

Analysis of *LAS* Scheduling for Job Size Distributions with High Variance *

Idris A. Rai, Guillaume Urvoy-Keller, Ernst W. Biersack
Institut Eurecom
2229, route des Crêtes
06904 Sophia-Antipolis, France
{rai,urvoy,erbi}@eurecom.fr

ABSTRACT

Recent studies of Internet traffic have shown that flow size distributions often exhibit a high variability property in the sense that most of the flows are short and more than half of the total load is constituted by a small percentage of the largest flows. In the light of this observation, it is interesting to revisit scheduling policies that are known to favor small jobs in order to quantify the benefit for small and the penalty for large jobs. Among all scheduling policies that do not require knowledge of job size, the *least attained service* (LAS) scheduling policy is known to favor small jobs the most. We investigate the M/G/1/LAS queue for both, load $\rho < 1$ and $\rho \geq 1$. Our analysis shows that for job size distributions with a high variability property, LAS favors short jobs with a negligible penalty to the few largest jobs, and that LAS achieves a mean response time over all jobs that is close to the mean response time achieved by SRPT.

Finally, we implement LAS in the ns-2 network simulator to study its performance benefits for TCP flows. When LAS is used to schedule packets over the bottleneck link, more than 99% of the shortest flows experience smaller mean response times under LAS than under FIFO and only the largest jobs observe a negligible increase in response time. The benefit of using LAS as compared to FIFO is most pronounced at high load.

Categories and Subject Descriptors

C.4 [Performance of Systems]: Performance attributes.

General Terms

Performance, Experimentation.

*Institut Eurécom's research is partially supported by its industrial members: Bouygues Télécom, Fondation d'entreprise Groupe Cegetel, Fondation Hasler, France Télécom, Hitachi, ST Microelectronics, Swisscom, Texas Instruments, Thales

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'03, June 10–14, 2003, San Diego, California, USA.
Copyright 2003 ACM 1-58113-664-1/03/0006 ...\$5.00.

Keywords

Size-based scheduling, least attained service, high variability property, Web objects response time.

1. INTRODUCTION

Internet traffic studies revealed that Internet traffic consists of many short flows and that a tiny fraction of the largest flows constitutes more than half of the total load [6, 21]. It is interesting to study the performance of scheduling policies in the light of these findings to see if one can improve the performance of short jobs without penalizing too much the long jobs. We say that a job size distribution exhibits the *high variability property* if less than 1% of the largest jobs account for more than half of the load. This property has also been referred to as *heavy-tail property* [7, 2], but it is not restricted to heavy-tail distributions [12]. In this paper, we analyze the impact of the high variability property of job size distributions on the performance of the LAS scheduling policy.

Response time and *slowdown* are commonly used as performance metrics to analyze scheduling policies. Response time is the overall time a job spends in the system. Slowdown is the normalized response, i.e., the ratio of the response time of a job to the size of that job. In particular, we use the **conditional mean response time** and the **conditional mean slowdown**, which (for a job of size x) are defined as $E[T(x)] \triangleq E[(T|X = x)]$ and $E[S(x)] \triangleq E[(S|X = x)]$ respectively. We also use the **mean response time** and the **mean slowdown** defined as $E[T] \triangleq \int_0^\infty E[T(x)]f(x)dx$ and $E[S] \triangleq \int_0^\infty E[S(x)]f(x)dx$ respectively. Slowdown is a useful metric to analyze fairness [2]. Processor Sharing (PS) is known to be a fair policy since it offers the same slowdown to all jobs.

The *shortest remaining processing time* (SRPT) scheduling policy is proven [20, 25] to be the optimal policy for minimizing the mean response time. It has been known for a long time that for negative exponentially distributed job sizes, SRPT severely penalizes large jobs. However, Bansal and Harchol-Balter [2] have recently shown that the performance of SRPT and the penalty experienced by the largest jobs very much depend on the characteristics of the job size distribution. In particular, for some load values and for job size distributions with the high variability property, SRPT does not increase at all the conditional mean response time of long jobs as compared to PS, i.e., all jobs have a lower conditional mean slowdown under SRPT than under PS. Nevertheless, SRPT has the drawback that it needs to know the size of the jobs. While the job size is known, for instance, in a Web server with static Web pages [14], it is generally not known in environments such as in Internet routers or Web servers with dynamic pages.

The *least attained service* (LAS) scheduling policy is a multi-level scheduling policy that favors small jobs without requiring knowledge about the job sizes. LAS is a preemptive scheduling policy that gives service to the job in the system that has received the least service. In the event of ties, the set of jobs having received the least service shares the processor in a processor-sharing mode. A newly arriving job always preempts the job (or jobs) currently in service and retains the processor until it departs, or until the next arrival appears, or until it has obtained an amount of service equal to that received by the job(s) preempted on arrival, whichever occurs first. An implementation of LAS needs to know the amount of service it has delivered to each job, which can be easily obtained from the server. LAS is also known in the literature under the names of *foreground-background* (FB) [17] or *shortest elapsed time* (SET) first scheduling [4].

The expression for the conditional mean response time $E[T(x)]$ for an M/G/1 queue served using LAS has been originally derived in [24], and were re-derived in [28, 4, 18] as well. However, the expression for $E[T(x)]$ is complex and difficult to evaluate numerically, and therefore very little has been done to evaluate the performance of LAS with respect to the variability of the job size distribution. The variability of a job size distribution can be expressed in terms of its **coefficient of variation** (C), which is defined as *the ratio of the standard deviation to the mean of the distribution*. For a given mean, the higher the variance, the higher the C value and so the higher the variability of the distribution. Coffman ([4], pp. 188-187), gives an intuitive result that compares the mean response times for LAS and PS: The mean response time of LAS is lower than the mean response time of PS for job size distributions with a C greater than 1, and it is higher for job size distributions with a C less than 1. Kleinrock ([17], pp. 196-197) compares the mean waiting times of LAS and FIFO scheduling for an M/M/1 queue and observes that LAS offers lower waiting times to short jobs while large jobs see significantly higher waiting times. The analysis of *M/G/1/LAS* queue at steady state was done by Schassberger [22, 23]. However, Schassberger did not evaluate the impact of the variance of the service time distribution. In a recent work, Balter et al. [15] show that for an M/G/1 queue, the conditional mean slowdown of all work-conserving scheduling policies, which includes LAS, asymptotically converges to the conditional mean slowdown of PS, i.e. $\lim_{x \rightarrow \infty} E[S(x)]_{LAS} = 1/(1 - \rho)$.

This paper investigates the performance of LAS considering the variability of job size distributions. We first compare LAS to PS to analyze its unfairness. We derive an upper bound on the unfairness of LAS for a general job size distribution (Theorem 1). In particular, for job size distributions that exhibit the high variability property, we observe that more than 99% of the jobs see their conditional mean slowdown significantly reduced under LAS and less than 1% of the largest jobs experience a negligible increase of their conditional mean slowdown under LAS. We also derive an upper bound that compares the mean response times of LAS and nonpreemptive policies (NPP) (Lemma 2).

While both, LAS and SRPT, favor short jobs, it is not known how the performance of LAS compares to the optimal policy SRPT. We provide mathematical expressions that compare these two policies and show that for job size distributions with the high variability property, both policies offer very similar conditional mean response times to all jobs (Theorem 4).

We also consider the overload case, i.e. $\rho \geq 1$, and prove the existence of a job size $x_{LAS}(\lambda)$ such that all jobs strictly smaller in size than $x_{LAS}(\lambda)$ will have a finite response time under LAS (Theorem 5). We then present a closed form expression for the conditional mean response time of LAS under overload (Theorem

6). We evaluate this expression by comparing the performance of LAS with PS and SRPT for $\rho \geq 1$. PS is known to exhibit an infinite mean response time under overload as opposed to LAS, which offers finite mean response time for job sizes less than $x_{LAS}(\lambda)$. The results are intriguing; for the job size distribution with the high variability property considered in this paper, at overload of $\rho = 2$, jobs of size up to the 99th percentile have a finite conditional mean response time under LAS.

The analysis conducted for M/G/1/LAS uses the notion of a job, which is defined as the entity that requests service from system *at once*. In packet networks we are interested in the performance of flows. However, the packets of a flow are spaced in time and are statistically multiplexed with packets from other flows. A flow therefore does not arrive at once in the system. To evaluate the performance of LAS in a packet network, we implement LAS in the ns-2 network simulator [16] and analyze it for Web-traffic with Pareto-distributed flow sizes. The simulation results for different load values show that more than 99% of the short flows generated during the simulation benefit from LAS as compared to FIFO and a negligible percentage of the largest flows experience a higher response time under LAS than under FIFO. Also, the difference in the mean response time between LAS and FIFO increases with increasing load.

The paper is organized as follows: In the next section we discuss the mathematical background necessary for the analysis carried out in the rest of the paper. We analytically compare LAS to PS, to nonpreemptive (NPP) policies, and to SRPT in Section 3. In Section 4, we investigate the performance of LAS under overload. In Section 5 we study the performance of LAS in a bottleneck router and compare the response time of flows under LAS and FIFO. We conclude the paper in Section 6.

2. MATHEMATICAL BACKGROUND

In this section we first present mathematical expressions of the conditional mean response time and the conditional mean slowdown for different scheduling policies. Then we discuss the reasons for the choice of the particular job size distributions used in this paper.

2.1 Expressions for the Performance Metrics

Let the average job arrival rate be λ . Assume a job size distribution X with a probability mass function $f(x)$. The abbreviation c.f.m.f.v is used to denote continuous, finite mean, and finite variance. Given the cumulative distribution function as $F(x) \triangleq \int_0^x f(t)dt$, we denote the survivor function of X as $F^c(x) \triangleq 1 - F(x)$. We define $m_n(x)$ as $m_n(x) \triangleq \int_0^x t^n f(t)dt$. Thus $m_1 \triangleq m_1(\infty)$ is the mean and $m_2 \triangleq m_2(\infty)$ is the second moment of the job size distribution. The load of jobs with size less than or equal to x is given as $\rho(x) \triangleq \lambda \int_0^x tf(t)dt$, and $\rho \triangleq \rho(\infty)$ is the total load in the system.

In most cases, this paper assumes an M/G/1 queue, where G is a c.f.m.f.v distribution. The expression of $E[T(x)]$ for M/G/1/SRPT [26] can be decomposed into the conditional mean waiting time $E[W(x)]$ (the time between the instant when a job arrives at the system until it receives the service for the first time) and the conditional mean residence time $E[R(x)]$ (the time a job takes to complete service once the server has started serving it). Hence,

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

$$E[W(x)]_{SRPT} = \frac{\lambda(m_2(x) + x^2 F^c(x))}{2(1 - \rho(x))^2} \quad (1)$$

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

Similarly, we decompose the expression of the $E[T(x)]$ for $M/G/1/LAS$ into two terms. We denote the first term as $E[\tilde{W}(x)]_{LAS}$ and the second term as $E[\tilde{R}(x)]_{LAS}$. Note that these terms do not denote the conditional mean waiting and residence times for LAS, rather they are introduced to ease the analysis in Section 3.3 where we compare LAS to SRPT. The formulas for $E[T(x)]_{LAS}$ are given in [24],

$$E[T(x)]_{LAS} = E[\tilde{W}(x)]_{LAS} + E[\tilde{R}(x)]_{LAS} \quad (3)$$

$$E[\tilde{W}(x)]_{LAS} = \frac{\lambda(m_2(x) + x^2 F^c(x))}{2(1 - \rho(x) - \lambda x F^c(x))^2} \quad (3)$$

$$E[\tilde{R}(x)]_{LAS} = \frac{x}{1 - \rho(x) - \lambda x F^c(x)} \quad (4)$$

Finally, the formulas of $E[T(x)]$ for the $M/G/1/PS$ and $M/G/1/NPP$ [5] (where NPP stands for any non-preemptive policy) are:

$$E[T(x)]_{PS} = \frac{x}{1 - \rho} \quad (5)$$

$$E[T(x)]_{NPP} = \frac{\lambda m_2}{2(1 - \rho)} + x \quad (6)$$

The definition of $E[S(x)]$ reveals that for two scheduling policies A and B, the ratio $\frac{E[T(x)]_A}{E[T(x)]_B} = \frac{E[S(x)]_A/x}{E[S(x)]_B/x} = \frac{E[S(x)]_A}{E[S(x)]_B}$.

2.2 The Choice of the Job Size Distribution

We will analyze LAS by using a general c.f.m.f.v job size distribution and some specific job size distributions. The choice of specific job size distributions is motivated by the characteristics of the Internet traffic where *most of the flows are short and more than half of the load is due to a tiny fraction of the largest flows*. Different distributions have been shown to model the empirical Internet traffic given their coefficient of variation is larger than 1 [1, 6, 10]. These distributions include Pareto, bounded Pareto, lognormal distributions, hyper-exponential, Weibull, inverse Gaussian, and hybrid of lognormal and Pareto distributions. The bounded Pareto (BP) distribution is commonly used in analysis because it can take a high variance and thus it can exhibit the high variability property as observed in the Internet traffic and also because the maximum job size can be set to mimic the largest Internet flow size [29, 7, 2]. In this paper, we also use the bounded Pareto job size distribution for the same reasons.

We denote the bounded Pareto distribution by $BP(k, p, \alpha)$, where k and p are the minimum and the maximum job sizes and α is the exponent of the power law. We will also consider exponentially distributed job sizes to see the performance difference when LAS is analyzed under $M/M/1$ queue model. Let X be a job size distribution, the probability density functions of the bounded Pareto ($f(x)_{BP}$) and the exponential $f(x)_{Exp}$ job sizes are given as:

$$f(x)_{BP} = \frac{\alpha k^\alpha}{1 - (k/p)^\alpha} x^{-\alpha-1}, \quad k \leq x \leq p, \quad 0 \leq \alpha \leq 2 \quad (7)$$

$$f(x)_{Exp} = \mu e^{-\mu x}, \quad x \geq 0, \quad \mu \geq 0$$

We also denote the exponential distribution with mean of $1/\mu$ by $Exp(1/\mu)$. The cumulative distribution function $F(x)$ and the

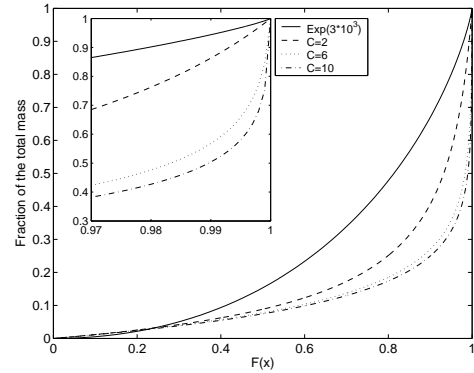
n -th moment m_n of the $BP(k, p, \alpha)$ distribution are:

$$F(x) = \frac{1}{1 - (k/p)^\alpha} [1 - (k/x)^\alpha], \quad k \leq x \leq p, \quad 0 \leq \alpha \leq 2$$

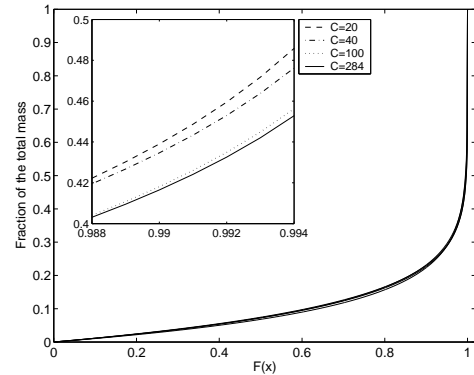
$$m_n = \frac{\alpha}{(n - \alpha)(p^\alpha - k^\alpha)} (p^n k^\alpha - k^n p^\alpha)$$

Hence, the expression for coefficient of variation is $C \triangleq \frac{\sqrt{m_2 - m_1^2}}{m_1}$. To obtain different values of C for $BP(k, p, \alpha)$, one needs to change one or more of the parameters of the BP distribution, i.e., k , α , and p . The C value of exponential distribution is always 1.

The variability of a distribution can be determined by using the *mass-weighted function* ($M_w(x)$) [8], which (for a job of size x) is defined as the fraction of the total load constituted by jobs of size less than or equal to x or $M_w(x) \triangleq \frac{\rho(x)}{\rho}$. We plot $M_w(x)$ as a function of $F(x)$ to see the fraction of total mass constituted by jobs of size less than or equal to x . The mass-weighted function will help us to see if a job size distribution exhibits the high variability property.



(a) $Exp(3 * 10^3)$ and BP with $C = 2, 6, 10$



(b) BP with $C = 20, 40, 100, 284$

Figure 1: Mass-weighted functions for $BP(k, p, \alpha)$ and $Exp(3 * 10^3)$ job size distributions.

Figure 1 shows the mass-weighted functions for the BP job size distribution with different C values and the exponential job size distribution, each with the same mean value of $3 * 10^3$. Figure 1(a) shows that for the exponential distribution, the largest 1% of the jobs constitute only about 5% of the total load, and that for BP the variability increases with increasing C value. We observe that the load constituted by the largest 1% of the jobs increases with increasing C from 15% of the total load for $C = 2$ to close to 50% for $C = 10$. On the other hand, Figure 1(b) shows that all

BP distributions with $C \geq 20$ exhibit the high variability property since 50% of the total load is constituted by less than 1% of the largest jobs.

In this paper, we shall use the BP job size distribution $BP(332, 10^{10}, 1.1)$ with $C = 284$ as a typical example of the BP distribution that exhibits the high variability property. This distribution was also used in [2] to analyze the unfairness of SRPT. BP distributions with similar parameters are also used in [29, 7]. In addition to the BP with $C = 284$, we also use other BP distributions with $C < 284$ and the exponential distribution $Exp(3 * 10^3)$ to see the impact of degree of variability on the performance of LAS.

3. ANALYTICAL RESULTS

Previously, no analysis of M/G/1/LAS has been carried out for different values of C. The analysis for M/G/1/LAS is tedious as it involves nested integrals of complicated functions. Here, we provide elements to compare M/G/1/LAS to M/G/1/PS, and M/G/1/SRPT systems.

3.1 Comparison for a general job size distribution

Processor sharing is a well known fair scheduling policy that at any load $\rho < 1$ gives the same conditional mean slowdown to all jobs, i.e., $E[S(x)]_{PS} = \frac{1}{(1-\rho)} \forall x$. In this section, we assume a general c.f.m.f.v job size distribution to investigate the unfairness of the LAS policy by comparing its slowdown to the slowdown of PS. Nonpreemptive (NPP) policies include FIFO scheduling, which is commonly implemented in today's routers. We also compare LAS with NPP policies to investigate the performance benefits offered under LAS in terms of mean response time with varying C and ρ values.

3.1.1 Comparison between LAS and PS

We define the function ϕ as $\phi(x) \triangleq \lambda \int_0^x t f(t) dt + \lambda x F^c(x)$. If we integrate $\int_0^x t f(t) dt$ by parts in the definition of ϕ we obtain $\phi(x) = \lambda \int_0^x F^c(t) dt$. The following result holds for the conditional mean response time of a job of size x :

Theorem 1. For all c.f.m.f.v job size distributions and load $\rho < 1$,

$$E[T(x)]_{LAS} \leq (1-\rho) \frac{2-\phi(x)}{2(1-\phi(x))^2} E[T(x)]_{PS} \quad (8)$$

$$E[S(x)]_{LAS} \leq (1-\rho) \frac{2-\phi(x)}{2(1-\phi(x))^2} E[S(x)]_{PS} \quad (9)$$

PROOF.

$$\begin{aligned} E[T(x)]_{LAS} &= \frac{\lambda \int_0^x t^2 f(t) dt + \lambda x^2 F^c(x)}{2(1-\phi(x))^2} + \frac{x}{1-\phi(x)} \\ &\leq \frac{\lambda x \int_0^x t f(t) dt + \lambda x^2 F^c(x)}{2(1-\phi(x))^2} + \frac{x}{1-\phi(x)} \\ &\quad \text{since } \int_0^x t^2 f(t) dt \leq x \int_0^x t f(t) dt \\ &= E[T(x)]_{PS} \frac{1-\rho}{1-\phi(x)} \left(\frac{\lambda \int_0^x t f(t) dt + \lambda x F^c(x)}{2(1-\phi(x))} + 1 \right) \\ &\quad \text{since } E[T(x)]_{PS} = \frac{x}{1-\rho} \\ &= E[T(x)]_{PS} \frac{1-\rho}{1-\phi(x)} \left(\frac{\phi(x)}{2(1-\phi(x))} + 1 \right) \end{aligned}$$

By definition of $\phi(x)$

$$= (1-\rho) \frac{2-\phi(x)}{2(1-\phi(x))^2} E[T(x)]_{PS} \quad (10)$$

Equation (9) follows directly by dividing both sides of Equation (10) by x . \square

We use Theorem 1 to elaborate on the comparison between the mean slowdown of LAS and PS. We first introduce a lemma that we need in the proof of Theorem 2.

Lemma 1. For any load $\rho < 1$,

$$\frac{2-\phi(x)}{2(1-\phi(x))^2} \leq \frac{2-\rho}{2(1-\rho)^2} \quad (11)$$

PROOF. Function $\phi(x)$ is an increasing function for $x \in [0, \infty)$ since $\frac{d\phi(x)}{dx} = \lambda F^c(x) \geq 0$. Also, function $\frac{2-u}{2(1-u)^2}$ is an increasing function for $u \in [0, 1]$. Hence, $\frac{2-\phi(x)}{2(1-\phi(x))^2}$ is an increasing function of x and upper bounded by $\frac{2-\rho}{2(1-\rho)^2}$ since $\phi(x) \in [0, \rho] \subset [0, 1]$. \square

Theorem 2. For all c.f.m.f.v job size distributions and load $\rho < 1$,

$$E[S]_{LAS} \leq \frac{2-\rho}{2(1-\rho)} E[S]_{PS} \quad (12)$$

PROOF.

$$\begin{aligned} E[S]_{LAS} &= \int_0^{+\infty} \frac{E[T(x)]_{LAS}}{x} f(x) dx \\ &\leq \int_0^{+\infty} \frac{2-\phi(x)}{2(1-\phi(x))^2} f(x) dx \quad \text{Theorem 1} \quad (13) \\ &\leq \frac{2-\rho}{2(1-\rho)^2} \int_0^{+\infty} f(x) dx \quad \text{Lemma 1} \\ &= \frac{2-\rho}{2(1-\rho)^2} \\ &= \frac{2-\rho}{2(1-\rho)} E[S]_{PS} \quad \text{Since } E[S]_{PS} = \frac{1}{1-\rho} \quad (14) \end{aligned}$$

\square

Corollary 1. For all c.f.m.f.v job size distributions and load $\rho < 1$,

$$E[T]_{LAS} \leq \frac{2-\rho}{2(1-\rho)} E[T]_{PS} \quad (15)$$

PROOF. The proof is similar to the proof of Theorem 2. \square

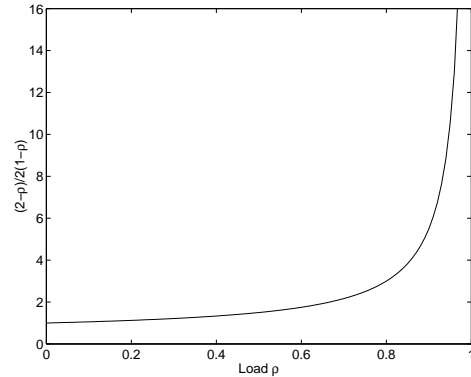


Figure 2: Upper bound on the mean slowdown ratio $\frac{E[S]_{LAS}}{E[S]_{PS}}$.

We illustrate the result of Theorem 2 in Figure 2. We can see that the ratio between the mean slowdown of LAS and PS is bounded by a value that is less than 2 for $\rho \leq 0.66$ and less than 6 for $\rho \leq 0.9$. This result clearly supports the fact that using LAS instead of PS results in a better slowdown for small jobs without affecting too much larger jobs since the average slowdown over small and large jobs under PS and LAS remains fairly close for most system loads. The result also indicates that for low and medium load the mean slowdown of LAS remains close to the mean slowdown for PS. Note that the real ratio between the average slowdown of LAS and PS is below the value of $\frac{2-\rho}{2(1-\rho)}$ used in Figure 2 due to the rather crude majorization of $\phi(x)$ applied in Lemma 1.

3.1.2 Comparison between LAS and Nonpreemptive policies (NPP)

We now compare the performance of LAS to NPP policies in terms of mean response time for job size distributions with varying C and for different values of load $\rho < 1$. Examples of nonpreemptive policies include FIFO, LIFO, and RANDOM. We derive an upper bound of the mean response time that is a function of C and ρ values. The bound shows that the mean response time offered by LAS improves in increasing C values. From Equations (5) and (6), we get the expressions of mean response times for PS and NPP policies as:

$$E[T]_{PS} = \frac{m_1}{(1-\rho)} \quad (16)$$

$$E[T]_{NPP} = \frac{\lambda m_2}{2(1-\rho)} + m_1 \quad (17)$$

Equation (17) can also be expressed as:

$$\begin{aligned} E[T]_{NPP} &= \frac{m_1}{(1-\rho)} + \frac{\lambda m_2}{2(1-\rho)} + m_1 - \frac{m_1}{(1-\rho)} \\ &= \frac{m_1}{(1-\rho)} + \frac{\lambda m_2}{2(1-\rho)} - \frac{\rho m_1}{(1-\rho)} \\ &= \frac{m_1}{(1-\rho)} + \frac{\lambda m_2}{2(1-\rho)} - \frac{\lambda m_1^2}{(1-\rho)} \\ &\quad \text{from } \rho = \lambda m_1 \\ &= \frac{m_1}{(1-\rho)} + \frac{\rho[C^2 - 1]}{2(1-\rho)} m_1, \\ &\quad \text{from } m_1^2 C^2 = m_2 - m_1^2 \text{ and } \lambda = \rho/m_1 \\ &= E[T]_{PS} + \frac{\rho[C^2 - 1]}{2(1-\rho)} m_1, \end{aligned} \quad (18)$$

Lemma 2. For a c.f.m.f.v. job size distribution with mean m_1 and coefficient of variation C , the mean response time under LAS at load ρ ($E[T]_{LAS}$) is upper bounded as:

$$E[T]_{LAS} \leq \frac{(2-\rho)}{2(1-\rho)} E[T]_{NPP} - \frac{\rho(2-\rho)[C^2 - 1]}{4(1-\rho)^2} m_1 \quad (19)$$

PROOF. The proof follows directly when substituting $E[T]_{PS}$ obtained from Equation (18) in Corollary 1. \square

The bound on the mean response time of LAS (Equation (19)) is a function of C and load ρ , since $E[T]_{NPP}$ itself is a function of C and load ρ (see Equation (18)) for a given value of m_1 . This bound is interesting since it enables us to compare the performance of LAS relative to that of any nonpreemptive policy for job size distributions with a large range of C values. Figure 3 shows the upper bound on the ratio of the mean response time of LAS to the mean response time of a nonpreemptive policy as a function of $C \geq 1$ and load $\rho < 1$. Observe that LAS has a higher mean response

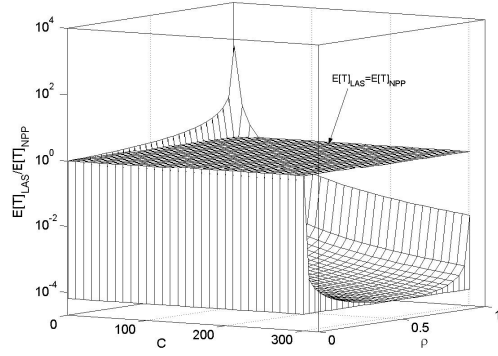


Figure 3: Upper bound on the mean response time ratio $\frac{E[T]_{LAS}}{E[T]_{NPP}}$ as a function of load ρ and coefficient of variation C .

time than nonpreemptive policies only for distributions with C values close to 1. In this case, we also observe that the mean response time ratio increases to large values with increasing load. On the other hand, the mean response time of LAS is lower than that of nonpreemptive policies for distributions with higher C values at all load values $\rho < 1$. Generally, for a given load ρ , the ratio $\frac{E[T]_{LAS}}{E[T]_{NPP}}$ decreases with increasing C .

In summary, we see that LAS achieves lower mean response time than any NPP policy, such as the FIFO policy, for job size distribution with a high variance. This is to be expected because in contrast to LAS, the service of a job under FIFO is not interrupted until the job leaves the system and so large jobs are favored over small jobs.

3.2 Distribution dependent comparison

In this section, we compare LAS to PS for job size distributions with varying C values. A comparison of LAS with PS helps to analyze the degree of unfairness (penalty) seen by the largest jobs under LAS. We start with general results for exponentially distributed job sizes. We show that for the case of the exponential distribution, the average slowdown of jobs under LAS is always better than the average slowdown of the jobs under PS.

Theorem 3. For an exponential job size distribution and load $\rho < 1$,

$$E[S]_{LAS} \leq E[S]_{PS} \quad (20)$$

PROOF. Following the same reasoning as in Theorem 2, one obtains:

$$E[S]_{LAS} = \int_0^{+\infty} \frac{E[T(x)]_{LAS}}{x} f(x) dx$$

Using Equation (13) in Equation (21), we get,

$$E[S]_{LAS} \leq \int_0^{+\infty} \frac{x(2-\phi(x))}{2(1-\phi(x))^2} \frac{f(x)}{x} dx \quad (21)$$

For exponential job sizes, we have $\phi(x) = \int_0^x \lambda F^c(t) dt = \rho(1 - e^{-\mu x}) = \rho F(x)$ and $f(x) = \mu e^{-\mu x}$. Using these facts and plugging $h(\phi(x)) \triangleq \frac{2-\phi(x)}{2(1-\phi(x))^2}$ in Equation (21) we obtain:

$$E[S]_{LAS} \leq \frac{1}{\rho} \int_0^{+\infty} h(\rho F(t)) \rho f(t) dt$$

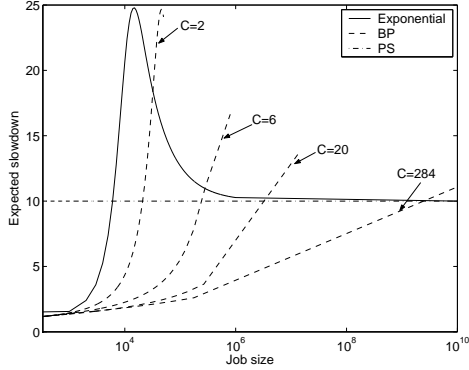
Replacing $\rho F(t)$ by u in the above equation we get:

$$\begin{aligned}
E[S]_{LAS} &\leq \frac{1}{\rho} \int_0^{+\rho} h(u) du \\
&= \frac{1}{2\rho} \left(-\ln(1-\rho) + \frac{\rho}{1-\rho} \right) \\
&= \frac{1}{2\rho} \left(-(1-\rho)\ln(1-\rho) + \rho \right) E[S]_{PS} \quad (22) \\
&\text{since } E[S]_{PS} = \frac{1}{1-\rho}
\end{aligned}$$

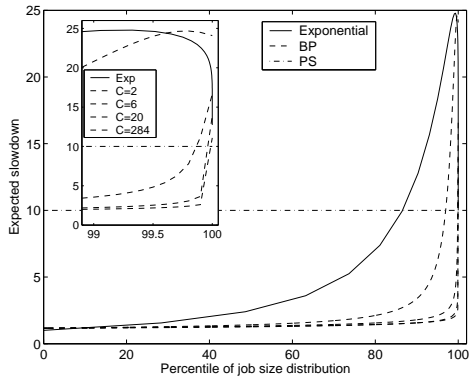
A straightforward study of the derivative of ψ defined as $\psi(\rho) \triangleq \frac{(\rho - (1-\rho)\ln(1-\rho))}{2\rho}$ indicates that ψ is a decreasing function and $\psi(\rho) \sim_{\rho \rightarrow 0} \frac{2\rho - \rho^2}{2\rho} \rightarrow 1$. Thus, $\forall \rho < 1, \psi(\rho) < 1$. \square

This result shows that even for the case of exponentially distributed job sizes, where the high variability property does not hold, the mean slowdown is at most $\frac{1}{(1-\rho)}$. We could not derive bound similar to the one of theorem 3 for the BP distribution. However, we expect LAS to perform better in terms of the mean slowdown for job size distributions with $C > 1$ than for exponential job size distribution.

Next, we investigate the unfairness of LAS for job size distributions with varying C values. We use the exponential distribution (with $C = 1$) and the BP distributions with C values 2, 6, 20, and 284 all with the same mean value of $3 * 10^3$. To obtain BP job size distributions with different C values and the same mean we varied the values of α and p .



(a) Job size



(b) Percentile of job size distribution

Figure 4: Expected slowdown under LAS and PS for different job size distributions and different C values.

Figure 4(a) and 4(b) show the conditional mean slowdown as a function of job size and percentile respectively. We see from both figures that the percentage of the largest jobs that experience a penalty under LAS (i.e., have higher mean slowdown under LAS than under PS) and the degree of penalty decreases in increasing C value. For the BP distribution with $C \geq 6$, less than 0.5% of the largest jobs suffer a small penalty and the performance difference between $C = 20$ and $C = 284$ is minor. We observe that as the C value increases the penalty experienced by the largest jobs and the percentage of these largest jobs decrease. In the remainder of this paper we will only consider $Exp(3 * 10^3)$ and the $BP(332, 10^{10}, 1.1)$ distribution when we numerically analyze LAS.

3.3 Quantitative comparison between LAS and SRPT

It is stated in [2] that for any job size distribution and at any load $\rho < 1$ every job x has higher mean response time ($E[T(x)]$) under LAS than under SRPT. However, it was not elaborated how much higher the mean slowdown of a job under LAS can be as compared to SRPT. We now compare LAS and SRPT quantitatively and show that at load $\rho < 1$, the conditional mean response time of a job under LAS is quite close to the mean response time under SRPT.

Theorem 4. Let $\phi(x) \triangleq \rho(x) + x\lambda F^c(x)$, then for all c.f.m.f.v job size distributions and at load $\rho < 1$,

$$E[T(x)]_{SRPT} \leq E[T(x)]_{LAS} \leq \left(\frac{1-\rho(x)}{1-\phi(x)} \right)^2 E[T(x)]_{SRPT} \quad (23)$$

PROOF. Since $\rho(x) \leq \phi(x) \leq \rho$, we obtain directly from Equations (1) and (2) and Equations (3) and (4): $E[\tilde{R}(x)]_{LAS} \geq E[R(x)]_{SRPT}$ and $E[\tilde{W}(x)]_{LAS} \geq E[W(x)]_{SRPT}$. Hence the left-hand side inequality. We begin the proof of the right-hand side inequality by studying $E[\tilde{R}(x)]_{LAS} - E[R(x)]_{SRPT}$:

$$\begin{aligned}
E[\tilde{R}(x)]_{LAS} - E[R(x)]_{SRPT} &= \frac{x}{1-\phi(x)} - \int_0^x \frac{dt}{1-\rho(t)} \\
&= \int_0^x \frac{dt}{1-\phi(x)} - \int_0^x \frac{dt}{1-\rho(t)} \\
&= \int_0^x \frac{\phi(x) - \rho(t)}{1-\phi(x)} \frac{1}{1-\rho(t)} dt
\end{aligned}$$

Applying the Chebyshev integral inequality [19] to $f(t) = \frac{\phi(x) - \rho(t)}{1-\phi(x)}$, which is a decreasing function of t and $g(t) = \frac{1}{1-\rho(t)}$, which is an increasing function of t , we get $E[\tilde{R}(x)]_{LAS} - E[R(x)]_{SRPT}$

$$\leq \frac{1}{x} \int_0^x \frac{\phi(x) - \rho(t)}{1-\phi(x)} dt \underbrace{\int_0^x \frac{dt}{1-\rho(t)}}_{E[R(x)]_{SRPT}} \quad (24)$$

Besides:

$$\begin{aligned}
\frac{1}{x} \int_0^x \frac{\phi(x) - \rho(t)}{1 - \phi(x)} dt &= \left([t(\phi(x) - \rho(t))]_0^x + \int_0^x \lambda t^2 f(t) dt \right) \\
&= \frac{1}{x(1 - \phi(x))} (x \lambda x F^c(x) \\
&\quad + \lambda m_2(x)) \\
&= \frac{\lambda}{x(1 - \phi(x))} (x^2 F^c(x) + m_2(x)) \\
&\leq \frac{\lambda}{x(1 - \phi(x))} (x^2 F^c(x) \\
&\quad + x \int_0^x t f(t) dt) \\
&\text{since } \int_0^x t^2 f(t) dt \leq x \int_0^x t f(t) dt \\
&= \frac{x \phi(x)}{x(1 - \phi(x))} \\
&= \frac{\phi(x)}{1 - \phi(x)} \tag{25}
\end{aligned}$$

Using Equation (25) in Equation (24), one obtains:

$$\begin{aligned}
E[\tilde{R}(x)]_{LAS} &\leq \left(1 + \frac{\phi(x)}{(1 - \phi(x))} \right) E[R(x)]_{SRPT} \\
&= \frac{1}{1 - \phi(x)} E[R(x)]_{SRPT} \tag{26}
\end{aligned}$$

Consider now $E[\tilde{W}(x)]_{LAS}$ and $E[W(x)]_{SRPT}$, we have:

$$E[\tilde{W}(x)]_{LAS} = \frac{(1 - \rho(x))^2}{(1 - \phi(x))^2} E[W(x)]_{SRPT} \tag{27}$$

Adding Equations (26) and (27), we obtain:

$$\begin{aligned}
E[T(x)]_{LAS} &\leq \max \left(\frac{(1 - \rho(x))^2}{(1 - \phi(x))^2}, \frac{1}{1 - \phi(x)} \right) E[T(x)]_{SRPT} \\
&= \frac{1}{1 - \phi(x)} \max \left(\frac{(1 - \rho(x))^2}{1 - \phi(x)}, 1 \right) E[T(x)]_{SRPT}
\end{aligned}$$

Since $(1 - \rho(x))^2 = 1 - 2\rho(x) + \rho(x)^2 = 1 - \rho(x) + \underbrace{\rho(x)^2 - \rho(x)}_{=\rho(x)(1 - \rho(x)) \geq 0}$

and $1 - \phi(x) = 1 - \rho(x) - \underbrace{x F^c(x)}_{\geq 0}$, then $\frac{(1 - \rho(x))^2}{1 - \phi(x)} \geq 1$ and

the right-hand side of the Theorem follows directly from Equation (28). \square

Figure 5 illustrates the result of Theorem 4 for the BP with $C = 284$ and the exponential distribution $Exp(3 * 10^3)$ as a function of percentiles of job sizes to obtain results for a job size distribution that exhibits the high variability property and for a job size distribution that does not exhibit the high variability property. We observe from Figure 5 that the ratio between $E[T(x)]_{LAS}$ and $E[T(x)]_{SRPT}$ highly depends on the variability of the job size distribution. We see that for the bounded Pareto distribution, the ratio $\frac{E[T(x)]_{LAS}}{E[T(x)]_{SRPT}}$ is about 1.25 for all jobs even under very high load, which shows that the conditional mean response times $E[T(x)]$ of jobs under LAS are quite close to ones under SRPT. For BP distributions with lower values of $284 > C \geq 20$, we observed that the value of $\frac{E[T(x)]_{LAS}}{E[T(x)]_{SRPT}}$ is quite close to the value for BP with $C = 284$ for all load values.

Jobs whose size is exponentially distributed experience under LAS a slightly higher higher conditional mean response time

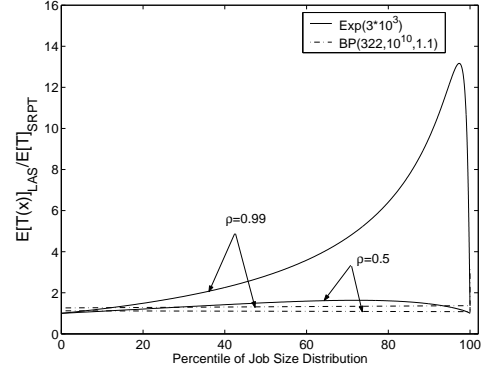


Figure 5: $\frac{E[T(x)]_{LAS}}{E[T(x)]_{SRPT}}$ as a function of percentile of job size distribution.

$E[T(x)]_{LAS}$ than under SRPT for medium load $\rho = 0.5$. However, for very high load $\rho = 0.99$, the difference is much more pronounced and the ratio $\frac{E[T(x)]_{LAS}}{E[T(x)]_{SRPT}}$ approaches 13 for some large jobs. Note also that as $x \rightarrow 100th$ percentile job size, the ratio $\frac{E[T(x)]_{LAS}}{E[T(x)]_{SRPT}} \rightarrow 1$ because $\lim_{x \rightarrow \infty} \phi(x) = \rho(x)$.

We conclude that for job size distributions with the high variability property, LAS scheduling is an attractive policy to be used, for instance, in Web-servers handling requests for both, static and dynamic pages. As opposed to SRPT, LAS can handle requests for dynamic pages since it does not need to know the job size and yet offers a response time performance very similar to SRPT.

4. LAS UNDER OVERLOAD

In real systems, it may happen that jobs arrive to the system at a higher rate than the rate at which they are serviced. This situation is referred to as overload condition where $\rho \geq 1$. To the best of our knowledge, no analysis of LAS under overload has been conducted. Here, we show that at load $\rho \geq 1$ LAS is stable for all small jobs up to a certain job size and we derive the formulas of the conditional mean response time for LAS under overload. SRPT was also proven to be stable under overload in [2] for a range of short jobs. Hence in this section, we also compare LAS to SRPT under overload.

Theorem 5. Let $\theta(x) \triangleq m_1(x) + x F^c(x)$, λ be the job mean arrival rate, and $f(t)$ be a service time distribution with mean m_1 . Then, for any load $\rho = \lambda m_1 \geq 1$, every job size $x < x_{LAS}(\lambda)$ with $x_{LAS}(\lambda) \triangleq \max\{x \mid \theta(x) \leq \frac{1}{\lambda}\}$ has a finite conditional mean response time under LAS, whereas every job of size $x \geq x_{LAS}(\lambda)$ experiences an infinite response time.

PROOF. The load offered to the server when LAS system is under overload is $\rho = \lambda m_1 \geq 1$. However, the effective load $\rho_{\text{effective}}$ that corresponds to the work serviced by the server is equal to 1. Let $\theta_{\text{effective}} \triangleq \frac{\rho_{\text{effective}}}{\lambda} = \frac{1}{\lambda}$. Then, $\theta_{\text{effective}}$ is the expected service offered to the set of jobs that access the server under overload. This set depends on the policy. With LAS, every newly arriving job in the system gets an immediate access to the server. On the average a job receives a service of $\theta_{\text{effective}}$. But, since some jobs require less than $\theta_{\text{effective}}$, the jobs of size strictly smaller than $x_{LAS}(\lambda)$ may receive more service than $\theta_{\text{effective}}$, where $x_{LAS}(\lambda)$ defined as:

$$x_{LAS}(\lambda) \triangleq \max\{x \mid \theta(x) \leq \frac{1}{\lambda}\} \tag{28}$$

Hence, for LAS under overload, one obtains $\rho_{\text{effective}} = \lambda \theta(x_{LAS}) = 1$. \square

Corollary 2. For SRPT under overload, every job of size $x < x_{SRPT}(\lambda)$ with $x_{SRPT}(\lambda) \triangleq \max\{x \mid m_1(x) \leq \frac{1}{\lambda}\}$ has a finite conditional mean response time, whereas every job of size $x \geq x_{SRPT}(\lambda)$ experiences an infinite response time

PROOF. The reasoning for SRPT is different from the one for LAS. Note that the service of a job under SRPT is not affected by the arrival of jobs with larger sizes. Hence, a set of jobs with size less than or equal to x contributes a system load of $\rho(x) = \lambda \int_0^x t f(t) dt$. Thus under overload, only jobs with size strictly less than $x_{SRPT}(\lambda)$ receive service, where $x_{SRPT}(\lambda) = \max\{x \mid \rho(x) \leq \rho_{\text{effective}} = 1\}$, which is equivalent to $x_{SRPT}(\lambda) = \max\{x \mid m_1(x) \leq \frac{1}{\lambda}\}$. The jobs with size greater than or equal to $x_{SRPT}(\lambda)$ can not access the server since the jobs with size less than $x_{SRPT}(\lambda)$ always preempt them to maintain $\rho_{\text{effective}} = 1$. \square

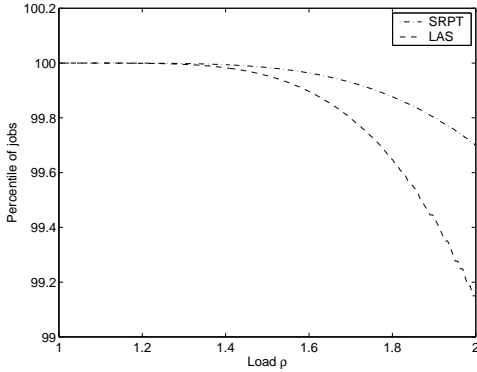


Figure 6: Percentile of the largest job sizes that can be serviced under overload for $BP(332, 10^{10}, 1.1)$ job size distribution.

Note that $x_{SRPT}(\lambda) \geq x_{LAS}(\lambda)$ since with SRPT under overload, a job accesses the server only if its service can be entirely completed, which may not be the case with LAS.

We observe in Figure 6 that for $BP(332, 10^{10}, 1.1)$, the maximum job size that SRPT can service under overload is always larger than or equal to the maximum job size that LAS can service at a given load. It is worth mentioning that for the overload values considered, the maximum job that is serviced by both policies is above the 99th percentile of the job size distribution. This result is to be expected since for the job size distribution considered, more than 99% of the jobs are small jobs and the fraction of load that these small jobs contribute to the system load is less than 50% of the total load.

We now compute the formulas for the conditional mean response time of the jobs under overload. For the SRPT policy, the formulas in case of underload and case of overload are different [2]. This may be explained by the facts that all the jobs of size $x < x_{SRPT}(\lambda)$ with $\rho(x_{SRPT}) \triangleq \int_0^{x_{SRPT}(\lambda)} t f(t) dt = 1$ have a finite response time, and all the jobs of size $x \geq x_{SRPT}(\lambda)$ receive no service at all. Therefore, under overload, the SRPT system works as if there are no jobs of size $x \geq x_{SRPT}(\lambda)$. Hence, the moment² $m_2(x) + x^2 F^c(x)$ that appears in the formulas of $E[T(x)]_{SRPT}$ is computed only for the jobs of size $x < x_{SRPT}(\lambda)$, whereas in underload, $E[T(x)]_{SRPT}$ is computed considering all job sizes.

² $m_2(x) + x^2 F^c(x)$ and $m_1(x) + x F^c(x)$ are the moments of a truncated distribution $f_x(y)$ (with $f_x(y) = f(y)$ if $y < x$, $f_x(y) = F^c(x)$ if $y = x$, and $f_x(y) = 0$ if $y > x$) that account for the contribution of all jobs of size y to the response time of the job of size x (see ([17], pp. 173).

For LAS under overload, all jobs can receive up to $x_{LAS}^-(\lambda) \triangleq \max\{x \mid \theta(x) < \frac{1}{\lambda}\}$. Thus, if the initial service requirement of a job is larger than $x_{LAS}^-(\lambda)$, it receives only $x_{LAS}^-(\lambda)$. Therefore, the moments² $m_2(x) + x^2 F^c(x)$ and $m_1(x) + x F^c(x)$ that appear in the formulas for $E[T(x)]_{LAS}$ must be computed considering all job sizes. As a consequence, the formulas for the conditional mean response time in overload and underload are identical:

Theorem 6. The conditional mean response time of a job size x for LAS under overload is:

$$E[T(x)]_{LAS} = \begin{cases} \frac{\lambda(m_2(x) + x^2(1 - F(x)))}{2(1 - \rho(x) - \lambda x(1 - F(x)))^2} + \frac{x}{1 - \rho(x) - \lambda x(1 - F(x))} & \text{if } x < x_{LAS}(\lambda) \\ +\infty & \text{if } x \geq x_{LAS}(\lambda) \end{cases}$$

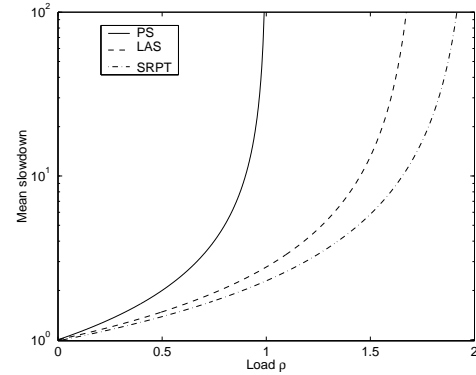


Figure 7: Mean slowdown for the 99th percentile job for $BP(332, 10^{10}, 1.1)$ as a function of load.

Figure 7 shows the mean slowdown of the 99th percentile job of the $BP(332, 10^{10}, 1.1)$ job size distribution under PS, LAS, and SRPT under overload. At load $\rho = 1.4$, the mean slowdown of the job under PS is infinity, whereas it is 10 and 4 under LAS and under SRPT respectively. It is obvious that LAS becomes unstable at lower load values than does SRPT. However, we see that LAS is quite close to the optimal policy SRPT in terms of the mean slowdown of a job even under overload. As seen in the Figure, LAS is unstable under overload for the job size considered. In general, it is well known that the mean response times of all jobs under PS $E[T(x)]_{PS}$ is undefined for load $\rho \geq 1$.

5. EVALUATION OF LAS IN A PACKET NETWORK

We analyzed the M/G/1/LAS queue using a notion of a job. In a packet switched network the entity that arrives at the router at once is a packet. However, we are not interested in the performance (response time and slowdown) of a packet but the performance of a flow. The results obtained for jobs cannot be directly applied to flows since a job is an entity that arrives to the system at once, whereas a flow consists of a sequence of packets that are spaced in time and are statistically multiplexed with packets from other flows. To investigate the performance of LAS in packet networks we implement LAS using the ns-2 simulator. We compare the mean response time of flows with varying sizes under LAS to their mean response time under FIFO scheduling with drop-tail buffer management policy, which drops newly arriving packets that arrive when the buffer is full.

LAS is implemented such that the packet that will be served (transmitted) next belongs to the flow that has received the least amount of service. If two or more flows have received an equal amount of service, they share the system resources fairly. The implementation of LAS involves maintaining a single priority queue and inserting each incoming packet at its appropriate position in that queue. The less service a flow has received so far, the closer to the head of the queue its arriving packets will be inserted. When a packet arrives and the queue is full, LAS first “inserts” the newly arriving packet at its appropriate position in the queue and then drops the packet that is at the end of the queue. Hence, LAS guarantees service priority to short flows even during congestion and will drop packets from the largest flows that currently have a packet in the queue.

The idea of giving preferential treatment to short flows has been considered by other researchers [3, 13], who propose DiffServ like models where *edge* routers mark packets as belonging to short or long flows and *core* routers utilize the marking to give preferential treatment to short flows. The first proposal is referred to as PS-*w* [13]. In PS-*w*, an edge router maintains a counter for every active flow and assigns a high priority to an incoming packet if the counter value is less than the predefined threshold otherwise a packet is assigned a low priority. In the core routers, low priority packets are serviced only if no high priority packets are backlogged. During congestion, low priority packets are more likely to be dropped than high priority ones. The simulation results in [13] show that with appropriate tuning of the drop rate of high priority and low priority packets, PS-*w* can reduce the response time of medium-sized flows that have 50-200 packets by up to 80% compared to *random early detection* (RED) [11] queue management policy without significantly penalizing the large flows. The response times of flows of sizes less than 20 packets are similar for PS-*w* and RED.

In [3], a short flow differentiating (SFD) algorithm is proposed and implemented as DiffServ policy in the ns-2 simulator. Similarly to PS-*w*, edge routers in SFD mark packets as IN or OUT based on the amount of data a flow has already sent. In the core routers, SFD applies the RED buffer management discipline to handle IN and OUT packets by maintaining a single FIFO queue with two virtual RED queues or two separate RED queues for IN and OUT packets with preferential dropping to OUT packets. For the case of physical RED queues, strict priority packet scheduler is used. The simulation results [3] show that SFD reduces the response time of about 90% of the shortest flows on average by about 30% as compared to RED and only 1% of the largest flows see higher response time under SFD.

Both studies show that it is feasible to implement LAS-like policies in the core routers and achieve scalability by identifying (marking) flows only at the edge of the network. In the next section, we present simulation results that compare LAS to FIFO. The results show that LAS offers much higher performance improvement for short flows, in terms of reducing their mean response times, than PS-*w* or SFD.

5.1 Simulation Results

We simulate LAS using the network topology shown in Figure 8 where C1-C5 are clients initiating a series of Web sessions, each retrieving some Web pages from a server randomly chosen from a pool of S1-S5 as proposed in the ns-2 Web model in [9]. Each Web page contains a certain number of objects. Each time an object is requested, a new TCP connection is established. We refer to the packets that belong to one TCP connection as a flow. We set the Web parameters as shown in Table 1. These values are based on the findings of [27] and summarized by [3]. The density function of

the Pareto distribution is obtained from Equation (7) when $p \rightarrow \infty$. Here, we use the Pareto distribution with mean of 12, $\alpha = 1.2$, and $k = 1$. The coefficient of variation of this Pareto distribution is infinite, since $\alpha < 2$. We use fixed size data packets of 1400 bytes in simulations. To obtain different load conditions we adjust the number of sessions and the session interarrival times as shown in Table 1.

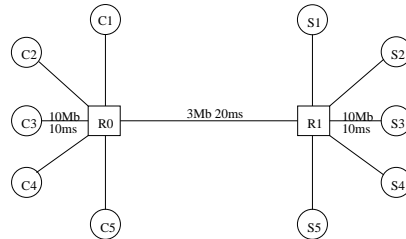


Figure 8: Simulated network topology.

We study three load conditions: load $\rho = 0.55, 0.73$, and 0.97 . We run each simulation for 36000 seconds and start collecting performance statistics after an initial period of 6000 seconds (to avoid transient effects). The results are obtained by first averaging the response times for all objects of a given size, then further averaging the mean response times of objects of sizes greater than 30 with linearly increasing bins starting with a bin of size 1.

The simulation results are shown in Figures 9 and 10. For all three load values, at least 99.5% of the shortest Web flows benefit from LAS as compared to FIFO. We see that for all loads, the mean response time of short flows is smaller under LAS than under FIFO. For heavy load $\rho = 0.97$ there is a difference of a factor of 40 for 99% of the shortest flows while 0.001% of the largest flows have higher mean response time under LAS than under FIFO by a factor of about 3. These results agree with the general numerical and analytical results observed previously. For lower load values, the difference in mean response time between LAS and FIFO for short objects decreases and the mean response time for the largest objects is similar under both policies.

The reduction in response time for short flows under LAS is due to two factors. (i) The packets of short flows are always inserted close to the head of the queue, which reduces their waiting time as compared to FIFO. (ii) Under high load, there will be congestion resulting in buffer overflow: For FIFO with drop-tail, both, short and long flows will experience loss. For LAS, however, it is only the long flows that suffer loss. These two effects explain why there is such a large difference in response time for short flows between LAS and FIFO. For LAS at high load, the response time for very long flows is higher than under FIFO since under LAS, as long flows who have received a certain amount of service will get service only when there is no packet in the queue from a shorter flow.

Load ρ	Loss Rate		$\frac{E[T(x)]_{FIFO}}{E[T(x)]_{LAS}}$ for the 99% shortest flows
	FIFO	LAS	
0.55	1.2%	0.9%	1.5
0.74	4.2%	3.66%	2.75
0.97	20.7%	4.9%	40

Table 2: Loss and response time performance for different loads.

In summary, we can conclude that LAS in a packet network improves the response time for short flows without penalizing much the long flows. Another advantage of LAS over FIFO is that it reduces the loss rate for a given load and assures that short flows will

Web model elements	Element attributes	Distribution	Parameters
Web session	Number of sessions	Constant	{3400, 3600, 4000}
	Interval between sessions (sec)	Exponential	mean: {20, 15, 10}
	Number of web pages per session	Exponential	mean: 100
Web page	Interval between pages (sec)	Exponential	mean: 10
	Number of web objects per page	Exponential	mean: 3
Web object	Interval between objects (sec)	Exponential	mean : 0.001
	Object size (packets)	Pareto	mean : 12, shape: 1.2

Table 1: Web Traffic Attributes and Corresponding Distributions.

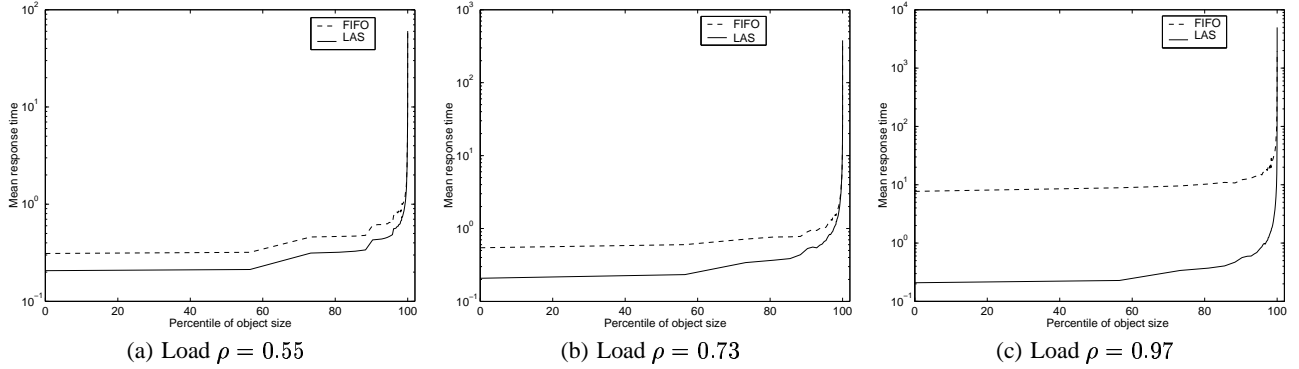


Figure 9: Mean response time as a function of percentile of object size (in packets).

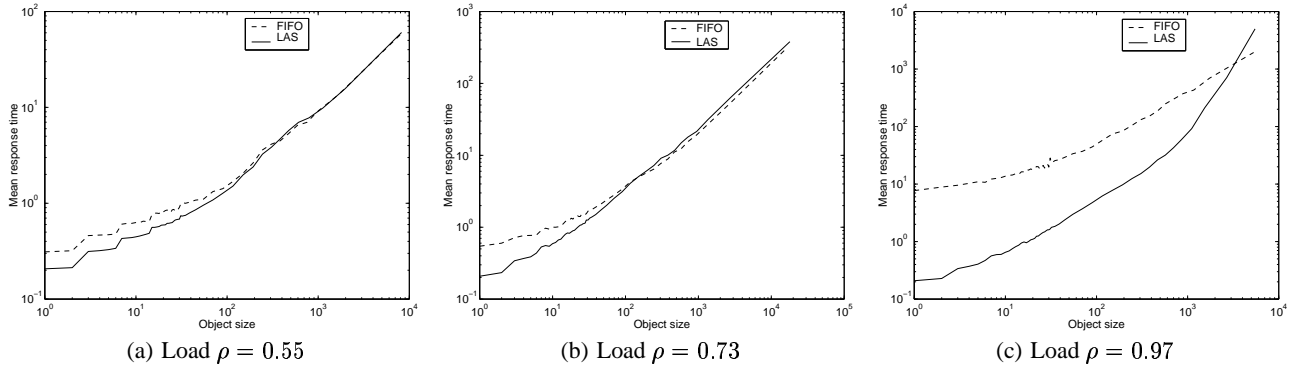


Figure 10: Mean response time as a function of object size (in packets).

see no or very little loss. The difference in loss rate between LAS and FIFO increases with increasing load as seen in Table 2. The results obtained indicate that only the *very few largest flows* suffer a penalty under LAS, while all the other flows see their response time reduced. For this reason, we expect that even under HTTP/1.1 with persistent connections, the size of most of these connections in terms of the number of packets will see their response time reduced under LAS. Table 1 shows that the mean size of a Web page is 36 packets. Hence, even if all data on the page is sent in one connection under HTTP/1.1, our simulation results indicate that on average the connection still benefits from LAS as compared with FIFO. Therefore, only those persistent connections that contain a request for a very large object will see an increase in their response time.

6. CONCLUSION

We analyzed the $M/G/1/LAS$ queuing model to evaluate the performance of the LAS for job size distributions with different degrees of variability. We showed through analysis and numerical evaluation that the variability of a job size distribution is important in determining the performance of LAS in terms of response time and slowdown. In particular, we saw that the percentage of jobs that have higher slowdowns under LAS than under PS is smaller for job size distributions with the high variability property than for job size distributions with values of C close to 1. For the case of the exponential job size distribution, we proved that the mean slowdown for LAS is always less than or equal to the mean slowdown for PS. Even for the case of general job size distribution, we showed that for moderate load values, the mean slowdown of LAS remains fairly close to the mean slowdown of PS.

The comparison of LAS to nonpreemptive policies (NPP) reveals that the benefits of LAS over NPP in terms of low mean response time improves with increasing variance of job size dis-

tribution. Similarly, we compared LAS to the optimal policy SRPT and demonstrated that $E[T(x)]_{LAS}$ is closer to the $E[T(x)]_{SRPT}$ for job size distributions with high variability property.

We also proved that LAS is stable under overload for a subset of small jobs and obtained the expression for the conditional mean response time $E[T(x)]$ for LAS under overload. The ability of LAS scheduling to schedule jobs under overload does not apply for PS and FIFO scheduling.

We used simulation to evaluate the benefit of LAS over FIFO for scheduling the transmission of packets over a bottleneck link and saw that most flows experienced a reduction of their response time that can be very significant under high load. These results illustrate the potential benefit of flow size aware scheduling policies such as LAS in routers. However, more work needs to be done to come up with a complete architecture for flow size aware scheduling in the Internet that specifies the functionalities required in the edge routers and the core routers.

Acknowledgement

We would like to thank Ken Sevcik and the anonymous Sigmetrics reviewers of this paper for their constructive comments.

7. REFERENCES

- [1] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira, "Characterizing Reference Locality in the WWW", In *Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems (PDIS'96)*, pp. 92–103, 1996.
- [2] N. Bansal and M. Harchol-Balter, "Analysis of SRPT Scheduling: Investigating Unfairness", In *Sigmetrics 2001 / Performance 2001*, pp. 279–290, June 2001.
- [3] X. Chen and J. Heidemann, "Preferential Treatment for Short Flows to Reduce Web Latency", *To appear in Computer Networks*, 2003.
- [4] E. G. Coffman and P. J. Denning, *Operating Systems Theory*, Prentice-Hall Inc., 1973.
- [5] R. W. Conway, W. L. Maxwell, and L. W. Miller, *Theory of Scheduling*, Addison-Wesley Publishing Company, 1967.
- [6] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", *IEEE/ACM Transactions on Networking*, pp. 835–846, December 1997.
- [7] M. Crovella, R. Frangioso, and M. Harchol-Balter, "Connection Scheduling in Web Servers", In *USENIX Symposium on Internet Technologies and Systems (USITS '99)*, pp. 243–254, Boulder, Colorado, October 1999.
- [8] M. E. Crovella, "Performance Evaluation with Heavy Tailed Distributions", In *Job Scheduling Strategies for Parallel Processing 2001 (JSSPP)*, pp. 1–10, 2001.
- [9] A. Feldmann, A. Gilbert, P. Huang, and W. Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control", In *Proc. of the ACM/SIGCOMM'99*, August 1999.
- [10] M. J. Fischer, D. Gross, D. M. B. Masi, and J. F. Shortle, "Analyzing the Waiting Time Process in Internet Queueing Systems With the Transform Approximation Method", *The Telecommunications Review*, pp. 21–32, 2001.
- [11] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance", *IEEE/ACM Transactions on Networking*, 1(4):397–413, August 1993.
- [12] W. Gong, Y. Liu, V. Misra, and D. Towsley, "On the Tails of Web File Size Distributions", In *Proc. of 39-th Allerton Conference on Communication, Control, and Computing*, October 2001.
- [13] L. Guo and I. Matta, "Differentiated Control of Web Traffic: A Numerical Analysis", In *SPIE ITCOM'2002: Scalability and Traffic Control in IP Networks*, August 2002.
- [14] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal, "Size-based Scheduling to Improve Web Performance", *ACM Transaction on Computer Systems*, (May), 2003.
- [15] M. Harchol-Balter, K. Sigman, and A. Wierman, "Asymptotic Convergence of Scheduling Policies with respect to Slowdown", *Performance Evaluation*, 49:241–256, September 2002.
- [16] <http://www.isi.edu/nsnam/ns/>, "The Network Simulator ns2",
- [17] L. Kleinrock, *Queueing Systems, Volume II: Computer Applications*, Wiley, New York, 1976.
- [18] H. Krayl, E. J. Neuhold, and C. Unger, *Grundlagen der Betriebssysteme*, Walter de Gruyter, Berlin, New York, 1975.
- [19] D. S. Mitrinovic, *Analytic Inequalities*, Springer-Verlag, 1970.
- [20] S. Muthukrishnan, R. Rajaraman, A. Shaheen, and J. E. Gehrke, "Online Scheduling to Minimize Average Stretch", In *40th Annual Symposium on Foundation of Computer Science*, pp. 433–442, 1999.
- [21] V. Paxson and S. Floyd, "Wide-Area Traffic: The failure of Poisson Modelling", *IEEE/ACM Transactions on Networking*, 3:226–244, June 1995.
- [22] R. Schassberger, "A Detailed Steady State Analysis of the M/G/1 Queue under various Time Sharing Disciplines", In P. J. C. In O.J. Boxma G. Iazeolla, editor, *Computer Performance and Reliability*, pp. 431–442, North Holland, 1987.
- [23] R. Schassberger, "The Steady State Distribution of Spent Service Times Present in the M/G/1 Foreground-Background Processor-sharing Queue", *Journal of Applied Probability*, 25(7):194–203, 1988.
- [24] L. E. Schrage, "The queue M/G/1 with feedback to lower priority queues", *Management Science*, 13(7):466–474, 1967.
- [25] L. E. Schrage, "A Proof of the Optimality of the Shortest Remaining Service Time Discipline", *Operations Research*, 16:670–690, 1968.
- [26] L. E. Schrage and L. W. Miller, "The queue M/G/1 with the shortest processing remaining time discipline", *Operations Research*, 14:670–684, 1966.
- [27] F. D. Smith, F. H. Campos, K. Jeffay, and D. Ott, "What TCP/IP protocol headers can tell us about the web", In *Proc. ACM SIGMETRICS*, pp. 245–256, June 2001.
- [28] R. W. Wolff, "Time Sharing with Priorities", *SIAM Journal of Applied Mathematics*, 19(3):566–574, 1970.
- [29] S. Yang and G. Veciana, "Size-based Adaptive bandwidth Allocation: Optimizing the Average QoS for Elastic Flows", In *Infocom 2002*, 2002.