are due to reply packets of the server being dropped inside the server's kernel. We will study the impact of this effect in greater detail in Section 6.

[5]Note that this is not quite as bad as for persistent overload because a job arriving into the overload period in transient overload, at worst, has to wait until the low-load period to receive service, so its expected response time is lower than under persistent overload.

The big improvements in mean response time are not too surprising, given the opportunistic nature of SRPT: schedule to minimize the number of connections. A more interesting question is what price large requests have to pay to achieve good mean performance.

This question is answered in Figure 8(d), which shows the mean response times as a function of the request size. We see that, surprisingly, even the big requests hardly do worse under SRPT. The very biggest request has a mean response time of 19.4 seconds under SRPT, compared to 17.8 seconds under FAIR. If we look at the biggest 1 percent of all requests, we find that their average response time is 2.8 seconds under SRPT compared to 2.9 seconds under FAIR, so these requests perform better on average under SRPT. We will explain this counter-intuitive result in Section 6.

(*B*) *WAN Delays.* The two most frequently used tools for WAN emulation are probably NistNet [National Institute of Standards and Technology] and Dummynet [Rizzo 1997]. NistNet is a separate package available for Linux that can drop, delay, or bandwidth-limit incoming packets. Dummynet applies delays and drops to both incoming and outgoing packets, hence allowing the user to create symmetric losses and delays. Since Dummynet is currently available for FreeBSD only, we implement Dummynet functionality in the form of a separate module for the Linux kernel. More precisely, we change the `ip_rcv()` and the `ip_output()` function in the Linux TCP-IP stack (to intercept in- and out-going packets) to create losses and delays.

In order to delay packets, we use the `add_timer()` facility to schedule the transmission of delayed packets. We recompile the kernel with `HZ = 1000` to get a finer-grained millisecond timer resolution. In order to drop packets, we use an independent, uniform random loss model (as in Dummynet) which can be configured to a specified probability.

We experiment with delays between 0 and 150 milliseconds and drop probabilities between 0 and 15%. This range of values was chosen to cover values used in related work [Nahum et al. 2001] and values reported in actual live Internet measurements [Internet Traffic Report 2004]. For example, for the month of October 2004, the Internet Traffic Report documented maximum round-trip-times of 140 milliseconds and maximum loss rates of 3.5%.

Figure 9(B) (left, middle) shows the effect of adding WAN delay (setup (B) in Table II). This assumes a baseline setup, with an RTT (round-trip time) delay of 100 milliseconds. While FAIR's mean response time is hardly affected since it is large compared to the additional delay, the mean response time of SRPT more than doubles.

Figure 9(B) (right) shows the mean response times for a range of RTTs from 0 to 150 milliseconds. Observe that adding WAN delays increases response times by a constant additive factor on the order of a few RTTs. SRPT improves upon FAIR by at least a factor of 2.5 for all RTTs considered.

In the previous experiments, we assumed that all clients experience the same WAN delays. We also experimented with a heterogeneous environment, where each of the five client machines emulates a different WAN delay ranging from 0 milliseconds to 150 milliseconds, where 150 milliseconds represents our notion

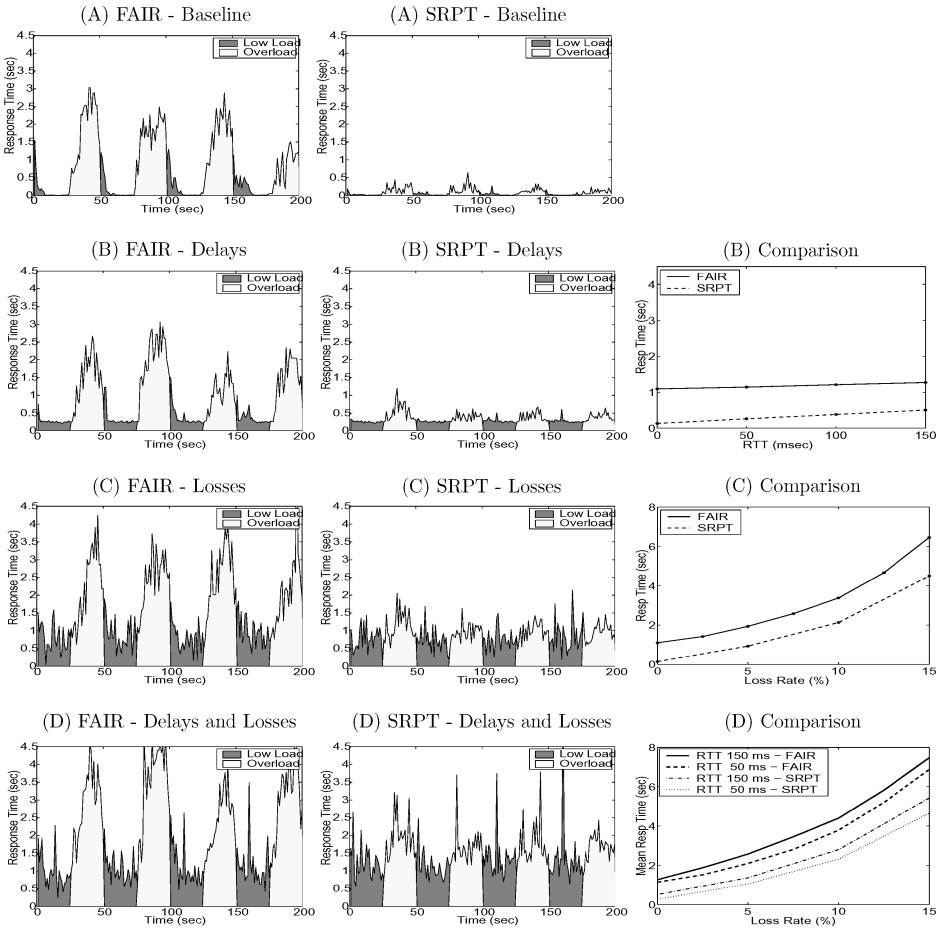


Fig. 9. Each row above compares SRPT and FAIR under workload W1 for one of the first four setups from Table II. The left and middle columns show the response times over time for the specific values given in Table II, column 3. The right column evaluates the range of values given in column 4 of Table II.

of the maximum end-to-end network delay. We found that the mean response time for a client with a given WAN delay equals that of an experiment where all clients emulate the same delay.

While one might think that in a heterogeneous environment high-delay clients might interfere with low-delay clients, thus depriving them of the full benefits of SRPT scheduling, this is not the case for two reasons. First, under overload, the time a connection stays alive at the server is mostly dominated by server delays rather than by WAN delays. Second, and most importantly, a slow connection will not block a fast connection; data for different connections is kept completely separate until after TCP/IP processing when it enters one of the priority queues shared by all connections (recall Figure 5). The data packets of a slow connection will therefore make it to the priority queues only after receiving the corresponding TCP ACKs (i.e., when the data is ready to be sent),

and can therefore not slow down the fast connections (which will receive their ACKs at a fast rate). Thus, the difference between SRPT and FAIR is minimized when all clients have a WAN delay of 150 milliseconds, and even here, the improvement factor of SRPT over FAIR is 2.5.

(*C*) *Network Losses.* Figure 9(C) (left, middle) shows the mean response times over time for the setup in Table II row (C), a loss rate of 5%. In this case, for both FAIR and SRPT the response times increase notably compared to the baseline case. FAIR's overall response time increases by almost a factor of 2 (from 1.1 seconds to 1.9 seconds). SRPT's response time increases from less than 140 milliseconds in the baseline case to around 930 milliseconds.

Figure 9(C) (right) shows the mean response times for loss rates ranging from 0 to 15%. Note that the response times don't grow linearly with the loss rate. This is expected, since the TCPs throughput is inversely proportional to the square root of the loss [Padhye et al. 2000].

Introducing frequent losses can increase FAIR's response times to more than 6 seconds and SRPT's response times to more than 4 seconds (as in the case of a 15% loss). Even in this case SRPT still improves upon FAIR by about a factor of 1.5.

(*D*) *Combination of delays and losses.* Finally, we look at the combination of losses and delays. Figure 9(D) (left, middle) shows the results for the setup in Table II, row (D), an RTT of 100 milliseconds with a loss rate of 5%. Here SRPT improves upon FAIR by a factor of 2. Figure 9(D) (right) shows the response times for various combinations of loss rates and delays. We observe that the negative effect of a given RTT is accentuated under high loss rates. The reason is that higher RTTs make loss recovery more expensive, since timeouts depend on the (estimated) RTT.

(*E*) *Persistent connections.* Next we explore how the response times change if multiple requests are permitted to use a single serial, persistent connection [Mogul 1995]. Figure 10(E) (left, middle) shows the results for the setup in Table II, row (E), where every connection is reused five times. Figure 10(E) (right) shows the response time as a function of the number of requests per connection, ranging from 0 to ten. We see that using persistent connections greatly improves the response times of FAIR. For example, reusing a connection five times reduces the response time by a factor of 2. SRPT, on the other, hand is hardly affected by using persistent connections. To see why FAIR benefits more than SRPT, observe that reusing an existing connection avoids the connection setup overhead; this overhead is bigger under FAIR, mainly because it suffers from drops inside the kernel,[6] SRPT doesn't see this improvement since it experiences hardly any packet loss in the kernel.

Nevertheless, we observe that SRPT still improves upon FAIR by a factor of 3, even if up to ten requests use the same connection.

[6]These drops occur when attempting to feed packets into an already full transmit queue (see Figure 5).

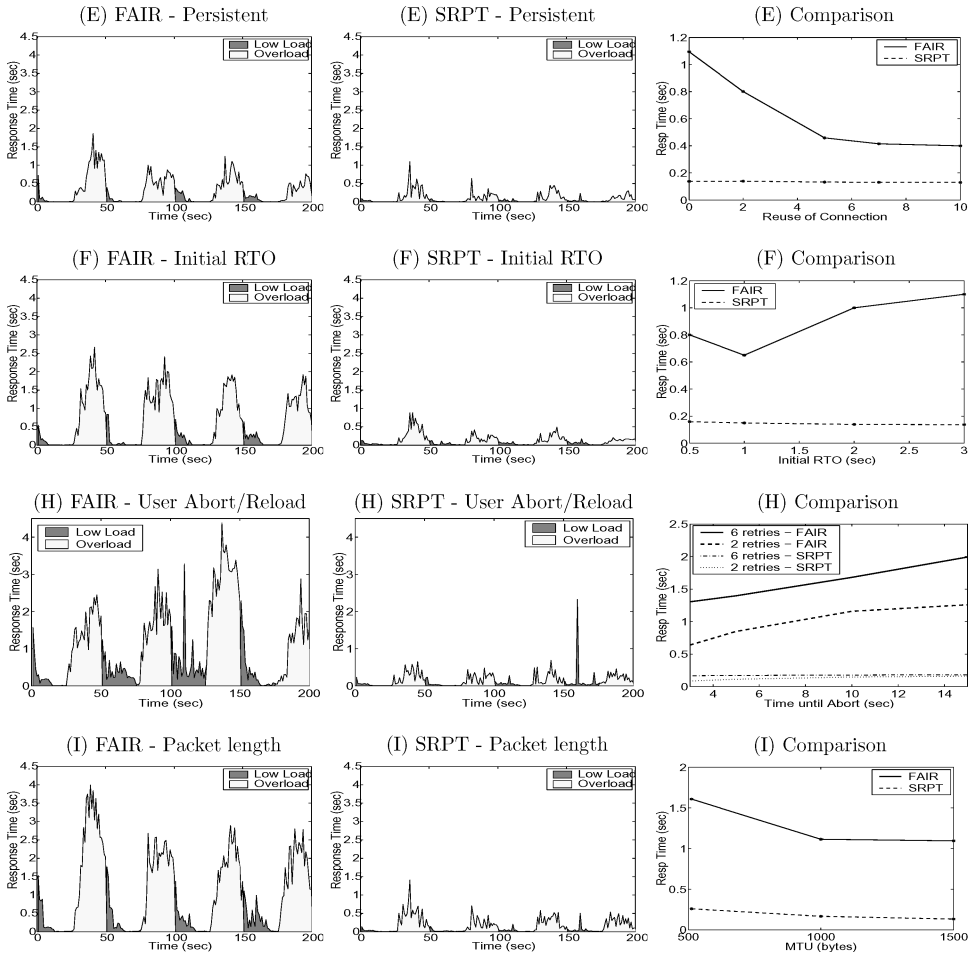


Fig. 10. Each row above compares SRPT and FAIR under workload W1 for one of setups (E), (F), (H), and (I) from Table II. The left and middle columns show the response times over time for the specific values given in Table II, column 3. The right column evaluates the range of values given in column 4 of Table II.

(*F*) *Initial TCP RTO Value.* We observed previously that packets that are lost in the connection setup phase incur very long delays which we attributed to the conservative initial RTO value of 3 seconds. We now ask how much of the total delay a client experiences in the FAIR server is due to the high initial RTO. To answer this, we change the RTO value in Linux's TCP implementation. Figure 10(F) (left, middle) shows the results for the setup in Table II, row (F), an RTO of 500 milliseconds. Figure 10(F) (right) explores the range from 500 milliseconds up to the standard 3 seconds.Lowering the initial RTO can reduce FAIR's mean response time from originally 1.1 seconds (for the standard RTO of 3 seconds) to 0.65 seconds (for an initial RTO of 1 second). Reducing the initial RTO below 1 second introduces too much overhead due to spurious retransmissions and therefore doesn't improve performance any further. SRPT's

response times don't improve for lower initial RTOs since SRPT has little loss in the kernel. Nevertheless, for all initial RTO values considered SRPT always improves upon FAIR by a factor of at least 4.

Having observed that the mean response times of a (standard) FAIR server can be significantly reduced by reducing the TCP's initial RTO, we are not suggesting to change this value in current TCP implementations, since there are reasons it is set conservatively [Paxson and Allman 2000].

(*G*) *SYN Cookies.* Recall that one of the problems under overload is the dropping of incoming packets due to a full SYN-queue. This leads us to the idea of using SYN cookies [Bernstein 1997] in overload experiments. SYN cookies were originally developed to avoid denial of service attacks, however, they have the added side-effect of eliminating the SYN-queue (when using SYN cookies the server makes the SYN-ACK contents purely a function of the SYN contents. This way the SYN contents can be recomputed upon receipts of the next ACK, thereby avoiding the need for maintaining a SYN-queue). Our hope is that by getting rid of the SYN-queue we will also eliminate the problems involving a full SYN-queue.

It turns out that the use of SYN cookies hardly affects the response times under workload W1 since in workload W1 the SYN-queue never fills up. In other transient workloads which exhibit SYN drops due to a full SYN-queue, the response times do improve, but only slightly (by 2–5%). The reason is that now, instead of the incoming SYN being dropped, it is the ACK for the SYN-ACK that is dropped due to a full ACK-queue.

Since SRPT does not lose incoming SYNs, its performance is not affected by using SYN cookies.

(*H*) *User Abort/Reload.* So far we have assumed that a user patiently waits until all the bytes for a request have been received. In practice, users abort requests after a while and hit the reload button of their browser. We model this behavior by repeatedly aborting a connection if it hasn't finished after a certain number of seconds and then opening a new connection.

Figure 10(H) (left, middle) shows the response times in the case where a user waits for at most 10 seconds before hitting reload and retrying this procedure up to 6 times before giving up. Figure 10(H) (right) shows mean response times if connections are aborted after 3 to 15 seconds and for either 2 or 6 retries.

We observe that taking user behavior into account can, depending on the parameters, sometimes increase and sometimes decrease response times. If users are impatient, abort connections quickly, and retry only very few times, the response times can decrease by up to 10%. This is because there is now a (low) upper bound on the maximum response times and also the work at the server is reduced, since clients might give up before the server has even invested any resources in the request. However, this reduced mean response time does not necessarily mean greater average user satisfaction, since some number of requests never complete. For example, if users abort a connection after 3 seconds and retry, at most, 2 times, more than 5 percent of the requests under FAIR never complete for workload W1.

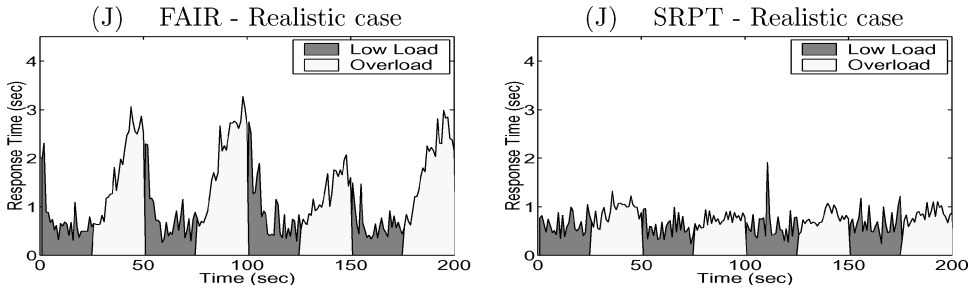


Fig. 11. Comparison of FAIR (left) and SRPT (right) for the realistic setup (J) for workload W1.

If users are willing to wait for longer periods of time and retry more often, response times significantly increase. The reason is that response times are long both for those users that go through many retries and also for other users, since frequent aborts and reloads increase the work at the server. For example, if users retry up to 6 times and abort a connection only after 15 seconds, the response times under FAIR almost double compared to the baseline case that doesn't take user behavior into account. On the other hand, the number of incomplete requests is very small in this case—under 0.02%.

For all choices of parameters for user behavior, the response times under SRPT improve upon those under FAIR by at least a factor of 8. Also, the number of incomplete requests is always smaller under SRPT, by as much as a factor of 7. The reason is that SRPT is smarter than FAIR. Because SRPT favors small requests, the small requests (the majority) have no reason to abort. Only the few big requests are harmed under SRPT. Nevertheless, even requests for the longest file suffer no more incompletes under SRPT than under FAIR.

(*I*) *Packet Length.* Next we explore the effect of the maximum packet length (MSS). Two different packet lengths are commonly observed on the Internet [CAIDA 1999]: 1500 bytes, since this is the MTU (maximum transmission unit) for Ethernet, and 536 bytes, which is used by some TCP implementations that don't perform MTU discovery [Kurose and Ross 2001, p. 319].

Figure 10 (I) (left, middle) shows the results for setup (I) in Table II, where the packet length is changed from 1500 bytes in the baseline case to 536 bytes. As expected, the mean response time increases, since for a smaller packet length more RTTs are necessary to complete the same transfer. FAIR's response time increases by almost 50% to 1.6 seconds and SRPT's response time doubles to 260 millisecond.

Figure 10 (I) (right) shows the mean response times for different packet lengths ranging from 500 bytes to 1500 bytes. For all packet lengths considered, SRPT improves upon FAIR by at least a factor of 6.

(*J*) *A Realistic Scenario.* So far, we have looked at several factors affecting Web performance in isolation. Figure 11 shows the results for an experiment that combines all the factors: We assume an RTT of 100 millisecond, a loss rate of 5%, no SYN-cookies, and the use of persistent connections (5 requests per connection). We leave the RTO at the standard 3 seconds and the MTU at 1500 bytes. We assume users abort after 7 seconds and retry up to 3 times.

We observe that the mean response time of both SRPT and FAIR increases notably compared to the baseline case. It is now 1.48 seconds under FAIR and 764 milliseconds under SRPT—a factor 2 improvement of SRPT over FAIR. Observe that both these numbers are still better than those in experiment (D) where we combined only losses and delays and saw a response time of 2.5 seconds and 1.2 seconds, respectively. The reason is that the use of persistent connections alleviates the negative effects of losses and delays during connection setup.

The largest 1% of requests (counting only those that complete) have a response time of 2.23 seconds under FAIR and 2.28 seconds under SRPT. Even the response time of the very biggest request is higher under FAIR, 13.2 seconds under FAIR and only 12.8 seconds under SRPT. The total number of incomplete requests (those aborted the full three times) is the same under FAIR and SRPT—about 0.2%.

Observe that it makes sense that unfairness to large requests is more pronounced in the baseline case because server delay dominates response time under the baseline case, in contrast to the realistic case where external factors can dominate.

Under workload W1 there were no SYN drops, neither for the baseline nor the realistic setup.

5.2 Other Transient Workloads

Figure 12 gives the mean response times for all ten workloads from Table I for the baseline case (Table II row (A)). While Figure 12 depicts *mean* response times over the entire duration of the experiment, it is important to notice that *peak* response times are actually much higher. For example, for workload W1, when considering the response times averaged over 1 second intervals (as in Figure 8(a) and (b)), we witnessed response times three times higher than that shown in Figure 12 for workload W1. This underscores the need for SRPT scheduling.

We choose to show performance for the baseline case rather than the realistic case because this way the effects of the different workloads are not blurred by external factors. Also, the starvation of large requests is by definition greater for the baseline case than for the realistic case, as explained above. In the discussion that follow, however, we will include the performance numbers for both the baseline and the realistic case.

5.2.1 *Mean Response Time.* For the baseline case, the mean response time of SRPT improves upon that of FAIR by a factor of 2–8 across the ten different workloads. More specifically, FAIR ranges from 300 millisecond–7.2 seconds, while SRPT ranges from 150 millisecond–1.2 seconds.

Under the realistic case (not shown), SRPT improves upon FAIR by a factor of 1.4–4 across workloads W1, W2, W4, W5, W7, W8, W9, and W10. Note that we exclude workloads W3 and W6 throughout the discussion of the realistic scenario. This is because these two workloads have such a long overload period that in the realistic scenario, their behavior mimics persistent overload. To see this, observe that under the realistic scenario, which involves user aborts, there is an increased load due to multiple aborts and retries which are especially

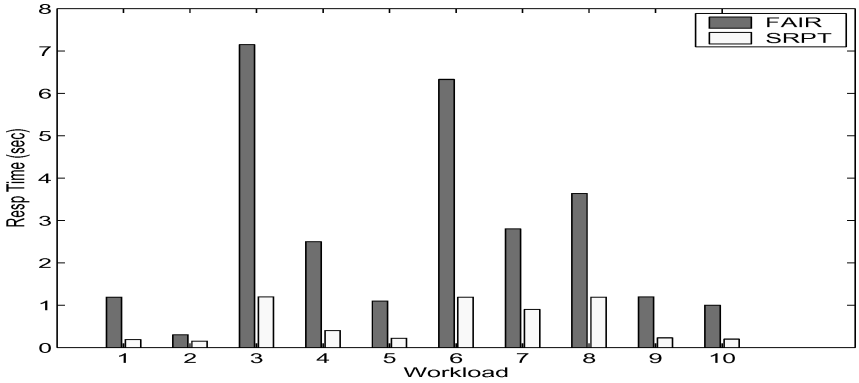


Fig. 12. Comparison of mean response times under FAIR and SRPT in the baseline case for the workloads in Table I. Peak response times are far worse than mean response times.

prevalent in W3 and W6 because of their long overload period. This increased load causes the time-average load in these workloads to rise from 0.9 to over 1.0. Hence, the system is running in persistent overload.

For both the baseline and realistic case, the mean performance under each workload is affected by: (1) the mean system load and (2) the length of the overload and low load periods. The latter should be clear from looking at Figure 12, where workloads W3 and W6, which have the longest periods of overload, both stand out. Point (2) is also justified by considering workloads W1 and W2 which only differ in the length of their overload periods, resulting in a factor 5 difference in their mean response time, or workloads W4 and W5 which also differ only in the length of their overload period, again resulting in a factor 2 difference in their mean response times. The length of the overload period and the overall load also affect the number of SYN drops, and consequently, the response times. While about half the workloads have zero SYN drops under FAIR and SRPT, in both the baseline and the realistic setups workload W3, which has 50 seconds of overload, results in 50% SYN drops under FAIR, and consequently very high mean response times. SYN drops do not occur under SRPT under any of our workloads.

Performance of Large Requests. Again, an obvious question to ask is whether this improvement in the mean response time comes at the price of longer response times for large file requests. In particular, for the workloads with higher mean system load and longer overload periods it might seem likely that SRPT leads to a higher penalty for the long requests.

We find that these concerns are unfounded. The mean response time of only the biggest 1% of all requests is never more than 10% higher under SRPT than under FAIR for any workload in the baseline case and is often lower under SRPT. This penalty is further substantially diminished under the realistic scenario.

When considering the performance of large requests in the realistic setup, it is important to also look at the number of incomplete requests (incomplete

requests are only a consequence of user aborts). We observe that the lack of unfairness under SRPT in the realistic setup is not a consequence of a large number of incomplete large requests. The overall fraction of incomplete requests is only 0.2% for both FAIR and SRPT when load is 0.7, and ranges from 10–15% for both FAIR and SRPT when load is 0.9 (again excluding workloads W3 and W6). Looking only at the largest 1% of requests, the fraction of incomplete requests is much more variable, ranging from 3–25% under SRPT and 3–30% under FAIR, but is typically smaller under SRPT than under FAIR.

Finally, we observe that for both the baseline and the realistic setup, increasing the length of the overload period or the mean system load does *not* result in more starvation. This agrees with the theoretical M/GI/1 results in [Bansal and Harchol-Balter 2001b].

6. WHY DOES SRPT WORK?

In this section, we will look in more detail at where SRPT's performance gains come from and we explain why there is no starvation of long jobs.

6.1 Where Do Mean Gains Come From?

Recall that we identified in Section 4 and Section 5 three reasons for the poor performance of a standard (FAIR) server under overload:

(1) high queueing delays at the server due to high numbers of connections sharing the bandwidth (see Figures 6 and 7);
(2) drops of SYNs because of a full SYN queue;
(3) loss of packets inside the kernel.

SRPT alleviates all these problems. It reduces queueing delays by scheduling connections more efficiently. It has a lower number of dropped SYNs since it takes longer for the SYN-queue to fill. Finally and less obviously, an SRPT server also sees less loss inside the kernel. The reason is that subdividing the transmit queue into separate shorter queues allows SRPT to isolate the short requests, which are most requests. These short requests go to a queue which is drained more quickly, and thus experience no loss in their transmit queue.

The question we address in this section is how many of SRPT's performance gains can be attributed to solving each of the above three problems.

We begin by looking at how many of the performance improvements under SRPT stem from alleviating the SYN drop problem. In workload W1, used throughout the article, no incoming SYNs were dropped. Hence SRPT's improvement was not due to alleviating SYN drops. Workload W4 did exhibit SYN drops (1% in the baseline case and 20% in the realistic case). We eliminate the SYN-drop advantage by increasing the length of the SYN-queue to the point where SYN drops are eliminated. This only improves FAIR by less than 5% for the baseline case and 30% for the realistic case—not enough alone to account for SRPT's big improvement over FAIR.

The remaining question is how many of SRPT's benefits are due to reducing problem (1) (queueing delays) versus problem (3) (packet drops inside the kernel). Observe that problem (3) is mainly a problem because of the high initial

RTO. We can mitigate the effect of problem (3) by dropping the initial RTO to, say, 500 millisecond. The result is shown in Figure 10 (F). Observe that, even when problem (3) is removed, the improvement of SRPT over FAIR is still a factor of 7.[7]

We therefore conclude that problem (1) is the main reason why SRPT improves upon FAIR. By timesharing among many requests, the FAIR scheduling policy ends up slowing down all the requests.

6.2 Why Are Long Requests Not Hurt?

In this section, we give some insight as to why, in the case of Web workloads, SRPT does not unfairly penalize large requests.

To understand what happens under transient overload, first consider the overload period. Under FAIR, all jobs suffer during the overload period. Under SRPT, all jobs of size $< x$ complete and all other jobs receive no service, where x is defined such that the load comprised of jobs of size $< x$ equals 1. While it is true that jobs of size $> x$ receive no service under SRPT during the overload period, they also receive little service under FAIR during the overload period because the number of connections with which they must share under FAIR increases so quickly (see Figure 6). Next consider the low load period. At the start of the low load, there are many more jobs present under FAIR; only large jobs are present under SRPT. These large jobs have received zero service under SRPT and negligible service under FAIR until now. Thus, the large jobs actually finish at about the same time under SRPT and FAIR.

These effects are accentuated by heavy-tailed request sizes for two reasons: (1) Under heavy-tailed workloads, a very small fraction of requests make up half the load, and therefore, the fraction of large jobs ($> x$) receiving no service during overload under SRPT is very small; (2) The little service that large jobs receive during overload under FAIR is even less significant because the large jobs are so large that, proportionately, the service appears small.

6.3 Theoretical Validation

As a final step in understanding the performance of FAIR vs. SRPT, we consider an M/GI/1 queue with alternating periods of overload and low load and derive an approximation of the expected response time as a function of request size for this model under FAIR and SRPT. The derivation is too involved to include hereof, but the interested reader should look at [Bansal and Harchol-Balter 2001b] for full details. We choose to evaluate our results when the service-requirement distribution, G, is a Bounded-Pareto distribution with α-parameter of 1.1, as has been shown to be representative of Web workloads [Barford and Crovella 1998].

[7]Problem (3) may seem to be a design flaw in Linux that could be solved by either adding a feedback mechanism when writing to the transmit queue, or by increasing the length of the transmit queue. However, this will increase the queueing delays that a packet might experience. Our experiments show that increasing the transmit queue up to a point slightly decreases response time, but beyond that actually hurts response time.

Although the M/GI/1 queue is at best a rough approximation of our implementation setup, we nevertheless find the same trends in analysis as we have witnessed in this article. In particular, we find that the buildup in the number of requests is much greater under FAIR than under SRPT, and consequently, response times are also far higher under FAIR than under SRPT. Likewise we find that the server recuperates more slowly when load is dropped under FAIR than under SRPT. Also similar to our experiments, we find that in analysis the very largest requests see approximately the same mean response time under SRPT as compared with FAIR.

7. RELATED WORK

Our work is new in two ways: It is the first to apply connection scheduling to improve performance of Web servers under overload; and it is the first to provide an evaluation of the effects, of external factors (Table II) on the performance of an overloaded Web server.

Following we describe related work, grouped into three different categories: (1) work on improving the performance of (nonoverloaded) Web servers, since this work is copious, we limit the description to scheduling-related work; (2) work on improving the performance of Web servers under overload; (3) studies of Web server performance and the effects of external factors.

Scheduling Solutions—Non-Overloaded Case. We summarize related work on scheduling in Web servers. It is important to note that none of this prior work deals with overloaded servers. Moreover, most previous work uses simulation rather than a real implementation.

The only work that implements scheduling for Web servers rather than simulating or analytically evaluating it, is our own work [Crovella et al. 1999; Harchol-Balter et al. 2003] and recent work by [Rawat and Kshemkayani 2003]. In Crovella et al. [1999] we experiment with connection scheduling at the application level. Our experimental Web server improves mean response times, but at the cost of a drop in throughput by a factor of almost two. The problem is that application-level scheduling does not provide fine enough control over the order in which packets enter the network. In Harchol-Balter et al. [2003], we implement connection scheduling at the kernel-level. This eliminates the drop in throughput and offers much larger performance improvements than Crovella et al. [1999].

The authors in Rawat and Kshemkayani [2003] extend the work in Harchol-Balter et al. [2003] by proposing and implementing a scheduling algorithm, which takes into account both the size of the request and the distance of the client from the server. They show that this new policy can improve the performance of large-sized files by 2.5 to 10%.

In addition to implementation work, there are simulation studies of scheduling algorithms for Web servers [Gong and Williamson 2003; Murta and Corlassoli 2003]. Gong and Williamson [2003] identify two different types of unfairness: endogenous unfairness that a job may suffer because of its own size; and exogenous unfairness that a job suffers as a consequence of the other jobs it sees when it arrives. They then proceed to evaluate SRPT and other

policies with respect to these types of unfairness. Murta and Corlassoli [2003] develop and simulate an extension to SRPT scheduling called FCF (fastest connection first). Similar to Rawat and Kshemkayani [2003], FCF considers WAN conditions in addition to request size when making scheduling decisions.

Finally, Friedman and Henderson [2003] suggest a new protocol called FSP (fair sojourn protocol) for use in Web servers. They show through analysis and simulation that FSP always outperforms processor sharing. Their simulation results suggest that FSP performs better than SRPT for large requests, while the opposite is true for small requests. However, they are not able to quantify these effects analytically.

Our work is different from all the above in that it focuses on the behavior of Web servers under overload and is the first to show that size-based scheduling can be applied to overloaded servers without overly penalizing large requests.

Solutions for Web Servers Under Overload. There is a large body of work on systems to support high traffic Web sites. Most of this work focuses on improving performance by reducing the load on the overloaded devices in the system. This is typically done in one of five ways: increasing the capacity of the system (for example, by using server farms or multiprocessor machines [Colajanni et al. 1998; Dias et al. 1996]); using caches either on the client or on the server side [Gwertzman and Seltzer 1994; Braun and Claffy 1994; Bestavros et al. 1995] designing more efficient software both at the OS level [Mogul and Ramakrishnan 1996; Banga et al. 1999; Druschel and Banga 1996; Kaashoek et al. 1996] and the application level [Pai et al. 1999], admission control [Cherkasova and Phaal 1998; Iyer et al. 2000; Voigt et al. 2001; Voigt and Gunnigberg 2001; Welsh and Culler 2003], or deployment of content distribution networks [Day et al. 2002; Krishnamurthy et al. 2001; Johnson et al. 2000]. Other means of avoiding overload are content adaptation [Abdelzaher and Bhatti 1999] and offloading work to the client [Andresen and Yang 1997].

Our work differs significantly from all these approaches in that we do not attempt to reduce the load on the overloaded device. Rather, our goal is to improve the performance of the system while it is overloaded. To accomplish this goal, we employ SRPT scheduling.

Studies of Factors Affecting Web Server Performance. While there exist relatively few studies on servers running under overload, there are many studies of Web server performance in general (not overloaded). These studies typically concentrate on the effects that network protocols/conditions have on Web server performance. We list just a few here.

In Aron and Druschel [1999] and Brakmo and Peterson [1995], the authors find that the TCP RTO value has a large impact on server performance under FreeBSD. This agrees with our study.

In Nahum et al. [2001], the authors study the effect of WAN conditions and find that losses and delays can affect response times. They use a different workload from ours (a surge workload) but have similar findings.

The benefits of persistent connections are evaluated by Mogul [1995] and Barford and Crovella [1999] in a LAN environment.

There are also several articles which study real Web servers in action, rather than in a controlled lab setting, for example, Mogul [1995] and Seshan et al. [1998].

The work by Banga and Druschel [1999] studies a Web server under overload, but the only external factors considered are WAN delays. Moreover, this work focuses only on the capacity of the Web server, that is, the maximum sustainable throughput, and does not evaluate user experience or performance under fluctuating load.

8. CONCLUSION

The purpose of this article is to demonstrate the effectiveness of SRPT connection scheduling in improving the performance of Web servers during transient periods of overload.

We implement SRPT in an Apache Web server running on Linux by scheduling the bandwidth on the server's uplink. This is done by modifying the order that socket buffers are drained within the kernel. We find that SRPT significantly improves both server stability and client experience persistent as well as transient overload conditions. Under persistent overload, the number of connections at the FAIR server grows quickly compared with the buildup of connections under SRPT, and consequently, the FAIR server is also quick to reach the point where incoming SYNs are dropped. As a result the client experience, in terms of both mean response time and the time until the first byte is received, is greatly improved under the SRPT server compared to the FAIR server. With respect to transient overload, we find that SRPT improves mean response times by factors of 1.5—8 over traditional FAIR scheduling, across ten different transient overload workloads. This is significant, since mean response times under FAIR can get quite high and peak response time are many-fold higher than mean response times.

Performance improvements are measured under a vast range of environmental factors, including a range of RTT's, loss rates, RTO's, persistent connections, user behaviors, packet sizes, and Web server configurations. Results are consistent across different traces. Importantly, we find that requests for large files are not penalized by SRPT scheduling under transient overload. In fact, for the largest 1% of requests the response time under SRPT is very close to that under FAIR.

We conclude by noting the broader general applicability of this work. First, the approximation of SRPT implemented in this article actually covers an entire family of algorithms, rather than one particular algorithm. The kernel-level implementation we have chosen is limited to a fixed number of priority classes, where we pick the size cutoffs between them so as to approximate SRPT as closely as possible. Instead one could also use more conservative cutoffs and/or fewer priority classes to achieve performance closer to a FAIR system, including better performance for very large requests.

Second, while this article focuses on static Web workloads where the bottleneck resource to be scheduled is the bandwidth on the uplink, we believe that the basic principles will extend to many other scenarios. We are currently

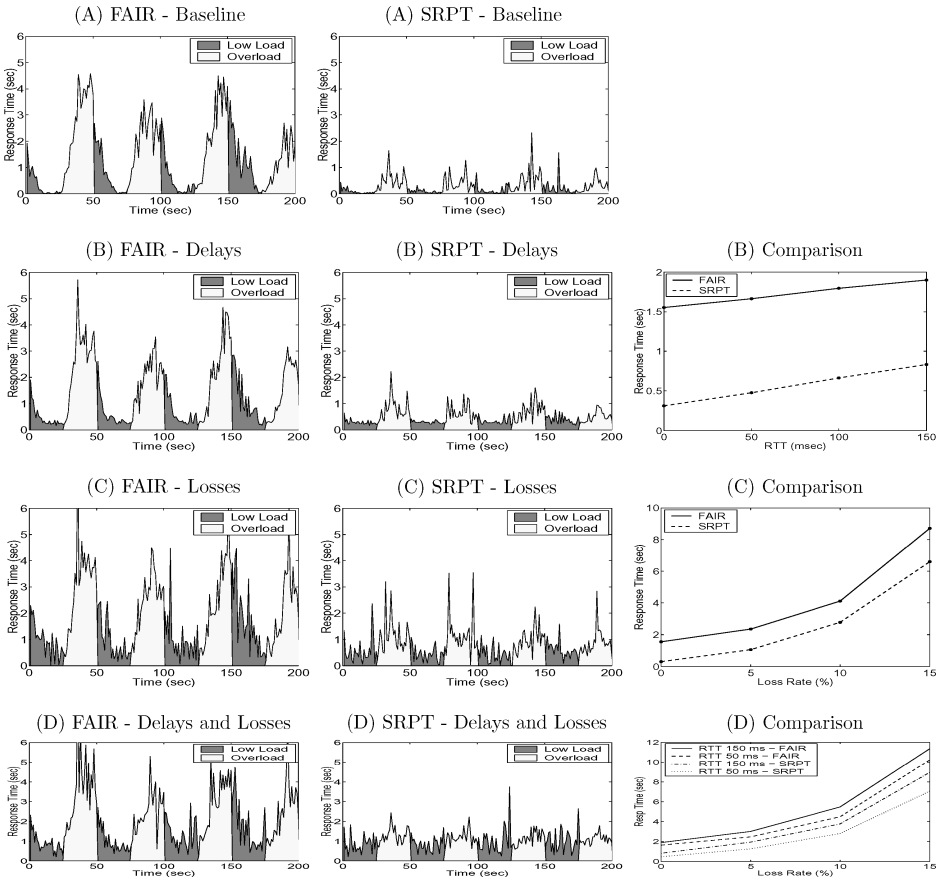


Fig. 13. Results under the NASA trace log for setups (E) to (F) from Table II. The left and middle columns show the response times over time for the specific values given in Table II, column 3. The right column evaluates the range of values given in the column 4 of Table II.

investigating scheduling of Web server backends serving *dynamic content* involving accesses to a database backend. Here, our results (see [McWherter et al. 2004]) show that, for database management systems (DBMS) with two-phase locking, a transaction's life is dominated by the time spent waiting in the internal database lock queues. Thus, rather than scheduling bandwidth as we have in this article, for applications with dynamic content we instead apply prioritized scheduling policies at the internal DBMS lock queues. However, many of the same ideas from this article apply.

The Web is in a perpetual state of evolution with new techniques for improving performance being developed every day. When proposing a new solution for reducing Web server response times, one therefore needs to consider how this solution may interact with other existing or future techniques. While we have evaluated the effects of many parameters on the effectiveness of our proposed solution, we have not been able to evaluate all parameters. In particular, the effects of caches or CDNs are left for future work.

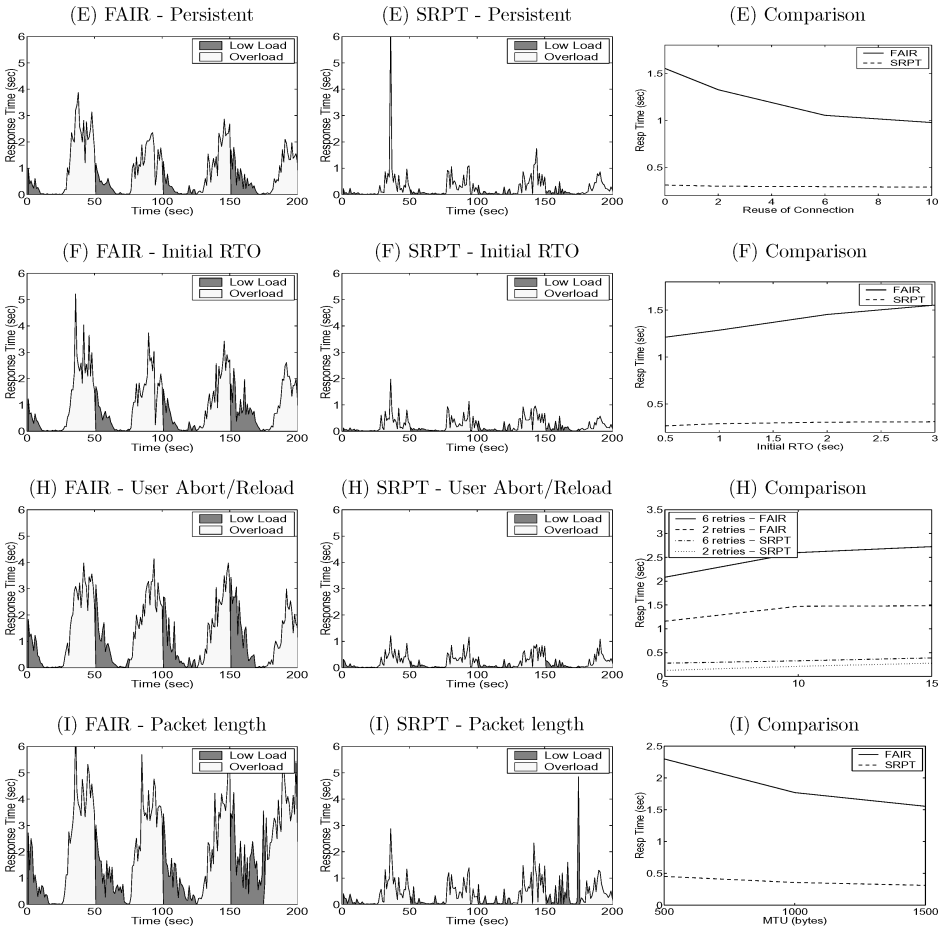
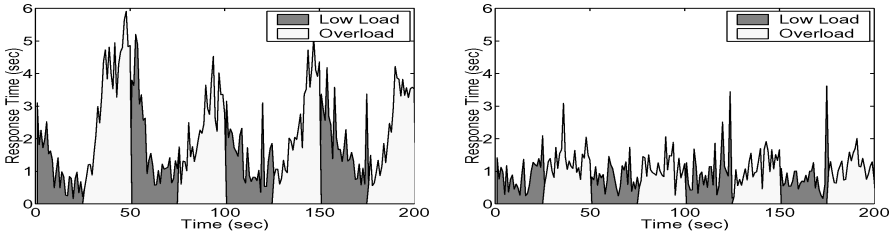


Fig. 14. Results under the NASA trace log for setups (E) to (F) from Table II. The left and middle columns show the response times over time for the specific values given in Table II, column 3. The right column evaluates the range of values given in the column 4 of Table II.

We believe that the scheduling ideas presented in this work are general enough to be used in conjunction with other methods, or incorporated into systems other than Web servers. We have already witnessed in this article how SRPT scheduling can be used in conjunction with persistent connections. Likewise it is probable that SRPT-like scheduling can be used in conjunction with caching schemes at a proxy server or can be used at an edge router. For example, Rai et al. [2003] have recently proposed least-attained-service (LAS) scheduling for routers, where LAS is a variant of SRPT scheduling that doesn't require a priori knowledge of the service demand.

APPENDIX: Another Trace

In this appendix, we consider one more trace and run all experiments on the new trace. We find that the results are very similar to those shown in the



(J) FAIR - Realistic case (J) SRPT - Realistic case

Fig. 15. Comparison of FAIR (left) and SRPT (right) in realistic setup for workload W1, under the NASA trace log.

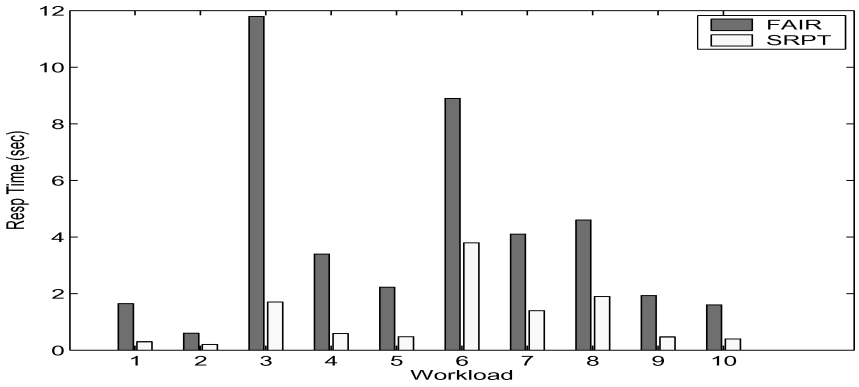


Fig. 16. Mean response time comparison of FAIR and SRPT in the baseline case for the workloads in Table I, under the NASA trace log.

body of the article, both with respect to comparative mean response times for the different scenarios (A) through (J) and with respect to unfairness issues.

The log used here was collected from a NASA Web server and is also available through the Internet Traffic Archive. We use several hours on one busy day of this log consisting of around 100,000 mostly static requests. The minimum file size is 50 bytes, the maximum file size is 1.93 Mbytes. The largest 2.5% of all requests makes up 50% of the total load, exhibiting a strong heavy-tailed property. The primary statistical difference between the NASA log and the Soccer World Cup log (used in body of the article) is the mean request size: the NASA log shows a mean file size of 19 Kbytes while for the World Cup log it was only around 5 Kbytes.

Results for the NASA log are shown in Figures 13, 14, 15, and 16. They are extremely similar to the corresponding Figures 9, 10, 11, and 12 for the World Cup trace. The only difference is that response times are higher under the NASA log compared with the World Cup log for both FAIR and SRPT. The relative performance gains of SRPT and FAIR are similar. The increase in response times under the NASA log may be attributed to the higher mean file size.

ACKNOWLEDGMENTS

Thanks to Srini Seshan for help with porting dummynet to Linux; to Christos Gkantsidis for providing detailed information on `libwww`; to Mukesh Agrawal for helping with initial overload experiments; and to Jennifer Rexford, John Wilkes, and Erich Nahum for proofreading the article.

REFERENCES

ABDELZAHER, T. F. AND BHATTI, N. T. 1999. Web content adaptation to improve server overload behavior. *WWW8/Comput. Networks 31,* 11-16, 1563–1577.

ALMESBERGER, W. Linux network traffic control—implementation overview. Available at http://lrcwww.epfl.ch/linux-diffserv/.

ANDRESEN, D. AND YANG, T. 1997. Multiprocessor scheduling with client resources to improve the response time of WWW applications. In *the International Conference on Supercomputing*. 92–99.

ARLITT, M. AND JIN, T. 2000. Workload characterization of the 1998 world cup Web site. *IEEE Network 14,* 3, 30–37.

ARON, M. AND DRUSCHEL, P. 1999. TCP implementation enhancements for improving webserver performance. Tech. Rep. TR99-335, Rice University.

BANGA, G. AND DRUSCHEL, P. 1999. Measuring the capacity of a web server under realistic loads. *World Wide Web 2,* 1-2, 69–83.

BANGA, G., DRUSCHEL, P., AND MOGUL, J. C. 1999. Resource containers: A new facility for resource management in server systems. In *Proceedings of OSDI '99*. 45–58.

BANSAL, N. AND HARCHOL-BALTER, M. 2001a. Analysis of SRPT scheduling: Investigating unfairness. In *Proceedings of ACM SIGMETRICS '01*.

BANSAL, N. AND HARCHOL-BALTER, M. 2001b. Scheduling solutions for coping with transient overload. Tech. rep. CMU-CS-01-134, Carnegie Mellon University.

BARFORD, P. AND CROVELLA, M. E. 1998. Generating representative Web workloads for network and server performance evaluation. In *Proceedings of SIGMETRICS '98* (July) 151–160.

BARFORD, P. AND CROVELLA, M. E. 1999. A performance evaluation of hyper text transfer protocols. In *Proceedings of ACM SIGMETRICS '99* (May) 188–179.

BENDER, M., CHAKRABARTI, S., AND MUTHUKRISHNAN, S. 1998. Flow and stretch metrics for scheduling continuous job streams. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*.

BERNSTEIN, D. J. 1997. Syn cookies. http://cr.yp.to/syncookies.html.

BESTAVROS, A., CARTER, R. L., CROVELLA, M. E., CUNHA, C. R., HEDDAYA, A., AND MIRDAD, S. A. 1995. Application-level document caching in the Internet. In *Proceedings of the 2nd International Workshop on Services in Distributed and Networked Environments (SDNE'95)* (June).

BRAKMO, L. AND PETERSON, L. 1995. Performance problems in 4.4 BSD TCP. *ACM Comput. Comm. Rev. 25,* 5.

BRAUN, H. AND CLAFFY, K. 1994. Web traffic characterization: An assessment of the impact of caching documents from NCSA's Web server. In *Proceedings of the 2nd International WWW Conference*.

CHERKASOVA, L. AND PHAAL, P. 1998. Session-based admission control: A mechanism for improving the performance of an overloaded web server. Tech. Rep. HPL-98-119, Hewlett Packard Laboratories.

COCKCROFT, A. 1996. Watching your web server. The Unix Insider at http://www.unixinsider.com (April).

COLAJANNI, M., YU, P. S., AND DIAS, D. M. 1998. Analysis of task assignment policies in scalable distributed Web-server systems. *IEEE Trans. Parall. Distrib. Syst. 9,* 6, 585–699.

Cooperative Association for Internet Data Analysis (CAIDA). 1999. Packet length distributions. http://www.caida.org/analysis/AIX/plen_hist.

CROVELLA, M., FRANGIOSO, R., AND HARCHOL-BALTER, M. 1999. Connection scheduling in web servers. In *USENIX Symposium on Internet Technologies and Systems* (Oct).

DAY, M., CAIN, B., TOMLINSON, G., AND RZEWSKI, P. 2002. A model for content internetworking (cdi). Internet Draft (draft-ietf-cdi-model-02.txt) (May).

DIAS, D. M., KISH, W., MUKHERJEE, R., AND TEWARI, R. 1996. A scalable and highly available web server. In *COMPCON*. 85–92.

DRUSCHEL, P. AND BANGA, G. 1996. Lazy receiver processing (LRP): A network subsystem architecture for server systems. In *Proceedings of OSDI '96* (Oct) 261–275.

ELNIKETY, S., NAHUM, E. M., TRACEY, J., AND ZWAENEPOEL, W. 2004. A method for transparent admission control and request scheduling in dynamic e-commerce Web sites. In *International World-Wide Web Conference (WWW'04)* (May) New York, NY.

FELDMANN, A. Web performance characteristics. IETF (Nov). http://www.research.att.com/anja/feldmann/papers.html.

FRIEDMAN, E. J. AND HENDERSON, S. G. 2003. Fairness and efficiency in web server protocols. In *Proceedings of the ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems* (May).

GONG, M. AND WILLIAMSON, C. 2003. Quantifying the properties of SRPT scheduling. In *Proceedings of IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*.

GWERTZMAN, J. AND SELTZER, M. 1994. The case for geographical push-caching. In *Proceedings of HotOS '94* (May).

HARCHOL-BALTER, M., SCHROEDER, B., AGRAWAL, M., AND BANSAL, N. 2003. Size-based scheduling to improve web performance. *ACM Trans. Comput. Syst. 21,* 2 (May).

Internet Town Hall. The Internet traffic archives. Available at `http://town.hall.org/Archives-/pub/ITA/`.

Internet Traffic Report. 2004. http://www.internettrafficreport.com.

IRCache Home. 2004. The trace files. http://www.ircache.net/Traces/.

IYER, R., TEWARI, V., AND KANT, K. 2000. Overload control mechanisms for web servers. In *Workshop on Performance and QoS of Next Generation Networks* (Nov).

JOHNSON, K. L., CARR, J. F., DAY, M. S., AND KAASHOEK, M. F. 2000. The measured performance of content distribution networks. In *Proceedings of the 5th International Web Caching and Content Delivery Workshop*.

KAASHOEK, M., ENGLER, D., WALLACH, D., AND GANGER, G. 1996. Server operating systems. In *SIGOPS European Workshop '96*. 141–148.

KRISHNAMURTHY, B. AND REXFORD, J. 2001. *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement*. Addison-Wesley.

KRISHNAMURTHY, B., WILLS, C., AND ZHANG, Y. 2001. The use and performance of content distribution networks. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement.*

KUROSE, J. AND ROSS, K. W. 2001. *Computer Networking: A Top-Down Approach Featuring the Internet*. Addison Wesley Longman, Inc.

LEFEBVRE, W. 2002 Cnn.com: Facing a world crisis. *The USENIX Technical Conference*, (June).

MAGGS, B. VICE PRESIDENT OF RESEARCH, 2001. Akamai Technologies. Personal communication.

MANLEY, S. AND SELTZER, M. 1997. Web facts and fantasy. In *Proceedings of the 1997 USITS*.

MCWHERTER, D., SCHROEDER, B., AILAMAKI, A., AND HARCHOL-BALTER, M. 2004. Priority mechanisms for OLTP and transactional web applications. In *the 20th International Conference on Data Engineering (ICDE'04)*.

Microsoft TechNet Insights AND ANSWER FOR IT PROFESSIONAL 2001. The arts and science of web server tuning with internet information services 5.0. http://www.microsoft.com/technet/.

MOGUL, J. C. 1995. The case for persistent-connection HTTP. In *Proceedings of ACM SIGCOMM '95* (Oct). 299–313.

MOGUL, J. C. 1995. Network behavior of a busy Web server and its clients. Tech. Rep. 95/5, Digital Western Research Laboratory (Oct).

MOGUL, J. C. AND RAMAKRISHNAN, K. K. 1996. Eliminating receive livelock in an interrupt-driven kernel. In *Proceedings of USENIX Technical Conference*. 99–111.

MURTA, C. D. AND CORLASSOLI, T. P. 2003. Fastest connection first: A new scheduling policy for web servers. In *Proceedings of the 18th International Teletraffic Congress (ITC-18)* (Sept).

NAHUM, E., ROSU, M., SESHAN, S., AND ALMEIDA, J. 2001. The effects of wide-area conditions on www server performance. In *Proceedings of ACM SIGMETRICS'01*. 257–267.

National Institute Standards and Technology. Nistnet. `http://snad.ncsl.nist.gov/itg/nistnet/`.

PADHYE, J., FIROIU, V., TOWSLEY, D. F., AND KUROSE, J. F. 2000. Modeling tcp reno performance: A simple model and its empirical validation. *IEEE/ACM Trans. Netw. 8,* 2, 133–145.

PAI, V. S., DRUSCHEL, P., AND ZWAENEPOEL, W. 1999. Flash: An efficient and portable web server. In *Proceedings of USENIX 1999* (June).

PAXSON, V. AND ALLMAN, M. 2000. Computing TCP's retransmission timer. RFC 2988, `http://www.faqs.org/rfcs/rfc2988.html`.

RAI, I. A., URVOY-KELLER, G., AND BIERSACK, E. 2003. Analysis of LAS scheduling for job size distributions with high variance. In *Proceedings of the ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)* (June).

RAWAT, M. AND KSHEMKAYANI, A. 2003. SWIFT: Scheduling in web servers for fast response time. In *the 2nd IEEE International Symposium on Network Computing and Applications* (April).

RIZZO, L. 1997. Dummynet: A simple approach to the evaluation of network protocols. *ACM Comput. Commun. Rev. 27,* 1.

SCHRAGE, L. E. AND MILLER, L. W. 1966. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Resear. 14*, 670–684.

SESHAN, S., BALAKRISHNAN, H., PADMANABHAN, V. N., STEMM, M., AND KATZ, R. 1998. TCP behavior of a busy internet server: Analysis and improvements. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*. 252–262.

SILBERSCHATZ, A., GALVIN, P., AND GAGNE, G. 2002. *Operating System Concepts, 6th Ed*. John Wiley & Sons.

STALLINGS, W. 2001. *Operating Systems, 4th Ed*. Prentice Hall.

The World Wide Web Consortium (W3C). Libwww—the W3C protocol library. `http://www.w3.org`.

VOIGT, T. AND GUNNIGBERG, P. 2001. Kernel-based control of persistent web server connections. *ACM SIGMETRICS Performance Evaluation Review 29,* 2, 20–25.

VOIGT, T., TEWARI, R., FREIMUTH, D., AND MEHRA, A. 2001. Kernel mechanisms for service differentiation in overloaded web servers. In *Proceedings of the USENIX Annual Technical Conference* (June) Boston, MA.

WELSH, M. AND CULLER, D. 2003. Adaptive overload control for busy internet servers. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*.

WIERMAN, A. AND HARCHOL-BALTER, M. 2003. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *ACM Sigmetrics International Conference on Measurement and Modeling of Computer Systems*.

Received June 2004; revised December 2004, January 2005; accepted January 2005