Classifying Scheduling Policies with Respect to Higher Moments of Conditional Response Time*

Adam Wierman
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
acw@cs.cmu.edu

Mor Harchol-Balter
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
harchol@cs.cmu.edu

ABSTRACT

In addition to providing small mean response times, modern applications seek to provide users predictable service and, in some cases, Quality of Service (QoS) guarantees. In order to understand the predictability of response times under a range of scheduling policies, we study the conditional variance in response times seen by jobs of different sizes. We define a metric and a criterion that distinguish between contrasting functional behaviors of conditional variance, and we then classify large groups of scheduling policies.

In addition to studying the conditional variance of response times, we also derive metrics appropriate for comparing higher conditional moments of response time across job sizes. We illustrate that common statistics such as raw and central moments are not appropriate when comparing higher conditional moments of response time. Instead, we find that cumulant moments should be used.

Categories and Subject Descriptors

F.2.2 [**Nonnumerical Algorithms and Problems**]: Sequencing and Scheduling; G.3 [**Probability and Statistics**]: Queueing Theory; C.4 [**Performance of Systems**]: Performance Attributes

General Terms

Performance, Algorithms

Keywords

Scheduling; response time; predictability; variance; cumulants; M/G/1; FB; LAS; SET; foreground-background; least attained service; PS; processor sharing; SRPT; shortest remaining processing time; PSJF; preemptive shortest job first

*Supported by NSF Career Grant CCR-0133077, NSF Theory CCR-0311383, NSF ITR CCR-0313148, IBM Corporation via Pittsburgh Digital Greenhouse Grant 2003, and a NSF Graduate Research Fellowship.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGMETRICS'05, June 6–10, 2005, Banff, Alberta, Canada.
Copyright 2005 ACM 1-59593-022-1/05/0006 ...$5.00.

1. INTRODUCTION

As size based policies have become prevalent in modern applications including routers [17, 18], web servers [9, 19], and transport protocols [33], scheduling research has shifted to questioning the "fairness" of such policies [1, 5, 7, 10, 17, 20, 31]. For example, is a policy that biases towards small job sizes as fair to large jobs as a policy without bias? To answer this question, researchers have studied the mean conditional response time experienced by a job of size x under a policy P, $E[T(x)]^P$. The typical setting for these studies is an M/GI/1 queue with load $\rho = \lambda E[X] < 1$, where λ is the mean arrival rate and X is a random variable distributed according to the service (job size) distribution. One popular criterion for fairness that has emerged is:

DEFINITION 1.1. *A scheduling policy, P, is **fair** under service distribution X and load ρ if for all x, $E[T(x)]^P/x \le 1/(1-\rho)$. Otherwise P is **unfair**.*

Definition 1.1 was introduced in [1], and has served as the basis for the work in [5, 7, 10, 17, 31]. The definition compares the mean response time of jobs with different sizes using the *metric* $E[T(x)]^P/x$ and then uses the *criterion* $1/(1-\rho)$ to distinguish between fundamentally different fairness behaviors.

In this paper, we extend the approach used to study $E[T(x)]$ in order to investigate the conditional variance in response time seen by a job of size x under policy P, $Var[T(x)]^P$, and higher conditional moments of response time across x under a wide range of scheduling policies. There has been a significant amount of prior literature *deriving* $Var[T(x)]$ under many common policies [26, 35, 12, 13]. However, possibly due to the complicated nature of these formulas, little work has studied the *behavior* of $Var[T(x)]$ across x. Recently, $Var[T(x)]$ has been investigated under a few common policies using simulation techniques [7]; however prior to that, investigations focused on providing customers estimates of response time as a function of the *full system state* at arrival under First-Come-First-Served (FCFS) and Processor-Sharing (PS) [28, 29, 30].

We choose to study $Var[T(x)]$ because we envision a situation where users know the size of the job they are submitting and would like to minimize the difference between their *experienced* response time, $T(x)$, and their *expected* response time, $E[T(x)]$; thus maximizing "predictability." Reducing "unpredictability" in response times can be more important to users than reducing the response times themselves because waiting much longer than expected causes far more user frustration than simply waiting longer on average [3, 36]. Note that $Var[T(x)]$ provides a better measure of user-perceived "predictability" than does $Var[T]$ in the situation where the size of the job is known by the user. Further, many QoS guarantees are of the form "90% of the time a job of size x will have response time $< g(x)$," for some function $g(\cdot)$. Such guarantees can be phrased as bounding $Var[T(x)]$ by applying Chebyshev's Inequality (see Section 2).

We define a notion of "predictability" by scaling $Var[T(x)]$ as follows.

DEFINITION 1.2. *A job size x is treated **predictably** under policy P, service distribution X, and load ρ if*

$$\frac{Var[T(x)]^P}{x} \leq \frac{\lambda E[X^2]}{(1-\rho)^3}$$

*Otherwise a job size x is treated **unpredictably**. A scheduling policy P is **predictable** if every job size is treated predictably. Otherwise P is **unpredictable**.*

It may not be immediately obvious why the appropriate *metric* for our definition is $Var[T(x)]/x$ or why the appropriate *criterion* is $\lambda E[X^2]/(1-\rho)^3$. We will discuss this in detail in Section 2.

We will show that scheduling policies have many different patterns of predictability. While some policies have monotonically increasing, but bounded, $Var[T(x)]/x$ under all loads and service distributions; others exhibit non-monotonic behavior where some range of sizes is overly penalized under some or all loads and service distributions. We introduce the following three classes of scheduling policies in order to distinguish between these patterns of predictability.

DEFINITION 1.3. *A scheduling policy P is: (i) **Always Predictable** if P is predictable under all loads and service distributions; (ii) **Sometimes Predictable** if P is predictable under some loads and service distributions; and unpredictable under other loads and service distributions or (iii) **Always Unpredictable** if P is unpredictable under all loads and service distributions.*

Introducing these three classes allows us to analyze large groups of policies with respect to predictability instead of focusing on any particular individual policy. This focus provides an understanding of the effects of *scheduling mechanisms and heuristics* on the functional behavior of $Var[T(x)]^P$ and thus is useful beyond the scope of common idealized policies. For example, we find that non-preemptive policies can be either Sometimes Predictable or Always Unpredictable; whereas preemptive policies can fall into any of the three classes (see Figure 1). We show that PS and Preemptive-Last-Come-First-Served (PLCFS) are Always Predictable. Further, we concentrate on various forms of prioritization: (a) size based, (b) age based, and (c) remaining size based. We show that all policies in (a) are Always Unpredictable, while policies in (b) and (c) may be Sometimes Predictable or Always Unpredictable.

After developing a classification for $Var[T(x)]/x$, we then pose the question of whether similar classifications exist for higher conditional moments of response time. The difficulty is that for higher moments the appropriate metric and criterion are even more unclear. For the ith moment, we will find that many common statistics such as raw moments, central moments, and moments of slowdown, $S(x) = T(x)/x$, do *not* provide appropriate metrics. Instead, we discover that little used *cumulant moments* facilitate the comparison of higher conditional moments of response time. This allows us to generalize Definitions 1.1 and 1.2 and define a metric with which to compare the higher moments of conditional response time across job sizes. Further, we motivate a conjecture that the constant $\lambda E[B^i]$, where B is an M/GI/1 busy period, will provide a criterion for the ith cumulant that distinguishes between fundamentally different functional behaviors.

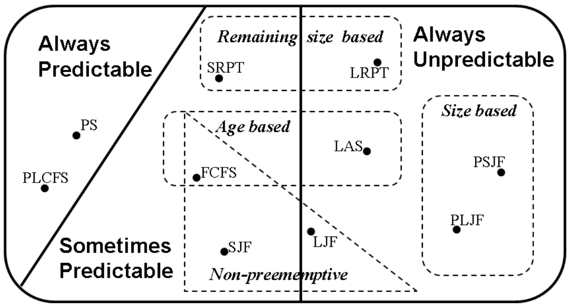


Figure 1: A diagram of the main results proved about the classification of predictability. A few examples of common policies in each class are shown.

Throughout this paper we will consider a work conserving, preempt-resume M/GI/1 system with a continuous service distribution having a finite third moment. We let $T(x)$ be the steady-state response time for a job of size x, where the response time is the time from when a job enters the system until it completes service. Let $\rho < 1$ be the system load. That is $\rho = \lambda E[X]$, where λ is the arrival rate of the system and X is a random variable distributed according to the service (job size) distribution $F(x)$ having density function $f(x)$ defined for all $x \geq 0$. Let $\overline{F}(x) = 1 - F(x)$. Define the slowdown for a job of size x, $S(x) = T(x)/x$. Define $m_i(x) = \int_0^x t^i f(t)dt$ and $\widetilde{m}_i(x) = i\int_0^x t^{i-1}\overline{F}(t)dt$. Notice that $m_i(x)/F(x) = E[X^i|X < x]$ and $\widetilde{m}_i(x)$ is the ith moment of $X_x = \min(X, x)$. Further, define $\rho(x) = \lambda m_1(x)$ and $\widetilde{\rho}(x) = \lambda\widetilde{m}_1(x)$. As introduced in [16], define the normalized k-th moment for $k = 2, 3$ of X to be $M_2[X] = \frac{E[X^2]}{E[X]^2}$ and $M_3[X] = \frac{E[X^3]}{E[X^2]E[X]}$. Notice that $M_2[X] = C^2 + 1$ where C is the coefficient of variation and $M_3[X]$ is closely related to skewness. Finally, we let B be the duration of a busy period, and $B(x)$ be the duration of a busy period started by a job of size x.

2. DEFINING PREDICTABILITY

It is clear that $Var[T(x)]^P$ is related to the "predictability" of a scheduling policy P; however the motivations for the metric and criteria in Definition 1.2 are not obvious. We will first illustrate that Definition 1.2 is mathematically grounded and that it parallels Definition 1.1. We will then show that Definition 1.2 is also motivated by the goal of providing QoS guarantees.

Relating predictability and fairness

Recall that Definition 1.1 for *fairness* stems from two motivations. First, intuitively, $E[T(x)]^P$ should be proportional to x since small jobs should have small response times and large jobs should have large response times. PS accomplishes this since $E[T(x)]^{PS} = x/(1-\rho)$. Further, PS is typically thought of as a fair policy because at every instant every job in the system receives an equal share of the server. Thus, a scheduling policy P can be viewed as unfair if jobs of some size x have $E[T(x)]^P > E[T(x)]^{PS} = x/(1-\rho)$.

Second, more formally, when comparing $E[T(x)]^P$ across x, we want a *metric* that scales $E[T(x)]^P$ appropriately to allow for comparison of $E[T(x)]^P$ between small and large x. For $E[T(x)]^P$, it is clear that $1/x$ is an appropriate scaling factor because $E[T(x)]^P = \Theta(x)$ under all work conserving scheduling policies [10], and thus we need to normalize by the growth rate. The criterion $1/(1 -$

$\rho)$ stems from two formal motivations [31]. First, it provides a min-max notion of fairness: $\min_P \max_x E[T(x)]^P/x = 1/(1-\rho)$. Second, $1/(1-\rho)$ provides a criterion that distinguishes between patterns of behavior of policies with respect to the metric $E[T(x)]^P/x$. The defined metric and criterion for fairness together allow a classification of scheduling policies as one of Always Fair, Sometimes Fair, or Always Unfair [31].

In defining predictability, Definition 1.2, while not related to the performance of PS (as was the case with Definition 1.1 for fairness), does have other properties that parallel Definition 1.1. The scaling factor for $Var[T(x)]^P$ in our definition of predictability is still $1/x$. This is motivated by the growth rate of $Var[T(x)]^P$, which is $\Theta(x)$ for common preemptive policies and $O(x)$ for all work conserving policies (see Theorem 2.1). Hence, scaling by $1/x$ makes sense; whereas using a stronger scaling such as $1/x^2$ would cause $Var[T(x)]^P/x^2 \to 0$ as $x \to \infty$.

THEOREM 2.1. *Under all work conserving scheduling policies* P, $\lim_{x\to\infty} Var[T(x)]^P/x \leq \lambda E[X^2]/(1-\rho)^3$. *Equality holds for* $P \in \{PSJF, LAS, SRPT, PLCFS, PS\}$.

This result is a special case of Theorem 6.2.

The *criterion* $\lambda E[X^2]/(1-\rho)^3$ in Definition 1.2 is also motivated by Theorem 2.1. Just as the criterion $1/(1-\rho)$ used in Definition 1.1 has the property that $\lim_{x\to\infty} E[T(x)]^P/x = 1/(1-\rho)$ under many common policies, Theorem 2.1 illustrates that the criterion in Definition 1.2 also serves as the limit for $Var[T(x)]^P/x$ under many common scheduling policies. Further, the results in this paper will illustrate that the criterion proves to be empirically useful because it differentiates between contrasting $Var[T(x)]^P/x$ behaviors. Specifically, when size based policies are unpredictable it is because $Var[T(x)]^P/x$ has a non-monotonic "hump" behavior – where some mid-range job sizes are treated the most unpredictably. On the other hand, when policies behave predictably it is because $Var[T(x)]^P/x$ is monotonically increasing.

It is important to observe that the criteria for fairness and predictability both derive from a *busy period*, they are $E[B(x)]/x$ and $Var[B(x)]/x$ respectively. In Section 6, we use this observation to present metrics and criteria for all higher moments that generalize fairness and predictability.

Relating predictability and QoS

Intuitively, the notion of "predictability" conveys the idea that $T(x)^P - E[T(x)]^P$ is never too large. Many QoS guarantees take the form "90% of the time $T(x) - E[T(x)] < g(x)$," or equivalently $P(T(x) - E[T(x)] \geq g(x)) \leq 10\%$. Chebyshev's Inequality [22] gives us a bound of the form

$$P(T(x)^P - E[T(x)]^P \geq g(x))) \leq \frac{Var[T(x)]^P}{g(x)^2} \qquad (1)$$

Thus, we can provide the desired QoS guarantee by ensuring that $Var[T(x)]/g(x)^2$ is not too large.[1] Looking more closely at Equation 1, we need to ask "what is the smallest value of $g(x)$ that allows $Var[T(x)]/g(x)^2$ to be bounded by a constant (10% in the above example) for all x?"

Suppose that $g(x) = kx^i$ for some k independent of x and some constant i. Then, we need to choose the smallest i that allows $Var[T(x)]/g(x)^2$ to be bounded by a constant. Notice that we can immediately rule out $i > 1$ because $T(x)^P$ and $E[T(x)]^P$ grow linearly in x for all P; thus it does not make sense to bound $T(x)^P - E[T(x)]^P$ by something growing superlinearly. We can also rule out $i < 1/2$ because for such i, $Var[T(x)]^P/x^{2i} \to \infty$ as $x \to \infty$ under all P. This leaves $i \in [1/2, 1]$, where $i = 1/2$ is the most desirable because it provides the tightest bound on $T(x) - E[T(x)]$ as x grows.

Definition 1.2 uses the metric $Var[T(x)]^P/x$, which corresponds to choosing $i = 1/2$. This choice makes sense because $Var[T(x)]^P/x$ is $O(1)$ under all work conserving policies P. Thus, any policy that is predictable will allow a QoS bound that is constant across x. Note that choosing $i \in (1/2, 1]$ is also reasonable; however the results are less interesting.[2]

3. ALWAYS PREDICTABLE

We start to develop a classification of predictability by studying the class of Always Predictable policies, policies where every job size is treated predictably under all service distributions and system loads. Two well known policies that are Always Predictable are PLCFS and PS. It is immediate to see that PLCFS is Always Predictable since $T(x)^{PLCFS} = B(x)$, and thus

$$Var[T(x)]^{PLCFS} = Var[B(x)] = \frac{\lambda x E[X^2]}{(1-\rho)^3}$$

However, understanding the variance of PS is more difficult. Working from the transform, [35] presents the following useful representation for $Var[T(x)]^{PS}$:

$$Var[T(x)]^{PS} \quad = \quad \frac{2}{(1-\rho)^2}\int_0^x (x-t)\overline{R}(t)dt$$

where $\overline{R}(t) = 1 - R(t)$ and $R(t) = (1-\rho)\sum_{n=0}^{\infty}\rho^n F^{*n}(t)$ with $F^{*n}(t) = \int_0^\infty F^{*(n-1)}(t-s)dF^{*1}(s)$,

$F^{*1}(t) = \frac{1}{E[X]}\int_0^t (1-F(s))ds$, and $F^{*0}(t) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$.

The complexity of this formula has led to mainly asymptotic analysis of the conditional variance of PS. However, we will be able to exploit this asymptotic information in order to show that PS is predictable for all x.

THEOREM 3.1. PS *is Always Predictable. Further,* $Var[T(x)]^{PS}/x$ *is strictly monotonically increasing in* x.

PROOF. We will prove the result by showing that $\frac{d}{dx}\left(Var[T(x)]^{PS}/x\right) > 0$ for all x. In combination with Theorem 2.1 this will complete the proof.

$$\frac{d}{dx}\frac{Var[T(x)]^{PS}}{x} = \frac{2}{(1-\rho)^2}\frac{d}{dx}\left(\int_0^x \overline{R}(t)dt - \frac{1}{x}\int_0^x t\overline{R}(t)dt\right)$$
$$= \quad \frac{2}{(1-\rho)^2}\left(\overline{R}(x) - \overline{R}(x) + \frac{1}{x^2}\int_0^x t\overline{R}(t)dt\right) > 0$$

□

It is interesting that $Var[T(x)]^{PS}/x$ is monotonically increasing in x under all service distributions. This is different than $E[T(x)]^{PS}/x = 1/(1-\rho)$, which is constant across x, and illustrates why $Var[T(x)]^{PS}/x$ is not an appropriate criterion for a definition of predictability.

[1]Note that a more complex bound including other information about the distribution of $T(x)$ could be used to provide QoS guarantees in practice. However, the simple calculation of Equation 1 provides intuition for an appropriate metric with which to study $Var[T(x)]$.

[2]For $i \in (1/2, 1]$, $Var[T(x)]^P/x^{2i} \to 0$ as $x \to \infty$ under all P. As a result, it can quickly be seen that policies fall into one of two classes based the bevaior of $Var[T(x)]^P$ as $x \to 0$, i.e whether $\lim_{x\to 0} Var[T(x)]^P/x^{2i} < \infty$. This makes intuitive sense because the bound on $T(x)^P - E[T(x)]^P$ is much looser as x grows and thus the performance of the small jobs dominates the QoS bound.

It is important to point out that the predictability of PS has been studied in much more detail by Ward and Whitt [28]. While we assume no knowledge of the system state in order to study how well response times will match with prior user experience, Ward and Whitt study how well $T(x)^{PS}$ can be predicted given knowledge of the system state (e.g. the number of jobs in the system upon arrival, N). They look at the question analytically as $N \to \infty$ and $x \to \infty$ and prove that predictions can be made quite accurately when either x or N is large.

4. ALWAYS UNPREDICTABLE

In this section we show that a large number of preemptive policies are Always Unpredictable, i.e. guaranteed under all system loads and all service distributions to treat some job size unpredictably. The policies in the Always Unpredictable class exhibit fundamentally different behavior with respect to $Var[T(x)]/x$ than those in the Always Predictable class. While the policies in the Always Predictable class have $Var[T(x)]/x$ that is either monotonically increasing or constant in x, the policies we study here all exhibit non-monotonic behavior – there is a "hump" in $Var[T(x)]/x$ where a small range of x has higher $Var[T(x)]/x$ than all other x (see Figure 2).

4.1 PSJF

Preemptive-Shortest-Job-First (PSJF) is the canonical example of a policy that prioritizes based on size, and it will serve as the building block for the analysis of all size based policies. Under PSJF at every moment in time, the server is processing the job with the smallest initial size. PSJF significantly improves on the mean response time of PS, and has recently been shown to be near optimal with respect to mean response time in a very strong sense [32]. Further, PSJF has the practical property that priorities can be set upon arrival and then do not need to be updated; thus implementation of PSJF is simple. The variance for a job of size x is [26]:

$$Var[T(x)]^{PSJF} = \frac{\lambda x m_2(x)}{(1-\rho(x))^3} + \frac{\lambda m_3(x)}{3(1-\rho(x))^3} + \frac{3}{4}\left(\frac{\lambda m_2(x)}{(1-\rho(x))^2}\right)^2$$

In this section, we will first prove that PSJF exhibits non-monotonic behavior in $Var[T(x)]^{PSJF}/x$, where mid-range job sizes are treated the most unpredictably. Then, we will bound the position and size of this "hump."

THEOREM 4.1. *PSJF is Always Unpredictable. Further, under all service distributions and all loads there exists some L such that all $x \geq L$ are treated unpredictably.*

PROOF. We separate this result into two cases. First, when the service distribution has an upper bound L, and second when the service distribution has no such upper bound. In the case of a bounded service distribution, it is straightforward to see that jobs of size L will be treated unpredictably. The case of unbounded service distributions is more complicated however. Observe that $Var[T(x)]^{PSJF}/x$ is increasing in x for small x. Also, recall that from Theorem 2.1 that $Var[T(x)]^{PSJF}/x \to \lambda x E[X^2]/(1-\rho)^3$ as $x \to \infty$. Hence, if we can show that the limit is approached from above, rather than below, we will have exhibited non-monotonic behavior. We accomplish this by showing that $\frac{d}{dx}\left(Var[T(x)]^{PSJF}/x\right)$ approaches 0 from below as $x \to \infty$. By observing that

$$\frac{d}{dx}\frac{Var[T(x)]^{PSJF}}{x} = \frac{x\frac{d}{dx}Var[T(x)]^{PSJF} - Var[T(x)]^{PSJF}}{x^2}$$

our goal reduces to showing that as $x \to \infty$

$$x\frac{d}{dx}Var[T(x)]^{PSJF} - Var[T(x)]^{PSJF} < 0 \qquad (2)$$

Computation yields that for any distribution with finite third moment:

$$\begin{aligned} & x\frac{d}{dx}Var[T(x)]^{PSJF} - Var[T(x)]^{PSJF} \\ = & \frac{\lambda x m_2(x)}{(1-\rho(x))^3} + O(x^4 f(x)) - Var[T(x)]^{PSJF} < 0 \text{ as } x \to \infty \end{aligned}$$

Thus, PSJF is unpredictable for all loads and all unbounded service distributions. □

Although there are always some sizes that are treated unpredictably under PSJF, most sizes receive predictable response times.

THEOREM 4.2. *Let K_1 be a constant such that $m_3(x) \leq K_1 x m_2(x)$. Then $Var[T(x)]^{PSJF} \leq Var[B(x)]h_1(\rho, x)^{PSJF}$ where*

$$h_1(\rho,x)^{PSJF} = \frac{(1-\rho)^3}{(1-\rho(x))^4}\left\{\left(1+\frac{K_1}{3}\right)+\left(\frac{5K_1}{12}-1\right)\rho(x)\right\}$$

Further, noting that $K_1 \leq 1$ for all service distributions, we have that $h_1(\rho,x) \leq \frac{(1-\rho)^3}{(1-\rho(x))^4}\left\{\frac{4}{3} - \frac{7}{12}\rho(x)\right\}$.

The proof of this result follows from direct calculation.

Notice that this bound guarantees that a large percentage of job sizes will be treated predictably. In particular, all job sizes such that $\rho(x) \leq 1 - \left(\frac{4}{3}(1-\rho)^3\right)^{1/4}$. For example, if the load is 0.8, all job sizes x such that $\rho(x) \leq 0.678$ will be treated predictably. If the job size distribution is highly variable, this is nearly all jobs (since a small percentage of the largest jobs make up half the load).

EXAMPLE 4.1. *Consider $X \sim Exp(1)$. Thus, $f(x) = e^{-x}$. Then, $\rho(x) = \rho\left(1 - e^{-x} - xe^{-x}\right)$. So, $\rho(x) \leq 1 - \left(\frac{4}{3}(1-\rho)^3\right)^{1/4}$ when $e^{-x} + xe^{-x} \geq 1 - \frac{1-\left(\frac{4}{3}(1-\rho)^3\right)^{1/4}}{\rho}$. This says that when $\rho = 0.8$, PSJF will be predictable for at least jobs of size $x \leq 3.3$. Thus, PSJF will be predictable for at least 96.3% of the jobs.*

Further, an even larger percentage of job sizes can be shown to be treated predictably if K_1 is bounded below 1.[3]

Theorem 4.2 shows that small (and in fact most) job sizes receive predictable service, but the question still remains as to how unpredictably the large jobs can be treated. The dependence of Theorem 4.2 on the bound $m_3(x) \leq K_1 x m_2(x)$ leads to an overestimate of $Var[T(x)]^{PSJF}$ for large job sizes. Thus, we must take a different approach in order to obtain a tighter bound for the large jobs.

THEOREM 4.3. *For jobs of size $x > K_2 E[X]$, $Var[T(x)]^{PSJF} \leq Var[B(x)]h_2(\rho)$ where $h_2(\rho) = \left(1 + \frac{M_3[X]}{3K_2}\right) + \frac{3\rho M_2[X]}{4K_2(1-\rho)}$.*

The proof of this Theorem follows from direct calculation.

The combination of the Theorems 4.2 and 4.3 provides a technique for determining both (i) which job sizes are treated unpredictably and (ii) how unpredictably they can be treated. We illustrate this process in the next example.

EXAMPLE 4.2. *Returning to the case of $X \sim Exp(1)$ we can use our prior calculation to set $K_2 = 3.3$ in the case where $\rho = 0.8$ in our PSJF system. Now, noting that $M_3[X] = 3$ and $M_2[X] = 2$ in the case of the exponential, we have $Var[T(x)]^{PSJF} \leq 3.1 Var[B(x)]$. Thus, although PSJF is Always Unpredictable, even in the case of an exponential service distribution with $\rho = 0.8$, PSJF is only unpredictable for at most 4% of jobs and this small fraction of jobs only receives a factor of 3.1 higher variance. This agrees with the behavior shown in Figure 2.*

[3]For instance, if $f(x)$ is decreasing, K_1 can be set to 3/4.

4.2 Preemptive size based policies

In this section we build on the analysis of PSJF and show that all size based policies are Always Unpredictable.

DEFINITION 4.1. *Under a* ***preemptive size based policy****, the priority of a job is assigned based on a fixed priority function that is a bounded bijection from job sizes to priorities. Priorities are assigned upon arrival and cannot be adjusted. The job with the highest priority is run at all instants, and if two jobs of the same size (and thus priority) are in the system, then the job that arrived first is given higher priority.*

Notice the generality of the definition of preemptive size based policies. The definition includes PSJF, but it also includes Preemptive-Longest-Job-First (PLJF) and many hybrid policies that bias towards small jobs but also give high priority towards some larger jobs to curb unfairness.

Although this group of policies is quite broad, there are some limitations to the definition of preemptive size based policies that hopefully can be addressed in future research. The class of preemptive size based policies does not include policies where jobs of different sizes all have equivalent priorities. Further, the results in this section do not include randomized policies. Thus, there may be a randomized size based policy from being that is predictable under all service distributions and all loads – though the randomization procedure will likely need to depend on the service distribution.

THEOREM 4.4. *All preemptive size based policies are Always Unpredictable.*

PROOF. We separate the proof into two cases. First, the case where a finite job size receives the lowest priority; and second the case where no finite job size receives the lowest priority.

First, let P be a preemptive size based policy where a finite size s has the lowest priority. Let W be the work in the system seen by an arrival. Then

$$Var[T(s)]^P \quad = \quad Var[B(s+W)] > Var[B(s)] = \frac{\lambda s E[X^2]}{(1-\rho)^3}$$

So s is always treated unpredictably under such a policy.

Next, let P be a preemptive size based policy where no finite job size s has the lowest priority. In this case there must be a sequence of sizes with decreasing priorities $\{s_i\}$ such that for some i, the priority of s_i is less than the priority of any $x \notin \{s_i\}$. Note that as $N \to \infty$, $\sum_{i>N} \rho(s_i) \to 0$ because our service distribution is continuous. Now there are three cases to deal with. The limit of this sequence could be 0, some finite s, or infinity. (If the limit does not exist, we can apply the same arguments to any of the points it oscillates between.)

First consider the subcase where the limit of the sequence is zero. Then there exists an infinite decreasing sequence of $\{s_i\}$ such that for all $x > s_i$, x has priority over s_i. As $i \to \infty$ we see that $\frac{Var[T(s_i)]^P}{s_i} \to \frac{Var[B(W)]}{s_i} = \infty$, which completes this case.

The subcase where the limit approaches some finite s can be reduced to the earlier case of a finite size s having the lowest priority.

Finally, we consider the subcase where the limit of the sequence is infinity. Pick an s_i such that jobs of size s_i are treated unpredictably under PSJF, and jobs of size s_i have lower priority than jobs of size t for all $t < s_i$. Note that we can always find such an s_i. Further, since jobs of size $x > s_i$ may also have higher priority than s_i we have $Var[T(s_i)]^P \geq Var[T(s_i)]^{PSJF} > \frac{\lambda s_i E[X^2]}{(1-\rho)^3}$. □

4.3 LAS

The Least-Attained-Service (LAS) policy[4] is the canonical example of a policy that prioritizes based on age. Under LAS, the job with the least attained service gets the processor to itself. If several jobs all have the least attained service they timeshare the server via PS. This is a practical policy since a job's age is always known, though its size may not be known. LAS improves upon PS with respect to mean response time and mean slowdown when the job size distribution has a decreasing failure rate (DFR) [21] and closely approximates the optimal policy for mean response time, SRPT, under DFR distributions. Recently a stream of research has suggested that LAS can provide significant improvements for routers [17, 18]. We have [34]:

$$Var[T(x)]^{LAS} \quad = \quad \frac{\lambda x \widetilde{m}_2(x)}{(1-\widetilde{\rho}(x))^3} + \frac{\lambda \widetilde{m}_3(x)}{3(1-\widetilde{\rho}(x))^3} + \frac{3}{4}\left(\frac{\lambda \widetilde{m}_2(x)}{(1-\widetilde{\rho}(x))^2}\right)^2$$

In this section, we will first prove that LAS exhibits non-monotonic behavior in $Var[T(x)]^{LAS}/x$, where large, but not the largest, job sizes are treated the most unpredictably. We will then bound the position and size of this "hump" through bounds on $Var[T(x)]^{LAS}$.

LEMMA 4.1. *For all x, $Var[T(x)]^{PSJF} \leq Var[T(x)]^{LAS}$*

Combining Lemma 4.1 with Theorem 4.1, we have:

COROLLARY 4.1. LAS *is Always Unpredictable. Further, under all service distributions and all loads there exists some L such that all $x > L$ are treated unpredictably.*

There are always some job sizes that are treated unpredictably under LAS, however most job sizes receive predictable response times.

THEOREM 4.5. *Let K_1 be a constant such that $m_3(x) \leq K_1 x m_2(x)$. Then $Var[T(x)]^{LAS} \leq Var[B(x)] h_1(\rho, x)^{LAS}$ where*

$$h_1(\rho, x)^{LAS} = \frac{(1-\rho)^3}{(1-\widetilde{\rho}(x))^4}\left\{\left(1+\frac{K_1}{3}\right) + \left(\frac{2K_1}{3} - 1\right)\widetilde{\rho}(x)\right\}$$

Further, noting that $K_1 \leq 1$ for all service distributions we have that $h_1(\rho, x)^{LAS} \leq \frac{(1-\rho)^3}{(1-\widetilde{\rho}(x))^4}\left\{\frac{4}{3} - \frac{1}{3}\widetilde{\rho}(x)\right\}$.

The proof follows using Lemmas A.1 and A.2.

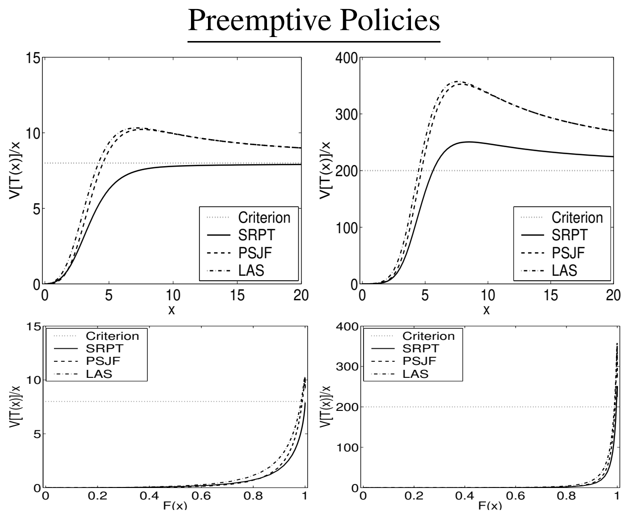
This bound guarantees that a large percentage of job sizes will be treated predictably. In particular, all job sizes such that $\widetilde{\rho}(x) \leq 1 - \left(\frac{4}{3}(1-\rho)^3\right)^{1/4}$. Thus, if $\rho = 0.8$, all jobs such that $\widetilde{\rho}(x) \leq 0.678$ will be treated predictably. However, the question still remains as to how unpredictably the large jobs can be treated.

THEOREM 4.6. *For jobs of size $x > K_2 E[X]$, $Var[T(x)]^{LAS} \leq E[B(x)] h_2(\rho)$, where $h_2(\rho) = \left(1 + \frac{M_3[X]}{3K_2}\right) + \frac{3\rho M_2[X]}{4K_2(1-\rho)}$.*

Note that this is the same bound on the hump size as under PSJF. The difference will come in the application because the bound on the position of the hump is in terms of $\widetilde{\rho}(x)$ under LAS instead of $\rho(x)$ as under PSJF, so K_2 will be smaller. We illustrate this using our running example.

---

[4]Note that LAS is sometimes referred to by two other names: Foreground-background (FB) and Shortest-Elapsed-Time (SET).



(a) Low load ($\rho = 0.5$) (b) High load ($\rho = 0.8$)

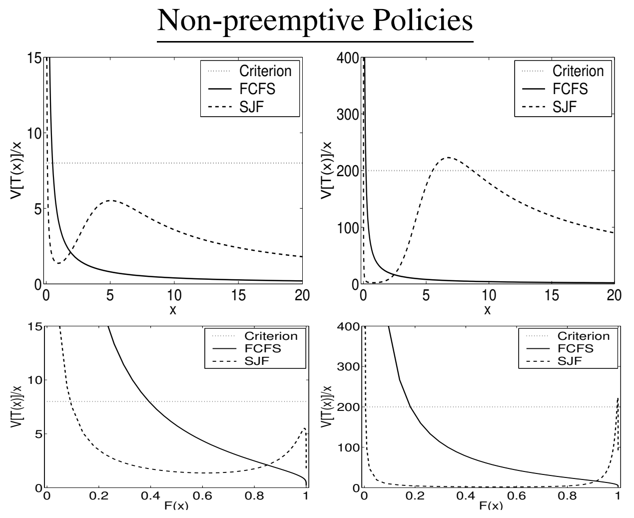
Figure 2: The conditional variance of PLCFS, SRPT, PSJF, and LAS are shown. The service distribution is exponential with mean 1. The dotted line shows the criterion for predictability. Notice that when load is low (left column), SRPT is predictable, but when load is high (right column) SRPT is unpredictable. In contrast PSJF and LAS are Always Unpredictable. However, as seen in the bottom row, they are only unpredictable to a small percentage of the large jobs.

EXAMPLE 4.3. *Again consider $X \sim Exp(1)$. Then, $\widetilde{\rho}(x) = \rho\left(1 - e^{-x}\right)$. So, $\widetilde{\rho}(x) \leq 1 - \left(\frac{4}{3}(1-\rho)^3\right)^{1/4}$ when $e^{-x} \geq 1 - \frac{1-\left(\frac{4}{3}(1-\rho)^3\right)^{1/4}}{\rho}$. This says that when $\rho = 0.8$, LAS will be predictable for at least jobs of size $x \leq 1.8$. Thus, LAS will be predictable for at least 83.4% of the jobs.*

We can use this result to set $K_2 = 1.8$ in the case where $\rho = 0.8$, which gives $Var[T(x)]^{PSJF} \leq 4.9Var[B(x)]$. Thus, although LAS is Always Unpredictable, when $\rho = 0.8$, LAS is only unpredictable for at most 17% of jobs and this fraction of jobs only receives at most a factor of 5 higher variance. Note that although this is not nearly as good as what we saw under PSJF, LAS is operating without knowledge of job sizes. This agrees with the behavior shown in Figure 2.

5. SOMETIMES PREDICTABLE

In this section we show that many policies (e.g. SRPT, FCFS) fall into the Sometimes Predictable class. That is many policies can be predictable for all job sizes under some loads and service distributions and unpredictable for some job size under other loads and service distributions. The policies that are Sometimes Predictable have more complicated behavior with respect to $Var[T(x)]$ than we observed in the cases of the Always Predictable and Always Unpredictable classes. For instance, we show that under all service distributions, SRPT maintains monotonic $Var[T(x)]/x$ for low loads similarly to policies in the Always Predictable class; but under high enough load, SRPT exhibits the same non-monotonic behavior seen under PSJF and LAS. Interestingly, load has the opposite effect for non-size based non-preemptive policies such as FCFS, which are predictable under high loads and unpredictable under low loads. These behaviors are illustrated in Figures 2 and 3.



(a) Low load ($\rho = 0.5$) (b) High load ($\rho = 0.8$)

Figure 3: The conditional variance of PLCFS, FCFS, and SJF are shown. The service distribution is exponential with mean 1. The dotted line shows the criterion for predictability. Notice that when load is low, the hump in SJF stays below the criterion for predictability, but when load is high the jobs in the hump of $Var[T(x)]^{SJF}/x$ are treated unpredictably. In contrast, $Var[T(x)]^{FCFS}/x$ is always monotonically decreasing. However, as seen in the bottom row, FCFS treats a significant percentage of small jobs unpredictably; whereas, especially under high load, SJF only treats a small percentage of jobs unpredictably.

5.1 Preemptive age based policies

In this section we build on the analysis of LAS and show that age based policies are either Sometimes Predictable or Always Unpredictable.

DEFINITION 5.1. *Under a **preemptive age based policy**, the priority of a job is assigned based on a fixed priority function that is a bounded bijection from ages to priorities. The priority of a job is updated as the age (attained service) of the job changes. The job with the highest priority is preemptively given service, and if two jobs have the same age (and thus priority), the job that attained that age first is given higher priority.*

It is important to point out the generality of the definition of preemptive age based policies. Not only does this definition include LAS, but it also includes FCFS and an array of hybrid policies that bias towards small ages but also give some larger ages high priority in order to curb unfairness. As with the definition of size based policies in Section 4.2, there are some limitations to the definition of age based policies that are left for future work.

THEOREM 5.1. *All preemptive age based policies are either Sometimes Predictable or Always Unpredictable. Further, all age based policies where no finite age receives the lowest priority are Always Unpredictable.*

PROOF. We again separate the proof into two cases. First, the case where a finite age receives the lowest priority; and second, the case where no finite age receives the lowest priority.

Let P be a preemptive age based policy where a finite age $a > 0$ has the lowest priority. Now, consider a job j_s of size $s = a + \varepsilon$

where $\varepsilon \to 0$. First notice that all of the jobs in the system when j_s arrives will complete or achieve age at least s while j_s is in the system, since j_s will get stuck with age a. Further, all jobs that arrive while j_s is in the system will either complete or get worked on up to at least age a while j_s is in the system. Notice that $Var[T(s)]$ in this system is larger than $Var[T(s)]^{LAS}$ in a system having a distribution with finite support truncated at s. Further, $Var[T(s)]^{LAS}$ in the system with finite support is worse than $Var[T(s)]^{PSJF}$ in the system with the same service distribution. Finally, note that we have already shown that a PSJF system where a finite sized job receives the lowest priority is Always Unpredictable, thus we can conclude that P is unpredictable in this case.

Note that this proof technique fails for the case where $a = 0$ because when we truncate the service distribution we are left with a degenerate distribution, for which our prior results for PSJF do not apply. To handle the case of $a = 0$, note that all P such that jobs with zero age have the lowest priority are non-preemptive. Finally, we show in Theorem 5.5 that all non-preemptive policies are unpredictable under service distributions that are defined on a neighborhood around zero.

Next let P be a preemptive age based policy where no finite age has lowest priority. This case can be dealt with symmetrically to the argument used in Theorem 4.4. ☐

5.2 SRPT

SRPT is perhaps the most important of the remaining size based policies due to the fact that it has been shown to be optimal with respect to mean response time [23]. Under SRPT, at every moment in time, the server is processing the job with the smallest remaining processing time. Recently SRPT has received a lot of attention [1, 15, 31, 17, 7] due to results showing that using SRPT in web servers can decrease user response times dramatically [9, 19]. However, in this stream of research the behavior of $Var[T(x)]^{SRPT}$ has only been evaluated using trace-based simulation [7]. Thus, we believe this paper represents the first analytic study of the behavioral properties of the conditional variance of response time under SRPT. The variance of response time for a job of size x under SRPT is [24]:

$$\begin{aligned} Var[T(x)]^{SRPT} &= \int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^3} dt + \frac{\lambda \widetilde{m}_3(x)}{3(1-\rho(x))^3} \\ &+ \frac{3}{4}\left(\frac{\lambda \widetilde{m}_2(x)}{(1-\rho(x))^2}\right)^2 - \frac{\lambda^2 x^2 \widetilde{m}_2(x)\overline{F}(x)}{(1-\rho(x))^4} \end{aligned}$$

We will start the section by showing that SRPT provides predictable response times for all job sizes at low load, regardless of the service distribution. Then, we show that under any service distribution, when the load is high enough, SRPT will be unpredictable to some job size. Finally, we show that, even when SRPT might not provide predictable response times for all job sizes, only a tiny percentage of the jobs receive unpredictable response times, and this unpredictability is not too bad.

THEOREM 5.2. *Let K_1 be a constant such that $m_3(x) \le K_1 x m_2(x)$. Under all service distributions* SRPT *is predictable when $\rho < 0.4$. Further, for x such that $\rho(x) > 0.4$, $Var[T(x)]^{SRPT} \le Var[B(x)] h_1(\rho, x)^{SRPT}$ where*

$$h_1(\rho, x)^{SRPT} = \frac{(1-\rho)^3}{(1-\rho(x))^4}\left\{\left(1 - \frac{2}{3}K_1\right) + \left(\frac{5}{3}K_1 - 1\right)\rho(x)\right\}$$

Noting that for all distributions $m_3(x) \le x m_2(x)$, we can set $K_1 = 1$ and obtain $h_1(\rho, x) \le \frac{(1-\rho)^3}{(1-\rho(x))^4}\left\{\frac{1}{3} + \frac{2}{3}\rho(x)\right\}$.

PROOF. Most of the proof is purely algebraic calculation, so we will only present the major steps. First, we upper bound $Var[T(x)]^{SRPT}$ using Lemmas A.1 and A.3

$$Var[T(x)]^{SRPT} \quad \le \quad Var[B(x)]^P \left(1 + \frac{\widetilde{m}_3(x)(5\rho(x) - 2)}{3xE[X^2](1-\rho(x))}\right)$$

From this, we see that $Var[T(x)]^{SRPT} \le Var[B(x)]$ for all x such that $5\rho(x) - 2 < 0$, i.e. $\rho(x) \le 0.4$. Then, we apply Lemma A.2 in the case when $\rho(x) > 0.4$ to finish the proof. ☐

Using this theorem, we can see that most job sizes will be treated predictably under SRPT even under high load. For x such that $\rho(x) > 0.4$, $Var[T(x)]^{SRPT} \le Var[B(x)]$ whenever $\rho(x) \le 1 - (1-\rho)^{3/4}$. Notice that this gives a much better range than the $\rho(x) < 0.4$ when ρ is high. When $\rho = 0.8$, SRPT is predictable for all job sizes x that have $\rho(x) \le 0.7$ regardless of the service distribution.

We now show that, though SRPT can provide predictable response times for all job sizes under low loads, SRPT will be unpredictable for some job size under high enough load.

THEOREM 5.3. SRPT *is Sometimes Predictable. For every service distribution, there exists some ρ_{crit} and L such that, for all $\rho > \rho_{crit}$,* SRPT *is unpredictable for all jobs of size $x \ge L$.*

PROOF. We will prove the result only in the case of an unbounded service distribution. The proof of the bounded case is similar. In what follows, define $\delta_x = \lambda m_2(x)/x = \rho(x)\frac{m_2(x)}{x m_1(x)}$.

We will prove the result by taking advantage of the Lemmas A.4 and A.5. Define $\varepsilon_x > 0$ as

$$\varepsilon_x = \frac{\lambda x E[X^2]}{(1-\rho)^3} - \frac{\lambda x m_2(x)}{(1-\rho(x))^3} + \frac{\lambda^2 x^2 \widetilde{m}_2(x)\overline{F}(x)}{(1-\rho(x))^4}$$

Jobs of size x are treated unpredictably if the following formula is negative. Using Lemmas A.4 and A.5 we have:

$$\begin{aligned} & Var[B(x)] - Var[T(x)]^{SRPT} \\ = \quad & \frac{\lambda x m_2(x)}{(1-\rho(x))^3} - \int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^3} dt \\ & - \frac{\lambda \widetilde{m}_3(x)}{3(1-\rho(x))^3} - \frac{3\lambda^2 \widetilde{m}_2(x)^2}{4(1-\rho(x))^4} + \varepsilon_x \\ \le \quad & \left(\frac{3\lambda^2 m_2(x)^2}{(1-\rho(x)+\delta_x)^3(1-\rho(x))^3} - \frac{3\lambda^2 m_2(x)^2}{4(1-\rho(x))^4} + \varepsilon_x\right) \\ & + \left(\frac{\lambda m_3(x)}{(1-\rho(x)+\delta_x)^3} - \frac{\lambda m_3(x)}{3(1-\rho(x))^3}\right) \end{aligned} \tag{3}$$

Now, we will show that as $x \to \infty$ the above equation approaches 0 from below when ρ is higher than some $\rho_{crit} < 1$. This will complete the proof because it will guarantee the existence of a ρ_{crit} such that, for all $\rho > \rho_{crit}$, all x larger than some L will be treated unpredictably. The L comes from the fact that we show the limit converges from below as $x \to \infty$, so there must exist a L such that all $x > L$ are treated unpredictably when $\rho > \rho_{crit}$.

In what remains, the following notation will be used to simplify the calculations. Let $\varepsilon_x^* = \frac{\varepsilon_x(1-\rho(x))(1-\rho(x)+\delta_x)^3}{3\lambda^2 m_2(x)^2} > 0$. It will be important that $x\varepsilon_x^* \to 0$ as $x \to \infty$, so we show this in Lemma A.6. We apply ε_x^* in order to continue our main calculations from Equation 3. We will start by showing the first term approaches 0 from below, and then move to the second term.

Working with the first term, we have

$$\begin{aligned} \frac{3\lambda^2 m_2(x)^2}{(1-\rho(x)+\delta_x)^3(1-\rho(x))^3} - \frac{3\lambda^2 m_2(x)^2}{4(1-\rho(x))^4} + \varepsilon_x < 0 \\ 4(1-\rho(x)) + \varepsilon_x^* < (1-\rho(x)+\delta_x)^3 \end{aligned}$$

Noting that $\delta_x = \lambda m_2(x)/x \le \rho(x)$, and thus $(1-\rho(x)+\delta_x) \le 1$, we can work with the simpler formula

$$
\begin{aligned}
3(1-\rho(x)) + \varepsilon_x^* &< \rho(x)\frac{m_2(x)}{xm_1(x)} \\
\left(\frac{E[X]}{m_1(x)}\right)\left(\frac{3x + x\varepsilon_x^*}{3x + \frac{m_2(x)}{m_1(x)}}\right) &< \rho \\
\left(1 + \frac{x\int_x^\infty tf(t)dt}{xm_1(x)}\right)\left(1 - \frac{\frac{m_2(x)}{m_1(x)} - x\varepsilon_x^*}{3x + \frac{m_2(x)}{m_1(x)}}\right) &< \rho
\end{aligned}
$$

We will now show that the Left Hand Side (LHS) approaches 1 from below as $x \to \infty$. And thus show that, for large enough ρ, the first term in Equation 3 is negative.

We can see that the LHS approaches 1 from below by realizing that, $\lim_{x\to\infty} x\int_x^\infty tf(t)dt \le \lim_{x\to\infty}\int_x^\infty t^2 f(t) = 0$ while $\lim_{x\to\infty}\frac{m_2(x)}{m_1(x)} - x\varepsilon_x^* = \frac{E[X^2]}{E[X]}$.

Thus, for large enough x the LHS approaches 1 from below because the first piece of the LHS converges to 1 from above with rate $o(1/x)$ while the second piece converges to 1 from below with rate $\Theta(1/x)$. Thus, the limit of the product is 1 and it is approached from below.

We now analyze the second term in Equation 3.

$$
\begin{aligned}
\frac{\lambda m_3(x)}{(1-\rho(x)+\delta_x)^3} - \frac{\lambda m_3(x)}{3(1-\rho(x))^3} &< 0 \\
3^{1/3}(1-\rho(x)) &< (1-\rho(x)+\delta_x)
\end{aligned}
$$

Noting that $3^{1/3} < 2$, we can work with the simpler equation

$$
2(1-\rho(x)) < \left(1 - \rho(x) + \frac{m_2(x)\rho(x)}{xm_1(x)}\right)
$$

$$
\left(1 + \frac{x\int_0^x tf(t)dt}{xm_1(x)}\right)\left(1 - \frac{\frac{m_2(x)}{m_1(x)}}{x + \frac{m_2(x)}{m_1(x)}}\right) < \rho
$$

Thus, to complete the proof we need to show that the LHS approaches 1 from below as $x \to \infty$. We again see that the first piece of the LHS converges to 1 from above with rate $o(1/x)$ while the second piece converges to 1 from below with rate $\Theta(1/x)$. Thus, the limit of the product is 1 and it is approached from below.

Putting the two calculations together, we see that as $x \to \infty$ there exists a $\rho_{crit} < 1$ and an L such that for all $\rho > \rho_{crit}$ jobs of size $x > L$ are treated unpredictably. $\square$

The prior theorems give bounds on the position and existence of the hump in $Var[T(x)]^{SRPT}/x$; to bound the height of the hump it turns out to be effective to use the same bound that we have used for PSJF and LAS.

LEMMA 5.1. *For all x, $Var[T(x)]^{SRPT} \le Var[T(x)]^{LAS}$*

Lemma 5.1 allows us to use the bound already derived for LAS in Theorem 4.6. As in the cases of PSJF and LAS, the combination of the above theorems provides tight bounds on the position and size of the hump in $Var[T(x)]^{SRPT}/x$.

EXAMPLE 5.1. *Again consider $X \sim Exp(1)$. $\rho(x) \le 1 - (1-\rho)^{3/4}$ when $e^{-x} + xe^{-x} \ge 1 - \frac{1-(1-\rho)^{3/4}}{\rho}$. This says that when $\rho = 0.8$, SRPT will be predictable for at least jobs of size $x \le 3.6$, which is at least 97.2% of the jobs.*

We can use this result to set $K_2 = 3.6$ in the case where $\rho = 0.8$ which gives $Var[T(x)]^{SRPT} \le 2.9Var[B(x)]$. Thus, although SRPT can be unpredictable, in the case of an exponential service distribution with $\rho = 0.8$, SRPT is only unpredictable for at most 3% of jobs and this fraction of jobs only receives at most a factor of 3 higher variance. Note both of these bounds are better than were obtained for either PSJF or LAS.

5.3 Preemptive remaining size based policies

In this section we build on the analysis of SRPT and show that all remaining size based policies are either Sometimes Unpredictable or Always Unpredictable.

DEFINITION 5.2. *Under a **preemptive remaining size based policy**, the priority of a job is assigned based on a fixed priority function that is a bounded bijection from remaining sizes to priorities. The priority of a job is updated as the remaining size of the job changes, and the job with the highest priority is preemptively given service. If two jobs have the same remaining size, the job that attained that remaining size first is given higher priority.*

Again it is important to point out the breadth of this definition. Not only does this definition include policies such as SRPT and Longest-Remaining-Processing-Time (LRPT), it also includes many hybrid policies where small remaining sizes receive high priority and some large remaining sizes also receive high priority in order to curb unfairness.

THEOREM 5.4. *All preemptive remaining size based policies are Sometimes Predictable or Always Unpredictable.*

PROOF. We again separate the proof into two cases. First, the case where a finite remaining size receives the lowest priority. Let P be a preemptive remaining size based policy such that a non-zero remaining size r receives the lowest priority. We will return to the case where there is no such r. We will consider a service distribution having upper bound r. Consider a job j_r of original size r. Then, while j_r is in the system, at least all jobs that arrived earlier and have original size less than r will complete, since j_r will be stuck at remaining size r. Further, when j_r has remaining size t at least all arrivals of size $< t$ will complete before j_r. Thus, the system has higher $Var[T(r)]$ than an SRPT system with a service distribution truncated at r. Finally, we saw that there are situations where SRPT will give jobs of size r unpredictable service. Thus, P is unpredictable in this case.

Second, the case where there is no finite job size that receives the lowest priority can be dealt with in the same manner as in the proof of Theorem 4.4. $\square$

5.4 Non-preemptive policies

We now move to a discussion of the predictability under non-preemptive policies.[5] Non-preemptive policies have very different behavior than the preemptive policies we have considered in this work so far. We will see in Section 6.3 that large job sizes see nearly deterministic response times under non-preemptive policies, because once they begin service they cannot be interrupted. However, one result of this bias towards large job sizes is that small job sizes can receive extremely variable service because they may have to wait behind the excess of a much larger job.

In fact, whenever the service distribution includes arbitrarily small jobs, these small jobs will receive unpredictable response times under non-preemptive policies.

THEOREM 5.5. *Non-preemptive policies are either Sometimes Predictable or Always Unpredictable. All non-preemptive policies are unpredictable for all loads if the service distribution includes arbitrarily small job sizes.*

[5] Note that there is some overlap between non-preemptive policies and age based policies, e.g. FCFS is in both groups.

PROOF. Let P be a work conserving non-preemptive policy. The response time of a job j_x of size x under P is the sum of the work in the system that will serve ahead of j_x, W_{j_x}, and all arrivals while j_x is in the system that serve ahead of j_x. This second piece can be viewed as a busy period, $B_{j_x}(W_{j_x})$. We can bound W_{j_x} from below by the excess of the job at the server upon the arrival of j_x, $\mathcal{E}$. Further, we can bound $Var[B_{j_x}(W_{j_x})] \geq Var[W_{j_x}] \geq Var[\mathcal{E}]$. Finally, we can complete the proof by observing that $\lim_{x\to 0} \frac{Var[T(x)]^P}{x} \geq \lim_{x\to 0} \frac{Var[\mathcal{E}]}{x} = \infty$. □

However, in many real world cases there is some lower bound that can be placed on the size of a service request. In this case, non-preemptive policies *can* provide predictable service. We illustrate this using the examples of FCFS and non-preemptive Shortest-Job-First (SJF). Note that [26]

$$\begin{aligned} Var[T(x)]^{FCFS} &= \frac{\lambda E[X^3]}{3(1-\rho)} + \frac{\lambda^2 E[X^2]^2}{4(1-\rho)^2} \\ Var[T(x)]^{SJF} &= \frac{\lambda E[X^3]}{3(1-\rho(x))^3} + \frac{\lambda^2 m_2(x) E[X^2]}{(1-\rho(x))^4} - \frac{\lambda^2 E[X^2]^2}{4(1-\rho(x))^4} \end{aligned}$$

THEOREM 5.6. FCFS *is Sometimes Predictable. (i) For all service distributions with no non-zero lower bound,* FCFS *is unpredictable. (ii) For all service distributions with lower bound $L \neq 0$, there exists a ρ_{crit} such that for all $\rho \in (\rho_{crit}, 1)$* FCFS *is predictable.*

THEOREM 5.7. SJF *is Sometimes Predictable. (i) For all service distribution with no non-zero lower bound,* SJF *is unpredictable. (ii) For service distributions with lower bound $L \neq 0$,* SJF *is predictable when $\frac{M_3[X]}{3} + \frac{3\rho M_2[X]}{4(1-\rho)} \leq \frac{L}{E[X]}$.*

The proofs of these theorems are straightforward, and are therefore omitted.

These two examples illustrate the strange effects of size based prioritization. While FCFS and all non-size based non-preemptive policies have $Var[T(x)]/x$ that is strictly decreasing in x, size based non-preemptive policies, such as SJF, exhibit non-monotonic behavior similar to that seen under preemptive policies such as SRPT, LAS, and PSJF.

6. HIGHER MOMENTS

The similarities between the metrics and criteria for fairness and predictability beg the question of how higher conditional moments of response times vary across x. In this section we begin to ask the question of how to generalize the metrics and criteria for fairness and predictability to higher moments.

We study the limiting case of $x \to \infty$ due to the role it played in developing Definitions 1.1 and 1.2. This limiting case provides insight into how conditional moments scale with x under a range of scheduling policies, and this scaling factor will motivate an appropriate metric and criterion for comparing higher conditional moments across x. However, this case is also interesting in its own right because of intuitive worries that large jobs receive larger, more variable response times under policies that bias towards small jobs [2, 25, 27].

There has been prior work on the question of analyzing the limiting distribution of $T(x)$ as $x \to \infty$. Motivated by fairness concerns, some of this has focused on the metric of slowdown and showed that under all work conserving policies, the asymptotic slowdown of large jobs is bounded almost surely by $1/(1-\rho)$ [10]. Slowdown was considered because the focus was on unfairness; however, in terms of understanding the distribution of the response times of large jobs, we will illustrate that the scaling factor in the slowdown metric is too heavy handed and hides all information about the variability of the limiting distribution.

Our goal is to find a scaling factor that provides information about the variability (and all higher moments) in the limiting distribution of $T(x)$ as $x \to \infty$.

6.1 Busy periods

In order to illustrate the issues in finding the appropriate scaling factor for the limit as $x \to \infty$, we will begin by looking at the asymptotic behavior of $B(x)$. Busy periods are fundamental to the analysis of many size based scheduling policies, and we will find that the correct scaling factor for $B(x)$ will match the scaling necessary for response times under many policies.

The Laplace transform of $B(x)$, $\mathcal{L}_{B(x)}(s)$, is: $\mathcal{L}_{B(x)}(s) = e^{-x(s+\lambda-\lambda\mathcal{L}_B(s))}$ where $\mathcal{L}_B(s)$ is the Laplace transform of a standard M/GI/1 busy period. We can proceed to calculate the moments of $B(x)$ using the following notation: $h(s) = -x(s+\lambda-\lambda\mathcal{L}_B(s))$. Thus, $h'(0) = -\frac{x}{1-\rho}$, $h^{(i)}(0) = (-1)^i \lambda x E[B^i]$. It is important to notice that in each of these terms, x has degree one since $E[B^i]$ does not depend on x. Using $h(s)$, we can derive the moments of $B(x)$. $E[B(x)] = h'(0)$ and $E[B(x)^2] = h''(0) + h'(0)^2$. This illustrates the heavy handedness of the slowdown metric because we can see that $E[B(x)^i/x^i] = \frac{E[B(x)^i]}{x^i}$, which leads to a degenerate limiting distribution: $\lim_{x\to\infty} Var[B(x)/x] = \lim_{x\to\infty} Var[B(x)]/x^2 = 0$.

Instead of using slowdown, another natural suggestion is to try to normalize the raw moments of $B(x)$. However, as can be seen through differentiation of the Laplace transform, $E[B(x)^i] = \Theta(x^i)$. Thus, only scaling by x^i is enough to keep the limit as $x \to \infty$ from going to ∞; however this scaling leads to a degenerate limiting distribution.

A third natural suggestion for an appropriate scaling factor is to consider the central moments of $B(x)$, $E[(B(x)-E[B(x)])^i]$. Up until the third central moment, it seems that central moments can be scaled appropriately using $E[(B(x) - E[B(x)])^i]/x$. However, beyond the third central moment the central moments become convoluted, and it becomes apparent that there is no simple, appropriate scaling factor for the central moments either. For $i = 2, 3$, $E[(B(X)-E[B(X)])^i] = \lambda x E[B^i]$; however for $i = 4$, $E[(B(X) - E[B(X)])^4] = \lambda x \left(E[B^4]\right) + 3(\lambda x E[B^2])^2$.

6.2 Introducing cumulants

The observation that the first three central moments are well behaved is important however. It hints that cumulants might provide the correct asymptotic metric. Cumulants have appeared sporadically in queueing [4, 6, 14], tending to be used in large deviation limits. Cumulants are a descriptive statistic similar to moments. Formally, the cumulant moments of a random variable X, $\kappa_i[X]$ $i = 1, 2, \ldots$, are defined in terms of the moments of X, $E[X^i]$, as follows:

$$e^{\kappa_1[X]t + \frac{\kappa_2[X]t^2}{2!} + \cdots} = 1 + E[X]t + \frac{E[X^2]t^2}{2!} + \ldots$$

From this definition it follows that the cumulants of X can be generated from the cumulant generating function, $\mathcal{K}_X(s) = \log(\mathcal{L}_X(s))$. That is, $(-1)^i \mathcal{K}_X^{(i)}(0) = \kappa_i[X]$.

Although not immediately evident from the definition, cumulants have many properties that both raw and central moments lack. For instance, letting c be a constant, $\kappa_1[X + c] = \kappa_1[X] + c$ but for $i \geq 2$, $\kappa_i[X+c] = \kappa_i[X]$. Thus the first cumulant is shift-equivariant, but all others are shift-invariant. Other nice properties of cumulants include homogeneity and additivity. Homogeneity states that $\kappa_i[cX] = c^i \kappa_i[X]$. Additivity states that for independent random variables X and Y, $\kappa_i[X+Y] = \kappa_i[X] + \kappa_i[Y]$. These properties make cumulants very attractive.

Practically, the cumulants capture many of the standard descriptive statistics. Each of the first four cumulants has a useful interpretation. The first cumulant is the mean; the second cumulant is the variance; the third cumulant measures the skewness of the distribution; and the fourth cumulant measures the kurtosis of the distribution. See [11] for tables of the relationships between cumulants, moments, and central moments.

6.3 Asymptotic convergence

In contrast to raw and central moments, the cumulants of $B(x)$ have a very simple form.

$$\mathcal{K}_{B(x)}(s) = \log(\mathcal{L}_{B(x)}(s)) = -x(s + \lambda - \lambda \mathcal{L}_B(s))$$

Calculating the cumulant moments through differentiation:

$$\kappa_i[B(x)] = \begin{cases} x/(1-\rho) & \text{for } i = 1 \\ \lambda x E[B^i] & \text{for } i > 1 \end{cases}$$

Thus, using κ_i/x, it is possible to capture the variability in the limiting distribution of response time.

We will now see that this scaling factor is appropriate for a large number of preemptive scheduling policies.

We will first prove an upper bound on the asymptotic ith cumulant moment of $T(x)$ that holds for all work conserving scheduling policies. We will then show that this bound is tight, and that many common scheduling policies have limiting response times that match this bound. We will next illustrate that there are however policies that have lower asymptotic cumulants, e.g. all non-preemptive policies.

THEOREM 6.1. *Under any work conserving policy P,*

$$\lim_{x \to \infty} \frac{\kappa_i[T(x)]^P}{x} \leq \begin{cases} 1/(1-\rho) & \text{for } i = 1 \\ \lambda E[B^i] & \text{for } i > 1 \end{cases}$$

PROOF. Let P be a work conserving policy. Then, $T(x)^P \leq B(x+V)$ because $B(x+V) = B(x) + B(V)$ corresponds to the time it would take to finish all the work in the system when x arrived in addition to all the arriving work while x is in the system. Thus, as $x \to \infty$

$$\begin{aligned} \mathcal{K}_{B(x+V)}(s)/x &= \log(\mathcal{L}_{B(x+V)}(s))/x \\ &= \log(\mathcal{L}_{B(x)}(s))/x + \log(\mathcal{L}_{B(V)}(s))/x \\ &\to s + \lambda - \lambda \mathcal{L}_B(s) \end{aligned}$$

which yields $\lim_{x\to\infty} \kappa_1[B(x+V)]/x = \frac{1}{1-\rho}$ and $\lim_{x\to\infty} \kappa_i[B(x+V)]/x = \lambda E[B^i]$ for $i > 1$; from which the result follows. □

Next, we illustrate that this upper bound is tight and that many common policies have limiting response times that match the bound.

THEOREM 6.2. *For $P \in \{PSJF, LAS, SRPT, PLCFS\}$,*

$$\lim_{x \to \infty} \frac{\kappa_i[T(x)]^P}{x} = \begin{cases} 1/(1-\rho) & \text{for } i = 1 \\ \lambda E[B^i] & \text{for } i > 1 \end{cases}$$

The proof of this theorem is a sequence of straightforward calculations using the cumulant generating functions (c.g.f.) for each policy. Normalizing the c.g.f. by x and letting $x \to \infty$ shows that the c.g.f. of each of these policies converges to the c.g.f of $B(x)$.

REMARK 6.1. *We conjecture that* PS *has the same limiting behavior as the above policies; however known asymptotics are only tight enough to show the convergence of the first and second cumulants. In particular, it is known that [37]: $E[T(x)^i]^{PS} = \frac{x^i}{(1-\rho)^i} + \frac{\lambda x^{i-1} E[X^2] i(i-1)}{2(1-\rho)^{i+1}} + o(x^{i-1})$, which proves the result for $\kappa_1[T(x)]^{PS}$ and $\kappa_2[T(x)]^{PS}$. However, information about higher cumulants is lost in the $o(x^{i-1})$ term.*

The combination of Theorems 6.1 and 6.2 serves to motivate the metrics and criteria in Definitions 1.1 and 1.2 for fairness and predictability. Further, these theorems suggest that similar metrics and criteria exist for higher conditional cumulants as well. In particular, we conjecture that $\lambda E[B^i]$ will serve as criterion for $\kappa_i[T(x)]/x$ that distinguishes between fundamentally different functional behaviors. The similarities between the classifications for $\kappa_1[T(x)]/x$ (fairness) and $\kappa_2[T(x)]/x$ (predictability) suggest that similar classifications exist for higher cumulants.

Although many common policies have equivalent distributions for $T(x)$ as $x \to \infty$, the limit of Theorem 6.2 is not the only possibility.

THEOREM 6.3. *Under any non-preemptive work conserving policy P,*

$$\lim_{x \to \infty} \frac{\kappa_i[T(x)]^P}{x} = \begin{cases} 1 & \text{for } i = 1 \\ 0 & \text{for } i > 1 \end{cases}$$

The proof of this result mimics the proof of Theorem 6.1.

We have now seen examples of two possible limiting behaviors; however these are not the only possibilities. It is straightforward to show that class based preemptive priority policies can achieve arbitrary $\kappa_i[T(x)]/x$ less than $\lambda E[B^i]$.

7. CONCLUSION

In many modern computer systems improving the predictability of response times is more important than improving response times on average. This is because users expect certain response times based on past experience and become frustrated if they must wait longer than expected. So, an important goal for a scheduling policy is to provide identical jobs nearly identical response times. In order to understand how "predictable" scheduling polices are, we introduce a two part definition of predictability (Definition 1.2) that uses the metric $Var[T(x)]/x$ and the criterion $\lambda E[X^2]/(1-\rho)^3$ in order to classify which policies provide all job sizes predictable response times. Definition 1.2 parallels the definition of fairness in prior work and is further motivated by the goal of providing QoS guarantees.

We build on Definition 1.2 to develop a classification of predictability (see Figure 1). Interestingly, the classification of predictability that we derive has many parallels to the classification of fairness in [31]. For instance, PS and PLCFS are both Always Fair and Always Predictable. Similarly, SRPT is both Sometimes Fair and Sometimes Predictable and exhibits the same interesting non-monotonic (hump shaped) behavior under both measures. In fact, the entire class of remaining size based policies receives a parallel classification under the two measures. Further, size based policies are both Always Unfair and Always Unpredictable.

Although there are many similarities between the predictability and fairness classifications, there are also some important differences. Both age based and non-preemptive non-size based policies can be Sometimes Predictable but are Always Unfair. Further, although PS is both Always Fair and Always Predictable, it has much better predictability than fairness properties – while $E[T(x)]^{PS}/x$

is constant, $Var[T(x)]^{PS}/x$ is monotonically increasing which means that PS provides less variable response times for small job sizes without increasing the variability of the large job sizes.

In classifying scheduling policies with respect to predictability, we find that $Var[T(x)]^{P}/x$ can exhibit four different patterns of functional behavior (see Figures 2 and 3). Some policies, e.g. PS, have $Var[T(x)]^{P}/x$ that grows monotonically and is bounded by a constant across x; whereas other policies, e.g. FCFS, have $Var[T(x)]^{P}$ that decreases monotonically in x and is unbounded as $x \to 0$. Further, it seems that prioritization, be it age based, size based or remaining size based, leads to non-monotonic behavior in normalized conditional response times. In particular, under PSJF, LAS, and SRPT mid-range job sizes have the largest $Var[T(x)]^{P}/x$. Further, SJF has a similar hump behavior for mid-range jobs; however the smallest job sizes still receive unbounded $Var[T(x)]^{P}/x$. Our work illustrates that the criterion $\lambda E[X^2]/(1-\rho)^3$ in Definition 1.2 for predictability distinguishes between these functional behaviors. If a policy has monotonically increasing, bounded $Var[T(x)]^{P}/x$ under some service distributions and loads then the policy is Always or Sometimes Predictable; otherwise the policy is Always Unpredictable because under all service distributions and loads either $Var[T(x)]^{P}/x$ is unbounded or some mid-range job sizes receive significantly worse $Var[T(x)]/x$ than other job sizes.

The parallels between the classifications of fairness and predictability beg the question of whether similar classifications exist for higher conditional moments. In this work, we take the first step towards answering this question by studying the higher conditional moments of $T(x)$ as $x \to \infty$ in order to derive appropriate metrics and criteria. We find that the natural extension to the definitions used for fairness and predictability are the little used cumulant moments, in particular $\kappa_i[T(x)]/x$. Further, we find that $\kappa_i[T(x)]/x \to \lambda E[B^i]$ as $x \to \infty$ for all $i > 1$ (recall that $E[B^i]$ is the ith moment of a busy period). This suggests that $\kappa_i[T(x)]/x$ will serve as *metrics* and $\lambda E[B^i]$ will serve as a *criteria* in definitions of classifications for higher conditional moments. We conjecture that these definitions for higher moments will lead to classifications that parallel the work in this paper.

8. REFERENCES

[1] N. Bansal and M. Harchol-Balter. Analysis of SRPT scheduling: Investigating unfairness. In *Proceedings of ACM Sigmetrics*, 2001.

[2] 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.

[3] B. Dellart. How tolerable is delay? consumers evaluations of internet web sites after waiting. *J. of Interactive Marketing*, 13:41–54, 1999.

[4] N. Duffield, W. Massey, and W. Whitt. A nonstationary offered-load model for packet networks. *Telecommunication Systems*, 13:271–296, 2001.

[5] E. Friedman and S. Henderson. Fairness and efficiency in web server protocols. In *Proceedings of ACM Sigmetrics*, 2003.

[6] G.L.Choudhury and W. Whitt. Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/G/1 queue. *Stochastic Models*, 10:453–498, 1994.

[7] M. Gong and C. Williamson. Quantifying the properties of SRPT scheduling. In *IEEE/ACM Symposium on Mod., Anal., and Sim. of Comp. and Telecomm. Sys. (MASCOTS)*, 2003.

[8] I. Gradshteyn and I. Ryzhik. *Tables of Integrals, Series, and Products*. Academic Press, 2000.

[9] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. Implementation of SRPT scheduling in web servers. *ACM Transactions on Computer Systems*, 21(2), May 2003.

[10] M. Harchol-Balter, K. Sigman, and A. Wierman. Asymptotic convergence of scheduling policies with respect to slowdown. *Performance Evaluation*, 49(1-4):241–256, 2002.

[11] M. Kendall. *The Advanced Theory of Statistics*. Griffin, London, 1945.

[12] L. Kleinrock. *Queueing Systems*, volume I. Theory. John Wiley & Sons, 1975.

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

[14] T. Matis and R. Feldman. Using cumulant functions in queueing theory. *Queueing Systems*, 40:341–353, 2002.

[15] R. Nunez-Queija. Queues with equally heavy sojourn time and service requirement distributions. *Ann. Oper. Res*, 113:101–117, 2002.

[16] T. Osogami and M. Harchol-Balter. A closed-form solution for mapping general distributions to minimal PH distributions. In *Modelling Tools and Techniques for Comp. and Comm. System Perf. Eval.*, 2003.

[17] I. Rai, G. Urvoy-Keller, and E. Biersack. Analysis of LAS scheduling for job size distributions with high variance. In *Proceedings of ACM Sigmetrics*, 2003.

[18] I. A. Rai, G. Urvoy-Keller, M. Vernon, and E. W. Biersack. Performance modeling of LAS based scheduling in packet switched networks. In *Proc. of ACM Sigmetrics-Performance*, 2004.

[19] M. Rawat and A. Kshemkalyani. SWIFT: Scheduling in web servers for fast response time. In *Symp. on Network Computing and App.*, 2003.

[20] D. Raz, H. Levy, and B. Avi-Itzhak. A resource-allocation queueing fairness measure. In *Proc. of ACM Sigmetrics-Performance*, 2004.

[21] R. Righter, J. Shanthikumar, and G. Yamazaki. On external service disciplines in single stage queueing systems. *J. of Applied Probability*, 27:409–416, 1990.

[22] S. Ross. *Introduction to Probability Models*. Academic Press, 1997.

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

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

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

[26] H. Takagi. *Queueing Analysis: Volume 1: Vacation and Priority Systems*. North-Holland, 1991.

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

[28] A. Ward and W. Whitt. Predicting response times in processor-sharing queues. In *Proc. of the Fields Institute Conf. on Comm. Networks*, 2000.

[29] W. Whitt. Improving service by informing jobs about anticipated delays. *Managemet Science*, 45:870–888, 1999.

[30] W. Whitt. Predicting queueing delays. *Management Science*, 45:192–207, 1999.

[31] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *Proceedings of ACM Sigmetrics*, 2003.

[32] A. Wierman and M. Harchol-Balter. Nearly insensitive bounds on SMART scheduling. In *Proc. of ACM Sigmetrics*, 2005.

[33] S. Yang and G. de Veciana. Enhancing both network and user performance for networks supporting best effort traffic. volume 12, pages 349–360, 2004.

[34] S. Yashkov. Processor sharing queues: Some progress in analysis. *Queueing Systems*, 2:1–17, 1987.

[35] S. Yashkov. Mathematical problems in the theory of shared-processor systems. *J. of Soviet Mathematics*, 58:101–147, 1992.

[36] M. Zhou and L. Zhou. How does waiting duration information influence customers' reactions to waiting for services. *J. of Applied Social Psychology*, 26:1702–1717, 1996.

[37] A. Zwart and O. Boxma. Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Sys. Thry. and App.*, 35:141–166, 2000.

APPENDIX

A. USEFUL LEMMAS

LEMMA A.1. $\lambda^2\widetilde{m}_2(x)^2 \leq \frac{4}{3}\lambda\widetilde{m}_3(x)\widetilde{\rho}(x)$

PROOF.

$$\begin{aligned}\lambda^2\widetilde{m}_2(x)^2 &\leq 4\lambda^2\left(\int_0^x (t\overline{F}(t)^{1/2})^2 dt\right)\left(\int_0^x (\overline{F}(t)^{1/2})^2 dt\right)\\ &= \frac{4}{3}\lambda\widetilde{m}_3(x)\widetilde{\rho}(x)\end{aligned}$$

□

LEMMA A.2. *Let K_1 be such that $m_3(x) \leq K_1 x m_2(x)$. Then $\widetilde{m}_3(x) \leq K_1 x E[X^2]$.*

PROOF.

$$\begin{aligned}\widetilde{m}_3(x) &= m_3(x) + x^3\overline{F}(x)\\ &\leq K_1 x m_2(x)\left(1 + \frac{x^2\int_x^\infty f(t)dt}{m_2(x)}\right)\\ &\leq K_1 x m_2(x)\left(1 + \frac{\int_x^\infty t^2 f(t)dt}{m_2(x)}\right)\\ &= K_1 x m_2(x)\left(1 + \frac{E[X^2] - m_2(x)}{m_2(x)}\right) = K_1 x E[X^2]\end{aligned}$$

□

LEMMA A.3.

$$Var[R(x)]^{SRPT} = \int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^3}dt \quad \leq \quad \frac{\lambda x E[X^2]}{(1-\rho(x))^3} - \frac{\lambda\widetilde{m}_3(x)}{(1-\rho(x))^3}$$

PROOF.

$$\begin{aligned}\int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^3}dt &\leq \frac{\int_0^x \lambda m_2(t)dt}{(1-\rho(x))^3} = \frac{\lambda x m_2(x) - \lambda m_3(x)}{(1-\rho(x))^3}\\ &\leq \frac{\lambda x E[X^2]}{(1-\rho(x))^3} - \frac{\lambda x^3\overline{F}(x)}{(1-\rho(x))^3} - \frac{\lambda m_3(x)}{(1-\rho(x))^3}\\ &= \frac{\lambda x E[X^2]}{(1-\rho(x))^3} - \frac{\lambda\widetilde{m}_3(x)}{(1-\rho(x))^3}\end{aligned}$$

□

LEMMA A.4.

$$Var[R(x)]^{SRPT} = \int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^3}dt \quad \geq \quad \frac{\lambda x m_2(x) - \lambda m_3(x)}{\left(1 - \rho(x)\left(1 - \frac{m_2(x)}{x m_1(x)}\right)\right)^3}$$

PROOF. We show this using Chebyshev's Integral Inequality [8]. The following holds for $i = 1, 2, 3$.

$$\left(\int_0^x 1-\rho(t)dt\right)\left(\int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^i}dt\right) \quad \geq \quad x\int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^{i-1}}dt$$

Thus,

$$\begin{aligned}\int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^3}dt &\geq \frac{x^3\int_0^x \lambda m_2(t)dt}{\left(\int_0^x 1-\rho(t)dt\right)^3}\\ &= \frac{\lambda x m_2(x) - \lambda m_3(x)}{\left(1 - \rho(x)\left(1 - \frac{m_2(x)}{x m_1(x)}\right)\right)^3}\end{aligned}$$

□

LEMMA A.5. *Define $\delta_x = \lambda m_2(x)/x = \rho(x)\frac{m_2(x)}{x m_1(x)}$ Then*

$$\begin{aligned}&\frac{\lambda x m_2(x)}{(1-\rho(x))^3} - \int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^3}dt\\ \leq\ &\frac{3\lambda^2 m_2(x)^2}{(1-\rho(x)+\delta_x)^3(1-\rho(x))^3} + \frac{\lambda m_3(x)}{(1-\rho(x)+\delta_x)^3}\end{aligned}$$

PROOF. Let $\gamma = \frac{\lambda m_3(x)}{(1-\rho(x)+\delta_x)^3}$. Then

$$\begin{aligned}&\frac{\lambda x m_2(x)}{(1-\rho(x))^3} - \int_0^x \frac{\lambda m_2(t)}{(1-\rho(t))^3}dt\\ \leq\ &\frac{\lambda x m_2(x)}{(1-\rho(x))^3} - \frac{\lambda x m_2(x) - \lambda m_3(x)}{\left(1 - \rho(x)\left(1 - \frac{m_2(x)}{x m_1(x)}\right)\right)^3}\\ =\ &\frac{\lambda x m_2(x)}{(1-\rho(x)+\delta_x)^3}\\ &\left\{\frac{(1-\rho(x))^3 + 3(1-\rho(x))^2\delta_x + 3(1-\rho(x))\delta_x^2 + \delta_x^3}{(1-\rho(x))^3} - 1\right\} + \gamma\\ =\ &\frac{\lambda x m_2(x)}{(1-\rho(x)+\delta_x)^3}\left\{\frac{3\delta_x}{(1-\rho(x))} + \frac{3\delta_x^2}{(1-\rho(x))^2} + \frac{\delta_x^3}{(1-\rho(x))^3}\right\} + \gamma\\ =\ &\frac{3\lambda^2 m_2(x)^2\left\{(1-\rho(x))^2 + \delta_x(1-\rho(x)) + \frac{\delta_x^2}{3}\right\}}{(1-\rho(x)+\delta_x)^3(1-\rho(x))^3} + \gamma\end{aligned}$$

We can complete the proof by noting that $\delta_x = \lambda m_2(x)/x \leq \rho(x)$, which gives us that

$$\begin{aligned}&(1-\rho(x))^2 + \delta_x(1-\rho(x)) + \frac{\delta_x^2}{3}\\ \leq\ &1 - 2\rho(x) + \rho(x)^2 + \rho(x) - \rho(x)^2 + \frac{\rho(x)^2}{3} \leq 1\end{aligned}$$

□

LEMMA A.6. $\lim_{x\to\infty} x\varepsilon_x^* = 0$.

PROOF.

$$\begin{aligned}x\varepsilon_x^* &= \frac{x\varepsilon_x(1-\rho(x))(1-\rho(x)+\delta_x)^3}{3\lambda^2 m_2(x)^2} \leq \frac{x\varepsilon_x}{3\lambda^2 m_2(x)^2}\\ &= \frac{1}{3\lambda^2 m_2(x)^2}\left(\frac{\lambda x^2 E[X^2]}{(1-\rho)^3} - \frac{\lambda x^2 m_2(x)}{(1-\rho(x))^3} + \frac{\lambda^2 x^2\widetilde{m}_2(x)\overline{F}(x)}{(1-\rho(x))^4}\right)\\ &= \frac{1}{3\lambda^2 m_2(x)^2}\left(\frac{\lambda x^2\int_x^\infty t^2 f(t)dt}{(1-\rho)^3}\right.\\ &\quad \left. + \frac{\lambda x^2 m_2(x)}{(1-\rho)^3} - \frac{\lambda x^2 m_2(x)}{(1-\rho(x))^3} + \frac{\lambda^2 x^2\widetilde{m}_2(x)\overline{F}(x)}{(1-\rho(x))^4}\right)\end{aligned}$$

Thus, as $x \to \infty$ it is clear that the last term disappears because the service distribution is taken to have a finite third moment. Further, letting $\gamma_x = \lambda\int_x^\infty tf(t)dt$ we can that the 2nd and 3rd terms cancel.

$$\begin{aligned}&\lim_{x\to\infty} x^2\left(\frac{1}{(1-\rho)^3} - \frac{1}{(1-\rho(x))^3}\right)\\ =\ &\lim_{x\to\infty} x^2\left(\frac{(1-\rho+\gamma_x)^3 - (1-\rho)^3}{(1-\rho)^3(1-\rho(x))^3}\right)\\ =\ &\lim_{x\to\infty} x^2\gamma_x\left(\frac{3(1-\rho)^2 + 3(1-\rho)\gamma_x + \gamma_x^2}{(1-\rho)^3(1-\rho(x))^3}\right) = 0\end{aligned}$$

where the last equality follows using L'Hopital's Rule

$$\begin{aligned}\lim_{x\to\infty} x^2\gamma_x &= \lim_{x\to\infty}\frac{\int_x^\infty t^2 f(t)dt}{x^{-2}}\\ &= \lim_{x\to\infty}\frac{x^2 f(x)}{2x}\\ &= \lim_{x\to\infty} xf(x) = 0\end{aligned}$$

Finally, the limit of the first term can seen to be 0 using a similar application of L'Hopital's Rule as above. □