Asymptotic Evaluation of Delay in the SRPT Scheduler

Chang Woo Yang and Sanjay Shakkottai

Abstract

In this paper, we consider the Shortest-Remaining-Processing-Time (SRPT) scheduling algorithm. We consider the SRPT scheduling rule for a discrete-time queueing system that is accessed by a large number of flows (a many flows regime). In such an asymptotic regime (large capacity and large number of flows), we derive expressions for the packet delay distributions for batch arrival processes, and with bounded packet sizes.

Using these results, we compare the delay asymptote (i.e., for any finite delay, and asymptotic in the number of flows) of the SRPT scheduler with that of a FIFO (First-In-First-Out) scheduler, when there is a mix of packet sizes. Our analysis holds for any finite mix of packet sizes. We apply the result to a system accessed by packets which are one of two sizes: 1 or M, and the arrival process is i.i.d. across flows. We show that the *difference* in rate function of the delay asymptote between SRPT and FIFO for *the size M packet* decays as $O(\frac{1}{M^\gamma})$ for any $0 < \gamma < 1$ and M sufficiently large. Thus, *for large packets, the delay distributions under FIFO and SRPT look similar.* On the other hand, for the *size 1 packet*, the delay rate function under SRPT is *invariant with M.* However for FIFO, the delay rate function for the *size 1 packet decays as* $O(\frac{1}{M^\gamma})$ *for any* $0 < \gamma < 1$ *and M large.* This shows that for size 1 packets, SRPT performs increasingly better as the range in packet size increases. Thus, these results indicate that SRPT is a good policy to implement for web-servers, where empirical evidence suggests a large variability in packet sizes.

Key words: Large deviations, Rate function, queueing system, SRPT, FIFO

I. Introduction

With web services becoming increasingly popular, today's web servers handle loads that could range from hundreds to thousands of simultaneous connections. These connections request file

This research has been supported by NSF Grants ACI-0305644, CNS-0325788, and CNS-0347400. The authors are with the Wireless Networking and Communications Group (WNCG), Department of Electrical and Computer Engineering, The University of Texas at Austin, {cyang,shakkott}@ece.utexas.edu.

transfers that range from a small download of a few kilobytes to very large downloads (tens of megabytes), and exhibit heavy tail behavior [1]. These requests are served by the web server by means of a scheduler, which "prioritizes" the requests, possibly based on request arrival times as well as request sizes.

There have been many scheduling policies that have been proposed and implemented. Among these policies, it has been long known that Shortest-Remaining-Processing-Time (SRPT) exhibits the minimum mean delay. Despite this advantage, most of the existing scheduling policies opt toward simpler policies such as First-In-First-Out (FIFO), or fairness oriented schedulers where the resources are equally shared among all connections. The main reason behind the lack of attention to SRPT is that it is believed that in the process of optimizing the mean delay, fairness among files requests[1] of different sizes might suffer. More specifically, it is believed that larger file requests will "starve" under the SRPT scheduling policy [16]. Intuitively this looks obvious – by giving priority to smaller file requests, the delay experienced by larger file requests will increase, thus leading to unfairness.

However, recent results have shown that this intuition is not necessarily correct. In [2], the authors show that under arrivals with heavy-tailed properties, the unfairness is quite small. The authors compare SRPT with the Processor-Sharing (PS) scheduling policy. Using an M/G/1 queuing model and heavy tail arrival characteristics, the authors have shown that regardless of the load, at most 1% of the packets have bigger expected response time under SRPT than under PS. Further, the authors provide an upper bound on how much worse the SRPT performs in the context of expected response time when compared to the PS scheduling discipline. Motivated by these results, the authors have implemented an SRPT scheduler for web servers [10].

Such observations show that SRPT with its optimal mean delay characteristics is a promising alternative to the prevalent scheduling policies. To realize its full potential, delay analysis for packets will be necessary in order to characterize the behavior of this policy. In this paper, we focus on a probabilistic approach which captures statistical multiplexing effects as well. More specifically, we will be looking at probabilities of the packet delay to exceed some threshold, when the web server is accessed by a large number of requests, using a large deviations

[1]In this paper, we use the words 'file requests' and 'packets' interchangeably, the reason being that file requests correspond to 'packets' in a queue with the SRPT discipline.

formulation.

A large deviations framework provides asymptotic results on delays distributions. These asymptotes have been based on either the large buffer framework [3], [7], [13], or the many sources framework [5], [6], [12]. The *large buffer* framework deals with a single queue and single arrival, and the overflow probability is derived as the buffer size goes to infinity. On the other hand, *the many sources regime* appropriately scales the service capacity with the number of arrivals. The asymptote is formulated as the scaling factor goes to infinity. Thus, a many sources asymptote can be used to study the probability that the delay exceeds any fixed value, in the regime where the number of flows are large. Given the number of requests for current web servers, many sources large deviations seems to be an appropriate regime to study the SRPT scheduler.

Related works includes [4] where SRPT system is analyzed in the large buffer regime, and [11] where it has been shown that various scheduling policies including SRPT and PS have the same *expected* slowdown[2] for the largest sized packets. However, there does not seem to be any corresponding work in the many sources regime. This paper focuses on the many sources regime, and compares the SRPT scheduler with a FIFO scheduler.

A. Main Contributions

(i) In Section IV, we derive the packet delay asymptotes of SRPT in the large number of flows regime.

(ii) Using these results, we compare the delay asymptote (i.e., for any finite delay, and asymptotic in the number of flows) of the SRPT scheduler with that of a FIFO (First In First Out) scheduler, when there is a mix of packet sizes. Our analysis holds for any finite mix of packet sizes. Next, we apply the analysis to a system accessed by packets which are one of two sizes: 1 or M, with mutually independent arrival processes. We show that the *difference* in rate function of the delay asymptote between SRPT and FIFO for *the size M packet* decays as $O(\frac{1}{M^\gamma})$ for any $0 < \gamma < 1$ and M large. Thus, *for large packets, the delay distributions under FIFO and SRPT look similar.* On the other hand, for the *size 1 packet,* the delay rate function under SRPT is *invariant with M*. However for FIFO, the delay distributions for the *size 1 packet decays as* $O(\frac{1}{M^\gamma})$ *for some* $0 < \gamma < 1$ *and M large.*

[2]Slowdown is defined as delay divided by the packet size.

This indicates that for size 1 packets, SRPT performs increasingly better as the range in packet size increases. Thus, our results complement the findings in [2], and indicate that SRPT is a good policy to implement for web-servers where empirical evidence suggests a large variability in packet sizes [1].

II. System Description

Let us consider a queueing system with a single queue and a single server. The system operates in discrete time, i.e., batch of packets arrive at the beginning of each time slot and packets are serviced at the end of the slot. The queue state is measured immediately after the service, and just before the arrivals of the next time-slot. The server services the packets with the least remaining-processing-time first. Thus, in each time-slot, a batch of packets are served, and partially served packets (if a packet cannot be completely served within a time-slot) are stored to be served in future time-slots. We comment that a partially served packet does not necessarily receive service in the next consecutive time-slot. This is because smaller sized packets could arrive in the next time-slot, thus preempting the partially served packet. Finally, we also assume that the size of the packets are restricted to multiples of a unit size and the set of possible packet sizes is $\mathscr{M} = \{1,2,3,\ldots,M\}$.

For each packet size $k \in \mathscr{M}$, we assume that N independent and identically distributed stationary, ergodic arrival processes access the queue. Thus, packet arrivals from a single stream of any given size can be correlated across time-slots. However, packets arrivals of different sizes form independent arrival processes. Equivalently, this model can represent a packet arrival process that can be decomposed into independent arrival processes according to the packet size. We define $A^N(a,b)$ as the total volume of arrivals by all the N arrival processes in the time-interval[3] (a,b). We define $A_k^N(a,b)$ as the total volume of packets of size k that arrive in the queue during time-interval (a,b) [4]. Thus, we have

$$A^N(a,b) \;=\; \sum_{k=1}^{M} A_k^N(a,b). \tag{1}$$

[3]The notation (a,b) corresponds to the time-slots $\{a,a+1,\ldots,b\}$.

[4]In other words, $A_k^N(a,b)$ corresponds to k times the number of size k arrivals.

Since we operate in many sources framework, we assume that the capacity of the server is scaled in proportion to the load, and at most NC unit data can be service at any time slot. We assume that the queueing system is stable, i.e., $E[A^N(0,0)] < NC$.

The *virtual delay* $V(0)$ that will be the focus of this paper is the delay seen by a fictitious packet that arrives at the end of a burst which arrives at $t = 0$, with the assumption that the system started at $t = -\infty$. The event $\{V(0) > m\}$ corresponds to the situation where a fictitious packet that arrives at the end of the batch of arrivals at time slot 0 did not depart the system until mth time slot. We define the rate function of the virtual delay as

$$I_V(m) = \lim_{N\to\infty} -\frac{1}{N}\log \Pr(V(0) > m).$$

Similar to [15], we can relate the virtual delay to the actual delay, but we skip the details here for simplicity.

III. Large Deviation

In this section we briefly describe the framework of large deviations, and state the assumption made in the paper. Large deviation quantifies the probability of a rare event occurring under some traffic mix and traffic resource restrictions. This paper develops the probability of delay exceeding some value when multiple arrival process access a single queue. In such a situation, the large deviation result state that under general conditions, the probability of a rare event (delay exceeding some value m) decays as $O(e^{-NI(m)})$ in the large N regime. The probability of delay exceeding some value decays exponentially in the asymptotic sense. The exponential "trend" of the probability, i.e., $I(m)$, is the rate function.

Formally, for any sequence of events $\mathbf{H}^N$, we define

$$\mathbf{J}(\mathbf{H}) = \lim_{N\to\infty} -\frac{1}{N}\log \Pr(\mathbf{H}^N). \tag{2}$$

We consider the rate function [14] of the random variables $\{A_k^N(-T,0)\}$, which is defined as follows. For each $\theta \in \mathscr{R}$, and for $k \in \{1,2,\ldots,M\}$, let us define

$$\Lambda_{A_k,T}(\theta) = \lim_{N\to\infty} \frac{1}{N}\log E(e^{\theta A_k^{(N)}(-T,0)}), \tag{3}$$

$$I_{A_k,T}(x) = \sup_{\theta} \left(\theta x - \Lambda_{A_k,T}(\theta)\right). \tag{4}$$

We assume the following about the event concerning arrival processes of size $k \in \{1,\ldots,M\}$ packets, see also [15].

Assumption 1: Fix any $\varepsilon > 0$, such that $E[\sum_{i=1}^{M} A_i^N(0,0)] < N(C-\varepsilon)$, and define

$$\mathbf{H}_T^{\varepsilon} = [\sum_{i=1}^{k-1} A_i^N(-T,m) + A_k^N(-T,0) > N(C-\varepsilon)(T+m+1)]. \tag{5}$$

We define $T^* = \arg \ inf_T \mathbf{J}(\mathbf{H}_T^{\varepsilon})$. The existence of T^* comes from the stability condition and we assume that it is unique. Furthermore, assume $\mathbf{J}(\mathbf{H}_T^{\varepsilon})$ satisfies the following condition.

$$\liminf_{T\to\infty} \frac{\mathbf{J}(\mathbf{H}_T^{\varepsilon})}{\log T} = \omega > 0 \tag{6}$$

$\mathbf{H}_T^{\varepsilon}$ is the event that the sum of the total arrivals (of all sizes $\leq k$) over the time interval $(-T,0)$ and the arrivals due to sizes $\{1,\ldots,k-1\}$ packets over the interval $(1,m)$ exceeds the service capacity of a system with capacity $N(C-\varepsilon)$ per time slot. The event, $\mathbf{H}_T^{\varepsilon}$ corresponds a possible "mode" by which a packet of size k can be delayed by more than a time-interval m.

IV. Delay Asymptote of SRPT

In this section, we derive the virtual delay asymptote of the SRPT system. The system in consideration is the system described in section II. We derive the virtual delay asymptote of the event that queueing delay exceeds the value m for packet size k, where $k \in \{1,\ldots,M\}$. First, we make the following observations which identifies an equivalent model to SRPT.

Consider a priority queueing system, where packets of different size are queued separately, and queues are assigned based on the packet size. Queues that are assigned to smaller sized packets have higher priority, while the queues that are assigned to larger sized packets have lower priority. For example, the queue that corresponds to size l packets (henceforth referred to as queue-l) has priority l, where queue-1 has the highest priority.

We make the observation that the operation of SRPT is equivalent to that of priority queueing system described above but with switching of packets during time slots. The essential operation of SRPT scheduling policy grants packets with less remaining service time higher priority compared to packets with larger remaining service time. Packets that have not been completely serviced in a time-slot will receive higher priority in future time-slot depending on the residual file size. Thus, the SRPT discipline can be interpreted as a priority queueing system, where the queue are not assigned by the packet size but the residual packet size. SRPT is equivalent to the priority queueing systems, but with partially served packets changing priority levels at the end of the time-slot. Note that an important difference from the priority queueing system from SRPT is

that, in the priority queueing discipline a partially served packet continues to remain in the same queue that it was originally in, and its priority level *does not change* in future time-slots.

Define $\overline{B}_k^N(a,b)$ as the volume of potential service (i.e., k times the number of packets that can be served) that the packets in queue-k can receive in interval (a,b) under SRPT. Potential service corresponds to the maximum amount of service that can be received if the corresponding queue is not empty. Note that size k packets in the queue is the result of arrivals of packets of size k packets that have not yet received any service, and the partially serviced packets (with an original size $> k$) that have remaining packet size of k.

Based on this observation, we next show that the rate function for SRPT is bounded by the corresponding rate functions of priority queues, and use results on priority queues from [15] to complete the proof.

Define $I_{\overline{V}}^{(k)}(m)$ as the rate function of the virtual delay of a size k packet under SRPT with total service rate NC. Similarly, let $I_{V_\mu}^{(k)}(m)$ denote the rate function of the virtual delay of a size k packet with the priority queueing system, and with total service rate $N\mu$.

Theorem 1: Fix any $\varepsilon > 0$. Then, for $k \in \{1,\ldots,M\}$ we have

$$I_{V_{C-\varepsilon}}^{(k)}(m) \leq I_{\overline{V}}^{(k)}(m), \tag{7}$$

$$I_{\overline{V}}^{(k)}(m) \leq I_{V_{C+\varepsilon}}^{(k)}(m). \tag{8}$$

Proof: First, we derive (7). Observe that if the virtual delay exceeds m, then we have that the queue length at time zero (i.e. $Q_k(0)$) over the time interval (1,m) is not served by time m [5]. In other words, $\{V(0) > m\} \subset \{Q_k(0) > \overline{B}_k^N(1,m)\}$, and consequently

$$\Pr(V(0) > m) \leq \Pr(Q_k(0) > \overline{B}_k^N(1,m)). \tag{9}$$

From Loyne's formula, we have

$$\Pr(Q_k(0) > \overline{B}_k^N(-T,m)) = \Pr(\sup_{T\geq 0}[A_k^N(-T,0) + \overline{S}_k^N(-T,0) - \overline{B}_k^N(-T,m)] \geq 0), \tag{10}$$

where $\overline{S}_k^N(-T,0)$ is the volume of arrivals to queue-k due to partially served packets arriving from lower priority queues in interval $(-T,0)$.

[5]The unit of the queue length is the volume of data. Thus, for queue-k, $Q_k(0)$ denotes k times the number of size k packets in the queue.

Let $-T^*$ be the first time in the past such that $Q_k(-T^*-1)=0$. Without loss of generality, we can show that $Q_l(-T^*-1)=0$, for all $l \le k$. (the proof is provided in Theorem 4.1 in [15] in the context of priority queues) Hence, we have

$$\begin{aligned}&\Pr(\sup_{T\ge 0}[A_k^N(-T,0)+\overline{S}_k^N(-T,0)-\overline{B}_k^N(-T,m)]\ge 0)\\ &=\Pr(A_k^N(-T^*,0)+\overline{S}_k^N(-T^*,0)-\overline{B}_k^N(-T^*,m)\ge 0).\end{aligned} \tag{11}$$

Since all queues-$\{1,\ldots,k\}$ are empty at time slot $-T^*-1$, the service available to queue-k in the interval $(-T^*,m)$ is lower bounded by the residual service after all arrivals to queues-$\{1,\ldots,k-1\}$ and arrivals to queues with priority higher than k generated by partially served packets are served in the same interval, i.e.,

$$\overline{B}_k^N(-T^*,m) \ge NC(T^*+m+1)-\sum_{i=1}^{k-1}A_i^N(-T^*,m)-(T^*+m+1)(k-1). \tag{12}$$

Note that in (12), the term $(T^*+m+1)(k-1)$ accounts for the worst case scenario where at every time slot in $(-T^*,m)$, a partially served packet arrives at queue-$(k-1)$ from queues with higher priority than k.

From (9), (10), (11), (12), and the fact that $\overline{S}_k^N(-T^*,0)\le k(T^*+1)$, we have

$$\begin{aligned}\Pr(V(0)>m) &\le \Pr[A_k^N(-T^*,0)+\overline{S}_k^N(-T^*,0)-\overline{B}_k^N(-T^*,m)>0]\\ &\le \Pr[A_k^N(-T^*,0)+\sum_{i=1}^{k-1}A_i^N(-T^*,m)-NC(T^*+m+1)\\ &\quad +(T^*+1)k+(T^*+m+1)(k-1)>0]\\ &\le \Pr[A_k^N(-T^*,0)+\sum_{i=1}^{k-1}A_i^N(-T^*,m)-NC(T^*+m+1)+2(T^*+m+1)k>0]\\ &\le \Pr[A_k^N(-T^*,0)+\sum_{i=1}^{k-1}A_i^N(-T^*,m)-N(C-\frac{2k}{N})(T^*+m+1)>0]\\ &\le \Pr[\bigcup_{T\ge 0}(A_k^N(-T,0)+\sum_{i=1}^{k-1}A_i^N(-T,m)-N(C-\frac{2k}{N})(T+m+1)>0)]\\ &\le \sum_{T\ge 0}\Pr[A_k^N(-T,0)+\sum_{i=1}^{k-1}A_i^N(-T,m)-N(C-\frac{2k}{N})(T+m+1)>0].\end{aligned} \tag{13}$$

Fix any $\varepsilon>0$. Observe that for N large enough, we have $(C-\frac{2k}{N})>(C-\varepsilon)$. Hence

$$\Pr(V(0)>m) \le \sum_{T\ge 0}\Pr[A_k^N(-T,0)+\sum_{i=1}^{k-1}A_i^N(-T,m)-N(C-\varepsilon)(T+m+1)>0]. \tag{14}$$

Note that (14) is the same expression for the rate function of size k packets in priority queues but for capacity $C-\varepsilon$. (see [8], [12]) Using similar techniques as in [8], [12], it follows that the lower bound of the rate function of $\Pr(Q_k(0) > \overline{B}_k^N(1,m))$ is $I^{(k)}_{V_{C-\varepsilon}}(m)$.

Specifically, by applying the contraction principle, the closed form expression of $I^{(k)}_{V_{C-\varepsilon}}(m)$ is

$$\inf_{T\geq 0} \Big[\inf_{\alpha(T+m+1)\leq x\leq (C-\varepsilon-\beta)(T+m+1)} \Big[\inf_{\{z_1,z_2,\ldots,z_{k-1}:\sum_{i=1}^{k-1} z_i = x\}} \{ \sum_{i=1}^{k-1} I_{A_i,T+m}(z_i) + I_{A_k,T}((C-\varepsilon)(T+m+1)-x)\}\Big]\Big], \qquad (15)$$

where $\alpha = E(\sum_{i=1}^{k-1} A_i(0,0))$ and $\beta = E(A_k(0,0))$.

Next, we derive (8). Since a lower bound on the probability is an upper bound of rate function, we concentrate on finding a lower bound on the probability of virtual delay of size k packets exceeding m. We do so by constructing a priority queueing based system which lower bounds the delay experienced in the SRPT system.

As a basis for comparison, we define the system PRI-0 to be a priority queueing system with capacity NC. This system consists of M queues, with packets of size k arriving to queue-k. Partially served packets in this system continue to reside in the same queue, i.e., no switching of packets occur.

Next, we consider a priority queueing system PRI-1, with capacity NC, where *all partially served packets completely leave the system*, instead of residing in the same queue or switching to a higher priority queue. By construction, this system has fewer arrivals to each queue, and at least as many departures from each queue as compared to the SRPT system. Thus, PRI-1 provides a lower bound on the delay experienced by a packet compared to the SRPT system.

Next, we fix an $\varepsilon > 0$, and consider the system PRI-2, which is a priority queueing system with capacity $N(C+\varepsilon)$. Thus, the operation of PRI-2 is identical to that of PRI-0, but with additional service capacity of $N\varepsilon$.

First, we compare PRI-1 and PRI-0. In any time slot, at most only one packet can be partially served. This implies that the maximum additional potential service that queue-k can receive in PRI-1 compared to PRI-0 is $(k-1)$. On the other hand, for any $N \geq M/\varepsilon$, the system PRI-2 will provide an additional service of $N\varepsilon \geq M \geq k$, compared to PRI-0. Thus, PRI-2 provides more potential service to queue-k compared to PRI-1. Further, note that PRI-2 has the same number of external arrivals as PRI-1. Thus, PRI-2 provides a lower bound on the virtual delay of a packet

compared to PRI-1, and consequently compared to that in the SRPT system.

We now describe the above argument in greater detail. Consider the case where there are two queues: size 1 and size k. Let the queues for PRI-1 be $Q_1^{(1)}(t)$ and $Q_k^{(1)}(t)$, and the queues for PRI-2 be $Q_1^{(2)}(t)$ and $Q_k^{(2)}(t)$. As described above, the arrival processes to PRI-1 and PRI-2 systems are the same. However, the potential service received by $Q_k^{(1)}(t)$ and $Q_k^{(2)}(t)$ are different. For PRI-1, we have that the potential service at time t to queue-k is upper bounded by the sum of $(NC - Q_1^{(1)}(t))$ and (possibly) the partially serviced size k packet. Thus, the upper bound on the potential service for $Q_k^{(1)}(t)$ is $(NC - Q_1^{(1)}(t) + k)$. On the other hand, $Q_k^{(2)}(t)$ (in system PRI-2) has potential service of $(NC - Q_1^{(2)}(t) + N\varepsilon)$. Further, we have that $Q_1^{(2)}(t) \leq Q_1^{(1)}(t)$. This is due to the following three facts: (i) the external arrivals to $Q_1^{(2)}(t)$ and $Q_1^{(1)}(t)$ are the same, (ii) packets of size 1 are fully served (i.e., there is no partially served size 1 packet), and (iii) PRI-2 has larger capacity than PRI-1. Combining the fact that $Q_1^{(2)}(t) \leq Q_1^{(1)}(t)$, and that $N\varepsilon \geq M \geq k$, we have that the potential service provided by PRI-1 is no more than PRI-2. This argument can be directly extended to case of multiple queues.

Thus, the delay experienced by a packet in a priority queueing system with capacity $N(C+\varepsilon)$ is a lower bound on the packet delay in PRI-1, and consequently a lower bound on the virtual delay in SRPT. In other words, we have that, $I_{\bar{V}}^{(k)}(m) \leq I_{V_{C+\varepsilon}}^{(k)}(m)$. Thus, the result follows. □

Following similar procedure, the delay asymptote for size 1 packets can be derived. We need to take into account that there is no queue with higher priority. Thus, in evaluating the lower bound, we do not consider the worst case partial packet arrival to queue-$(k-1)$. We have the delay asymptote for size 1 packets given by

$$I_{\bar{V}}^{(1)}(m) = \inf_{T \geq 0}[I_{A_1,T}(C(T+m+1))], \tag{16}$$

and the delay asymptote of packet size k is given by

$$\begin{aligned} I_{\bar{V}}^{(k)}(m) &= \inf_{T \geq 0}[\inf_{\alpha(T+m+1) \leq x \leq (C-\beta)(T+m+1)} \\ &\quad [\inf_{\{z_1,z_2,\ldots,z_{k-1}:\sum_{i=1}^{k-1} z_i = x\}} \{\sum_{i=1}^{k-1} I_{A_i,T+m}(z_i) + I_{A_k,T}(C(T+m+1)-x)\}]]. \end{aligned} \tag{17}$$

V. Comparison of SRPT and FIFO: Upper bounds and their Difference

In this section, we compare SRPT with the FIFO discipline in an asymptotic sense by comparing the difference of their rate functions. A FIFO (First in First out) queueing system

consists of a single server and queue, and packets are processed in the order they arrived. Using standard large deviation techniques, the rate function of packet delay for a FIFO queue is given by

$$\begin{aligned} I_{\hat{V}}(m) &= \inf_{T\geq 0}\Big[\inf_{\alpha(T+1)\leq x\leq C(T+m+1)-\beta(T+1)} \\ &\quad \Big[\inf_{\{z_1,z_2,\ldots,z_{k-1}:\sum_{i=1}^{k-1} z_i=x\}} \{\sum_{i=1}^{k-1} I_{A_i,T}(z_i)+I_{A_k,T}(C(T+m+1)-x)\}\Big]\Big]. \end{aligned} \tag{18}$$

We comment that the rate function is invariant to the size of the packet. In other words, the virtual delay seem by a size l packet is the same as any other.

We next derive an upper bound on the rate functions of SRPT and FIFO. The upper bound of the rate function of SRPT for the delay asymptote for size k packets is given by selecting specific values in the infimizing set,

$$I_{\bar{V}}^{(k)}(m) \leq \inf_{T\geq 0}[I_{A_k,T}((C-\alpha)(T+m+1))] \leq I_{A_k,0}((C-\alpha)(m+1)). \tag{19}$$

Using the same technique, the upper bound of the rate function of FIFO for the delay asymptote for size k packets is given by

$$I_{\hat{V}}(m) \leq I_{A_k,0}(C(m+1)-\alpha). \tag{20}$$

From (19) and (20), and the fact that the rate functions are non-negative, it follows that

$$|I_{\bar{V}}^{(k)}(m)-I_{\hat{V}}(m)| \leq \max\{I_{A_k,0}((C-\alpha)(m+1)), I_{A_k,0}(C(m+1)-\alpha)\}. \tag{21}$$

To explicitly compute the bounds described in (19) and (20), we make the following assumption. Lets denote $A_k(0,0)$ to be the volume of arrivals of size k packets in a single flow in time slot 0.

Assumption 2: We assume that $A_k(0,0)$ has the following property.

$$A_k(0,0) = \begin{cases} 0 & \textit{with probability p(k)} \\ k & \textit{with probability q(k) = 1-p(k)}, \end{cases} \tag{22}$$

where $q(k)$ satisfies: for any $\eta < 1$ *there exists constant $A_\eta \geq 1$, $K_\eta \geq 1$ such that for all $k \geq K_\eta$* (and $k \leq M$, where M is the largest packet size)

$$q(k)e^{k^\eta} \geq A_\eta. \tag{23}$$

Arrival process described in Assumption 2 include arrivals with truncated heavy-tail properties. Since (23) implies a sub-exponential decay in the arrival process and since we consider finite sized packets we have the truncation of heavy tail. For example, with $q(k) = 1/k$ (which has the property that the average volume of arrivals per time-slot for different sized arrivals are the same) is truncated heavy-tailed and satisfies the assumption. Other more general examples include $q(k) = \sum_{i=1}^{n} c_i/k^i$, for some n. Under such assumptions, we next derive an upper bound of $I_{A_k,0}(x)$.

Theorem 2: Fix any $0 < \gamma < 1$, *and* $x > E(A_k(0,0))$. *Then there exists* $\bar{K}(x)$ *such that*

$$I_{A_k,0}(x) \leq \frac{x^2(x+1)}{k^\gamma} \qquad \forall k \geq \bar{K}(x) \tag{24}$$

Proof: From (3) and the i.i.d. assumption across flows, we have $\Lambda_{A_k,0}(\theta) = \log E(e^{\theta A_k(0,0)})$. By (4), $I_{A_k,0}(x) = \sup_{\theta\in\mathbb{R}}(x\theta - \Lambda_{A_k,0}(\theta))$, where we denote θ_k^* as the supremizing value which depends on k. It is well known [9] that $\Lambda_{A_k,0}(\theta)$ is convex, with $\Lambda_{A_k,0}(0) = 0$ (see Figure 1(a)). Further, since $x > E(A_k(0,0))$, we can restrict the supremizing set of θ to $\{\theta \geq 0\}$.

We define $f_k(\theta)$ as follows. For some $0 < \gamma < 1$,

$$f_k(\theta) = \begin{cases} 0 & 0 \leq \theta \leq 1/k^\gamma \\ (x+1)\theta & \theta > 1/k^\gamma \end{cases} \qquad \forall k > \max(K_{1-\gamma}, 2(x+1)) \tag{25}$$

The function $f_k(\theta)$ is constructed such that for all $k > \max(K_{1-\gamma}, 2(x+1))$ it is a lower-bound of $\Lambda_{A_k,0}(\theta)$. (see Figure 1(a)) To show this, first observe that for all $\theta \leq 1/k^\gamma$, we have $0 = f_k(\theta) \leq \Lambda_{A_k,0}(\theta)$. Further, for $\theta = 1/k^\gamma$, we have

$$\begin{aligned} [\frac{d\Lambda_{A_k,0}(\theta)}{d\theta}]_{\theta=\frac{1}{k^\gamma}} &= \frac{kq(k)e^{k^{1-\gamma}}}{1+q(k)e^{k^{1-\gamma}}} = k\frac{1}{1+\frac{1}{q(k)e^{k^{1-\gamma}}}} \\ &\geq k\frac{1}{1+\frac{1}{A_{1-\gamma}}} \geq \frac{k}{2} \\ &\geq x+1, \qquad \forall k > \max(K_{1-\gamma}, 2(x+1)) \end{aligned} \tag{26}$$

where (26) follows from Assumption 2.

As $\Lambda_{A_k,0}(\theta)$ is convex, it follows that $\frac{d\Lambda_{A_k,0}(\theta)}{d\theta} \geq x+1$ for all $\theta \geq 1/k^\gamma$. Further, as $f_k(1/k^\gamma) \leq \Lambda_{A_k,0}(1/k^\gamma)$, it follows that $f_k(\theta) \leq \Lambda_{A_k,0}(\theta)$ for all $\theta \geq 0$.

We define $\bar{\theta}_k$ as the value that satisfies $x\theta = \Lambda_{A_k,0}(\theta)$, and $\hat{\theta}_k$ as the value that satisfies $x\theta = f_k(\theta)$. Note that $\bar{\theta}_k$ and $\hat{\theta}_k$ depends on k. Since $x\theta$ is affine and $\Lambda_{A_k,0}(\theta)$ is convex, we

have $\theta_k^* \leq \bar{\theta}_k$. Furthermore, as depicted in Figure 1(a), since $f_k(\theta) \leq \Lambda_{A_k,0}(\theta)$ for all $\theta \geq 0$, we have that $\bar{\theta}_k \leq \hat{\theta}_k$.

Thus, an upper bound on $I_{A_k,0}(x)$ is given by

$$I_{A_k,0}(x) = x\theta_k^* - \Lambda_{A_k,0}(\theta_k^*) \leq x\theta_k^* \leq x\bar{\theta}_k \leq x\hat{\theta}_k. \tag{27}$$

Computing $\hat{\theta}_k$, we have $\hat{\theta}_k = \frac{x(x+1)}{k^\gamma}$. Consequently

$$I_{A_k}(x) \leq \frac{x^2(x+1)}{k^\gamma} \qquad \forall k > \max(K_{1-\gamma}, 2(x+1)) = \bar{K}(x)\square \tag{28}$$

As a corollary of the above theorem, we derive the upper bound of the difference in rate functions of delay asymptote for size k packets in SRPT and FIFO.

Corollary 1: With arrival process that satisfy Assumption 2, the upper bound of the difference between the rate function of size k packets for SRPT and FIFO have the following characteristic for k sufficiently large. For any $0 < \gamma < 1$, *there exists* $\hat{K}(m)$ *such that*

$$|I_{\bar{V}}^{(k)}(m) - I_{\hat{V}}(m)| = O(\frac{1}{k^\gamma}) \quad \forall k \geq \hat{K}(m) \tag{29}$$

Proof: Combining Theorem 2 and the general upper bound of size k packets in SRPT and FIFO, given in (19) and (20), we have the following upper bound on the rate functions. Denote $c_1 = (C-\alpha)(m+1)$, $c_2 = C(m+1) - \alpha$ and note that $c_1, c_2 > \beta$, where $\beta = E[A_k(0,0)]$. For some fixed $0 < \gamma < 1$ there exists $\bar{K}_1(m)$ and $\bar{K}_2(m)$ such that

$$I_{\bar{V}}^{(k)}(m) \leq \frac{{c_1}^2(c_1+1)}{k^\gamma} \quad \forall k \geq \bar{K}_1(m) \quad \text{and} \quad I_{\hat{V}}(m) \leq \frac{{c_2}^2(c_2+1)}{k^\gamma} \quad \forall k \geq \bar{K}_2(m). \tag{30}$$

Applying (21), we have that the upper bound on the difference between the rate functions for k sized packets of delay asymptote for SRPT and FIFO is

$$|I_{\bar{V}}^{(k)}(m) - I_{\hat{V}}(m)| \leq \quad \max\{\tfrac{{c_1}^2(c_1+1)}{k^\gamma}, \tfrac{{c_2}^2(c_2+1)}{k^\gamma}\}, \quad \forall k \geq \max\{\bar{K}_1(m), \bar{K}_2(m)\} = \hat{K}(m) \tag{31}$$

which proves the corollary. $\square$

Remark: Suppose that the arrival process consists of only 2 sizes: size 1 or size M, where M is chosen large enough such that $M \geq \hat{K}(m)$. Then, we have $|I_{\bar{V}}^{(M)}(m) - I_{\hat{V}}(m)| = O(\frac{1}{M^\gamma})$, for size M packet, and $|I_{\bar{V}}^{(1)}(m) - I_{\hat{V}}(m)| = O(1)$ for size 1 packet.

Above result implies that *the difference of rate functions for SRPT and FIFO decays at least as fast as* $O(1/M^\gamma)$ *for M large*. Thus, even though the size of the packet is increasing, the delay performance of SRPT approaches that of FIFO.

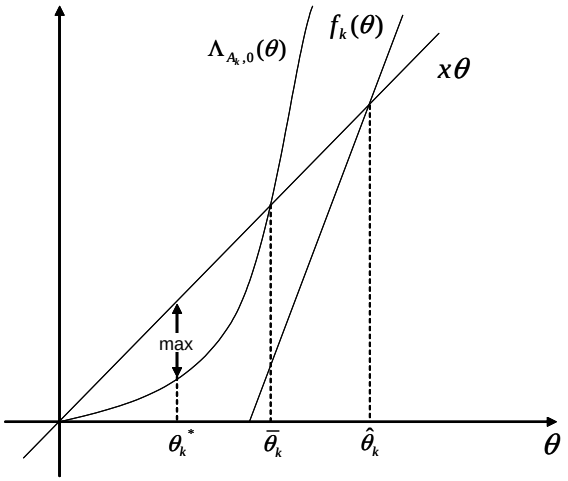

Figure 1(a) Bounds of θ^*.

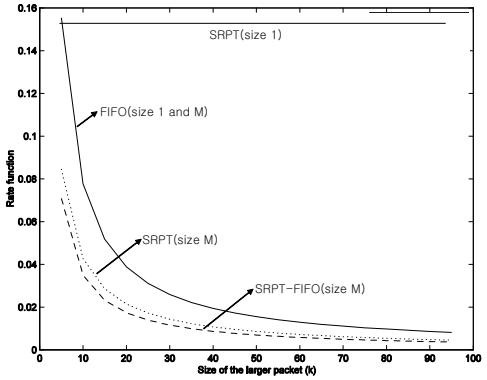

Figure 1(b) Rate functions of size 1 and M packets.

On the other hand, for the *size 1 packet*, the difference of rate functions remains constant. This is due to the fact that the delay distributions under SRPT is *invariant with k*. However for FIFO, the delay distributions for the *size 1 packet* decays as $O(\frac{1}{k^\gamma})$ for k large. Thus, these results indicate that SRPT is a good policy to implement for web-servers, where empirical evidence suggests a large variability in packet sizes [1].

VI. Numerical Results

In this section we compare the delay asymptotes of SRPT and FIFO by actual calculation of each rate function in a more specific scenario. We consider the system where the arrival process is assumed to be composed of independent ON-OFF processes which are one of two types: one that transmits size 1 packet with probability p, and one that transmits size M packets with a proportionally smaller probability p/M. The quantitative evaluation of the rate function is calculated for $C = 0.9, p = 0.4, m = 2$ using the closed expressions (16) and (17). The simulation results comparing the size 1 and size M packets were obtained and are shown in Figure 1(b).

Figure 1(b) reinforces the result of Corollary 1 which shows that the difference of the rate function of size M packets decay as M increase. We conclude that even in extreme scenarios, the effect of increased delay for of the larger packets in SRPT compared to FIFO becomes smaller in the asymptotic sense. In comparison, the rate function for size 1 packets of SRPT is superior of that of FIFO by at least factor of 100.

As shown, the difference in the delay asymptote between SRPT and FIFO indicates that the rate function approaches that of FIFO for increasingly larger packets while the delay performance for smaller packets remains much better than FIFO. It has been shown that web server requests exhibit heavy tail characteristics [1]. The two classes traffic model that we have studied approximates such a heavy traffic behavior for large M. Thus, the results indicate that SRPT is a promising scheduling policy which can be readily employed in web-servers.

REFERENCES

[1] M. F. Arlitt and C .L. Williamson. Web server workload characterization: The search for invariants. In *Proceeding of the ACM Sigmetrics*, Philadelphia, PA, April 1996.

[2] N. Bansal and M. Harchol-Balter. Analysis of SRPT scheduling: Investigating unfairness. In *Proceedings of ACM Sigmetrics/Performance*, pages 279–290, June 2001.

[3] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis. Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. *IEEE Transactions on Automatic Control*, 43:315–335, March 1998.

[4] S. C. Borst, O .J. Boxma, and et. al. The impact of the service discipline on delay asymptotics. *Performance Evaluation*, 54(2):175–206, October 2003.

[5] D.D. Botvich and N.G. Duffield. Large deviations, economies of scale, and the shape of the loss curve in large multiplexers. *Queueing Systems*, 20:293–320, 1995.

[6] C. Coubercoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Prob.*, 33(3):886–903, 1996.

[7] G. de Veciana and J. Walrand. Effective bandwidths: Call admission, traffic policing and filtering for ATM networks. *Queueing Systems Theory and Applications*, 20:37–59, 1995.

[8] S. Delas, R. Mazumdar, and C. Rosenberg. Cell loss asymptotics for buffers handling a large number of independent stationary sources. In *Proceedings of IEEE Infocom*, volume 2, pages 551–558, New York, NY, 1999.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2nd edition, New York, NY, 1998.

[10] Mor Harchol-Balter, Bianca Schroeder, Nikhil Bansal, and Mukesh Agrawal. Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems*, 21(2):207–233, May 2003.

[11] Mor Harchol-Balter, Karl Sigman, and Adam Wierman. Asymptotic convergence of scheduling policies with respect to slowdown. *Performance 2002. IFIP WG 7.3 International Symposium on Computer Modeling, Measurement and Evaluation*, 49(1):241–256, September 2002.

[12] N. Likhanov and R. Mazumdar. Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *Journal of Applied Probability*, 36:86–96, 1999.

[13] I. Paschalidis. Class-specific quality of service guarantees in multimedia communication networks. *Automatica, Special Issue on Control Methods for Communication Networks, V. Anantharam and J. Walrand, editors*, 35(12):1951–1969, 1999.

[14] A. Schwartz and A. Weiss. *Large deviations for performance analysis*. Chapman and Hall, London, UK, 1995.

[15] S. Shakkottai and R. Srikant. Many-sources delay asymptotics with applications to priority queues. *Queueing Systems: Theory and Applications*, 39:183–200, October 2001.

[16] W. Stalling. *Operating Systems*. Second Edition, Prentice Hall, 1995.