RC23782 (W0511-048) November 7, 2005
Mathematics

# IBM Research Report

## Dynamic Scheduling to Optimize Sojourn Time Moments and Tail Asymptotics 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**

# Dynamic Scheduling to Optimize Sojourn Time Moments and Tail Asymptotics 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 time is well known. However, higher-order statistical properties of customer sojourn times are also very important. We therefore consider alternative scheduling approaches in queueing systems with the goal of providing mean sojourn times close to those under SRPT while also providing better higher-order statistics. Our analysis includes deriving expressions for the mean, variance and tail asymptotics of the customer sojourn time distribution in these alternative queueing systems. This mathematical framework is then exploited to determine the control parameters of these alternative scheduling policies in order to optimize performance functions that combine the mean and higher-order statistics of sojourn times in a general and flexible manner. Our results show that one of the alternative scheduling policies provides the greatest flexibility with respect to optimizing these general sojourn time objective functions, and in particular can achieve the desired goal of mean sojourn times comparable to SRPT with superior higher-order statistics.

## 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 queueing systems; refer to [16, 44, 43, 45, 27, 52] 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 [43, 27, 52]. Some more recent studies have further argued that SRPT does not unfairly penalize large customers in order to benefit small customers, and hence these studies propose the use of SRPT to improve performance in Web sites (e.g., [14, 4, 20] and database systems (e.g., [33]).

On the other hand, as Schrage and Miller point out in their original study [44], 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. Normally, preemption also incurs 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 [18].

We therefore consider a corresponding multiclass (fixed) priority policy as an alternative to SRPT for scheduling the service of customers. Our objective is to alleviate some of the potential difficulties with SRPT while achieving mean sojourn times close to those obtained under SRPT. In fact, Schrage and Miller [44] consider scheduling policies related to multiclass (fixed) priority queues (with two classes) in order to alleviate some of the potential difficulties with SRPT while achieving mean sojourn times that attempt to approximate those under SRPT. This approach has the added advantage that the precise service time of each customer is not required. Instead, one only needs to be able to partition the workload into classes where the service times within each class are relatively similar and the service times across classes are relatively different. This partitioning of the workload into multiple classes also provides an additional form of control that can be used to determine the optimal multiclass priority policy and its parameters.

On the other hand, minimizing the mean sojourn time of customers is only one of several important scheduling objectives, and a priority policy that yields a small gain in first moment sojourn times can perform very poorly in terms of higher-order statistics [52], such as the second moment and tail distribution asymptotics. In particular, it often has been argued that a system with reasonable and predictable sojourn times may be more desirable than a system that is faster on average but exhibits high variability and/or large deviations from the mean [11, 28, 12, 46]. The original study of Schrage and Miller [44], however, does not consider any of these issues related to the higher-order statistics of customer sojourn times.

We therefore 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 higher-order statistical properties. Serving customers within each class according to a FCFS queueing discipline can minimize the sojourn time variance within each class (among disciplines that do not affect the per-class queue length distribution) [22], can maximize the sojourn time tail asymptotics within each class [48], and can reduce the preemptions among customers with fairly similar service times, whereas the priority discipline among the classes can yield a service ordering 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 improving (respectively, degrading) the mean sojourn time and degrading (respectively, improving) the higher-order sojourn time statistics, especially at heavy traffic intensities.

Specifically, we consider single-class SRPT M/G/1 queues and multiclass Fixed Priority (FP) M/G/1 queues, thus extending the comparison in [44] beyond the first moment and beyond two classes. The M/G/1 FP queue, however, is somewhat limited for our purposes to control the mean and higher-order statistics of customer sojourn times. Hence, we also consider multiclass M/G/1 priority queues under time-function scheduling (TFS) [17] in which the customers are scheduled according to general functions of the time they spend in the system. More precisely, the priority of each customer increases according to a monotonically nondecreasing per-class function of its time in system and the customer with the highest instantaneous priority value in the queue is selected for service at each scheduling epoch. The time-function parameters provide additional forms of control over higher-order sojourn time statistics. Our focus in this paper is on linear time-functions where the priority of each customer increases linearly with its time in system.

One of the main goals of this paper is to develop a theoretical foundation for alternative scheduling policies that provide mean sojourn times comparable to SRPT with superior higher-order statistical properties. We therefore derive expressions for the statistical properties of the per-class sojourn times in linear TFS (LTFS) M/G/1 queues, specifically providing results for the mean, variance and tail behavior of customer sojourn times. To the best of our knowledge, these are the first second moment sojourn time results for LTFS M/G/1 queues and the first large-deviations decay rate results for the sojourn time tail distribution in LTFS GI/G/1 queues to appear in the research literature. We also derive closed-form expressions and a linear algorithm to determine a set of LTFS control parameters that satisfy a given vector of sojourn times, which similarly provide the first such results to appear in the literature. The corresponding first two moments and large-deviations results for single-class SRPT and multiclass FP M/G/1 queues are presented and used for comparison, with some extensions for FP queues.

Our analysis demonstrates that the LTFS policy can satisfy the above goals by exploiting the results derived in this paper. In particular, by exploiting our results for performance measures and control parameters, we show that LTFS can provide mean sojourn times that are comparable to those obtained under SRPT while also providing superior second moment properties for customer sojourn time than those obtained under SRPT and FP. Our analysis also shows that LTFS provides a large-deviations decay rate of customer sojourn times that is superior (faster and better) to the corresponding large-deviations decay rate under SRPT and shares some of the characteristics of the maximal (fastest and best) decay rate. In fact, we establish an explicit ordering among the large-deviations decay rate under FCFS, LTFS, FP and SRPT scheduling policies.

Another main goal of this paper is to develop a theoretical foundation for the control of alternative scheduling policies to optimize general performance functions that combine the mean and higher-order statistics of customer sojourn times in a flexible manner to realize the goals of a broad spectrum of applications. Specifically, in determining the optimal scheduling policy, we formulate a utility optimization problem with utility defined as a function of the first two moments of customer sojourn times, 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 workloads to be used in the FP and LTFS M/G/1 queues. These results are obtained based in part on our derivations of the first two moments of the sojourn times in LTFS M/G/1 queues. It is important to note that this analysis is not restricted to the specific utility function considered herein and thus can be applied more generally.

Once again, our analysis demonstrates that the LTFS policy can satisfy the above goals by exploiting the results derived in this paper. Specifically, by exploiting our results for performance measures and control parameters, for a large range of parameters of the utility function, we can optimize the LTFS control parameters to yield superior utility values than under SRPT.

Our primary focus in this paper is on the theoretical foundations noted above. However, the theoretical results we derive and present throughout the paper within the context of M/G/1 (and, in some cases, GI/G/1) queues are also generally important in practice, even for a wide variety of non-M/G/1 (and non-GI/G/1) type environments. In particular, based on numerous simulations with traces from a large-scale production Web site, we find that our theoretical results are completely consistent with those obtained from these simulation experiments demonstrating that the trends observed from our theoretical results tend to hold more generally in practical systems. Moreover, a great deal of experience over the past few years where the LTFS policy has been deployed in a wide variety of production environments has demonstrated that the LTFS policy can be implemented efficiently with very low overhead while yielding performance characteristics that are completely consistent with our theoretical results presented in this paper [**?**].

Some aspects of our analysis are related to the so-called achievable region approach (e.g., refer to [12, 15]), and in fact LTFS is a general class of scheduling policies that correspond to interior points of the achievable polytopes. There are, however, important differences between the two paradigms. This includes the class of scheduling policies considered in the achievable region approach (e.g., [12, 2, 15]) versus those considered in our present study. Moreover, our mathematical framework directly considers higher-order statistical properties from first principles, whereas to the best of our knowledge the achievable region literature has only considered second moment properties as constraints; e.g., see [2].

The remainder of the paper is organized as follows. We first consider M/G/1 queues under the various scheduling policies of interest, summarizing known results and deriving new results of interest. §3 compares the statistical properties of these M/G/1 queues based on the results presented in §2, including our use of mean-variance objective functions. We then consider in §4 the optimal segmentation of the original SRPT single-class workload into the per-class service time distributions of the multiclass FP and LTFS workloads. §5 briefly examines our theoretical results based on simulations with large-scale production Web site traces. Concluding remarks are provided in §6.

## 2. MATHEMATICAL ANALYSIS

Consider the standard M/G/1 queue in which customers arrive according to an independent Poisson process $A(t)$ with finite rate $\lambda = 1/\mathbb{E}[A]$ 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. 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 (r.v.) for the overall customer sojourn time, $W$ the r.v. for the overall customer waiting time, and $R$ the r.v. for the overall 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 $\mathbb{E}[R^2] = \mathbb{E}[S^2]$. Let $\mathcal{M}_X(\theta) = \mathbb{E}[\mathrm{e}^{\theta X}]$ generally denote the moment generating function of a r.v. $X$, and let $f(n) \sim g(n)$ denote that $\lim_{n\to\infty} f(n)/g(n) = 1$ for any functions $f$ and $g$.

We also consider the standard multiclass M/G/1 queue in which class $k$ customers arrive according to an independent Poisson process $A_k(t)$ with finite rate $\lambda_k = 1/\mathbb{E}[A_k]$ and class $k$ customer service times are i.i.d. having 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 $\lambda = \sum_{k=1}^K \lambda_k$ and $\mu^{-1} = \sum_{k=1}^K \mu_k^{-1}(\lambda_k/\lambda)$. (Note in particular that this supports the so-called heavy-tailed property considered in [4].) Let $\rho_k = \lambda_k/\mu_k$ denote the traffic intensity for class $k$, and thus $\rho = \lambda/\mu = \sum_{k=1}^K \rho_k$. Customers within each class are served in a FCFS manner. Let $T_k$ denote the r.v. for the class $k$ sojourn time, $W_k$ the r.v. for the class $k$ waiting time, and $R_k$ the r.v. for the class $k$ residence time, where $T_k = W_k + R_k$. Note that when preemption is not allowed, then $R_k$ follows the service time distribution $F_k(\cdot)$, and thus $\mathbb{E}[R_k] = \mathbb{E}[S_k]$ and $\mathbb{E}[R_k^2] = \mathbb{E}[S_k^2]$. From the law of total probability, we then have

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

Our primary focus in this section is the first two moments and the tail behavior of the customer sojourn time distribution in M/G/1 priority queues (and GI/G/1 priority queues in the case of tail behavior results) under the SRPT, FP and LTFS scheduling policies. We shall assume throughout that $\rho < 1$.

## 2.1 Shortest Remaining Processing Time

The 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 [44], from which we obtain expressions for the first two moments of the customer sojourn times as follows

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

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

The large-deviations decay rate of the sojourn time distribution in GI/G/1 SRPT preemptive-resume queues has been recently obtained by Nuyens and Zwart [38], where the result depends upon whether or not the service time distribution has a mass at its right endpoint. Define $x_S \equiv \sup\{x : \mathbb{P}[S > x] > 0\}$. The corresponding results are summarized in the following two theorems.

THEOREM 1 (NUYENS,ZWART). *In GI/G/1 SRPT preemptive-resume queues with $\mathbb{P}[S = x_S] = 0$, we have as $x \to \infty$*

$$\log \mathbb{P}[T > x] \ \sim \ -\Lambda^* x, \qquad (6)$$

*where*

$$\begin{aligned}
\Lambda^* &= \sup_{\theta\geq 0}\{\theta - \Lambda(\theta)\}, && (7)\\
\Lambda(\theta) &= -\mathcal{M}_A^{-1}(1/\mathcal{M}_S(\theta)). && (8)
\end{aligned}$$

THEOREM 2 (NUYENS,ZWART). *In GI/G/1 SRPT preemptive-resume queues with $\mathbb{P}[S = x_S] > 0$, we have as $x \to \infty$*

$$\log \mathbb{P}[T > x] \ \sim \ -\Lambda^* x, \qquad (9)$$

*where*

$$\begin{aligned}
\Lambda^* &= \sup_{\theta\in[0,\hat{\Lambda}^*]}\{\theta - \Lambda(\theta)\}, && (10)\\
\Lambda(\theta) &= -\mathcal{M}_{\ddot{A}_1}^{-1}(1/\mathcal{M}_{\ddot{S}_1}(\theta)), && (11)\\
\hat{\Lambda}^* &= \sup\{\theta : \mathcal{M}_A(-\theta)\mathcal{M}_S(\theta) \leq 1\}, && (12)
\end{aligned}$$

*and (in this SRPT case) $\ddot{A}_1$ and $\ddot{S}_1$ are generic r.v.s for the inter-arrival and service times, respectively, of customers with service times strictly less than $x_S$.*

## 2.2 Fixed Priority

The FP scheduling policy, which gives priority to class $k$ customers over class $k'$ customers for all $1 \leq k < k' \leq K$, 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

$$\begin{aligned}
\mathbb{E}[T_k] &= \frac{\sum_{j=1}^k \lambda_j \mathbb{E}[S_j^2]}{2(1-\rho_{k-1}^+)(1-\rho_k^+)} + \frac{\mathbb{E}[S_k]}{1-\rho_{k-1}^+}, && (13)\\
\mathbb{E}[T_k^2] &= \frac{\sum_{j=1}^k \lambda_j \mathbb{E}[S_j^3]}{3(1-\rho_{k-1}^+)^2(1-\rho_k^+)} + \frac{\mathbb{E}[S_k^2]}{(1-\rho_{k-1}^+)^2} \ + \\
&\quad \left(\frac{\sum_{j=1}^{k-1} \lambda_j \mathbb{E}[S_j^2]}{(1-\rho_{k-1}^+)^2} + \frac{\sum_{j=1}^k \lambda_j \mathbb{E}[S_j^2]}{(1-\rho_{k-1}^+)(1-\rho_k^+)}\right)\mathbb{E}[T_k], && (14)
\end{aligned}$$

where $\rho_k^+ \equiv \sum_{j=1}^k \rho_j$. Variants of these results were first obtained by Miller [36], Takács [49] and Welch [51].

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

$$\begin{aligned}
\mathbb{E}[T_k] &= \frac{\sum_{j=1}^K \lambda_j \mathbb{E}[S_j^2]}{2(1-\rho_{k-1}^+)(1-\rho_k^+)} + \mathbb{E}[S_k], && (15)\\
\mathbb{E}[T_k^2] &= \frac{\sum_{j=1}^K \lambda_j \mathbb{E}[S_j^3]}{3(1-\rho_{k-1}^+)^2(1-\rho_k^+)} \ + \\
&\quad \frac{\left(\sum_{j=1}^k \lambda_j \mathbb{E}[S_j^2]\right)\left(\sum_{j=1}^K \lambda_j \mathbb{E}[S_j^2]\right)}{2(1-\rho_{k-1}^+)^2(1-\rho_k^+)^2} \ + \\
&\quad \frac{\left(\sum_{j=1}^{k-1} \lambda_j \mathbb{E}[S_j^2]\right)\left(\sum_{j=1}^K \lambda_j \mathbb{E}[S_j^2]\right)}{2(1-\rho_{k-1}^+)^3(1-\rho_k^+)} + \mathbb{E}[S_k]. && (16)
\end{aligned}$$

Variants of (15) were first given by Cobham [10], whereas variants of (16) were first obtained by Kesten and Runnenburg [21].

Turning to the asymptotic tail behavior, first observe that the large-deviations decay rate of the sojourn time distribution for class 1 customers in preemptive and nonpreemptive GI/G/1 FP queues is the same as the large-deviations decay rate of the sojourn time distribution in the corresponding single-class FCFS queue, which Stolyar and Ramanan have recently shown to be maximal among all work-conserving scheduling policies [48]. Further note that this asymptotic decay rate is not affected by whether the FP policy is preemptive or nonpreemptive, which is certainly not the case for the first two sojourn time moments as evident from the results above. This yields our desired tail asymptotic result for class 1.

THEOREM 3. *In preemptive and nonpreemptive GI/G/1 FP queues we have as* $x \to \infty$

$$\log \mathbb{P}[T_1 > x] \ \sim\ -\hat{\Lambda}^* x, \tag{17}$$

*where*

$$\hat{\Lambda}^* \ = \ \sup\{\theta : \mathcal{M}_A(-\theta)\mathcal{M}_S(\theta) \le 1\}. \tag{18}$$

The corresponding tail asymptotic result for customers of class 2 is presented in [1] for the M/G/1 queue and in [38] for the GI/G/1 queue.

THEOREM 4. *In preemptive and nonpreemptive GI/G/1 FP queues we have as* $x \to \infty$

$$\log \mathbb{P}[T_2 > x] \ \sim\ -\Lambda^* x, \tag{19}$$

*where*

$$\Lambda^* \ = \ \sup_{\theta \in [0,\hat{\Lambda}^*]} \{\theta - \Lambda(\theta)\}, \tag{20}$$

$$\Lambda(\theta) \ = \ -\mathcal{M}^{-1}_{A_1}(1/\mathcal{M}_{S_1}(\theta)), \tag{21}$$

*and* $\hat{\Lambda}^*$ *is as given in (18).*

Finally, we extend the above results to obtain the corresponding large-deviations decay rate for all customer classes and for the overall customer sojourn time distribution as follows.

THEOREM 5. *In preemptive and nonpreemptive GI/G/1 FP queues, we have as* $x \to \infty$

$$\log \mathbb{P}[T > x] \ \sim\ -\Lambda^* x, \tag{22}$$

$$\log \mathbb{P}[T_k > x] \ \sim\ -\Lambda^*_k x \tag{23}$$

*where*

$$\Lambda^* \ = \ \sum_{k=1}^{K} \Lambda^*_k \frac{\lambda_k}{\lambda}, \tag{24}$$

$$\Lambda^*_1 \ = \ \sup\{\theta : \mathcal{M}_{A_1}(-\theta)\mathcal{M}_{S_1}(\theta) \le 1\}, \tag{25}$$

$$\Lambda^*_k \ = \ \sup_{\theta \in [0,\hat{\Lambda}^*]} \{\theta - \Lambda_k(\theta)\}, \qquad k \ge 2, \tag{26}$$

$$\Lambda_k(\theta) \ = \ -\mathcal{M}^{-1}_{\hat{A}_k}(1/\mathcal{M}_{\hat{S}_k}(\theta)), \qquad k \ge 2, \tag{27}$$

*and* $\hat{A}_k$ *and* $\hat{S}_k$ *denote the aggregated arrival and service times for classes of* $1, 2, \cdots, k$.

PROOF. The results for classes 1 and 2 follow immediately from Thms. 3 and 4, respectively. To obtain the corresponding results for classes $k = 3, \ldots, K$, we consider the aggregation of classes $1, \ldots, k-1$ into a new (single) higher priority class with class $k$ as a new lower priority class. Since the large-deviations decay rate is not affected by whether the FP policy is preemptive or nonpreemptive, we will base our arguments on the preemptive case. In particular, the decay rate of the sojourn time tail distribution for class $k$ does not depend upon any lower class $j = k+1, \ldots, K$. Hence, upon applying Thm. 4 to the system with the aggregated higher priority class substituted for class 1 and the lower priority class $k$ substituted for class 2, we obtain (23),(26),(27) from (18) – (21). Eq. (24) then follows upon conditioning on class $k$. □

## 2.3 Linear Time-Function Scheduling

The corresponding sojourn time results for the LTFS 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 nonpreemptive or preemptive-resume fashion with ties broken in a FCFS manner, are much less well established in the research literature. Kleinrock [23, 24] derives expressions for the per-class mean sojourn time in nonpreemptive and preemptive-resume LTFS M/M/1 queues, and subsequently extends these first moment results for the corresponding nonpreemptive M/G/1 queue (although restriction to exponential service times is unnecessarily included in the derivation) [27]. However, to the best of our knowledge, there are no second moment sojourn time results for nonpreemptive and preemptive-resume LTFS M/G/1 queues in the research literature, as well as no previous first moment sojourn time results for preemptive-resume LTFS M/G/1 queues. Moreover, to the best of our knowledge, there are no results in the literature on the tail behavior of the sojourn time distribution in nonpreemptive and preemptive-resume LTFS GI/G/1 queues. The results of our present study fill these major gaps in the research literature.

### 2.3.1 Moments of Sojourn Times

Our goal is to derive expressions for $\mathbb{E}[T]$ and $\mathbb{E}[T^2]$ under LTFS. Due to space restrictions, we cannot present our results for both the preemptive and nonpreemptive LTFS queues. Hence, in this section we focus on our results for nonpreemptive LTFS M/G/1 (and, in some cases, GI/G/1) queues because of the arguments in the introduction and those in [44] that tend to favor nonpreemptive over preemptive-resume disciplines, which includes the additional overheads of the latter. 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 and Little's Law [52, 3]. Assume throughout that $b_1 \ge b_2 \ge \ldots \ge b_K \ge 0$.

Consider an arbitrary arrival at some time $t$ of a so-called tagged customer of class $k$ in a nonpreemptive LTFS M/G/1 queue. Let $N_{jk}$ be the r.v. denoting the number of class $j$ customers in the system at time $t$ that receive service before the tagged class $k$ customer, $M_{jk}$ the r.v. denoting the number of class $j$ customers that arrive after time $t$ and receive service before the tagged class $k$ customer, and $W_0$ the r.v. 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

$$W_k = W_0 + \sum_{j=1}^{K} \Bigg( \sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji} \Bigg), \tag{28}$$

for $k = 1, \ldots, K$, where $X_{j1}, X_{j2}, \ldots$ and $Y_{j1}, Y_{j2}, \ldots$ are sequences of i.i.d. r.v.s such that $X_{ji} \overset{\mathrm{d}}{=} Y_{ji} \overset{\mathrm{d}}{=} S_j$ for all $j = 1, \ldots, K$.

Our solutions for $\mathbb{E}[T_k]$ and $\mathbb{E}[T_k^2]$ are based on the derivation of solutions for the probability measures involved in the above equation. The solutions for these probability measures are presented in a sequence of results, theorems and lemmas that follow. Then, at the end of the section, we present our main results in two theorems based on this sequence of results derived below.

Kleinrock [23, 25] first established the following M/G/1 conservation law that will be useful for our purposes below.

THEOREM 6 (KLEINROCK). *For a nonpreemptive M/G/1 queue under any work-conserving scheduling policy, the mean per-class waiting times* $\mathbb{E}[W_k]$ *must satisfy*

$$\sum_{k=1}^{K} \rho_k \mathbb{E}[W_k] \ = \ \frac{\rho \mathbb{E}[W_0]}{1-\rho}. \tag{29}$$

From (28), we obtain the first two moments of the class $k$ sojourn times (recalling that $R_k = S_k$ in the nonpreemptive case)

$$\begin{aligned}\mathbb{E}[T_k] &= \mathbb{E}[W_k] + \mathbb{E}[S_k], &(30)\\ \mathbb{E}[T_k^2] &= \mathbb{E}[W_k^2] + 2\mathbb{E}[W_k]\mathbb{E}[S_k] + \mathbb{E}[S_k^2], &(31)\end{aligned}$$

in terms of

$$\begin{aligned}\mathbb{E}[W_k] &= \mathbb{E}[W_0] + \mathbb{E}\Big[\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big], &(32)\\ \mathbb{E}[W_k^2] &= \mathbb{E}[W_0^2] + 2\mathbb{E}\Big[W_0\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big]\\ &+ \mathbb{E}\Big[\Big\{\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big\}^2\Big]. &(33)\end{aligned}$$

Under the assumptions of a nonpreemptive 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 (e.g., see [26])

$$\mathbb{E}[W_0] = \sum_{k=1}^{K}\rho_k\frac{\mathbb{E}[S_k^2]}{2\mathbb{E}[S_k]}, \qquad (34)$$

$$\mathbb{E}[W_0^2] = \sum_{k=1}^{K}\rho_k\frac{\mathbb{E}[S_k^3]}{3\mathbb{E}[S_k]}. \qquad (35)$$

Upon multiplying Eq. (28) for any pair $k, k'$ and taking expectations, we obtain

$$\begin{aligned}\mathbb{E}[W_kW_{k'}] &= \mathbb{E}[W_0^2] + \mathbb{E}\Big[\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big]\\ &\mathbb{E}\Big[\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk'}} X_{ji} + \sum_{i=1}^{M_{jk'}} Y_{ji}\Big)\Big] +\\ &\mathbb{E}\Big[W_0\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big] +\\ &\mathbb{E}\Big[W_0\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk'}} X_{ji} + \sum_{i=1}^{M_{jk'}} Y_{ji}\Big)\Big]. &(36)\end{aligned}$$

Similarly, multiplying Eq. (28) for each $k$ by $W_0$ and taking expectations yields

$$\mathbb{E}[W_0W_k] = \mathbb{E}[W_0^2] + \mathbb{E}\Big[W_0\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big]. \qquad (37)$$

To complete our solution for $\mathbb{E}[T_k^2]$, we need to determine the measures $\mathbb{E}[W_k^2]$, $\mathbb{E}[W_kW_{k'}]$ and $\mathbb{E}[W_0W_k]$ provided in (33), (36) and (37), respectively, which yields a total of $K + \left(\begin{array}{c}K\\2\end{array}\right) + K = K(K+3)/2$ unknowns.

The next two lemmas provide our results for the first two moments of $N_{jk}$ and $M_{jk}$.

LEMMA 7. *For the r.v. $N_{jk}$, we have*

$$\begin{aligned}\mathbb{E}[N_{jk}] &= \lambda_j\mathbb{E}[W_j], \;\; \forall j \le k, &(38)\\ \mathbb{E}[N_{jk}^2] &= (\lambda_j^2/2)\mathbb{E}[W_j^2] + \mathbb{E}[N_{jk}], \;\; \forall j \le k, &(39)\\ \mathbb{E}[N_{jk}] &= \lambda_j\mathbb{E}[W_j]\frac{b_j}{b_k}, \;\; \forall j > k, &(40)\\ \mathbb{E}[N_{jk}^2] &= \Big(\lambda_j\mathbb{E}[W_j]\frac{b_j}{b_k}\Big)^2 + \lambda_j\mathbb{E}[W_j]\frac{b_j}{b_k}, \;\; \forall j > k, &(41)\end{aligned}$$

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

PROOF. Observe that all class $j \le k$ 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 $j \le k$ customers who arrived before time $t$. Eqs. (38) and (39) then follow from the PASTA property and distributional versions of Little's Law [30, 19, 7, 6], as the required properties for each customer class hold under our model assumptions. Now, based on the definition of $N_{jk}$, 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 are satisfied by a class $j > k$ customer provided that $t - t' < W_j \le t - t' + W_k$. The upper limit ensures that the priority of the class $j > k$ 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., $b_kW_k$). From $b_kW_k = b_j(t - t' + W_k)$ we obtain the relationship $t - t' + W_k = b_k/(b_k - b_j)(t - t')$. Since the arrivals of such class $j > k$ customers follow a non-homogeneous Poisson process, and the time dependent arrival moments can be expressed as functionals of $W_j$, we have

$$\begin{aligned}\mathbb{E}[N_{jk}] &= \int_0^{\infty}\lambda_j\mathbb{P}[y < W_j \le \frac{b_k}{b_k - b_j}y]dy,\\ &= \lambda_j\mathbb{E}[W_j] - \lambda_j\frac{b_k - b_j}{b_k}\mathbb{E}[W_j],\end{aligned}$$

which yields (40). Eq. (41) then follows because $N_{jk}$ can be treated as a Poisson r.v. □

LEMMA 8. *For the r.v. $M_{jk}$, we have*

$$\begin{aligned}\mathbb{E}[M_{jk}] &= \lambda_j[1 - (b_k/b_j)]\mathbb{E}[W_k], \;\; \forall j < k, &(42)\\ \mathbb{E}[M_{jk}^2] &= (\lambda_j^2/2)\mathbb{E}[W_k^2][1 - 2(b_k/b_j) + (b_k/b_j)^2]\\ &\quad + \mathbb{E}[M_{jk}], \;\; \forall j < k, &(43)\\ M_{jk} &= 0, \;\; \text{with probability 1}, \;\; \forall j \ge k, &(44)\end{aligned}$$

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

PROOF. Eq. (44) follows directly from the fact that no customer with an equal or smaller slope $b_j$ and arrival time after $t$ can overtake the tagged class $k$ customer for all $j \ge k$. When $j < k$, since the priority of the tagged class $k$ customer when it starts service is given by $b_kW_k$, then $M_{jk}$ is the number of class $j < k$ customers arriving in the time interval $(t, t + Z_j)$ such that $b_kW_k = b_j(W_k - Z_j)$, or equivalently $Z_j = [1 - (b_k/b_j)]W_k$. From Little's Law [30, 52, 3] we have $\mathbb{E}[M_{jk}] = \lambda_j\mathbb{E}[Z_j]$ which yields (42). Similarly, it follows from distributional versions of Little's Law [19, 7, 6] that $\mathbb{E}[M_{jk}^2] = (\lambda_j^2/2)\mathbb{E}[Z_j^2] + \mathbb{E}[M_{jk}]$, since the required properties for each customer class hold under our model assumptions. Finally, from $Z_j = [1 - (b_k/b_j)]W_k$ we obtain Eq. (43). □

Since $N_{jk}$ and $M_{jk}$ are stopping times for the i.i.d. sequences of r.v.s $X_{j1}, X_{j2}, \ldots$ and $Y_{j1}, Y_{j2}, \ldots$, respectively, both with finite first two moments, we have the following results from Wald's equation [9].

LEMMA 9.

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=1}^{N_{jk}} X_{ji}\Big] &= \mathbb{E}[N_{jk}]\mathbb{E}[S_j], &(45)\\ \mathbb{E}\Big[\sum_{i=1}^{M_{jk}} Y_{ji}\Big] &= \mathbb{E}[M_{jk}]\mathbb{E}[S_j], &(46)\\ \mathbb{E}\Big[\Big(\sum_{i=1}^{N_{jk}} X_{ji}\Big)^2\Big] &= \mathbb{E}[N_{jk}]\mathbb{E}[S_j^2] + \mathbb{E}[S_j]^2(\mathbb{E}[N_{jk}^2] - \mathbb{E}[N_{jk}]), &(47)\\ \mathbb{E}\Big[\Big(\sum_{i=1}^{M_{jk}} Y_{ji}\Big)^2\Big] &= \mathbb{E}[M_{jk}]\mathbb{E}[S_j^2] + \mathbb{E}[S_j]^2(\mathbb{E}[M_{jk}^2] - \mathbb{E}[M_{jk}]). &(48)\end{aligned}$$

PROOF. Eqs. (45) and (46) follow immediately from Wald's equation and the properties of the r.v.s involved. Similarly, (47) and (48) follow immediately from the second moment version of Wald's equation and the r.v. properties. □

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

$$\begin{aligned} W_k^2 &= \mathbb{E}[W_k^2|W_k] = \mathbb{E}[W_0^2|W_k] \\ &\quad + 2\mathbb{E}\Big[W_0 \sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big|W_k\Big] \\ &\quad + \mathbb{E}\Big[\Big\{\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big\}^2\Big|W_k\Big] \qquad (49) \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}\Big[\Big\{\sum_{j=1}^{K}\Big(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big\}^2\Big|W_k\Big] \\ = &\sum_{j,\ell=1, j\neq\ell}^{K}\Big\{\mathbb{E}\Big[\sum_{i=1}^{N_{jk}} X_{ji}\sum_{i=1}^{N_{\ell k}} X_{\ell i}\Big|W_k\Big] \\ &+ \mathbb{E}\Big[\sum_{i=1}^{M_{jk}} Y_{ji}\Big|W_k\Big]\mathbb{E}\Big[\sum_{i=1}^{M_{\ell k}} Y_{\ell i}\Big|W_k\Big] \\ &+ \mathbb{E}\Big[\sum_{i=1}^{N_{jk}} X_{ji}\Big|W_k\Big]\mathbb{E}\Big[\sum_{i=1}^{M_{\ell k}} Y_{\ell i}\Big|W_k\Big]\Big\} \\ &+ \sum_{j=1}^{K}\Big\{\mathbb{E}\Big[\Big(\sum_{i=1}^{N_{jk}} X_{ji}\Big)^2\Big|W_k\Big] + \mathbb{E}\Big[\Big(\sum_{i=1}^{M_{jk}} Y_{ji}\Big)^2\Big|W_k\Big] \\ &+ \mathbb{E}\Big[\Big(\sum_{i=1}^{N_{jk}} X_{ji}\Big)\Big(\sum_{i=1}^{M_{jk}} Y_{ji}\Big)\Big|W_k\Big]\Big\}. \qquad (50) \end{aligned}$$

Since $N_{jk}$ can be treated as a Poisson r.v. (c.f. proof of Lem. 7), we obtain for the product of two related compound Poisson r.v.s

$$\begin{aligned} &\mathbb{E}\Big[\sum_{i=1}^{N_{jk}} X_{ji}\sum_{i=1}^{N_{\ell k}} X_{\ell i}\Big|W_k\Big] \\ = &\ \mathbb{E}\Big[\sum_{i=1}^{N_{j\ell k}^{(0)}+N_{j\ell k}^{(j)}} X_{ji} \sum_{i=1}^{N_{j\ell k}^{(0)}+N_{j\ell k}^{(\ell)}} X_{\ell i}\Big|W_k\Big] \\ = &\ \mathbb{E}\Big[\mathbb{E}\Big[\sum_{i=1}^{N_{j\ell k}^{(0)}+N_{j\ell k}^{(j)}} X_{ji} \sum_{i=1}^{N_{j\ell k}^{(0)}+N_{j\ell k}^{(\ell)}} X_{\ell i}\Big|W_k, N_{j\ell k}^{(0)}\Big]\Big] \\ = &\ \mathbb{E}[S_j]\mathbb{E}[S_\ell]\mathbb{E}\Big[\mathbb{E}[(N_{j\ell k}^{(0)})^2|W_k] + \mathbb{E}[N_{j\ell k}^{(0)}|W_k]\mathbb{E}[N_{j\ell k}^{(j)}|W_k] \\ &+\mathbb{E}[N_{j\ell k}^{(0)}|W_k]\mathbb{E}[N_{j\ell k}^{(\ell)}|W_k] + \mathbb{E}[N_{j\ell k}^{(j)}|W_k]\mathbb{E}[N_{j\ell k}^{(\ell)}|W_k]\Big], \qquad (51) \end{aligned}$$

where $N_{j\ell k}^{(0)}$ and $N_{j\ell k}^{(j)}, N_{j\ell k}^{(\ell)}, N_{j\ell k}^{(k)}$ are the common and remaining components of the correlated Poisson r.v., respectively. Observe that the variance of $N_{j\ell k}^{(0)}$ 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 $\mathbb{E}[W_j W_\ell|W_k]$.

As a result of conditional independence, we have

$$\mathbb{E}[W_0 \sum_{i=1}^{M_{jk}} X_{ji}|W_k] \quad = \quad \rho_j[1 - b_k/b_j]\mathbb{E}[W_0]\mathbb{E}[W_k]. \qquad (52)$$

The continuation of our calculations yields, for $j > k$,

$$\begin{aligned} \mathbb{E}[W_0 \sum_{i=1}^{N_{jk}} X_{ji}|W_k] &= \mathbb{E}[\mathbb{E}[W_0 \sum_{i=1}^{N_{jk}} X_{ji}|W_k, W_0]|W_k] \\ &= \frac{\rho_j b_j}{b_k}\mathbb{E}[W_0[\mathbb{E}[W_j|W_0, W_k]|W_k] \\ &= \frac{\rho_j b_j}{b_k}\mathbb{E}[W_0 W_j|W_k], \qquad (53) \end{aligned}$$

and very similar calculations for $j \leq k$ yields

$$\mathbb{E}[W_0 \sum_{i=1}^{N_{jk}} X_{ji}|W_k] \quad = \quad \rho_j\mathbb{E}[W_0 W_j|W_k]. \qquad (54)$$

Moreover, we have for the product of two compound Poisson r.v.s

$$\begin{aligned} &\mathbb{E}\Big[\sum_{i=1}^{N_{jk}} X_{ji} \sum_{i=1}^{N_{\ell k'}} X_{\ell i}\Big] = \mathbb{E}\Big[\Big[\sum_{i=1}^{N_{jk}} X_{ji} \sum_{i=1}^{N_{\ell k'}} X_{\ell i}\Big]\Big|W_k, W_{k'}\Big] \\ &= \mathbb{E}[S_j]\mathbb{E}[S_\ell]\mathbb{E}[\mathbb{E}[N_{jk}N_{jk'}|W_k, W_{k'}]]. \qquad (55) \end{aligned}$$

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

Our main results on sojourn time moments under LTFS based on the foregoing derivations are then summarized in the next two theorems.

THEOREM 10. *In a nonpreemptive LTFS M/G/1 queue, the overall mean customer sojourn time can be expressed as*

$$\begin{aligned} \mathbb{E}[T] &= \sum_{k=1}^{K}(\mathbb{E}[W_k] + \mathbb{E}[S_k])(\lambda_k/\lambda), \qquad (56) \\ \mathbb{E}[W_k] &= \frac{(\mathbb{E}[W_0]/(1-\rho)) - \sum_{i=k+1}^{K}\rho_i\mathbb{E}[W_i](1 - b_i/b_k)}{1 - \sum_{i=1}^{k-1}\rho_i(1 - b_k/b_i)}, \qquad (57) \end{aligned}$$

PROOF. Upon applying (45),(46) from Lem. 9 to (32), substituting Eqs. (38),(40) from Lem. 7 and Eqs. (42),(44) from Lem. 8 into the resulting expression, and simplifying 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 (57) then follows from the conservation law given in (29) and straightforward algebra. Eq. (56) directly follows from the law of total probability, the definition of $T_k$, and the fact that $\mathbb{E}[R_k] = \mathbb{E}[S_k]$ in nonpreemptive queues. □

THEOREM 11. *The second moment of customer sojourn times in a nonpreemptive LTFS M/G/1 queue is given by*

$$\begin{aligned} \mathbb{E}[T^2] &= \sum_{k=1}^{K}\mathbb{E}[T_k^2](\lambda_k/\lambda), \qquad (58) \\ \mathbb{E}[T_k^2] &= \mathbb{E}[W_k^2] + 2\mathbb{E}[W_k]\mathbb{E}[S_k] + \mathbb{E}[S_k^2], \qquad (59) \\ \mathbb{E}[W_k^2] &= \mathbb{E}[W_0^2] + 2\mathbb{E}[W_0\sum_{j=1}^{K}(\sum_{i=1}^{N_{jk}} X_{ji}\sum_{i=1}^{M_{jk}} Y_{ji})] \\ &\quad + \mathbb{E}[\{\sum_{j=1}^{K}(\sum_{i=1}^{N_{jk}} X_{ji} + \sum_{i=1}^{M_{jk}} Y_{ji})\}^2], \qquad (60) \end{aligned}$$

*which can be efficiently computed from the expressions (29) – (57).*

PROOF. Eq. (58) directly follows from the law of total probability, whereas the definition of $T_k$, the fact that $\mathbb{E}[R_k] = \mathbb{E}[S_k]$ in nonpreemptive queues, and direct calculations yield (59). Upon squaring and taking expections of both sides of (28) and simplifying we obtain (60). To solve for $\mathbb{E}[T_k^2]$ in (59) and substitute the result in (58), 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 (60), (36) and (37), respectively. This yields a system of $K + \left( \begin{array}{c} K \\ 2 \end{array} \right) + K = K(K+3)/2$ equations of unknowns in terms of (29) – (57). It follows from the derivations above that all of the measures in this system of equations can be expressed as linear functions of the unknowns and that this system of linear equations has a coefficient matrix which is triangular and nonsingular. These properties then can be exploited to efficiently compute $\mathbb{E}[T_k^2]$ which in turn yield $\mathbb{E}[T^2]$ via (58). □

### 2.3.2 Tail Behavior of Sojourn Times

We now turn to consider the large-deviations decay rate of the sojourn time distribution in GI/G/1 LTFS queues. Let us first introduce some notation and preliminary results used in our analysis.

Recall that $A_k(t)$ denotes the arrival process for customers of class $k$. Denote by $A_{k1}, A_{k2}, \ldots$ and $S_{k1}, S_{k2}, \ldots$ the sequences of interarrival time and service time r.v.s, respectively, such that $A_{ki} \overset{\mathrm{d}}{=} A_k$ and $S_{ki} \overset{\mathrm{d}}{=} S_k$ for all $k = 1, \ldots, K$. Define $J_k(t) \equiv \sum_{i=1}^{A_k(t)} S_{ki}$ to be the amount of class $k$ work that has arrived to the system by time $t$, and $W$ the total amount of work in system observed by a customer arrival. Let $P_k^{\mathrm{V}}$ be a generic busy period of class $k$ customers and $P_k^{\mathrm{V}}(x)$ a busy period of class $k$ customers with an initial customer of size $x$, both in a GI/G/1 queue under policy $V$. We then have $P_k^{\mathrm{FP}}(x) \overset{\mathrm{d}}{=} \inf\{t \geq 0 : x + J_k(t) \leq t\}$. Let $W_k^{\mathrm{V}}$ denote the stationary waiting time experienced by class k customers in a GI/G/1 queue under policy $V$. Then the following result due to Nuyens and Zwart [38] will be convenient for our purposes below.

THEOREM 12 (NUYENS,ZWART). *As $x \to \infty$, $\log \mathbb{P}[W_2^{FP} > x] \sim -\Lambda_2^* x$ where $\Lambda_2^*$, $\Lambda(\theta)$, $\hat{\Lambda}^*$ are as given in (20), (21), (18), respectively.*

Our main result on the tail asymptotics of the sojourn time distribution now can be presented in the following theorem, which can be easily extended to handle preemptive GI/G/1 LTFS queues.

THEOREM 13. *For customers of class $k$ in nonpreemptive GI/G/1 LTFS queues, we have as $x \to \infty$*

$$\begin{aligned} \log \mathbb{P}[T > x] &\sim -\Lambda^* x, && (61) \\ \log \mathbb{P}[T_k > x] &\sim -\Lambda_k^* x, && (62) \end{aligned}$$

*where*

$$\begin{aligned} \Lambda^* &= \sum_{k=1}^{K} \Lambda_k^* \frac{\lambda_k}{\lambda}, && (63) \\ \Lambda_k^* &= \sup_{\theta \in [0, \hat{\Lambda}^*]} \{\theta - \Lambda_k(\theta)\}, && (64) \\ \Lambda_k(\theta) &= -\mathcal{M}_{\tilde{A}_k}^{-1} / (1/\mathcal{M}_{\tilde{S}_k}(\theta)), && (65) \end{aligned}$$

*and $\tilde{A}_k$ and $\tilde{S}_k$ denote the aggregation of arrival and service times for all classes of customers except those of class $k$.*

PROOF. Let us initially consider the two class case. Obviously, the busy period of class 1 customers in a FP GI/G/1 queue is longer than that in a LTFS GI/G/1 queue. Hence, under the LTFS policy,

$$\mathbb{P}[T_2 > x] \leq \mathbb{P}[W_2^{\mathrm{FP}} > x],$$

and from Thm. 12 we know that

$$\log \mathbb{P}[W_2^{\mathrm{FP}} > x] \sim -\Lambda_2^* x,$$

which provides an upper bound on the large-deviations decay rate.

Turning to the lower bound, we have for any fixed $a > 0$

$$\begin{aligned} \mathbb{P}[T_2^{\mathrm{FP}} > x] &\geq \mathbb{P}[W > ax]\mathbb{P}[P_1^{\mathrm{LTFS}}(ax) > x], \\ &= e^{-a\hat{\Lambda}^* x + o(x)}\mathbb{P}[P_1^{\mathrm{LTFS}}(ax) > x]. && (66) \end{aligned}$$

Now we proceed to estimate $\mathbb{P}[P_1^{\mathrm{LTFS}}(ax) > x]$ by using a fairly typical change-of-measure argument; e.g., refer to [3, 38]. Define

$$\begin{aligned} \mathbb{P}_\theta[A_{1i} \in dx] &\equiv e^{-\Lambda_2(\theta)x}\mathbb{P}[A_{i1} \in dx]/\mathcal{M}_{\tilde{A}_2}(-\Lambda_2(\theta)), \\ \mathbb{P}_\theta[S_{1i} \in dx] &\equiv e^{\theta x}\mathbb{P}[S_{i1} \in dx]/\mathcal{M}_{\tilde{S}_2}(\theta), \end{aligned}$$

where in the 2-class case $\tilde{A}_2$ and $\tilde{S}_2$ translate to $A_1$ and $S_1$, respectively, and $\Lambda_2(\theta)$ is as given in (65). For any $\epsilon < a$, we choose $\theta$ such that

$$\Lambda_2'(\theta) = 1 - a + \epsilon.$$

Define $Z_n^{A_1} \equiv A_{11} + \ldots + A_{1n}$ and $Z_n^{S_1} = S_{11} + \ldots + S_{1n}$, from which we know that

$$H_n^\epsilon = \exp(\Lambda_2(\theta)Z_n^{A_1} - \theta Z_n^{S_1})$$

is a martingale, and so is $1/H_n^\epsilon$. We apply option stopping to obtain

$$\begin{aligned} &\mathbb{P}[P_1^{\mathrm{LTFS}}(ax) > x] = \mathbb{E}_\theta[H_{\tau_1(x)}^\epsilon \mathbf{1}\{P_1^{\mathrm{LTFS}}(ax) > x\}], \\ &\mathbb{P}[P_1^{\mathrm{LTFS}}(ax) > x] \geq \mathbb{E}_\theta[H_{\tau_1(x)}^\epsilon \mathbf{1}\{P_1^{\mathrm{LTFS}}(ax) > x\}\mathbf{1}\{Z_{A_1(x)}^{S_1} \leq (1 - a + \epsilon)x\}], \end{aligned}$$

where $\tau_1(x)$ is a stopping time for the Borel $\sigma$-algebra generated by $A_{11}, \ldots, A_{1n}, S_{11}, \ldots, S_{1n}$. Since $\mathbb{P}_\theta[P_1^{\mathrm{LTFS}}(ax) > x,\ Z_{A_1(x)}^{S_1} \leq (1 - a + \epsilon)x]$ is bounded away from zero uniformly in $x$ for every $\epsilon > 0$ by the law of large numbers, we have

$$\liminf_{x \to \infty} x^{-1} \log \mathbb{P}[P_1^{\mathrm{LTFS}}(ax) > x] \geq -\theta(1 - a + \epsilon) + \Lambda_2(\theta),$$

which upon taking the limit as $\epsilon \longrightarrow 0$ yields

$$\log \mathbb{P}[P_1^{\mathrm{LTFS}}(ax) > x] \geq -x(\hat{\Lambda}^*(1 - a) - \Lambda_2(\hat{\Lambda}^*)) + o(x).$$

This in combination with (66) provides the desired lower bound, which together with the upper bound establishes the class 2 result.

It can be easily seen that the same arguments go through to establish the result for class 1, and then these arguments can be readily generalized to establish the result for a general number of classes $K$. Finally, Eq. (63) follows upon conditioning on class $k$. □

### 2.3.3 Setting of Control Parameters

The set of slopes $\{b_1, \ldots, b_K\}$ represent one set of control parameters available in nonpreemptive LTFS 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 and a linear algorithm to determine a set of control parameters $\{b_1, \ldots, b_K\}$ that satisfy a given objective 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 Eq. (57) 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 a much more expensive iterative scheme to obtain a solution for systems with more than two classes [29].

The mean waiting time for customers of class $k$ in nonpreemptive LTFS M/G/1 queues as a function of the per-class control parameters is given by (57). Observe the very simple dependence that $\mathbb{E}[W_k]$ has on the control parameters, namely the slopes $b_k$ only appear as ratios. 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 $(\mathbb{E}[W_1^*], \ldots, \mathbb{E}[W_K^*])$ can be readily verified as part of our recursive algorithm by ensuring the corresponding variables satisfy the obvious constraints.

Following [37, 47], we define

$$\alpha_k \equiv \sum_{i=1}^{K} \frac{\lambda_i \mathbb{E}[S_i^2]}{2(1-\rho)} - \sum_{i=k+1}^{K} \rho_i \mathbb{E}[W_i], \tag{67}$$

$$\beta_k \equiv 1 - \sum_{i=1}^{k-1} \rho_i, \tag{68}$$

$$C_k \equiv \sum_{i=k+1}^{K} \rho_i b_i \mathbb{E}[W_i], \tag{69}$$

$$D_k \equiv \sum_{i=1}^{k-1} \frac{\rho_i}{b_i}. \tag{70}$$

Our main results on closed-form expressions for LTFS control parameters are then summarized in the subsequent theorem.

THEOREM 14. *Consider a nonpreemptive LTFS M/G/1 queue and a feasible objective performance vector* $(\mathbb{E}[W_1^*] = \mathbb{E}[T_1^*] - \mathbb{E}[S_1], \ldots, \mathbb{E}[W_K^*] = \mathbb{E}[T_K^*] - \mathbb{E}[S_K])$. *Without loss of generality, suppose that* $b_1 \geq b_2 \geq \ldots \geq b_K = 1$. *Then any feasible objective performance vector must satisfy*

$$\mathbb{E}[W_0] \leq \mathbb{E}[W_1^*] \leq \mathbb{E}[W_2^*] \leq \ldots \leq \mathbb{E}[W_K^*], \tag{71}$$

*and a set of control parameters* $\{b_1, \ldots, b_K\}$ *that achieve the feasible objective vector* $(\mathbb{E}[W_1^*], \ldots, \mathbb{E}[W_K^*])$ *is given by*

$$b_k = \frac{-(\mathbb{E}[W_k]\beta_{k+1} - \alpha_k) + \sqrt{(\mathbb{E}[W_k]\beta_{k+1} - \alpha_k)^2 + 4C_k \mathbb{E}[W_k] D_{k+1}}}{2\mathbb{E}[W_k] D_{k+1}}, \tag{72}$$

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

PROOF. Eq. (71) follows directly from (57), the theorem supposition and straightforward calculations. Upon substituting the definitions (67) – (70) and the Eq. (34) into (57), we obtain

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

Substituting the relationships $\beta_{k+1} = \beta_k - \rho_k$, $D_k = D_{k+1} - \rho_k/b_k$ from (68), (70) and simplifying then yields Eq. (72) since $b_1 \geq b_2 \geq \ldots \geq b_K = 1$. □

The above theorem also forms the basis of our linear algorithm to obtain a set of control parameters $b_{K-1}, \ldots, b_1$ 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 nonpreemptive LTFS M/G/1 queue. In particular, observe that the value of $b_k$ in Eq. (72) depends only on the values of $b_1, \ldots, b_{k-1}$. We then have the following algorithm to compute in linear time the control parameters $b_{K-1}, \ldots, b_1$ to achieve a given objective vector of sojourn times $(\mathbb{E}[T_1^*], \ldots, \mathbb{E}[T_K^*])$.

**Initialization:** $\mathbb{E}[W_k^*] = \mathbb{E}[T_k^*] - \mathbb{E}[S_k]$, for all $k = 1, \ldots, K$; $C_K = 0$; $b_K = 1$; $D_K = (\mathbb{E}[W_0]/(1-\rho) - \beta_K \mathbb{E}[W_K^*])/(\mathbb{E}[W_K^*])$.

**Linear Recursion:** The corresponding variables for classes $k = K-1, K-2, \ldots, 2, 1$ are computed consecutively as follows:

$$C_k = C_{k+1} + \rho_{k+1} b_{k+1} \mathbb{E}[W_{k+1}^*]; \tag{73}$$

$$b_k = \text{RHS of Eq. (72) with } \mathbb{E}[W_k] = \mathbb{E}[W_k^*]; \tag{74}$$

$$D_k = \frac{\alpha_k + C_k/b_k - \beta_k \mathbb{E}[W_k^*]}{b_k \mathbb{E}[W_k^*]}. \tag{75}$$

# 3. SCHEDULING POLICY COMPARISON

Let us now consider the properties of customer sojourn time statistics in SRPT, FP and LTFS M/G/1 queues based on the results derived in the previous section. We first directly focus on the higher-order statistical properties of the different M/G/1 queues and then we turn to general functions that combine the various statistics of customer sojourn times in a general risk-based manner.

## 3.1 Statistical Properties

As originally suggested by Schrage and Miller [44] and extended herein to also consider higher-order statistical properties and the LTFS policy, one can attempt to closely approximate the mean sojourn times in M/G/1 SRPT preemptive-resume queues with an appropriately chosen multiclass M/G/1 priority queue. In fact, working in the opposite direction, Phipps [40] took a somewhat related approach to obtain an exact expression for the mean sojourn time under nonpreemptive SRPT (also known as shortest job first) by extending the analysis of Cobham [10] to consider an equivalent nonpreemptive M/G/1 FP queue with an infinite number of classes, indexed by the customer service time. In this section we present a direct comparison of the statistical properties of customer sojourn times under SRPT, FP and LTFS based on our analysis in §2.

Our goal is to compare $\mathbb{E}[T_{\text{SRPT}}^2]$ and $\Lambda_{\text{SRPT}}^*$ with the corresponding performance measures obtained under instances of FP and LTFS where $\mathbb{E}[T_{\text{FP}}] = \mathbb{E}[T_{\text{SRPT}}] + \epsilon$ and $\mathbb{E}[T_{\text{LTFS}}] = \mathbb{E}[T_{\text{SRPT}}] + \epsilon'$, $\epsilon, \epsilon' > 0$. As an initial starting point, and to gain the greatest analytical insights, we consider two-class M/G/1 priority queues under FP and LTFS, i.e., instances of these queues with $K = 2$; multiclass priority queues with larger values of $K$ are examined in §3.2 and 5. As a representative example of our results, we further consider the service time distribution to be hyperexponential with parameters $(\mu_1, \mu_2, p)$ where $\mathbb{E}[S] = 1/\mu_1 p + 1/\mu_2(1-p)$, $\mathbb{E}[S^2] = 2/\mu_1^2 p + 2/\mu_2^2(1-p)$, and $\mu_1 > \mu_2$. We postpone until §4 our optimal segmentation of the customer service times comprising the single-class workload in order to determine the multiclass workloads of the FP and LTFS M/G/1 priority queues, and for now we consider the multiclass workload consisting of exponential class 1 service times with rate $\mu_1$, exponential class 2 service times with rate $\mu_2$, and independent Poisson arrival processes with rates $p\lambda$ and $(1-p)\lambda$ for classes 1 and 2, respectively. Hence, $\mathbb{E}[S_k] = 1/\mu_k$, $\mathbb{E}[S_k^2] = 2/\mu_k^2$ and $\mathbb{E}[S_k^3] = 6/\mu_k^3$. Additional single-class and multiclass workloads are examined in §3.2 and 5.

The first two moments for the SRPT and FP M/G/1 queues are directly obtained from the formulas in §2.1 and 2.2, respectively. For the LTFS M/G/1 queue, we have from Thm. 10 the first moment results

$$\mathbb{E}[W_1] = \left(\frac{p\lambda}{\mu_1^2} + \frac{(1-p)\lambda}{\mu_2^2}\right)(1-\rho)^{-1} - \rho_2 \mathbb{E}[W_2]\left(1 - \frac{1}{b_1}\right), \tag{76}$$

$$\mathbb{E}[W_2] = \frac{(p\lambda/\mu_1^2 + (1-p)\lambda/\mu_2^2)/(1-\rho)}{1 - \rho_1(1 - 1/b_1)}, \tag{77}$$

and exploiting Thm. 11 we obtain the corresponding second moment results from the solution of the system of linear equations in terms of the 5 unknowns $\mathbb{E}[W_1^2]$, $\mathbb{E}[W_2^2]$, $\mathbb{E}[W_1 W_2]$, $\mathbb{E}[W_0 W_1]$, $\mathbb{E}[W_0 W_2]$, which can be numerically computed in an efficient manner since the coefficient matrix is triangular and nonsingular.

In Fig. 1, we illustrate the effect of the slope ratio on the second moment. In a two class LTFS queue, with arrival rate $\lambda_1 = 0.2$, $\lambda_2 = 0.2$, the service distribution for the two classes are exponential with rate 1 and 0.5 respectively. We let $b_2/b_1$ the ratio of the slopes from the two classes vary from 0.1 to 0.9. It can be observed that the the second moments change linearly according to the ratio. When the ratio becomes larger than 1, we can switch the index, and

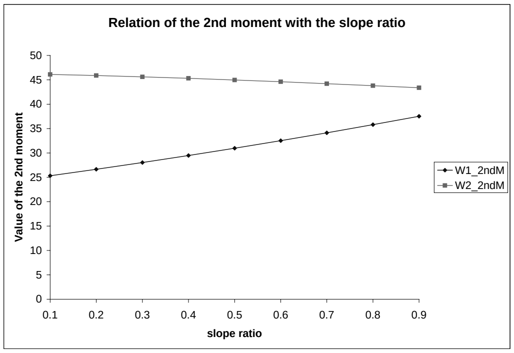

**Figure 1: The impact of LTFS slope ratio on second moments**

obtain the symmetric results, and hence we omit that part. When $b_2/b_1$ approaches 0, the LTFS queue tends toward a fixed priority queue, and as $b_2/b_1$ approaches 1, it tends toward a FCFS queue. The selection between $b_1$ and $b_2$, and slopes for general cases, provides great flexibility in achieving different levels of performance based upon both first two moments. Examples in the next section will further demonstrate this aspect of the LTFS policy.

Of course, given the optimality of SRPT, it is not possible to achieve the same mean sojourn time under LTFS as obtained under SRPT. However, based on numerical experiments with the theoretical foundation derived in §2, our results demonstrate that it is possible to approximate the mean sojourn time of SRPT quite well for reasonable values of $\epsilon$. Moreover, the performance measures of the M/G/1 queue under a LTFS policy tend to have the best higher-order statistics such as smaller variance properties and faster large-deviations decay rates. Given that these performance results for the workload considered in this section are completely consistent with the more general results considered later in this section (a subset of which are presented), we omit plots of these results in the interest of space. This consistency of results also extends to simulations with data from a large-scale production Web site (a subset of which are presented in §5). It is important to emphasize, however, that we consistently observed among the various policies which can achieve similar mean customer waiting times that the LTFS policy exhibits superior higher-order statistical properties over SRPT and FP. Furthermore, it is very important to note that increasing the number of classes, together with the results in §4, will certainly make it possible to obtain closer first-order statistics among the SRPT, FP and LTFS queues. This is clear from a theoretical perspective (e.g., [40]) and it is also demonstrated by our numerical results.

We now turn to compare the large-deviations decay rate of the sojourn time distribution in SRPT, FP and LTFS GI/G/1 queues. Our main results are summarized in the following theorem.

THEOREM 15.

$$\Lambda^*_{FCFS} \;>\; \Lambda^*_{LTFS}, \Lambda^*_{FP} \;>\; \Lambda^*_{SRPT^+} \;>\; \Lambda^*_{SRPT^0}. \tag{78}$$

*where SRPT*$^+$ *and SRPT*$^0$ *denotes the SRPT GI/G/1 queue with* $\mathbb{P}[S = x_S] > 0$ *and* $\mathbb{P}[S = x_S] = 0$*, respectively.*

PROOF. Stolyar and Ramanan [48] have shown $\Lambda^*_{\text{FCFS}}$ to be maximal among all work-conserving scheduling policies. Nuyens and Zwart [38] have established the last inequality between the two versions of SRPT queues. The first inequality with respect to $\Lambda^*_{\text{FP}}$ follows directly from Thms. 3 and 5, since $\Lambda^*_{\text{FP}}$ is of the same order as $\Lambda^*_{\text{FCFS}}$ only for class 1 and is otherwise smaller. From Eq. (11) of Thm. 2 and Eq. (21) of Thm. 4, we see that $\Lambda_{\text{FP}}(\theta) < \Lambda_{\text{SRPT}^+}(\theta)$, and thus $\Lambda^*_{\text{FP}} = \sup_{\theta\in[0,\hat{\Lambda}^*]}\{\theta - \Lambda_{\text{FP}}(\theta)\} > \Lambda^*_{\text{SRPT}^+} = \sup_{\theta\in[0,\hat{\Lambda}^*]}\{\theta - \Lambda_{\text{SRPT}^+}(\theta)\}$. This together with Thm. 5 yields the second inequality with respect to $\Lambda^*_{\text{FP}}$. It is readily verified from Thms. 5 and 13 that $\Lambda^*_{\text{FP}}$ and $\Lambda^*_{\text{LTFS}}$ are of the same order. □

## 3.2 Mean-Variance Utility Functions

Even though we have considerable formal and numerical evidence (together with the simulation evidence in §5) that the LTFS policy can satisfy desired mean sojourn times while also providing better variance properties, another fundamental and general objective of interest to us is based on utility 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 [52, 3]. 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 Markowitz popularized the basic idea; e.g., see [31, 32].

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 [34, 8, 42]. In particular, the following functional form for utility

$$U(\theta, \sigma) \;=\; \theta^a - \sigma^b, \tag{79}$$

where $\theta$ and $\sigma$ are the mean and standard deviation of the measure of interest and $a$ and $b$ are function parameters, has been proposed and empirically evaluated [42]. 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 [50] has recently shown the functional form in (79) 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. This functional form is also consistent with those considered in [35].

We therefore are interested in using the functional form in (79) with $\theta = \mathbb{E}[T]$ and $\sigma = (\mathbb{E}[T^2] - \mathbb{E}[T]^2)^{1/2}$. This yields the following equivalent form of (79) for our purposes:

$$U(\mathbb{E}[T], (\mathbb{E}[T^2] - \mathbb{E}[T]^2)^{1/2}) = \mathbb{E}[T]^a - (\mathbb{E}[T^2] - \mathbb{E}[T]^2)^{b/2}. \tag{80}$$

We note that a similar functional form can be used with the variance term replaced by a function of the tail of the sojourn time distribution, but this is not further explored herein due to space restrictions.

As a representative example of our results, we consider a three-class queueing system with Poisson arrivals where the proportions of customer classes 1, 2 and 3 are 80%, 14% and 6%, respectively. The mean service times for the three classes are 1, 20 and 1000. Numerical experiments were conducted for both SRPT and LTFS under different traffic intensities $\rho$, a small subset of which are presented in Figs. 2 and 3. To make the comparison fair, for each $\rho$, we set the LTFS control parameter slopes such that the mean waiting time is as close to that of SRPT as possible in each instance of the queueing system. In addition to LTFS providing smaller higher-order statistical properties as shown by our results from the previous section, we also observe the significance of the impact of the policy in terms of the utility function. We observe in Figs. 2 and 3 that when the traffic is light, understandably, there is no significant difference between the two policies; when the traffic grows heavier, 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 $(a, b)$. Once again, increasing the number of classes, together with the results in §4, will certainly provide greater optimal utility values in the FP queues and especially in the LTFS queues.

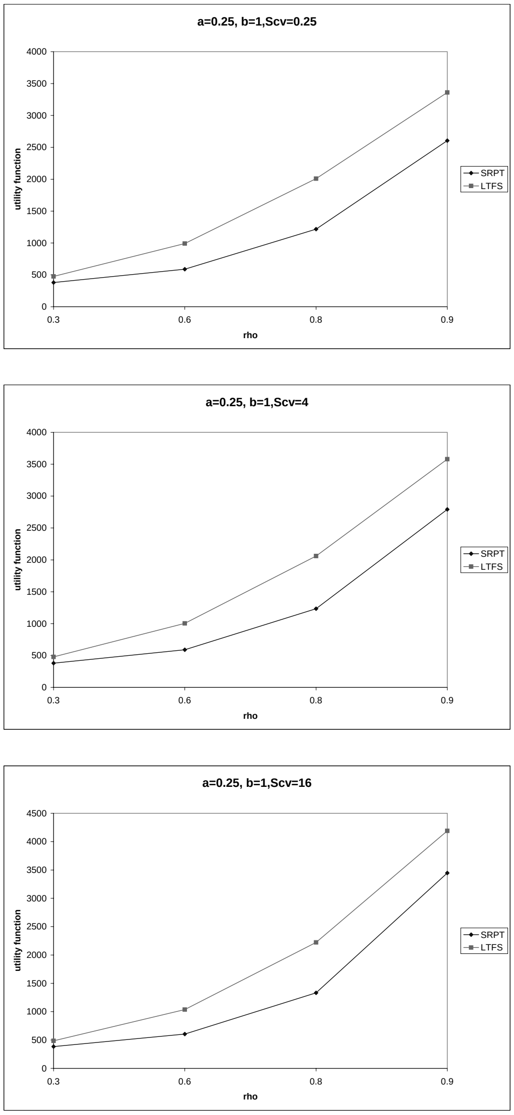

**Figure 2: Different utility function values under different service time SCV, a=0.25, b=1**

We also examine the general shape of the utility function when the relative values of $(a, b)$ are varied while fixing the queueing system parameters. These results, omitted due to space restrictions, show that the shape of the utility functions is generally concave with respect to $b$ and their derivatives with respect to $b$ have a descending trend as the traffic intensity $\rho$ increases.

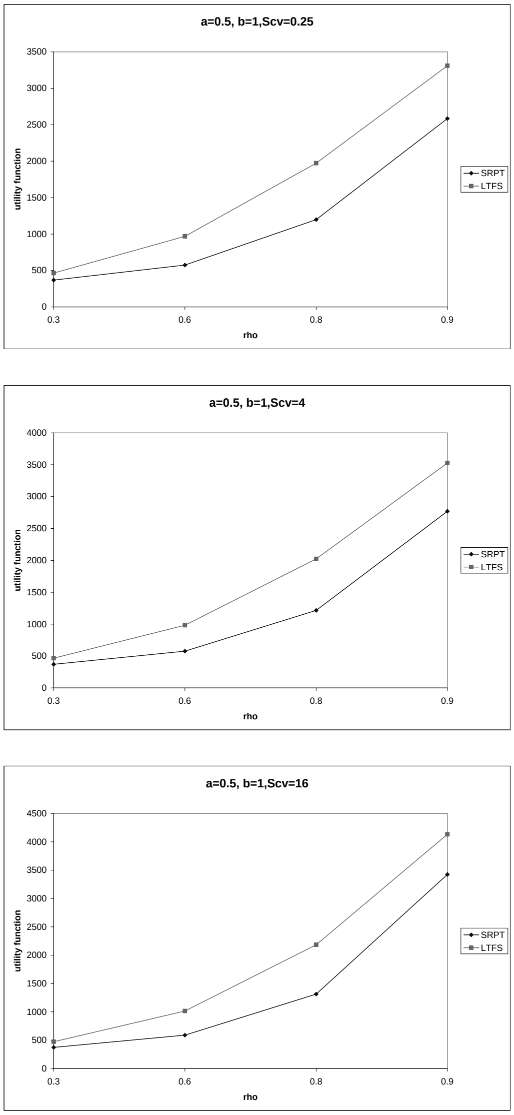

**Figure 3: Different utility function values under different service time SCV, a=0.5, b=1**

Once again, the controls in LTFS M/G/1 queues make it possible to optimize this utility function for different $a, b$ values which is much more limited in M/G/1 queues under SRPT and FP.

## 4. PARTITIONING OF SERVICE TIMES

Our analysis in §2 and 3 assumes that the workloads for the multiclass FP and LTFS 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 these 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 workloads to be used in the FP and LTFS 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 $B_L = p_0 \leq p_1 \leq \ldots \leq p_{K-1} \leq p_K = B_U$ 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 optimally partitioning the customer service times into multiclass workloads as part of determining the optimal FP and LTFS control parameters to maximize the overall utility of the system as a function of the first two moments of customer sojourn times. Using the functional form for utility in (80), we have

$$\max_{p_1,\cdots,p_{K-1}} \quad \mathbb{E}[T]^a - (\mathbb{E}[T^2] - \mathbb{E}[T]^2)^{b/2} \tag{81}$$

$$\text{s.t.} \quad B_L = p_0 \leq p_1 \leq \ldots \leq p_{K-1} \leq p_K = B_U. \tag{82}$$

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. Once again, we note that a similar functional form can be used for the objective in (81) with the variance term replaced by a function of the tail of the sojourn time distribution.

In general, the objective function is nonlinear in the decision variables, but the optimal solution can be efficiently computed using known methods in nonlinear optimization; e.g., see [5, 13]. However, in many cases of interest, the objective function is convex in the decision variables, and thus the optimal solution can be very efficiently computed using known methods in convex optimization; e.g., refer to [5]. Furthermore, it is important to note that the same approach can be exploited if the objective in (81) is replaced with a different function of $\mathbb{E}[T]$ and $(\mathbb{E}[T^2] - \mathbb{E}[T]^2)$.

## 5. SIMULATION EXPERIMENTS

While the main contributions of this paper consist of the foregoing collection of mathematical results, in this section we briefly consider the performance characteristics of a single-server queueing system under SRPT and LTFS in practice based on the workload from an Internet application environment. Specifically, we use simulation to estimate the performance of both queueing systems under the actual workload from each server of a large-scale production Web site. The arrival times of customers are obtained directly from the access logs of the Web site, whereas the corresponding service times are obtained from measurements on the real system.

The characteristics of this Web site are typical to what has been reported in the research literature for such Internet application environments. In particular, most of the pages are dynamic with the vast majority of the requests being for static objects that comprise these pages. The service times used in our simulation are obtained from measurements of the time to serve these dynamic pages and static objects, which we can identify directly from the contents of the access logs. Furthermore, we identify and focus on sufficiently long stationary intervals of traffic periods found in our analysis of the access logs from each server of the production 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 scheduling. 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, where there can be tens to hundreds of client requests within a second at each server during peak traffic periods for the production Web site.

The stationarity of the corresponding arrival and service processes extracted from the access logs of each server and system measurements is confirmed by the stationarity testing method recently proposed in [39]. Detailed statistical analysis of the workload data shows that the sequence of interarrival times is long-range dependent and that the service time distribution has high variability. These technical details are important because the system environment considered in our simulation experiment is very different from the M/G/1 (and, in some cases, GI/G/1) queues considered in the previous sections, but they are outside the scope of this paper; however, these technical details can be found in [41].

Space restrictions prevent us from including our simulation results. However, we can summarize that the results from our numerous simulation experiments are completely consistent with those presented in Figs. 2 and 3. The magnitude of the results vary for different values of $(a, b)$, but we consistently found that the trends observed from our theoretical results are identical for the large-scale production Web site, even with its correlations in the arrival stream and variability in the service times.

## 6. CONCLUSIONS

In this paper we have developed a theoretical foundation for a general class of scheduling policies, LTFS, that provide mean sojourn times comparable to SRPT with superior higher-order statistical properties and that provide control mechanisms to optimize general performance functions which combine the mean and higher-order statistics of customer sojourn times. Our mathematical analysis included deriving expressions for the mean, variance and tail asymptotics of the customer sojourn time distribution in LTFS queueing systems, providing the first higher-order statistical results to appear in the research literature. We also derived closed-form expressions and a linear algorithm to determine the corresponding control parameters that satisfy a given vector of sojourn times, which similarly provide the first such results to appear in the literature. As an additional set of control parameters, and based in part on our derivations of the corresponding performance measures, our analysis further included the optimal segmentation of the service time distribution for the original single-class workload into the per-class service time distributions of the LTFS multiclass workload.

Our study demonstrated that, by exploiting our performance measure and control parameter results, the LTFS policies can provide mean sojourn times that are comparable to those obtained under SRPT while also providing superior second moment properties for customer sojourn time than those obtained under SRPT. We also showed that this class of policies provides a large-deviations decay rate of customer sojourn times that is superior to the corresponding large-deviations decay rate under SRPT and shares some of the characteristics of the maximal decay rate. These results illustrated and quantified a fundamental performance tradeoff between improving the mean sojourn time and degrading the higher-order sojourn time statistics, and vice versa, especially at heavy traffic intensities. Our study further demonstrated that, by exploiting our performance measure and control parameter results, the LTFS policies can used to optimize general performance functions that combine the mean and higher-order statistics of customer sojourn times in terms of mean-variance utility functions based on recent results in portfolio theory, although our analysis is not restricted to these utility functions and can be applied more generally.

The theoretical foundation derived and presented in this paper is also generally important in practice for a broad spectrum of application environments. Based on numerous simulations with traces from a large-scale production Web site, we found that the trends observed from our theoretical results tend to hold in such Web site environments. Moreover, a great deal of experience over the past few years in a wide variety of production environments has demonstrated that the LTFS policies provide the same performance characteristics as shown by our theoretical results and can be implemented efficiently with very low overhead [**?**].

## 7. ACKNOWLEDGMENTS

We thank Ed Coffman for several helpful comments on an earlier draft of this work.

## 8. REFERENCES

[1] J. Abate and W. Whitt. Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems*, 25:173–223, 1997.

[2] P. Ansell, K. Glazebrook, I. Mitrani, and J. Nino-Mora. A semidefinite programming approach to the optimal control of a single-server queueing system with imposed second moment constraints. *Journal of the Operational Research Society*, 50(7):765–773, 1999.

[3] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, Second edition, 2003.

[4] 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.

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

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

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

[8] 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.

[9] Y. S. Chow and H. Teicher. *Probability Theory: Independence, Interchangeability, Martingales*. Springer-Verlag, Second edition, 1988.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] M. Dacre, K. Glazebrook, and J. Nino-Mora. The achievable region approach to the optimal control of stochastic systems. with discussion. *Journal of the Royal Statistical Society, Series B, Methodological*, 61(4):747–791, 1999.

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

[17] 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.

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

[19] R. Haji and G. F. Newell. A relation between stationary queue and waiting time distributions. *Journal of Applied Probability*, 8:617–620, 1971.

[20] 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.

[21] 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.

[22] 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.

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

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

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

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

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

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

[29] 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.

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

[31] H. M. Markowitz. Foundations of portfolio theory. *Journal of Finance*, 46(2):469–477, 1991.

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

[33] 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.

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

[35] J. A. V. Mieghem. Capacity management, investment, and hedging: Review and recent developments. *Manufacturing and Service Operations Management*, 5(4):269–302, 2003.

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

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

[38] M. Nuyens and B. Zwart. A large-deviations analysis of the

GI/GI/1 SRPT queue. Technical Report SPOR-Report 2005-06, Department of Mathematics and Computer Science, Eindhoven University of Technology, May 2005.

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

[40] T. E. Phipps, Jr. Machine repair as a waiting line problem. *Operations Research*, 4:76–86, 1956.

[41] A. Radovanovic, B. Ray, and M. S. Squillante. Statistical properties of traffic patterns in large-scale commercial web sites and their performance implications. Technical report, IBM Research Division, 2003.

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

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

[44] 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.

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

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

[47] 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.

[48] A. L. Stolyar and K. Ramanan. Largest weighted delay first scheduling: Large deviations and optimality. *Annals of Applied Probability*, 11(1):1–48, 2001.

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

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

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

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