Asymptotic Convergence of Scheduling Policies with Respect to Slowdown [⋆]

Mor Harchol-Balter [a] Karl Sigman [b] Adam Wierman [a]

[a] *Carnegie Mellon University, Computer Science Department. Pittsburgh, PA 15213, USA.*

[b] *Columbia University, Department of Industrial Engineering and Operations Research. New York, NY 10027, USA.*

Abstract

We explore the performance of an M/GI/1 queue under various scheduling policies from the perspective of a new metric: the *slowdown* experienced by largest jobs. We consider scheduling policies that bias against large jobs, towards large jobs, and those that are fair, e.g., Processor-Sharing. We prove that as job size increases to infinity, all work conserving policies converge almost surely with respect to this metric to no more than $1/(1-\rho)$, where ρ denotes load. We also find that the expected slowdown under any work conserving policy can be made arbitrarily close to that under Processor-Sharing, for all job sizes that are sufficiently large.

Key words: scheduling, conservation, large jobs, convergence, Shortest Remaining Processing Time, SRPT, Processor-Sharing

[⋆] This work was supported by NSF Career Grant CCR-0133077, NSF ITR Grant 99-167 ANI-0081396 and by Spinnaker Networks via Pittsburgh Digital Greenhouse Grant 01-1.

Email addresses: `harchol@cs.cmu.edu` (Mor Harchol-Balter), `sigman@ieor.columbia.edu` (Karl Sigman), `acw@cs.cmu.edu` (Adam Wierman).

1 Introduction

It is well-known that choosing the right scheduling algorithm can have a big impact on performance, both in theory and in practice. For example, changing the scheduling algorithm in a CPU from Processor-Sharing (PS) to a policy which biases towards small jobs, such as Shortest-Remaining-Processing-Time-First (SRPT), or to a policy which biases towards young jobs, such as Least-Attained-Service (LAS), can improve mean response time (a.k.a. sojourn time) dramatically.

However, less well understood is the performance impact of different scheduling policies on large jobs. For example, how does a policy which biases towards small jobs, such as SRPT, compare against a policy which biases towards large jobs, such as Longest-Remaining-Processing-Time-First (LRPT), when the performance metric is the response time of the large jobs?

In this paper we limit our discussion to an M/GI/1 queue. For the M/GI/1/PS queue with load ρ, all jobs (large or small) are slowed down by the same factor, $\frac{1}{1-\rho}$, in expectation. Because the slowdown (response time divided by job size) is the same for all job sizes, the PS policy is often referred to as the *fair policy*.

We will show that all *work conserving* scheduling policies have the same performance as PS with respect to large jobs. In particular, we show that the slowdown as job size tends to infinity under any work conserving policy is at most $\frac{1}{1-\rho}$ almost surely; even for policies which clearly bias against large jobs. We also consider the expected slowdown for jobs that are not the very largest. We show that all “sufficiently-large” jobs have slowdown arbitrarily close to that of PS, where “sufficiently-large” depends on ρ and includes most jobs provided ρ is not too high.

2 Previous work

Ever since the discovery that SRPT has the lowest mean response time of any scheduling policy (for any sequence of arrival times and job sizes) [20,24,21], the evaluation of various scheduling policies has intrigued system designers and queueing theorists. There exist over a hundred survey papers to date on the analysis of scheduling policies, as well as many wonderful books such as [6,12,17,5].

The SRPT policy in particular has received much attention. Schrage and Miller first derived the expressions for the response times in an M/GI/1/SRPT queue [21]. This was further generalized by Pechinkin *et al.* to disciplines where the remaining times are divided into intervals [15]. The steady-state appearance of the M/GI/1/SRPT queue was obtained by Schassberger [19]. Rajaraman et al. showed further that the mean slowdown under SRPT is at most twice the optimal mean slowdown for any sequence of job arrivals [9].

Though analytical formulas for the M/GI/1 queue with various scheduling policies have been known for a long time, they are difficult to evaluate numerically, due to their complex form (many nested integrals). Hence, there was little work on the relative comparison of different scheduling policies, until more recently. The following papers have compared the *mean response times* of various scheduling policies under specific job size distributions and specific loads, by plotting the known formulas: [16,22,13,18,21]. A 7-year long study at University of Aachen under Schreiber [16,22] involved extensive evaluation of SRPT for various job size distributions and loads, and showed that SRPT has significant mean response time improvements compared to policies like FCFS, LFCS and PS. The survey paper by Schreiber [22] summarizes the results.

The above mentioned results were based on plots for *specific* job size distributions and loads. Hence it is not clear whether the conclusions hold for more general job size distributions and loads. Furthermore the above studies examined *mean* response time and did not raise the problem of possible *unfairness* to long jobs.

It has often been cited that the superior performance of scheduling policies which bias towards small jobs may come at the cost of starving large jobs [3,25,26,23]. Usually, examples of adversarial arrival sequences where a particular job starves are given to justify this. However, such worst case examples do not reflect the behavior of these policies in the average case. The term “starvation” is also used by people to indicate *unfairness*. It is often argued that policies which favor small jobs should result in higher expected response time for long jobs than policies which are ”fair,” like PS. This argument is valid for scheduling policies that do not make use of size; see the famous Kleinrock Conservation Law for non-preemptive, work conserving policies [12, Page 114] and for preemptive, work conserving policies with exponential service time distributions [11, Page 82].

Very recently, several papers have appeared which try to evaluate the problem of *unfairness* analytically, and thus consider the behavior of scheduling policies as a function of the job size. Bender et al. consider the metric *max slowdown* of a job as an indication of unfairness [3]. They show, with an example, that SRPT can have an arbitrarily large *max slowdown*. However, *max slowdown* is not an appropriate metric to measure unfairness. A large job may have an exceptionally long response time in some cases, but it might do well most of the time.

Bansal and Harchol-Balter [2] compare the SRPT policy and the PS policy analytically for an M/GI/1 queue on a per-job-size basis. They prove that if the load ρ is less than $\frac{1}{2}$, then every job, including the very largest job, has a lower expected response time under SRPT than under PS, for every job size distribution. They also prove that for arbitrary load ρ, the expected response time of a job of size x under SRPT is no more than c times that under PS, where c is a function of $\frac{1}{1-\rho}$. This result nicely complements the result in this paper (Theorem 5) which states that for all ρ, for every job size distribution, all sufficiently large jobs have expected response time (and slowdown) under SRPT which is *arbitrary close* to that of PS.

There has also been work in the area of proposing new SRPT-like policies [4,14] which try to reduce the problem of unfairness, while still favoring the short jobs. These usually prioritize based on *both* the time a job has waited so far, and its remaining size. These policies are usually analytically intractable and have been evaluated by simulation only. However simulations show that they are promising.

Other related research involves tail asymptotics for steady-state delay; see for example [8], in which the emphasis is on heavy-tailed distributions such as subexponential distributions.

3 The slowdown metric, the fairness metric, and some initial notation

We will throughout be considering a stable M/GI/1 queue. The average arrival rate will be λ. A job's *size* (service requirement) will be denoted by the random variable X and will be chosen i.i.d. from a continuous distribution with *finite mean* and *finite variance*. The probability density function (pdf) of the job size distribution is $f(x)$, and the cumulative distribution function (cdf) is $F(x) = P(X \leq x),\ x \geq 0$. We will denote the tail, $1 - F(x)$, by $\overline{F}(x)$. We assume that $f(x) > 0,\ x > 0$; service times can be arbitrarily large. Throughout we distinguish between the "size of a job" and the "remaining size of a job." The former denotes the service requirement upon time of arrival (original size chosen from F). The latter denotes the leftover (remaining) service time at the time in question. The load (utilization), ρ, of the server is $\rho \overset{\text{def}}{=} \lambda E[X] = \lambda \int_0^\infty x f(x) dx$. *We always will assume that* $\rho < 1$; the queue is stable. The load made up by the jobs of size less than or equal to x, $\rho(x)$, is $\rho(x) \overset{\text{def}}{=} \lambda \int_0^x t f(t) dt$.

We will use T to denote the steady-state response time (a.k.a. sojourn time) and $T(x)$ to denote the steady-state response time for a job of size x; a customer arriving in steady-state bringing a service time of length x has a response time $T(x)$. By definition, T has the same distribution as $T(X)$, and $E[T] = \int_0^\infty E[T(x)] f(x) dx$ where X is chosen independent of T throughout this paper. Note that $\{T(x) : x \geq 0\}$ is a stochastic process. Formally, at time $t = 0$ we initially start the system in steady-state, and then for each x, we construct each $T(x)$ using the same initial state and future service and interarrival times (along each sample path).

Definition 3.1 *For any given policy, the slowdown, S, is defined as response time divided by job size, namely, $S = \frac{T(X)}{X}$. The slowdown for a job of size x, $S(x)$, is thus given by $S(x) = \frac{T(x)}{x}$. The expected slowdown for a job of size x, $E[S(x)]$, is given by $E[S(x)] = \frac{E[T(x)]}{x}$. The overall mean slowdown is given by $E[S] = \int_0^\infty E[S(x)] f(x) dx$.*

Our **primary metric of interest** in this paper is **slowdown**. Mean slowdown is often used as a measure of system performance as opposed to the more traditional mean response time for two reasons [7,1,10]. First, it is desirable that a job's response time be correlated with its size (processing requirement). In many cases we'd like small jobs to have small response times and big jobs to have big response times. Second, mean slowdown is more representative of the performance of a large fraction of jobs. Mean response time is dominated by the contribution from just a few large jobs, whereas under mean slowdown the dominating effect of the large jobs is removed by normalizing the response times using the job sizes.

It is well known that for an M/GI/1/PS queue, $E[S(x)]^{PS} = \frac{1}{1-\rho}$. This says that for any given load $\rho < 1$, under PS scheduling, all jobs have the same expected slowdown; hence PS is **"fair"**. In this paper we will consider policies that significantly improve upon PS with respect to mean slowdown by giving priority to short jobs, or to young jobs. We will ask whether the large jobs suffer as a consequence. Specifically, we will be interested in the slowdown for large jobs.

Definition 3.2 *For any given scheduling policy, the slowdown for large jobs is defined (when it exists) by* $\lim_{x\to\infty} S(x)$ *whereby the convergence is almost sure (a.s.) convergence, by which we mean with probability* 1. *The expected slowdown for large jobs is defined (when it exists) by* $\lim_{x\to\infty} E[S(x)]$.

4 Brief review of common scheduling policies

In this section we define several common scheduling policies and summarize known results for these policies under an M/GI/1 queue, with respect to the mean response time for a job of size x.

PS: Processor-Sharing

Under the PS policy the processor is shared fairly among all jobs currently in the system [27]:

$$E[T(x)]^{PS} = \frac{x}{1-\rho}$$

SRPT: Shortest-Remaining-Processing-Time-First

Under the SRPT policy, at every moment of time, the server is processing that job with the shortest remaining processing time. The SRPT policy is well-known to be optimal for minimizing mean response time [21]. The mean response time for a job of size x, $E[T(x)]^{SRPT}$, can be decomposed into a sum:

$$E[T(x)]^{SRPT} = E[W(x)]^{SRPT} + E[R(x)]^{SRPT}$$

where $E[W(x)]^{SRPT}$ is the expected waiting time for the job (the expected time for a job of size x from when it first arrives to when it receives service for the first time) and $E[R(x)]^{SRPT}$ is the expected residence time (the time it takes for a job of size x to complete service once it begins execution) [21].

$$E[W(x)]^{SRPT} = \frac{\frac{\lambda}{2}\int_0^x t^2 f(t)dt + \frac{\lambda}{2}x^2\overline{F}(x)}{(1-\rho(x))^2}, \tag{1}$$

$$E[R(x)]^{SRPT} = \int_0^x \frac{dt}{1-\rho(t)}. \tag{2}$$

P-LCFS: Preemptive-Last-Come-First-Served

Under P-LCFS, whenever a new arrival enters the system, it preempts the job in service. Only when that arrival completes does the preempted job resume service. A new arrival can be thought of as starting its own busy period, where the new arrival can't leave until this busy period completes. Letting $B(x)$ denote the length of a busy period started by a job of length x, we have [12]:

$$E[T(x)]^{P-LCFS} = E[B(x)] = \frac{x}{1-\rho} \tag{3}$$

LAS: Least-Attained-Service

Under LAS, the job with the least attained service gets the processor to itself. If several jobs all have the least attained service, they time-share the processor via PS. This is a very practical policy, since a job's *age* (attained service) is always known, although it's size may not be known. This policy is conjectured to improve upon PS with respect to mean response time and mean slowdown when the job size distribution has decreasing failure rate.

Both $E[T(x)]^{LAS}$ and the Laplace transform of $T(x)^{LAS}$ under LAS are known [12]. We need some preliminary notation: For $x \geq 0$, let $X_x = \min\{x, X\}$. Then

$$E[X_x] = \int_0^x yf(y)dy + x\overline{F}(x)$$
$$E[X_x^2] = \int_0^x y^2 f(y)dy + x^2\overline{F}(x)$$

Observe that X_x is similar to the R.V. X, except that all job sizes have been capped at a maximum of x. Given the above definitions and letting $\rho_x = \lambda E[X_x]$, we have:

$$E[T(x)]^{LAS} = \frac{x(1-\rho_x) + \frac{\lambda}{2}E[X_x^2]}{(1-\rho_x)^2} \tag{4}$$

LRPT: Longest-Remaining-Processing-Time

Under the LRPT policy, at every moment of time, the server is processing the job with the longest remaining processing time. If multiple jobs in the system have the same remaining processing time, they time-share the processor via PS. Since the LRPT policy biases towards the *longest* jobs, it is of little practical value. We couldn't locate an analysis of this policy for the M/GI/1 queue anywhere, although analyzing LRPT isn't difficult, and we do so later in the paper.

SJF: Shortest-Job-First

SJF is the non-preemptive variant of SRPT. Under SJF, when the server is free it chooses to run the shortest job [6]:

$$E[T(x)]^{SJF} = x + \frac{\rho E[X^2]}{2E[X]} \cdot \frac{1}{(1-\rho(x))^2}$$

Other policies not mentioned above

There are many other scheduling policies that we haven't mentioned. All non-preemptive policies that don't make use of a job's size, for example, FCFS (First-Come-First-Served), LCFS (non-preemptive Last Come First Served), or RANDOM will have the same mean response time, $E[T]$, and thus for all such policies,

$$E[T(x)] = E[T] - E[X] + x = \frac{\lambda E[X^2]}{2(1-\rho)} + x$$

Since these have the same performance with respect to $E[T(x)]$, we will discuss them as a group.

5 Convergence of scheduling policies in expectation

In this section, we evaluate the *expected slowdown* for the largest jobs under different scheduling policies. In Section 5.1 we consider five particular scheduling policies and show that they have the same expected slowdown as PS for the largest job. In Section 5.2 and Section 5.3 we generalize these results to all work conserving scheduling policies. Finally, in Section 5.4 we consider the broader problem of expected slowdown as a function of job size, for all job sizes. We find that for any work conserving policy, for sufficiently large jobs, the expected slowdown can be shown to be arbitrarily close to that of PS, where our definition of sufficiently large will typically include most jobs.

5.1 Convergence of five scheduling policies in expectation

This section will prove the following theorem:

Theorem 1 *As $x \to \infty$, expected slowdown for SRPT, P-LCFS, LAS, and LRPT is the same as for PS:*

$$\lim_{x\to\infty} E[S(x)]^{SRPT} = \lim_{x\to\infty} E[S(x)]^{P\text{-}LCFS} = \lim_{x\to\infty} E[S(x)]^{LAS} = \lim_{x\to\infty} E[S(x)]^{LRPT} = \frac{1}{1-\rho}$$

That is, the expected slowdown for the largest job is the same under policies that bias towards short jobs, policies that bias towards long jobs, and policies that treat all jobs fairly.

Proof for SRPT

We start by looking at the waiting time component of SRPT:

$$\begin{aligned} E[W(x)]^{SRPT} &= \frac{\frac{\lambda}{2}\int_0^x t^2 f(t)dt + \frac{\lambda}{2}x^2\overline{F}(x)}{(1-\rho(x))^2} = \frac{\lambda\int_0^x t\,\overline{F}(t)dt}{(1-\rho(x))^2} \\ \lim_{x\to\infty} E[W(x)]^{SRPT} &= \frac{\lambda\int_0^\infty t\,\overline{F}(t)dt}{(1-\rho)^2} < \infty \end{aligned}$$

where finiteness follows since the service time distribution F is assumed to have finite second moment.

Thus we have

$$\lim_{x\to\infty}\frac{E[W(x)]^{SRPT}}{x} = 0$$

We now complete the proof by considering the residence time component of SRPT.

$$\begin{aligned} \lim_{x\to\infty}\frac{E[R(x)]^{SRPT}}{x} &= \lim_{x\to\infty}\frac{1}{x}\int_0^x \frac{dt}{1-\rho(t)} = \lim_{x\to\infty}\frac{1}{1-\rho(x)} \text{ (by L'Hopital)} \\ &= \frac{1}{1-\rho} \end{aligned}$$

Proof for LAS

The limiting slowdown of large jobs is the same under LAS and SRPT as shown below:

$$\begin{aligned} \rho_x &= \lambda\int_0^x yf(y)dy + \lambda x\overline{F}(x) = \lambda\int_0^x \overline{F}(y)dy \\ \lim_{x\to\infty}\rho_x &= \lambda\int_0^\infty \overline{F}(y)dy = \lambda E[X] = \rho \end{aligned}$$

$$
\begin{aligned}
E[T(x)]^{LAS} &= \frac{x}{1-\rho_x} + \frac{\frac{\lambda}{2}\left(\int_0^x y^2 f(y)dy + x^2\overline{F}(x)\right)}{(1-\rho_x)^2} \\
&= \frac{x}{1-\rho_x} + \frac{\lambda \int_0^x \overline{F}(y)dy}{(1-\rho_x)^2} \\
\lim_{x\to\infty} E[S(x)]^{LAS} &= \lim_{x\to\infty} \frac{x}{1-\rho_x}\cdot\frac{1}{x} + \lim_{x\to\infty} \frac{\lambda \int_0^x \overline{F}(y)ydy}{(1-\rho_x)^2}\cdot\frac{1}{x} \\
&= \frac{1}{1-\rho} + \frac{\lambda\int_0^\infty \overline{F}(y)ydy}{(1-\rho)^2} \lim_{x\to\infty}\frac{1}{x}
\end{aligned}
$$

Again, by the finiteness of the second moment of F, $\lim_{x\to\infty} E[S(x)]^{LAS} = \frac{1}{1-\rho}$.

Proof for LRPT

We will use the following notation in this section and throughout the rest of the paper: B will denote the length of a regular busy period. $B(x)$ will denote the length of a busy period started by a job of size x (an exceptional first service busy period). $B(x)|_{\lambda'}$ will denote the length of a busy period started by a job of size x where the arrival rate is λ'.

If the job enters a busy system, then we can again take advantage of the above property to see that $T(x) = B(x+V)$, in distribution, where V is the amount of work in the system (in steady-state) seen by an arbitrary arrival.

Since LRPT is work conserving, and arrivals are Poisson, we know via PASTA that:

$$
E[V] = E[W(x)]^{FCFS} = \frac{\lambda E[X^2]}{2(1-\rho)},
$$

where $W(x)^{FCFS}$ is the steady-state delay in queue (not including service) in a FCFS queue. Note that $E[V]$ does not depend on x.

It is well known that $E[B(Y)] = \frac{E[Y]}{1-\rho}$ for any exceptional first service time Y. This holds for $Y = x$ and $Y = x + V$. Using this we obtain, as $x \to \infty$:

$$
E[S(x)]^{LRPT} = \frac{E[B(x+V)]}{x} = \frac{x+E(V)}{(1-\rho)x} = \frac{1}{1-\rho} + \frac{E(V)}{(1-\rho)x} \to \frac{1}{1-\rho} \quad (5)
$$

Proof for P-LCFS

For the `P-LCFS` policy it trivially follows from (3) that:

$$\lim_{x\to\infty} \frac{E[T(x)]^{P-LCFS}}{x} = \frac{1}{1-\rho}$$

5.2 Convergence of all work conserving scheduling policies in expectation

This section extends the analysis of the previous section. The goal is to to bound convergence in expectation of slowdown under *any work conserving policy*.

Theorem 2 *For any work conserving scheduling policy*

$$\lim_{x\to\infty} E[S(x)] \le \frac{1}{1-\rho}.$$

If the policy is also non-preemptive, then $E[S(x)] \to 1$ *as* $x \to \infty$.

Proof : The proof of the $\frac{1}{1-\rho}$ bound stems from the observation that `LRPT` provides an upper bound on $T(x)^P$ for any work conserving policy P. That is, under `LRPT`, every job finishes the moment the busy period the job arrived into ends, which is the last possible completion moment for any work conserving policy. So, the result follows from Equation (5). For any work conserving policy P:

$$\lim_{x\to\infty} E[S(x)]^P \le \lim_{x\to\infty} E[S(x)]^{LRPT} = \frac{1}{1-\rho}.$$

This proves the first half of the theorem.

Now we limit our discussion to non-preemptive work conserving policies. In this case $T(x) = W(x) + x$, and $W(x)$ is smaller than the length of a busy period started by a job of size equal to V. So $W(x) \le B(V)$ and $E[W(x)] \le \frac{E[V]}{1-\rho}$, and

$$E[S(x)] \le \frac{E[V]}{x(1-\rho)} + 1 \to 1, \text{ as } x \to \infty.$$

■

A few followup observations are in order regarding Theorem 2.

Remark 3 *Theorem 2 does not extend to policies that are not work conserving. In fact, for every* $z \in [1, \infty)$ *there is a non work conserving policy such that* $\lim_{x\to\infty} E[S(x)] = z$.

To see this, consider the policy that makes each job wait $(z-1)x$ time before it is allowed to enter the queue of a non-preemptive, work conserving system.

Remark 4 *The* $\frac{1}{1-\rho}$ *bound in Theorem 2 is tight. In fact, For every* $z \in [1, \frac{1}{1-\rho}]$ *there is a work conserving policy such that* $E[S(x)] \to z,$ *as* $x \to \infty$.

Proof : Consider a linear combination of the FCFS and P-LCFS policies. More specifically, consider the following scheduling policy, P: with probability q an arriving job preempts the job being serviced, and with probability $1-q$ an arriving job is placed at the back of a FCFS queue to await service.

We can quickly analyze this policy to find $E[S(x)]^P$. Consider an arrival that gets placed at the front of the queue. This arrival can only be bothered by other jobs that are allowed to preempt. Thus, for this job $T(x) = B(x)|_{\lambda'}$, where $\lambda' = q\lambda$ for $q \in [0,1]$. That is, $T(x)$ is the length of a busy period started by a job of size x where the arrival rate is λ'.

Now consider a job that gets placed in the back of the queue. If the system is idle when the job arrives, we again see that $T(x) = B(x)|_{\lambda'}$. However, if the system is busy at the time of the arrival $T(x) = B(x+V))|_{\lambda'}$, where V is the amount of work in system seen by an arbitrary arrival. Let $\rho' = \frac{\lambda'}{\mu}$. Then, putting these two pieces together, we see that as $x \to \infty$:

$$\begin{aligned} E[S(x)]^P &= q\frac{E[B(x)]|_{\lambda'}}{x} + (1-q)\frac{E[B(x+V)]|_{\lambda'}}{x} \\ &= q\frac{1}{1-\rho'} + (1-q)\frac{1+\frac{1}{x}\frac{\lambda E[X^2]}{2(1-\rho)}}{1-\rho'} \to \frac{1}{1-\rho'} \end{aligned}$$

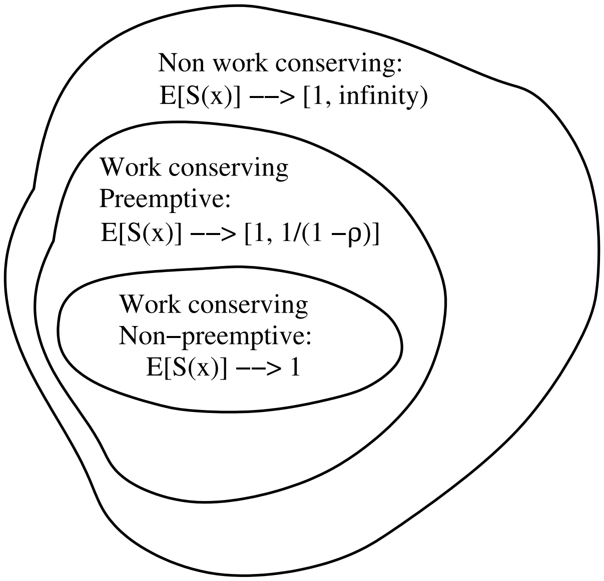

Fig. 1. *Taxonomy of scheduling policies defined by the metric* $\lim_{x\to\infty} E[S(x)]$.

Notice that since ρ' is an arbitrary number in $[0, \rho]$, we can make $\frac{1}{1-\rho'}$ any number in $[1, \frac{1}{1-\rho}]$. ■

The above remarks show that the metric $\lim_{x\to\infty} E[S(x)]$ defines a taxonomy on all scheduling policies, as shown in Figure 1. Non work conserving policies have a value in $[1, \infty)$ under this metric. Preemptive work conserving policies have a value in $[1, \frac{1}{1-\rho}]$ under this metric. Non-preemptive work conserving policies all have a value of 1 under this metric. Each class is complete in that for each value in the range, there exists a policy with that value.

5.4 Bounding all work conserving policies for sufficiently-large job sizes

Until now we have concentrated on the limiting behavior as the job size $x \to \infty$. We now show that we can easily prove an upper bound of $(1+\varepsilon)\frac{1}{1-\rho}$ for the expected slowdown of all "sufficiently large" jobs under all work conserving scheduling policies for any $\varepsilon > 0$.

Let V be the amount of work in the system when a job arrives. Recall that $E[V]$ is the same under all work conserving policies and for jobs of any size. In fact, $E[V] = E[W(x)]^{FCFS}$.

Theorem 5 *Fix $\varepsilon > 0$. Then under any work conserving scheduling policy P, if $x \geq \frac{1}{\varepsilon}E[V]$, then*

$$E[S(x)]^P \leq (1+\varepsilon)E[S(x)]^{PS} = (1+\varepsilon)\frac{1}{1-\rho}.$$

If the policy is also non-preemptive and $x \geq \frac{1}{\varepsilon(1-\rho)}E[V]$, then $E[S(x)]^P \leq 1+\varepsilon$.

Before we begin the proof, observe that provided ρ is not too high, the above theorem says that in fact many jobs are sufficiently large, since $E[W(x)]^{FCFS}$ will be low. As an example of using the theorem, if we consider $E[S(x)]$ under an M/M/1 with $\mu = 1$ and $\rho = .5$ we find that for a job x in the largest one percent of the service distribution $E[S(x)] \leq 2.4$, as compared with a limiting slowdown of 2.

Proof :

Recall that LRPT provides an upper bound on $S(x)^P$ for any work conserving policy P. That is, every job finishes at the last possible moment under LRPT, and so the slowdown of any other policy must be bounded by that of LRPT . Thus, we need simply show that for sufficiently large x, $E[S(x)]^{LRPT} \leq \frac{1+\varepsilon}{1-\rho}$.

Observing that $T(x)^{LRPT}$ has the same distribution (hence mean) as $B(x+V)$, we have

$$E[S(x)]^{LRPT} = \frac{1}{x}E[T(x)]^{LRPT} = \frac{1}{x}\cdot\frac{x+E(V)}{(1-\rho)} = \frac{E[V]}{x(1-\rho)} + \frac{1}{1-\rho}$$

Letting $x \geq \frac{1}{\varepsilon}E[V]$ gives us $E[S(x)]^P \leq E[S(x)]^{LRPT} \leq \frac{1+\varepsilon}{1-\rho}$.

Further, we can obtain a similar bound on convergence for non-preemptive, work conserving policies. Recall from the proof of Theorem 2 that for any non-preemptive, work conserving policy P, we have $E[S(x)]^P \leq \frac{E[V]}{1-\rho}\frac{1}{x} + 1$.

Thus, letting $x \geq \frac{1}{\varepsilon(1-\rho)}E[V]$ gives us $E[S(x)]^P \leq 1+\varepsilon$. ■

6 Almost sure convergence of scheduling policies

In this section, we extend the analysis of Theorem 2 in order to show that under any work conserving policy the performance of the largest jobs will be at most that of PS almost surely. Recall that:

Definition 6.1 *The sequence of random variables* $\{Y_n, n = 1, 2, \dots\}$ *is said to converge almost surely to a random variable* Y*, written* $Y_n \overset{a.s.}{\to} Y$ *as* $n \to \infty$*, if* $P(\lim_{n\to\infty} Y_n = Y) = 1$*. We equivalently say that* Y_n *converges to* Y *with probability* 1 *(w.p.1.).*

Theorem 6 *Under Processor-Sharing it holds a.s. that* $\lim_{x\to\infty} S(x)^{PS} = \frac{1}{1-\rho}$.

Proof :

We begin by introducing an alternative model that serves as an upper bound for PS, and an appropriate coupling. Under PS denote the number of jobs in system at time t by $X(t)$, and the remaining service times of the jobs by $Y_1(t), \dots, Y_{X(t)}(t)$.

Consider an alternative M/GI/1/PS model denoted by PS^1 in which whenever there are $n \geq 1$ jobs in the system, the server instead of giving capacity $1/n$ to each of the n jobs, gives the smaller amount $1/(n+1)$. This amounts to adding a fictitious job – called an *observer* – with service time $x = \infty$ to the PS system at time $t = 0$. The observer remains in the system forever using service capacity but is not counted as a real job. Denote the number of jobs in the PS^1 system at time t by $X^1(t)$, and the remaning service times of the jobs by $Y_1^1(t), \dots, Y_{X^1(t)}(t)$. Assume that job service times are brought by each arrival (instead of being handed out by the server). By using the same arrival sequence input (arrival times, service times) for both models it follows that if $X(0) = X^1(0) = 0$, then

$$X(t) \leq X^1(t), \ t \geq 0, \tag{6}$$

$$Y_i(t) \leq Y_i^1(t), \text{ for any job } i \text{ that is in } \textit{both} \text{ systems at time } t, \tag{7}$$

because PS^1 always serves each job at a slower rate (hence each job departs later from PS^1 than from PS). Thus letting $t \to \infty$, we obtain time-stationary and ergodic versions of both models, while retaining the relations (6) and (7). We assume from now on that this has been done so that at time $t = 0$ both are stationary (e.g., have their stationary distributions).

For the PS model, it is well known that the time-stationary distribution is given by $P(X(0) = 0) = 1 - \rho$,

$$P(X(0) = n, Y_1(0) \le x_1, \dots , Y_n(0) \le x_n) = (1-\rho)\rho^n F_e(x_1) \cdots F_e(x_n),\ n \ge 1,$$

where $G_e(x)$ denotes the equilibrium distribution function of F with density $f_e(x) = \mu\overline{F}(x)$.

PS^1 still operates under a "symmetric" service discipline (e.g., Theorem 26, Page 339 in Wolff [27]), and hence the steady-state distribution of $X^1(t)$ as $t \to \infty$ is insensitive to the service time distribution except through its mean $1/\mu$. Let P_n^1 denote the limiting probability that there are n jobs in the PS^1 system. Using exponential service times yields a Birth and Death model with balance equations

$$\lambda P_n^1 = \left(\frac{n+1}{n+2}\right) \mu P_{n+1}^1,\ n \ge 0,$$

and solution

$$P_n^1 = (n+1)\rho^n(1-\rho)^2,\ n \ge 0. \tag{8}$$

(Note that the stability condition remains $\rho < 1$ since $\frac{n+1}{n+2} \to 1,\ n \to \infty$.)

For general service time distribution F then, PS^1 has time-stationary distribution given by $P(X^1(0) = 0) = P_0^1$ and

$$P(X^1(0) = n, Y_1^1 \le x_1, \dots , Y_n^1 \le x_n) = P_n^1 F_e(x_1) \cdots F_e(x_n),\ n \ge 1.$$

A job of size x arriving to PS (at time $t = 0$ for simplicity via PASTA) will cause $X(0)$ to jump to $X(0)+1$, and then cause (during its sojourn time $T(x) = T(x)^{PS}$) all current and future jobs in the PS system to be treated as if in a PS^1 system; the x-job has the effect of an observer. Let $X_\infty(t),\ t \ge 0$ denote the number of jobs in a

PS^1 system started off with the stationary distribution of PS, i.e., $X_\infty(0) = X(0)$ and the $X(0)$ remaining service times are $Y_1(0), \dots Y_{X(0)}(0)$. It follows that the service capacity given to the x-job at time t in the PS system is given by $(1 + X_\infty(t))^{-1}$; thus sojourn time for the x-job in the PS system can be expressed as

$$T(x) = \min\{t > 0 : \int_0^t \frac{1}{1 + X_\infty(u)} du = x\} = B^{-1}(x),$$

where

$$B(t) = \int_0^t \frac{1}{1 + X_\infty(u)} du$$

is the amount of service that the x-job receives during the first t time units.

By construction $X(t) \leq X_\infty(t) \leq X^1(t),\ t \geq 0$, yielding the bounds

$$C^{-1}(x) \leq T(x) \leq A^{-1}(x),$$

where

$$A(t) = \int_0^t \frac{1}{1 + X^1(u)} du$$
$$C(t) = \int_0^t \frac{1}{1 + X(u)} du.$$

Whereas both $\{X(t)\}$ and $\{X^1(t)\}$ are stationary, $\{X_\infty(t)\}$ is not because of its initial condition but will become so as $t \to \infty$. In fact, for the random time

$$\tau = \min\{t \geq 0 : X^1(t) = 0\}, \tag{9}$$

$X_\infty(\tau) = 0$ (since $X_\infty(t) \leq X^1(t)$), and $X_\infty(\tau + t) = X^1(\tau + t),\ t \geq 0$, a.s., the two processes are identical a.s. from time τ onwards.

We now analyze slowdown under PS for job-x as $x \to \infty$:

Observe that $\{A(t)\}$ is strictly increasing and has stationary ergodic increments due to the stationary ergodicity of $\{X^1(t)\}$. Thus by Birkhoff's ergodic theorem,

$$\lim_{t\to\infty} \frac{A(t)}{t} = E[A(1)], \text{ a.s..} \tag{10}$$

By stationarity, non-negativity and (8),

$$\begin{aligned} E[A(1)] &= E\left[\int_0^1 \frac{1}{1+X^1(u)}du\right] = \int_0^1 E\left[\frac{1}{1+X^1(u)}\right] \\ &= \int_0^1 E\left[\frac{1}{1+X^1(0)}\right] \\ &= E\left[\frac{1}{X^1(0)+1}\right] \\ &= \sum_{n=0}^{\infty} \frac{1}{n+1}(n+1)\rho^n(1-\rho)^2 \\ &= (1-\rho)\sum_{n=0}^{\infty} \rho^n(1-\rho) = 1-\rho \end{aligned} \tag{11}$$

The inverse process

$$A^{-1}(x) = \min\{t > 0 : \int_0^t \frac{1}{X^1(u)+1}du = x\}$$

is strictly increasing to ∞ and by definition $A(A^{-1}(x)) = x$; thus from (10) and (11)

$$\lim_{x\to\infty} \frac{x}{A^{-1}(x)} = \lim_{x\to\infty} \frac{A(A^{-1}(x))}{A^{-1}(x)} = \lim_{t\to\infty} \frac{A(t)}{t} = 1-\rho, \text{ a.s.}$$

and we conclude that

$$\lim_{x\to\infty} \frac{A^{-1}(x)}{x} = (1-\rho)^{-1}, \text{ a.s.} \tag{12}$$

From (9), $A(t) - A(\tau) = B(t) - B(\tau),\ t \geq \tau$ yielding

$$\lim_{t\to\infty} \frac{B(t)}{t} = \lim_{t\to\infty} \frac{A(t)}{t}, \text{ a.s.}$$

and thus

$$\lim_{x\to\infty} \frac{B^{-1}(x)}{x} = \lim_{x\to\infty} \frac{A^{-1}(x)}{x}, \text{ a.s..}$$

Since $T(x) = B^{-1}(x)$, (12) yields $T(x)/x \to (1-\rho)^{-1}$, a.s.. ■

Theorem 7 *Under all work conserving scheduling policies it holds a.s. (assuming the limit exists) that*

$$\lim_{x\to\infty} S(x) \le \frac{1}{1-\rho}.$$

If the policy is also non-preemptive, then the limit does exists and $S(x) \overset{a.s.}{\to} 1$ *as* $x \to \infty$.

Proof : The proof for *non-preemptive,* work conserving policies is quick: Start with the observation that

$$P(S(x)^P \ge 1) = 1 \quad \forall x, \forall \text{ policies P}$$

This follows simply by definition of slowdown. By taking limits, a.s. it holds that

$$\liminf_{x\to\infty} S(x)^P \ge 1, \forall \text{ policies P}$$

Now, recall that we have a.s.

$$S(x)^P \le 1 + \frac{B(V)}{x} \quad \forall x, \forall \text{work conserving, non-preemptive policies P}$$

Taking limits we have a.s. that:

$$\limsup_{x\to\infty} S(x)^P \le 1, \forall \text{work conserving, non-preemptive policies P}$$

Thus for all work conserving, non-preemptive policies P the limit does exists and

$$S(x) \overset{a.s.}{\to} 1 \text{ as } x \to \infty.$$

The remainder of the proof will concentrate on work conserving policies that may allow for *preemption*.

We know that a.s.

$$T(x) \leq B(x+V),$$

where $B(y)$ is used to denote the length of a busy period started by a job of size y.

Thus

$$\begin{aligned} \lim_{x\to\infty} S(x) &= \lim_{x\to\infty} T(x)/x \leq \lim_{x\to\infty} \frac{B(x+V)}{x} \\ &= \lim_{x\to\infty} \frac{B(x)}{x} + \lim_{x\to\infty} \frac{B(V)}{x} \end{aligned}$$

We now make two observations. First observe that since V is finite w.p.1.

$$\lim_{x\to\infty} \frac{B(V)}{x} = 0$$

Second, observe further that if we let $\{B_{(i)} : i \geq 1\}$ denote an i.i.d. sequence of regular busy periods (non-exceptional), then $B(x)$ can be expressed as

$$B(x) = x + \sum_{i=1}^{N(x)} B_{(i)}$$

where $\{N(x) : x \geq 1\}$ is a Poisson process of rate λ independent of $\{B_{(i)} : i \geq 1\}$. We conclude that this version of $\{B(x) : x \geq 0\}$ is a compound Poisson process with a linear x term added on, so it has stationary and independent increments.

Thus, almost surely,

$$\begin{aligned} \lim_{x\to\infty} S(x) &= \lim_{x\to\infty} \frac{B(x)}{x} + 0 = \lim_{x\to\infty} \frac{1}{x} \sum_{i=1}^{x} B(1)_{(i)} = E[B(1)] \text{ (by S.L.L.N)} \\ &= \frac{1}{1-\rho} \end{aligned}$$

Notice that we assumed that x is integer valued, however the proof is valid even if this is not the case; the fractional remainder of x does not affect the limit. ■

7 Conclusion

In this paper we consider the performance metric “slowdown for the largest job” and we show that under this metric the performance of all work conserving scheduling policies is bounded by $\frac{1}{1-\rho}$ almost surely.

This metric is also interesting for another reason; it allows us to categorize all scheduling policies into 3 classes. We find that for *non work conserving policies*, the expected slowdown of the largest job can range from 1 to infinity (and in fact every value in between is achieved by some non work conserving policy). For *preemptive work conserving policies*, the expected slowdown of the largest job can range from 1 to $\frac{1}{1-\rho}$ (and again each value in between is achieved by some preemptive work conserving policy). Lastly, for non-preemptive work conserving policies, the expected slowdown of the largest job is always 1.

This paper also raises the question of how scheduling policies compare with respect to slowdown on job sizes other than the very largest. We find that for all “sufficiently large” jobs, the expected slowdown of these jobs under any work conserving policy can be made arbitrarily close to $\frac{1}{1-\rho}$, where the definition of “sufficiently large” depends on the degree of closeness and on the system load. When the system load is not too high, “sufficiently large” ends up including most jobs. The behavior of scheduling policies on jobs other than the largest job is an interesting question which will surely generate further research.

The proofs in this paper are varied, but all surprisingly simple, which should help others in extending this work. The proofs rely on a few key observations about subdividing busy periods and on some alternative formulations of scheduling formulas. Perhaps the most useful observation is that the Longest-Remaining-Processing-Time policy can be used to bound all other work conserving policies, and that it suffices to therefore to concentrate on this one policy.

References

[1] Baily, Foster, Hoang, Jette, Klingner, Kramer, Macaluso, Messina, Nielsen, Reed, Rudolph, Smith, Tomkins, Towns, and Vildibill. Valuation of ultra-scale computing systems. White Paper, 1999.

[2] Nikhil Bansal and Mor Harchol-Balter. Analysis of SRPT scheduling: Investigating unfairness. In *Proceedings of* Sigmetrics ’01, 2001.

[3] M. Bender, S. Chakrabarti, and S. Muthukrishnan. Flow and stretch metrics for scheduling continous job streams. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[4] L. Cherkasova. Scheduling strategies to improve response time for web applications. In *High-performance computing and networking: international conference and exhibition*, pages 305–314, 1998.

[5] E.G. Coffman and L. Kleinrock. Computer scheduling methods and their countermeasures. In *AFIPS conference proceedings*, volume 32, pages 11–21, 1968.

[6] Richard W. Conway, William L. Maxwell, and Louis W. Miller. *Theory of Scheduling*. Addison-Wesley Publishing Company, 1967.

[7] Allen B. Downey. A parallel workload model and its implications for processor allocation. In *Proceedings of High Performance Distributed Computing*, pages 112–123, August 1997.

[8] K. Sigman. (Guest Editor). Special volume on queues with heavy-tailed distribution. *QUESTA*, 33, 1999.

[9] J.E. Gehrke, S. Muthukrishnan, R. Rajaraman, and A. Shaheen. Scheduling to minimize average stretch online. In *40th Annual symposium on Foundation of Computer Science*, pages 433–422, 1999.

[10] M. Harchol-Balter and A. Downey. Exploiting process lifetime distributions for dynamic load balancing. *ACM Transactions on Computer Systems*, 15(3), 1997.

[11] L. Kleinrock. *Communication Nets*. McGraw-Hill Book Comp., 1964.

[12] L. Kleinrock. *Queueing Systems*, volume II. Computer Applications. John Wiley & Sons, 1976.

[13] M. Crovella M. Harchol-Balter and S. Park. The case for SRPT schduling in web servers. Technical Report MIT-LCS-TR-767, MIT Lab for Computer Science, Oct. 1998.

[14] E. Modiano. Scheduling algorithms for message transmission over a satellite broadcast system. In *Proceedings of IEEE MILCOM '97*, pages 628–634, 1997.

[15] A.V. Pechinkin, A.D. Solovyev, and S.F. Yashkov. A system with servicing discipline whereby the order of remaining length is serviced first. *Tekhnicheskaya Kibernetika*, 17:51–59, 1979.

[16] R. Perera. The variance of delay time in queueing system M/G/1 with optimal strategy SRPT. *Archiv fur Elektronik und Uebertragungstechnik*, 47:110–114, 1993.

[17] M. Pinedo. *On-line algorithms, Lecture Notes in Comp. Science*. Prentice Hall, 1995.

[18] J. Roberts and L. Massoulie. Bandwidth sharing and admission control for elastic traffic. In *ITC Specialist Seminar*, 1998.

[19] R. Schassberger. The steady-state appearance of the M/G/1 queue under the discipline of shortest remaining processing time. *Advances in Appl. Prob.*, 22:456–479, 1990.

[20] Linus E. Schrage. A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16:678–690, 1968.

[21] Linus E. Schrage and Louis W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14:670–684, 1966.

[22] F. Schreiber. Properties and applications of the optimal queueing strategy SRPT - a survey. *Archiv fur Elektronik und Uebertragungstechnik*, 47:372–378, 1993.

[23] A. Silberschatz and P. Galvin. *Operating System Concepts, 5th Edition*. John Wiley & Sons, 1998.

[24] D.R. Smith. A new proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 26:197–199, 1976.

[25] W. Stallings. *Operating Systems, 2nd Edition*. Prentice Hall, 1995.

[26] A.S. Tanenbaum. *Modern Operating Systems*. Prentice Hall, 1992.

[27] Ronald W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, 1989.