RC23415 (W0411-053) November 8, 2004
Mathematics

IBM Research Report

Scheduling to Minimize General Functions of the Mean and Variance of Sojourn Times in Queueing Systems

Yingdong Lu, Mark S. Squillante
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Scheduling to Minimize General Functions of the Mean and Variance of Sojourn Times in Queueing Systems

Yingdong Lu and Mark S. Squillante
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
{yingdong,mss}@watson.ibm.com

Abstract

The optimality of shortest remaining processing time (SRPT) and its variants with respect to minimizing mean sojourn times are well known. Some recent studies have further argued that SRPT does not unfairly penalize large customers in order to benefit small customers, and thus have proposed the use of SRPT to improve performance in computer systems under various applications such as Web sites and databases. On the other hand, the variance of customer sojourn times is another important property of performance in these systems. We therefore consider alternative approaches to scheduling customers in queueing systems with the goal of providing mean sojourn times relatively close to those obtained under SRPT while also providing better variance properties. Our analysis includes deriving expressions for the mean and variance of customer sojourn times in these queueing systems, as well as for the parameters of the alternative scheduling policies. These results illustrate and quantify a fundamental performance tradeoff between decreasing the mean sojourn time and increasing the sojourn time variance, and vice versa. Our mathematical framework is then exploited to determine scheduling policies and their control parameters in order to optimize general functions of the mean and variance of sojourn times in queueing systems.

1 Introduction

Stochastic models and related queueing-theoretic results have played a fundamental role in the design of scheduling strategies of both theoretical and practical interest. This has been especially the case in single-server queues; refer to [12, 34, 33, 35, 22, 41] and the references cited therein. In particular, it is well known that scheduling the service of customers according to the shortest remaining processing time (SRPT) policy and its variants minimizes the mean sojourn time of customers [33, 22, 41]. Some recent studies have further argued that SRPT does not unfairly penalize large customers in order to benefit small customers, and therefore these studies propose the use of SRPT to improve performance in Web sites [11, 1, 15] and database systems [27].

However, as Schrage and Miller point out in their original study [34], the SRPT policy can raise several difficulties for a number of important reasons. Such difficulties can arise from the inability to accurately predict service times, or the complicated nature of implementing the preemptive aspect of the SRPT policy which requires keeping track of the remaining service times of all waiting customers as well as of the customer in service. Preemption can also incur additional costs, and thus one might want to avoid the preemption of customers in service whose remaining service time is not much larger than that of a new arrival. The results of a recent study further suggests that the workloads found at various commercial Web sites consist of multiple classes of customers based on the different service requirements of these customers [14].

We therefore consider a corresponding multiclass priority queue as an alternative to SRPT for scheduling the service of customers, with the goal of providing mean sojourn times close to their optimal values. In fact, Schrage and Miller [34] have demonstrated that scheduling policies similar to multiclass priority queues can alleviate some of the potential difficulties with SRPT while achieving mean sojourn times that are nearly as good as those under SRPT. Moreover, the approach based on the multiclass priority queue has the added advantage of not requiring to know precisely the service time of each customer. Instead, one only needs to be able to partition the workload into different classes where the service times within each class are relatively close to each other and the service times across classes are relatively different. This also provides an additional dimension that can be used to determine the optimal multiclass priority and its control parameters.

On the other hand, minimizing the mean customer sojourn time is only one of several important scheduling properties, and a priority policy that yields a small gain in first moment sojourn times can perform very poorly in terms of other measures of performance such as higher moments [41]. In particular, it often has been argued that a system with reasonable and predictable sojourn time may be more desirable than a system that is faster on average but highly variable [8, 23, 9, 36]. The original study of Schrage and Miller [34], however, does not consider any issues related to the variance of customer sojourn times. Thus, we consider versions of the corresponding multiclass priority queue, using a first-come first-serve (FCFS) ordering within each class, as alternative approaches for scheduling the service of customers in queueing systems, with the goal of providing mean sojourn times relatively close to those obtained under SRPT while also providing better variance properties. Serving customers within each class according to an FCFS queueing discipline can reduce the sojourn time variance within each class (among disciplines that do not affect the per-class queue length distribution) [17] and can reduce the preemptions among customers with fairly similar service times, whereas the priority discipline among the classes can yield a service ordering somewhat close to SRPT, provided that the service time variability within each class is relatively low. As we shall demonstrate and quantify, there is an important tradeoff between decreasing (respectively, increasing) the mean sojourn time and increasing (respectively, decreasing) the sojourn time variance, especially at heavy traffic intensities.

For consistency with the Schrage and Miller study [34], we analyze both single-class SRPT M/G/1 queues and multiclass Fixed Priority (FP), or Head-of-the-Line (HOL), M/G/1 queues. The M/G/1 FP queue, however, is somewhat limited for our purposes to control the first two moments of customer sojourn times. We therefore consider in detail another multiclass M/G/1 priority queue in which customers are scheduled according to general functions of the time customers spend in the system. This priority queueing discipline is called time-function scheduling (TFS) [13] and the time-function parameters provide the ability to further control the first two moments of customer sojourn times. More precisely, the priority of each customer increases according to a monotonically nondecreasing per-class function of its time in the system and the customer with the highest instantaneous priority value in the queue is selected for service at each scheduling epoch. Under the assumption of linear time-functions, which shall be our focus in this paper, the priority of each customer increases linearly with its time in system. As part of our analysis, we derive expressions for the first two moments of the per-class sojourn times in non-preemptive linear TFS M/G/1 queues, which generalizes the limited first moment results in the research literature and provides for the first time second moment sojourn time results. We also derive closed-form expressions to determine a set of TFS control parameters that (exactly) satisfy a given vector of sojourn times, whereas to the best of our knowledge, the only previous results of this type published in the research literature are based on an iterative scheme to obtain a (approximate) solution for systems with more than two classes [24]. Note further that the study in [24] focuses on a very different problem and it does not consider second moment properties, although our results can be exploited within the context of the study in [24].

Our results demonstrate that the linear TFS policy can satisfy desired mean sojourn times while also providing better variance properties. Another more fundamental and general objective of interest to us is based on functions that combine both the mean and variance of customer sojourn times in a flexible manner so as to realize the goals of a broad spectrum of applications. Specifically, in determining the optimal scheduling policy, we formulate the problem as a function of the first two moments of customer sojourn times to maximize the overall utility of the system, where we exploit recent results in portfolio theory to obtain a general mean-variance utility function and use this to explore a spectrum of mean-variance objectives. Our analysis includes determining how to optimally segment the service time distribution of the single-class workload from the original M/G/1 SRPT preemptive-resume queue into the per-class service time distributions of the multiclass workload to be used in the FP and linear TFS M/G/1 queues. These results are obtained based on our derivations for the first two moments of the sojourn times in linear TFS M/G/1 queues. Finally, we use simulation together with traces from a large-scale production Web site to explore these same issues under non-M/G/1 type environments. Our results, however, are consistent with those obtained for the M/G/1 queues.

2 M/G/1 Queue Scheduling Policies

Consider the standard M/G/1 queue in which customers arrive according to an independent Poisson source with rate λ and customer service times are independent and identically distributed (i.i.d.) having a common distribution function $F(\cdot)$ with finite first two moments $\mathbb{E}[S] = \mu^{-1} = \int_0^\infty t\,dF(t)$ and $\mathbb{E}[S^2] = \int_0^\infty t^2\,dF(t)$. Let $\rho = \lambda/\mu$ denote the traffic intensity, where we assume throughout that $\rho < 1$. When preemption is allowed, we shall focus on preemptive-resume scheduling disciplines in which preempted customers resume service where they left off without any penalties. Let T denote the random variable for the customer sojourn time, W the random variable for the customer waiting time, and R the random variable for the customer residence time, where $T = W + R$. Note that when preemption is not allowed, then R follows the service time distribution $F(\cdot)$, and thus $\mathbb{E}[R] = \mathbb{E}[S]$ and $\mathbb{E}[R^2] = \mathbb{E}[S^2]$. Our primary focus in this section will be on obtaining expressions for the first two moments $\mathbb{E}[T]$ and $\mathbb{E}[T^2]$ of the customer sojourn times.

For multiclass versions of this M/G/1 queue, we shall use the index k to refer to class $k = 1, \ldots, K$. More specifically, the arrival stream of class k customers follows an independent Poisson process with rate λ_k such that $\lambda = \sum_{k=1}^{K} \lambda_k$. The class k customer service times are i.i.d. according to a common distribution function $F_k(\cdot)$ with finite first two moments $\mathbb{E}[S_k] = \mu_k^{-1} = \int_0^\infty t\, dF_k(t)$ and $\mathbb{E}[S_k^2] = \int_0^\infty t^2\, dF_k(t)$, where $\mu^{-1} = \sum_{k=1}^{K} \mu_k^{-1}(\lambda_k/\lambda)$. (Note in particular that this supports the so-called heavy-tailed property considered in [1].) Let $\rho_k = \lambda_k/\mu_k$ denote the traffic intensity for class k, and thus $\rho = \lambda/\mu = \sum_{k=1}^{K} \rho_k$. We shall consider priority scheduling disciplines among classes where customers within each class are served in an FCFS manner. Let T_k denote the random variable for the class k sojourn time, W_k the random variable for the class k waiting time, and R_k the random variable for the class k residence time, where $T_k = W_k + R_k$. From the total law of probability, we then have

$$\begin{aligned}
\mathbb{E}[T] &= \sum_{k=1}^{K} \mathbb{E}[T_k]\, \mathbb{P}[\text{ customer belongs to class } k\ ] \\
&= \sum_{k=1}^{K} \mathbb{E}[T_k]\, \frac{\lambda_k}{\lambda}, \qquad (1) \\
\mathbb{E}[T^2] &= \sum_{k=1}^{K} \mathbb{E}[T_k^2]\, \mathbb{P}[\text{ customer belongs to class } k\ ] \\
&= \sum_{k=1}^{K} \mathbb{E}[T_k^2]\, \frac{\lambda_k}{\lambda}. \qquad (2)
\end{aligned}$$

Kleinrock [18, 20] first established an M/G/1 conservation law that will be useful for our purposes below. Specifically, for a non-preemptive M/G/1 queue under any work-conserving scheduling policy, the mean per-class waiting times $\mathbb{E}[W_k]$ must satisfy

$$\sum_{k=1}^{K} \rho_k \mathbb{E}[W_k] = \frac{\rho \mathbb{E}[W_0]}{1-\rho}. \qquad (3)$$

2.1 Shortest Remaining Processing Time

The shortest remaining processing time (SRPT) policy schedules in a preemptive manner the customer with the smallest remaining processing time at every point in time. An analysis of M/G/1 SRPT preemptive-resume queues was first derived by Schrage and Miller [34], from which one can obtain expressions for the first two moments of the customer sojourn times as follows

$$
\begin{aligned}
\mathbb{E}[T] \;&=\; \mathbb{E}[R] \;+\; \mathbb{E}[W], && (4)\\
&=\; \int_0^\infty \frac{1-F(t)}{1-\rho(t)}dt \\
&+\; \frac{\lambda}{2}\int_0^\infty \left\{ \frac{\int_0^p t^2 dF(t) + p^2(1-F(p))}{(1-\rho(p))^2} \right\} dF(p), && (5)\\
\mathbb{E}[T^2] \;&=\; \mathbb{E}[R^2] \;+\; 2\mathbb{E}[R]\mathbb{E}[W] \;+\; \mathbb{E}[W^2], && (6)\\
&=\; \int_0^\infty \Bigg\{ \int_0^p \frac{\lambda \int_0^t y^2 dF(y)}{1-\rho(t)}dt \\
&+\; \left[\int_0^p \frac{dt}{1-\rho(t)}\right]^2 \Bigg\} dF(p) \\
&+\; 2\left(\int_0^\infty \frac{1-F(t)}{1-\rho(t)}dt\right) \\
&\quad\; \left(\frac{\lambda}{2}\int_0^\infty \left\{ \frac{\int_0^p t^2 dF(t) + p^2(1-F(p))}{(1-\rho(p))^2} \right\} dF(p)\right) \\
&+\; \lambda \int_0^\infty \frac{\int_0^p t^3 dF(t) + p^3(1-F(p))}{3(1-\rho(p))^3} \\
+\, 2\lambda^2 &\int_0^\infty \left\{ \frac{\left(\int_0^p t^2 dF(t) + p^2(1-F(p))\right)\int_0^p t^2 dF(t)}{2(1-\rho(p))^4} \right\}, && (7)
\end{aligned}
$$

where $\rho(p) = \lambda \int_0^p t dF(t)$.

2.2 Fixed Priority

The fixed priority (FP) scheduling policy (also known as the head-of-the-line (HOL) priority policy), in which the service of class k customers has priority over class k' customers for all $1 \le k < k' \le K$, has received considerable attention in the research literature. In particular, it is well-known that the first two moments of the class k sojourn times in M/G/1 FP preemptive-resume queues are given by

$$
\begin{aligned}
\mathbb{E}[T_k] \;&=\; \frac{\sum_{j=1}^k \lambda_j \mathbb{E}[S_j^2]}{2(1-\rho_{k-1}^+)(1-\rho_k^+)} \;+\; \frac{\mathbb{E}[S_k]}{1-\rho_{k-1}^+}, && (8)\\
\mathbb{E}[T_k^2] \;&=\; \frac{\sum_{j=1}^k \lambda_j \mathbb{E}[S_j^3]}{3(1-\rho_{k-1}^+)^2(1-\rho_k^+)} \;+\; \frac{\mathbb{E}[S_k^2]}{(1-\rho_{k-1}^+)^2} \\
&+\; \left(\frac{\sum_{j=1}^{k-1} \lambda_j \mathbb{E}[S_j^2]}{(1-\rho_{k-1}^+)^2} + \frac{\sum_{j=1}^k \lambda_j \mathbb{E}[S_j^2]}{(1-\rho_{k-1}^+)(1-\rho_k^+)} \right) \mathbb{E}[T_k], && (9)
\end{aligned}
$$

where $\rho_k^+ \equiv \sum_{j=1}^{k} \rho_j$. Variants of these results were first obtained by Miller [29], Takács [38] and Welch [40].

Similarly, the first two moments of the class k sojourn times in non-preemptive M/G/1 FP queues can be expressed as

$$\begin{aligned}
\mathbb{E}[T_k] &= \frac{\sum_{j=1}^{K} \lambda_j \mathbb{E}[S_j^2]}{2(1-\rho_{k-1}^+)(1-\rho_k^+)} + \mathbb{E}[S_k], \qquad (10) \\
\mathbb{E}[T_k^2] &= \frac{\sum_{j=1}^{K} \lambda_j \mathbb{E}[S_j^3]}{3(1-\rho_{k-1}^+)^2(1-\rho_k^+)} \\
&\quad + \frac{\left(\sum_{j=1}^{k} \lambda_j \mathbb{E}[S_j^2]\right)\left(\sum_{j=1}^{K} \lambda_j \mathbb{E}[S_j^2]\right)}{2(1-\rho_{k-1}^+)^2(1-\rho_k^+)^2} \\
&\quad + \frac{\left(\sum_{j=1}^{k-1} \lambda_j \mathbb{E}[S_j^2]\right)\left(\sum_{j=1}^{K} \lambda_j \mathbb{E}[S_j^2]\right)}{2(1-\rho_{k-1}^+)^3(1-\rho_k^+)} + \mathbb{E}[S_k]. \qquad (11)
\end{aligned}$$

Variants of equation (10) were first given by Cobham [7], whereas variants of equation (11) were first obtained by Kesten and Runnenburg [16].

2.3 Linear Time-Function Scheduling

The corresponding sojourn time results for the linear time-function scheduling (TFS) policy, where the priority of each customer increases according to a linear function of its time in system with slope b_k and offset zero and the highest priority customer among all classes is served in either a non-preemptive or preemptive-resume manner with ties broken in an FCFS manner, are much less well established in the research literature. Kleinrock [18, 19, 22] derives expressions for the mean sojourn time in the non-preemptive and preemptive-resume cases under exponential service time distributions. (Note in particular that, while the analysis in [22] starts with M/G/1 assumptions instead of the M/M/1 assumptions in [18, 19], Kleinrock subsequently adds the restriction of exponential service times in [21].) However, to the best of our knowledge, more general first moment sojourn time results are lacking and no second moment sojourn time results are available in the research literature.

In this section we derive expressions for the first two moments of the per-class sojourn times in non-preemptive linear TFS M/G/1 queues, thus filling major gaps in the research literature. Our approach is based on a generalization of classical approaches for decomposing the per-class sojourn times in multiclass M/G/1 priority queues that rely primarily on the PASTA (Poisson Arrivals See Time Averages) property [41] and Little's Law [25]. We focus on non-preemptive linear TFS in part because Schrage and Miller tend to favor non-preemptive over preemptive-resume disciplines due to the additional overhead of the latter. It is important to note, however, that our extensive simulation results with both preemptive and non-preemptive versions of all scheduling policies considered herein demonstrate the same trends under preemptive priority policies as those shown in this paper for non-preemptive priority policies. Furthermore, we are currently working to extend the results derived in this section to handle the case of preemptive-resume linear TFS M/G/1 queues. Assume throughout that $b_1 \geq b_2 \geq \ldots \geq b_K \geq 0$.

2.3.1 Moments of Sojourn Times

Consider an arbitrary arrival at some time t of a so-called tagged customer of class k in a non-preemptive linear TFS M/G/1 queue. Let N_{jk} be the random variable denoting the number of class j customers in the system at time t that receive service before the tagged class k customer, let M_{jk} be the random variable denoting the number of class j customers that arrive after time t and receive service before the tagged class k customer, and let W_0 be the random variable denoting the residual life of the customer in service at time t. Then the waiting time of the tagged class k customer can be expressed as

$$W_k \;=\; W_0 \,+\, \sum_{j=1}^{K} \left(\sum_{i=1}^{N_{jk}} X_{ji} \,+\, \sum_{i=1}^{M_{jk}} X_{ji} \right), \tag{12}$$

for $k = 1, \ldots, K$, from which we obtain the first two moments of the class k sojourn times

$$\begin{aligned} \mathbb{E}[T_k] &= \mathbb{E}[W_k] \,+\, \mathbb{E}[S_k], &(13)\\ \mathbb{E}[T_k^2] &= \mathbb{E}[W_k^2] \,+\, 2\mathbb{E}[W_k]\mathbb{E}[S_k] \,+\, \mathbb{E}[S_k^2], &(14) \end{aligned}$$

in terms of

$$\begin{aligned} \mathbb{E}[W_k] &= \mathbb{E}[W_0] + \mathbb{E}\left[\sum_{j=1}^{K} \left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji} \right) \right], &(15)\\ \mathbb{E}[W_k^2] &= \mathbb{E}[W_0^2] + 2\mathbb{E}\left[W_0 \sum_{j=1}^{K} \left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji} \right) \right] \\ &\quad + \mathbb{E}\left[\left\{ \sum_{j=1}^{K} \left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji} \right) \right\}^2 \right], &(16) \end{aligned}$$

where $X_{j1}, X_{j2}, \ldots$ is a sequence of i.i.d. random variables such that $X_{ji} \overset{\mathrm{d}}{=} S_j$ for all $j = 1, \ldots, K$.

Under the assumptions of a non-preemptive M/G/1 queue, it can be easily shown that the first two moments of the residual life of the customer in service at time t are given by

$$\mathbb{E}[W_0] = \sum_{k=1}^{K} \rho_k \frac{\mathbb{E}[S_k^2]}{2\mathbb{E}[S_k]}, \quad \mathbb{E}[W_0^2] = \sum_{k=1}^{K} \rho_k \frac{\mathbb{E}[S_k^3]}{3\mathbb{E}[S_k]}. \tag{17}$$

Upon multiplying equation (12) for any pair k, k' and taking expectations, we obtain

$$\begin{aligned}
\mathbb{E}[W_k W_{k'}] &= \mathbb{E}[W_0^2] + \mathbb{E}\left[\sum_{j=1}^{K}\left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji}\right)\right] \\
&\quad \mathbb{E}\left[\sum_{j=1}^{K}\left(\sum_{i=1}^{N_{jk'}} X_{ji} + \sum_{i=1}^{M_{jk'}} X_{ji}\right)\right] \\
&+ \mathbb{E}\left[W_0 \sum_{j=1}^{K}\left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji}\right)\right] \\
&+ \mathbb{E}\left[W_0 \sum_{j=1}^{K}\left(\sum_{i=1}^{N_{jk'}} X_{ji} + \sum_{i=1}^{M_{jk'}} X_{ji}\right)\right] \qquad (18)
\end{aligned}$$

Similarly, multiplying equation (12) for each k by W_0 and taking expectations yields

$$\mathbb{E}[W_0 W_k] = \mathbb{E}[W_0^2] + \mathbb{E}\left[W_0 \sum_{j=1}^{K}\left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji}\right)\right]. \qquad (19)$$

Observe that $M_{jk} = 0$ with probability 1 for all $j \geq k$ since no customer with an equal or smaller slope b_j and arrival time after t can overtake the tagged class k customer. Moreover, the priority of the tagged class k customer when it starts service is given by $b_k W_k$, and thus M_{jk} is the number of class j customers arriving in the interval $(t, t + Z_j)$ such that $b_k W_k = b_j(W_k - Z_j)$, or equivalently $Z_j = [1 - (b_k/b_j)]W_k$. It then follows from applications of the first 2 moments of Little's Law [25, 4, 3] under our model assumptions that

$$\begin{aligned}
\mathbb{E}[M_{jk}] &= \lambda_j \mathbb{E}[Z_j] \;=\; \lambda_j[1 - (b_k/b_j)]\mathbb{E}[W_k], \qquad (20) \\
\mathbb{E}[M_{jk}^2] &= (\lambda_j^2/2)\mathbb{E}[Z_j^2] + \mathbb{E}[M_{jk}] \\
&= (\lambda_j^2/2)\mathbb{E}[W_k^2][1 - 2(b_k/b_j) + (b_k/b_j)^2] \\
&\quad + \mathbb{E}[M_{jk}], \qquad (21)
\end{aligned}$$

for $j < k = 1, \ldots, K$.

We further observe that all class $j \leq k$ customers in the system at time t will receive service before the tagged class k customer, since its smaller or equal slope b_k and arrival time t prevents the tagged class k customer from overtaking those class j customers who arrived before time t. It therefore follows from the PASTA property and the first 2 moments of Little's Law [25, 4, 3] that

$$\begin{aligned}
\mathbb{E}[N_{jk}] &= \lambda_j \mathbb{E}[W_j], \qquad (22) \\
\mathbb{E}[N_{jk}^2] &= (\lambda_j^2/2)\mathbb{E}[W_j^2] + \mathbb{E}[N_{jk}], \qquad (23)
\end{aligned}$$

for all $j \leq k = 1, \ldots, K$. Now consider a class $j > k$ customer who arrives at time $t' < t$, is in the system at time t, and receives service before the tagged class k customer. These conditions based on the definition of N_{jk} for $j > k$ are satisfied by a class j customer provided that $t - t' < W_j \leq t - t' + W_k$. The upper limit is required to ensure that the priority of the class j customer at time $t - t' + W_k$ (i.e., $b_j(t - t' + W_k)$) is not less than the priority of the tagged class k customer at the same time (i.e., $b_k W_k$). From $b_k W_k = b_j(t - t' + W_k)$ we obtain the relationship $t - t' + W_k = b_k/(b_k - b_j)(t - t')$. Since the arrivals of such class j customers follow a non-homogeneous Poisson process, and the time dependent arrival moments can be expressed as functionals of W_j, we have

$$\begin{aligned}
\mathbb{E}[N_{jk}] &= \int_0^\infty \lambda_j \mathbb{P}[y < W_j \leq \frac{b_k}{b_k - b_j} y] dy, \\
&= \lambda_j \mathbb{E}[W_j] - \lambda_j \frac{b_k - b_j}{b_k} \mathbb{E}[W_j] \\
&= \lambda_j \mathbb{E}[W_j] \frac{b_j}{b_k}, \qquad (24) \\
\mathbb{E}[N_{jk}^2] &= \left(\lambda_j \mathbb{E}[W_j] \frac{b_j}{b_k} \right)^2 + \lambda_j \mathbb{E}[W_j] \frac{b_j}{b_k}, \qquad (25)
\end{aligned}$$

for all $j > k = 1, \ldots, K - 1$.

To complete our solution, we derive the remaining measures conditional on the waiting time for the tagged customer W_k. In particular, upon conditioning on W_k, it is easy to see that M_{jk} is conditionally independent of the other variables. We therefore have

$$\begin{aligned}
W_k^2 &= \mathbb{E}[W_k^2 | W_k] = \mathbb{E}[W_0^2 | W_k] \\
&+ 2\mathbb{E}\left[W_0 \sum_{j=1}^{K} \left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji} \right) \Big| W_k \right] \\
&+ \mathbb{E}\left[\left\{ \sum_{j=1}^{K} \left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji} \right) \right\}^2 \Big| W_k \right] \qquad (26)
\end{aligned}$$

and

$$
\begin{aligned}
& \mathbb{E}\left[\left\{\sum_{j=1}^{K}\left(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} X_{ji}\right)\right\}^2 \Big| W_k\right] \\
= \quad & \sum_{j,\ell=1,j\neq\ell}^{K}\left\{\mathbb{E}\left[\sum_{i=1}^{N_{jk}} X_{ji}\sum_{i=1}^{N_{\ell k}} X_{\ell i}\Big| W_k\right]\right. \\
& +\mathbb{E}\left[\sum_{i=1}^{M_{jk}} X_{ji}\Big| W_k\right]\mathbb{E}\left[\sum_{i=1}^{M_{\ell k}} X_{\ell i}\Big| W_k\right] \\
& \left.+\mathbb{E}\left[\sum_{i=1}^{N_{jk}} X_{ji}\Big| W_k\right]\mathbb{E}\left[\sum_{i=1}^{M_{\ell k}} X_{\ell i}\Big| W_k\right]\right\} \\
& +\sum_{j=1}^{K}\left\{\mathbb{E}\left[\left(\sum_{i=1}^{N_{jk}} X_{ji}\right)^2 \Big| W_k\right] + \mathbb{E}\left[\left(\sum_{i=1}^{M_{jk}} X_{ji}\right)^2 \Big| W_k\right]\right. \\
& \left.+\mathbb{E}\left[\left(\sum_{i=1}^{N_{jk}} X_{ji}\right)\left(\sum_{i=1}^{M_{jk}} X_{ji}\right)\Big| W_k\right]\right\}.
\end{aligned} \tag{27}
$$

Since N_{jk} can be treated as a Poisson random variable, we obtain

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{i=1}^{N_{jk}} X_{ji}\sum_{i=1}^{N_{\ell k}} X_{\ell i}\Big| W_k\right] \\
= \quad & \mathbb{E}\left[\sum_{i=1}^{N^0_{j\ell k}+N^j_{j\ell k}} X_{ji}\sum_{i=1}^{N^0_{j\ell k}+N^\ell_{j\ell k}} X_{\ell i}\Big| W_k\right] \\
= \quad & \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N^0_{j\ell k}+N^j_{j\ell k}} X_{ji}\sum_{i=1}^{N^0_{j\ell k}+N^\ell_{j\ell k}} X_{\ell i}\Big| W_k, N^0_{j\ell k}\right]\right] \\
= \quad & \mathbb{E}[S_j]\mathbb{E}[S_\ell]\mathbb{E}\Big[\mathbb{E}[(N^0_{j\ell k})^2|W_k] + \mathbb{E}[N^0_{j\ell k}|W_k]\mathbb{E}[N^j_{j\ell k}|W_k] \\
& +\mathbb{E}[N^0_{j\ell k}|W_k]\mathbb{E}[N^\ell_{j\ell k}|W_k] + \mathbb{E}[N^j_{j\ell k}|W_k]\mathbb{E}[N^\ell_{j\ell k}|W_k]\Big]
\end{aligned} \tag{28}
$$

Observe that the variance of $N^0_{j\ell k}$ is the same as the covariance of N_{jk} and $N_{\ell k}$, and thus it can be uniquely determined by the measures $\mathbb{E}[W_j|W_k]$, $\mathbb{E}[W_\ell|W_k]$ and $\mathbb{E}[W_j W_\ell|W_k]$.

The continuation of our calculations yields

$$\begin{aligned}\mathbb{E}[W_0\sum_{i=1}^{N_{jk}} X_{ji}|W_k] &= \mathbb{E}[\mathbb{E}[W_0\sum_{i=1}^{N_{jk}} X_{ji}|W_k, W_0]|W_k] \\ &= \frac{\mathbb{E}[S_j]b_j}{b_k}\mathbb{E}[W_0[\mathbb{E}[W_j|W_0, W_k]|W_k] \\ &= \frac{\mathbb{E}[S_j]b_j}{b_k}\mathbb{E}[W_0W_j|W_k] \qquad (29)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{N_{jk}} X_{ji}\sum_{i=1}^{N_{\ell k'}} X_{\ell i}\right] = \mathbb{E}\left[\left[\sum_{i=1}^{N_{jk}} X_{ji}\sum_{i=1}^{N_{\ell k'}} X_{\ell i}\right]\Big|W_k, W_{k'}\right] \\ &= \mathbb{E}[S_j]\mathbb{E}[S_\ell]\mathbb{E}[\mathbb{E}[N_{jk}N_{jk'}|W_k, W_{k'}]]. \qquad (30)\end{aligned}$$

Once again, given the knowledge that N_{jk} and $N_{jk'}$ are Poisson random variables, the same arguments as above can be used to finalize this result.

The next two theorems summarize our main results based on the foregoing derivations.

Theorem 1. *In a non-preemptive linear TFS M/G/1 queue, the mean waiting time of class* $k = 1, \ldots, K$ *customers can be expressed as*

$$\mathbb{E}[W_k] = \frac{(\mathbb{E}[W_0]/(1-\rho)) - \sum_{i=k+1}^{K}\rho_i\mathbb{E}[W_i](1-b_i/b_k)}{1-\sum_{i=1}^{k-1}\rho_i(1-b_k/b_i)}, \qquad (31)$$

with the overall mean sojourn time given by

$$\begin{aligned}\mathbb{E}[T] &= \sum_{k=1}^{K}(\\ &\quad \frac{(\mathbb{E}[W_0]/(1-\rho)) - \sum_{i=k+1}^{K}\rho_i\mathbb{E}[W_i](1-b_i/b_k)}{1-\sum_{i=1}^{k-1}\rho_i(1-b_k/b_i)} \\ &\quad +\mathbb{E}[S_k])\,\frac{\lambda_k}{\lambda}. \qquad (32)\end{aligned}$$

Proof. Upon substituting (20), (22) and (24) into the expression that results from applying Wald's equation [41] to (15), we have

$$\mathbb{E}[W_k] = \frac{\mathbb{E}[W_0] + \sum_{i=1}^{k}\rho_i\mathbb{E}[W_i] + \sum_{i=k+1}^{K}\rho_i\mathbb{E}[W_i](b_i/b_k)}{1-\sum_{i=1}^{k-1}\rho_i(1-b_k/b_i)}.$$

The triangular set of equations in (31) then follows from the conservation law given in (3) and straightforward algebra. Adding $\mathbb{E}[S_k]$ to (31) and substituting the result into (1) yields (32). □

Theorem 2. *The second moment of the customer sojourn times in a non-preemptive linear TFS M/G/1 queue is given by equations (2), (14) and (16) which can be efficiently computed from equations (17) – (30).*

Proof. To solve for $\mathbb{E}[T_k^2]$ in (14) and substitute the result in (2), we need to determine the measures $\mathbb{E}[W_k^2]$, $\mathbb{E}[W_k W_{k'}]$ and $\mathbb{E}[W_0 W_k]$ provided in (16), (18) and (19), respectively, which yields a system of $K(K+1)/2$ equations of unknowns. It follows from the above derivations that all of the terms in these equations can be expressed as linear functions of the unknowns and that the coefficient matrix is triangular and nonsingular. Hence, the second moment of the class k sojourn times can be obtained in a very efficient manner by exploiting these properties of the system of linear equations derived above. □

2.3.2 Setting of Policy Control Parameters

The set of slopes $\{b_1, \ldots, b_K\}$ represent the control parameters available to us in non-preemptive linear TFS M/G/1 queues to achieve any feasible vector of desired sojourn times $(\mathbb{E}[T_1^*], \ldots, \mathbb{E}[T_K^*])$. We therefore derive closed-form expressions to determine a set of control parameters $\{b_1, \ldots, b_K\}$ that satisfy a given vector $(\mathbb{E}[W_1^*] = \mathbb{E}[T_1^*] - \mathbb{E}[S_1], \ldots, \mathbb{E}[W_K^*] = \mathbb{E}[T_K^*] - \mathbb{E}[S_K])$ by inverting the mapping in (31) for any $K \geq 2$. It is important to note that, to the best of our knowledge, the only previous results of this type published in the research literature are based on an iterative scheme to obtain a solution for systems with more than two classes [24].

The mean waiting time for customers of class k in non-preemptive linear TFS M/G/1 queues as a function of the per-class control parameters is given by (31). Observe the very simple dependence that $\mathbb{E}[W_k]$ has on the control parameters, namely the slopes b_k only appear as ratios. It follows from equation (31) with $b_1 \geq b_2 \geq \ldots \geq b_K$ that any feasible objective vector must have $\mathbb{E}[W_0] \leq \mathbb{E}[W_1^*] \leq \mathbb{E}[W_2^*] \leq \ldots \leq \mathbb{E}[W_K^*]$. Further observe that the scheduling policy decisions are not changed upon scaling all control parameters by any fixed constant, and thus without loss of generality we set $b_K = 1$. Additional feasibility requirements for the vector $(W_1^*, \ldots, W_K^*)$ can be readily verified as part of the following recursive algorithm by ensuring the corresponding variables satisfy the obvious constraints.

Following [30, 37], we define

$$
\begin{aligned}
\alpha_k &\equiv \sum_{i=1}^{K} \frac{\lambda_i \mathbb{E}[S_i^2]}{2(1-\rho)} - \sum_{i=k+1}^{K} \rho_i \mathbb{E}[W_i], \\
\beta_k &\equiv 1 - \sum_{i=1}^{k-1} \rho_i, \\
C_k &\equiv \sum_{i=k+1}^{K} \rho_i b_i \mathbb{E}[W_i], \\
D_k &\equiv \sum_{i=1}^{k-1} \frac{\rho_i}{b_i}.
\end{aligned}
$$

Upon substituting (17) and these definitions into equation (31), we obtain

$$\mathbb{E}[W_k] = \frac{\alpha_k + C_k/b_k}{\beta_k + b_k D_k}.$$

Substituting the relationship $D_k = D_{k+1} - \rho_k/b_k$ and $\beta_{k+1} = \beta_k - \rho_k$ from the above definitions and simplifying then yields

$$b_k = \frac{-(\mathbb{E}[W_k]\beta_{k+1} - \alpha_k) \pm \sqrt{(\mathbb{E}[W_k]\beta_{k+1} - \alpha_k)^2 + 4C_k\mathbb{E}[W_k]D_{k+1}}}{2\mathbb{E}[W_k]D_{k+1}},$$

for $k = 1, \ldots, K-1$. Since $b_1 \geq b_2 \geq \ldots \geq b_K = 1$, we have

$$b_k = \frac{-(\mathbb{E}[W_k]\beta_{k+1} - \alpha_k) + \sqrt{(\mathbb{E}[W_k]\beta_{k+1} - \alpha_k)^2 + 4C_k\mathbb{E}[W_k]D_{k+1}}}{2\mathbb{E}[W_k]D_{k+1}}, \tag{33}$$

for $k = 1, \ldots, K-1$.

Observe that the value of b_k in equation (33) depends only on the values of $b_1, \ldots, b_{k-1}$. We then have the following algorithm to recursively obtain the control parameters b_{K-1}, . . ., b_1 that can be used to achieve any feasible vector of desired sojourn times $(\mathbb{E}[T_1^*], \ldots, \mathbb{E}[T_K^*])$ in the corresponding non-preemptive linear TFS M/G/1 queue. First initialize the class K variables: $C_K = 0$; $b_K = 1$; $D_K = \frac{\mathbb{E}[W_0]/(1-\rho) - \beta_K \mathbb{E}[W_K]}{\mathbb{E}[W_K]}$. The corresponding variables for classes $k = K-1, K-2, \ldots, 2, 1$ are then computed consecutively as follows:

$$C_k \;=\; C_{k+1} + \rho_{k+1} b_{k+1} \mathbb{E}[W_{k+1}] \tag{34}$$

$$b_k \;=$$

$$\frac{-(\mathbb{E}[W_k]\beta_{k+1} - \alpha_k) + \sqrt{(\mathbb{E}[W_k]\beta_{k+1} - \alpha_k)^2 + 4C_k \mathbb{E}[W_k] D_{k+1}}}{2\mathbb{E}[W_k] D_{k+1}} \tag{35}$$

$$D_k \;=\; \frac{\alpha_k + C_k/b_k - \beta_k \mathbb{E}[W_k]}{b_k \mathbb{E}[W_k]} \tag{36}$$

3 Comparison of Scheduling Policies

With the motivation and issues provided in the introduction, we now consider the properties of the first two moments of the customer sojourn times in SRPT, FP, and linear TFS M/G/1 queues based on the results derived in the previous section. We first focus on the variability properties of the different M/G/1 queues and then we turn to general functions of both the mean and variance of customer sojourn times.

3.1 Variance Properties

As originally suggested by Schrage and Miller [34] and extended herein to also consider second moment properties, one can attempt to approximate the mean sojourn times in M/G/1 SRPT preemptive-resume queues with an appropriately chosen multiclass M/G/1 FP queue. Based on numerical experiments with the results in Section 2 (some to be presented later in this section) and simulations with data from a large-scale production Web site (some to be presented in Section 5), both of which were conducted as part of our present study, we consistently observed that among the various policies which can achieve similar mean customer waiting times, the performance measures of the M/G/1 queue under a linear TFS policy tend to have the smallest variability properties.

To further investigate this idea more formally, let us consider a special case with only two classes of customers ($K = 2$) where we compare the performance of M/G/1 queues under linear TFS and random priority policies. A random priority policy with parameter $\alpha \in [0, 1]$ can be realized as follows: at the beginning of each busy period, we will decide whether to give fixed priority to class 1 or class 2 where, in the long run, the proportion of busy periods in which class 1 has priority over class 2 is α. It is easy to see that under this policy, both the first and the second moments of the waiting time will be the convex combination of the moments of the two fixed priority policies.

Turing to the linear TFS policy, consider equation (12) where, for our purposes here, only M_{12} and N_{21} need to be examined. From our derivation in Section 2.3, we know that upon conditioning on the proper σ-algebra, M_{12} follows a Poisson distribution with mean $A(1 - b_1/b_2)$ and N_{21} follows a Poisson distribution with mean $A(b_1/b_2)$, where A and B are linear combinations of unknowns. Hence, when we solve for the linear TFS control parameters to achieve the same mean waiting time as the random priority policy of interest, we will obtain an affine relationship between α and b_1/b_2. On the other hand, the second moment of the customer waiting times under this linear TFS policy will be a quadratic function of b_1/b_2, i.e., a convex function. In then follows from these two properties that the linear TFS policy can always provide lower customer waiting time variance than under the random priority policy while providing the same mean customer waiting time performance.

3.2 Mean-Variance Utility Functions

Even though we have considerable numerical, simulation and formal evidence that the linear TFS policy can satisfy desired mean sojourn times while also providing better variance properties, the more fundamental and general objective of interest to us is based on functions that combine both the mean and variance of customer sojourn times in a flexible manner so as to realize the goals of a broad spectrum of applications. Given the first moment of customer sojourn times as the natural candidate measure for performance, the corresponding second moment is usually associated with risks through the use of moment inequalities, such as Chebeschev's inequality [41]. Therefore, in determining the optimal scheduling policy, we can formulate the problem as a function of the first two moments of customer sojourn times to maximize the overall utility of the system. Similar practices have been widely adapted in the field of finance, ever since H. Markowitz popularized the basic idea; see [26].

This so-called mean-variance approach is quite popular in portfolio theory and its applications, and a wide range of specific functional forms for two-parameter preferences have been proposed and used; e.g., refer to [28, 6, 32]. In particular, the following functional form for utility

$$U(\theta, \sigma) \;=\; \theta^a - \sigma^b, \tag{37}$$

where θ and σ are the mean and standard deviation of the measure of interest and a and b are function parameters, has been proposed and empirically evaluated [32]. This utility function is able to exhibit a broad spectrum of risk attitudes by appropriately choosing values for the parameters $a > 0$ and $b \in \mathbb{R}$. For example, the choices $a > 1$, $a = 1$ and $a < 1$ respectively represent decreasing, constant and increasing absolute risk aversion, whereas the choices $a > b$, $a = b$ and $a < b$ respectively represent decreasing, constant and increasing relative risk aversion. Moreover, Wagener [39] has recently shown the functional form in (37) to be very efficient from a computational perspective and to have much greater flexibility in covering the wide range of risk attitudes of interest than other functional forms that are commonly used in practice.

We therefore use an equivalent form for the utility function in (37) where the variance of customer sojourn times is used instead of the standard deviation such that $\theta = \mathbb{E}[T]$, $\sigma = \mathbb{E}[T^2] - \mathbb{E}[T]^2$ and $b = b/2$. A three class queueing system with Poisson arrivals is considered. The proportions of customer class 1, 2 and 3 are 80%, 14% and 6% respectively. Meanwhile the mean service time for the three classes are 1, 20 and 1000. Experiments are conducted for both SRPT and linear time function(LTF) rule under different traffic intensity ρ. To make the comparison fair, for each ρ, we set the slopes in the time function schedule so that the mean waiting time is close to that of SRPT. Besides the fact that time function scheduling provides smaller variances as our results demonstrate above, we also observe the significance of the impact of the policy in terms of the utility function. In Figure 1, we observe that the impact of the policy change is visibly more significant than the changes of variability in the service time. In the figure, vertically, we compare the utility function for these two policies with different coefficients of variation for the service time; horizontally, we compare the impact of the different selection of (a, b). In Figure 2, we fixed the system behavior, and show the general shape of the utility function when the relative (a, b) value changes. We see that in general it is a concave function, and the derivatives has a descending trend when the traffic intensity ρ increases.

4 Partitioning of Service Times into Classes

Our analysis in Sections 2 and 3 assumes that the workloads for the multiclass FP and linear TFS M/G/1 queues have been previously determined. However, in order to completely determine the optimal scheduling policy and its control parameters, it is equally important to obtain the best segmentation of the customer service times to determine the multiclass workloads. In this section, we consider how to optimally segment the service time distribution of the single-class workload from the original M/G/1 SRPT preemptive-resume queue into the per-class service time distributions of the multiclass workload to be used in the FP and linear TFS M/G/1 queues.

More specifically, we consider the set of variables $\{p_0, p_1, \ldots, p_{K-1}, p_K\}$ used to designate the priority class of customers according to whether their service times are in the interval $(p_{k-1}, p_k]$, such that $B_L = p_0 \le p_1 \le \ldots \le p_{K-1} \le p_K = B_U$ where B_L and B_U are lower and upper bounds on the customer service times, respectively. Given such a partitioning, the class k customer service times are i.i.d. according to a common distribution function with finite first three moments $\mathbb{E}[S_k] = \int_{p_{k-1}}^{p_k} t\, dF(t)$, $\mathbb{E}[S_k^2] = \int_{p_{k-1}}^{p_k} t^2\, dF(t)$ and $\mathbb{E}[S_k^3] = \int_{p_{k-1}}^{p_k} t^3\, dF(t)$, respectively. We now can formulate the problem of the optimal partitioning of the customer service times into a multiclass workload as part of determining the optimal scheduling policy and its control parameters as a function of the first two moments of customer sojourn times to maximize the overall utility of the system.

$$\text{(OP)} \qquad \max_{p_1,\ldots,p_{K-1}} \quad \mathbb{E}[T]^a - (\mathbb{E}[T^2] - \mathbb{E}[T]^2)^{b/2} \tag{38}$$

$$\text{s.t.} \quad B_L = p_0 \le p_1 \le \ldots \le p_{K-1} \le p_K = B_U \tag{39}$$

The decision variables are the partition points $\{p_1, \ldots, p_{K-1}\}$, and the parameters a and b are chosen to weight the first two moments of the customer sojourn times according to the application area of interest. In general, the objective function is nonlinear in the decision variables but its solution can be efficiently computed using known methods in nonlinear optimization; e.g., see [2, 10]. However, in many cases of interest, the objective function is convex in the decision variables, and thus its solution can be very efficiently computed using known methods in convex optimization; e.g., refer to [2].

5 Simulation Experiments

While the main contribution of this paper is the foregoing collection of mathematical results, in this section we briefly consider the queueing behavior in an Internet environment to apply and extend our understanding of the relationship between SRPT and linear TFS. Specifically, we use the access log data from large-scale production Web sites. The access logs contain several pieces of useful information about each client request served by the corresponding server. This includes the arrival time epoch of the k^{th} request served on the i^{th} server of the Web site, and the number of bytes comprising this client request. The unit of time in the access logs available to us is one second, which is quite standard. There can be tens to hundreds of client requests within a second at each server during peak traffic periods for the production Web sites of interest.

Even though the Web sites considered in our study serve dynamic content, we use the byte size of each request as an estimate of the service time for the request. This is both reasonable and accurate for the production Web sites of interest for a number of important reasons. While most of the pages are dynamic, the vast majority of the requests found at each server are for static objects. These production Web sites can also exploit techniques that keep track of each database update and the dynamic pages affected by the update so that these dynamic pages can be prebuilt upon such a database update, and this in turn makes it possible to serve dynamic pages exactly like static content [5]. Moreover, measurements of the time to serve the static objects, as well as to serve the dynamic pages using these techniques, demonstrate that these service times fit extremely well to a linear function of the byte size of the object or page, and that applying this function to the byte sizes yields service times with the same stochastic properties as those shown herein for byte sizes [5]. We therefore focus in this section on the number of bytes comprising each client request, which has the added benefit of making our results comparable to those of many previous studies. In particular, we use a measurement-based function of the byte size of each client request as an accurate estimate of the service time for the request.

We identify and focus on sufficiently long stationary intervals of traffic periods found in our analysis of the access logs from each server of every Web site. Of particular interest are peak traffic periods, given the importance of such intervals in capacity planning, dynamic resource allocation and other applications of performance analysis and control. These stationary intervals of peak traffic are comprised of traffic periods whose lengths are on the order of several hours and consist of at least several hundred-thousand data points. Hence, the corresponding arrival and service processes extracted from the Web site access logs are stationary sequences, which is confirmed by the stationarity testing method recently proposed in [31]. Moreover, in the interest of space, we will henceforth focus on a representative access log from a specific server of a particular Web site.

In this set of experiments, arrival times and service times are extracted from the trace of a Web site, and then they are fed into a single server queue following the SRPT and TFS rules. The outcome is presented consistent with the mean-variance scheme we have discussed. The utility function with $a = 0.25, b = -1$ and $a = 0.25, b = 0.5$ respectively for different intensities(ρ) are plotted in Figure 3. From the figure on the right, we can see the impact of different scheduling policies. In the figure on the left, the impact can not be distinguished, we believe that this is the result of the correlations in the arrival stream and between arrival and service in the Web trace.

6 Conclusions

The optimality of shortest remaining processing time (SRPT) and its variants with respect to minimizing mean sojourn times are well known. Some recent studies have further argued that SRPT does not unfairly penalize large customers in order to benefit small customers, and thus have proposed the use of SRPT to improve performance in computer systems under various applications such as Web sites and databases. On the other hand, minimizing the mean customer sojourn time is only one of several important scheduling properties, and a priority policy that yields a small gain in first moment sojourn times can perform very poorly in terms of other measures of performance such as higher moments. In particular, it often has been argued that a system with reasonable and predictable sojourn time may be more desirable than a system that is faster on average but highly variable. We therefore considered alternative approaches to scheduling customers in queueing systems with the goal of providing mean sojourn times relatively close to those obtained under SRPT while also providing better variance properties. Our analysis included deriving expressions for the mean and variance of customer sojourn times in these queueing systems, as well as for the control parameters of the alternative scheduling policies. These results illustrated and quantified a fundamental performance tradeoff between decreasing the mean sojourn time and increasing the sojourn time variance, and vice versa. Our mathematical framework is then exploited to determine scheduling policies and their control parameters in order to optimize general functions of the mean and variance of sojourn times in queueing systems. More specifically, in determining the optimal scheduling policy, we formulated the problem as a function of the first two moments of customer sojourn times to maximize the overall utility of the system, where we exploited recent results in portfolio theory to obtain a general mean-variance utility function and used this to explore a spectrum of mean-variance objectives. Our analysis included determining how to optimally segment the service time distribution of the single-class workload from the original SRPT queue into the per-class service time distributions of the multiclass workloads to be used in our alternative scheduling policies.

References

[1] N. Bansal and M. Harchol-Balter. Analysis of SRPT scheduling: Investigating unfairness. In *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 279–290, June 2001.

[2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

[3] D. Bertsimas and D. Nakazato. The distributional Little's law and its applications. *Operations Research*, 43(2):298–310, 1995.

[4] S. L. Brumelle. A generalization of $L = \lambda W$ to moments of queue length and waiting times. *Operations Research*, 20:1127–1136, 1972.

[5] J. R. Challenger, P. Dantzig, A. Iyengar, M. S. Squillante, and L. Zhang. Efficiently serving dynamic data at highly accessed Web sites. *IEEE/ACM Transactions on Networking*, 12(2):233–246, April 2004.

[6] V. Chopra and W. T. Ziemba. The effect of erros in mean and co-variance estimates on optimal portfoilio choice. *Journal of Portfolio Management*, pages 6–11, 1993.

[7] A. Cobham. Priority assignment in waiting line problems. *Operations Research*, 2:70–76, 1954.

[8] E. G. Coffman, Jr. and L. Kleinrock. Computer scheduling methods and their countermeasures. In *Proceedings of AFIPS Spring Joint Computer Conference*, volume 32, pages 11–21, April 1968.

[9] E. G. Coffman, Jr. and I. H. Mitrani. A characterization of waiting-time performance achievable by single-server queues. *Operations Research*, 28(3):810–821, 1980.

[10] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust-Region Methods*. SIAM, 2000.

[11] M. E. Crovella, R. Frangioso, and M. Harchol-Balter. Connection scheduling in Web servers. In *Proceedings of USENIX Symposium on Internet Technologies and Systems*, pages 243–254, October 1999.

[12] D. W. Fife. Scheduling with random arrivals and linear loss functions. *Management Science*, 11(3):429–437, 1965.

[13] L. L. Fong and M. S. Squillante. Time-Function Scheduling: A general approach to controllable resource management. In *Proceedings of Symposium on Operating Systems Principles*, page 230, December 1995.

[14] S. Ghosh and M. S. Squillante. Analysis and control of correlated Web server queues. *Computer Communications*, 27(28):1771–1785, December 2004.

[15] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems*, 21(2):207–233, 2003.

[16] H. Kesten and J. T. Runnenburg. Priority in waiting line problems. In *Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen*, volume A60, pages 312–336, 1957.

[17] J. F. C. Kingman. The effect of queue discipline on waiting time variance. In *Proceedings of the Cambridge Philosophical Society*, volume 58, pages 163–164, 1962.

[18] L. Kleinrock. *Communication Nets: Stochastic Message Flow and Delay*. McGraw-Hill, 1964.

[19] L. Kleinrock. A delay dependent queue discipline. *Naval Research and Logistics Quarterly*, 11:329–341, 1964.

[20] L. Kleinrock. A conservation law for a wide class of queueing disciplines. *Naval Research and Logistics Quarterly*, 12:181–192, 1965.

[21] L. Kleinrock. *Queueing Systems Volume I: Theory*. John Wiley and Sons, 1975.

[22] L. Kleinrock. *Queueing Systems Volume II: Computer Applications*. John Wiley and Sons, 1976.

[23] B. W. Lampson. A scheduling philosophy for multiprocessing systems. *Communications of the ACM*, 11(5):347–360, May 1968.

[24] M. K. Leung, J. C. S. Lui, and D. K. Yau. Adaptive proportional delay differentiated services: Characterization and performance evaluation. *IEEE/ACM Transactions on Networking*, 9(6), 2001.

[25] J. D. C. Little. A proof of the queuing formula $L = \lambda W$. *Operations Research*, 9:383–387, 1961.

[26] H. M. Markowitz. The early history of portfolio theory: 1600 – 1960. *Financial Analysts Journal*, 55:5–16, 1999.

[27] D. McWherter, B. Schroeder, N. Ailamaki, and M. Harchol-Balter. Priority mechanisms for oltp and transactional web applications. In *Proceedings of the 20th International Conference on Data Engineering (ICDE 2004)*, April 2004.

[28] J. Meyer. Two-moment decision models and expected utility maximization. *American Economic Review*, 77:421–430, 1987.

[29] R. G. Miller, Jr. Priority queues. *The Annals of Mathematical Statistics*, 31(1):86–103, 1960.

[30] R. D. Nelson. Invertible mapping of waiting times in a M/G/1 queue with linear priorities. Unpublished Draft, June 1993.

[31] H. Ombao, J. Raz, R. von Sachs, and B. Malow. Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, 96, 2001.

[32] A. Saha. Risk preference estimation in the non-linear standard deviation approach. *Economic Inquiry*, 35:770–782, 1997.

[33] L. E. Schrage. A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16:687–690, 1968.

[34] L. E. Schrage and L. W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14(4):670–684, 1966.

[35] K. C. Sevcik. Scheduling for minimum total loss using service time distributions. *Journal of the ACM*, 21(1):66–75, 1974.

[36] A. Silberschatz, P. B. Galvin, and G. Gagne. *Operating System Concepts*. John Wiley and Sons, Sixth edition, 2004.

[37] M. S. Squillante, L. L. Fong, S. Liu, and S. K. Ryan. A control study of time-function scheduling: Part I. Technical Report RC 19765, IBM Research Division, September 1994.

[38] L. Takács. Priority queues. *Operations Research*, 12(1):63–74, 1964.

[39] A. Wagener. Linear risk tolerance and mean-variance utility functions. Technical report, Department of Economics, University of Vienna, July 2004.

[40] P. D. Welch. On preemptive resume priority queues. *The Annals of Mathematical Statistics*, 35(2):600–612, 1964.

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

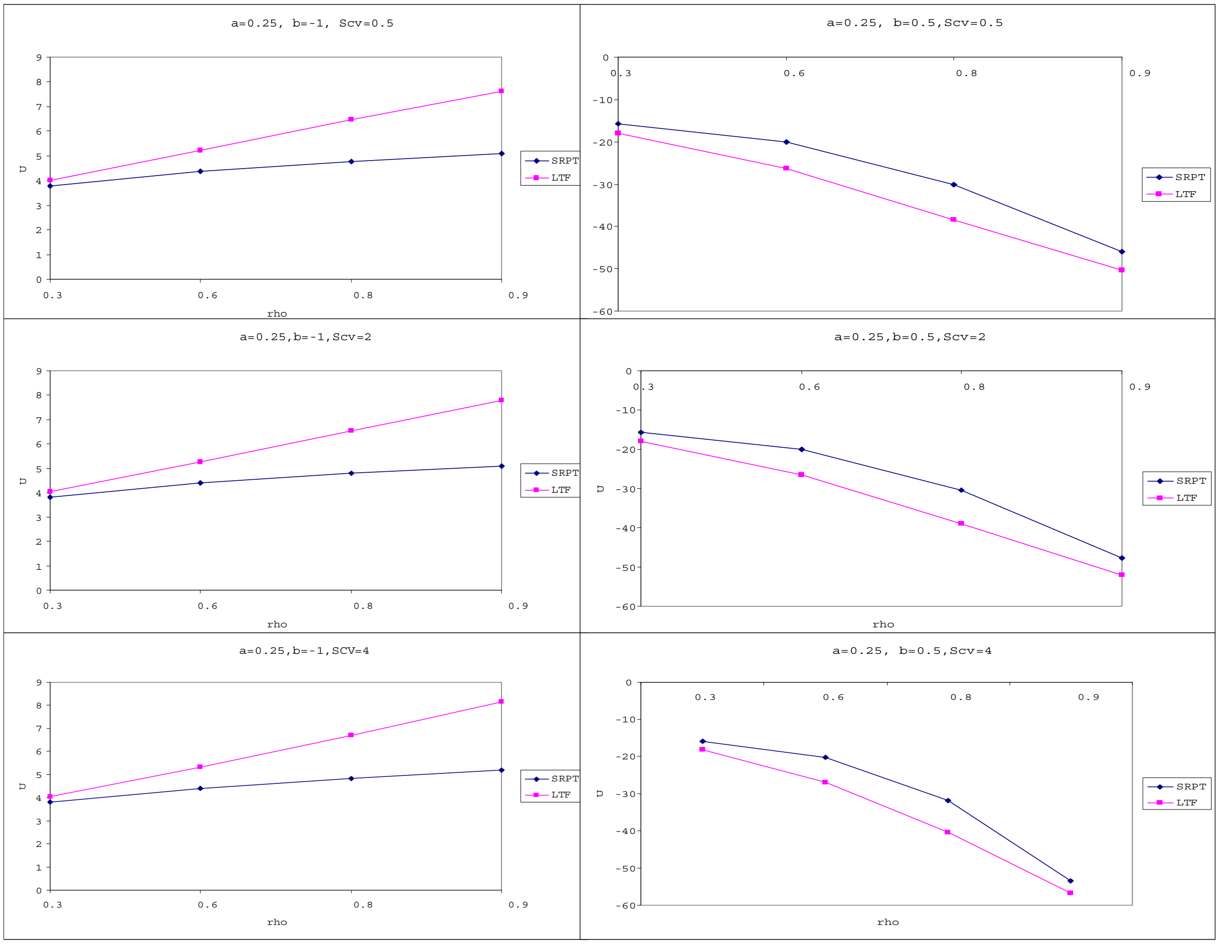

Fig. 1. Effects of variability vs. pol

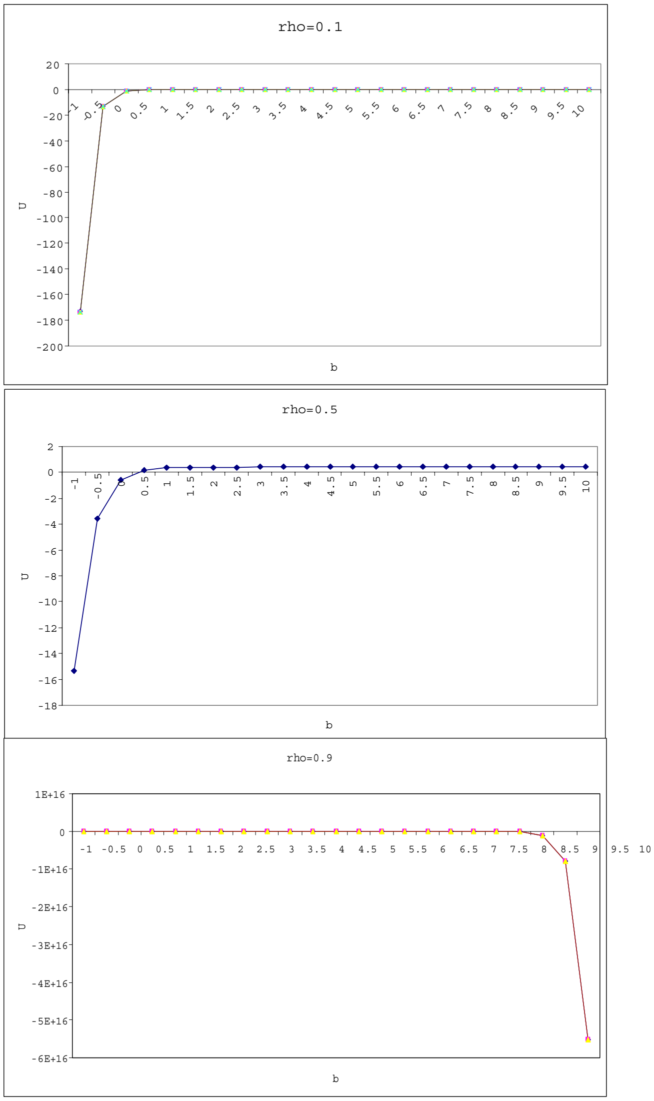

Fig.2 Behavior of the utility function under different traffic intensity

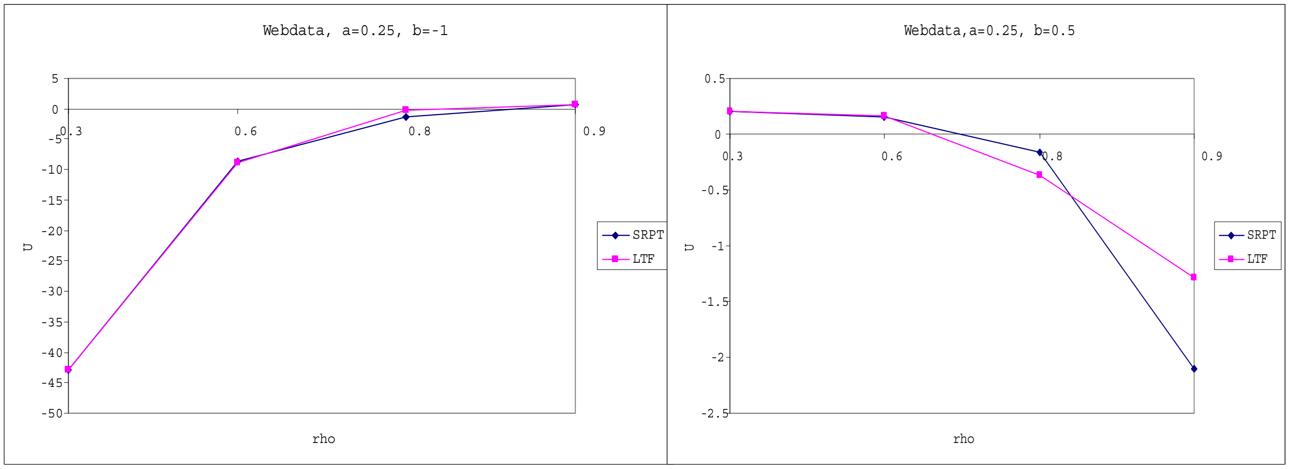

Fig. 3. Effects of policy for webdata