Heavy-Tailed Limits for Medium Size Jobs and Comparison Scheduling*

Predrag R. Jelenković Xiaozhu Kang Jian Tan
Department of Electrical Engineering
Columbia University, New York, NY 10027
{predrag, xiaozhu, jiantan}@ee.columbia.edu
Tel.: (212) 854 8174 Fax: (212) 932 9421

January 2007; revised June 2007

Abstract

We study the conditional sojourn time distributions of processor sharing (PS), foreground background processor sharing (FBPS) and shortest remaining processing time first (SRPT) scheduling disciplines on an event where the job size of a customer arriving in stationarity is smaller than exactly $k \geq 0$ out of the preceding $m \geq k$ arrivals. Then, conditioning on the preceding event, the sojourn time distribution of this newly arriving customer behaves asymptotically the same as if the customer were served in isolation with a server of rate $(1-\rho)/(k+1)$ for PS/FBPS, and $(1-\rho)$ for SRPT, respectively, where ρ is the traffic intensity. Hence, the introduced notion of conditional limits allows us to distinguish the asymptotic performance of the studied schedulers by showing that SRPT exhibits considerably better asymptotic behavior for relatively smaller jobs than PS/FBPS.

Inspired by the preceding results, we propose an approximation to the SRPT discipline based on a novel adaptive job grouping mechanism that uses relative size comparison of a newly arriving job to the preceding m arrivals. Specifically, if the newly arriving job is smaller than k and larger than $m-k$ of the previous m jobs, it is routed into class k. Then, the classes of smaller jobs are served with higher priorities using the static priority scheduling. The good performance of this mechanism, even for a small number of classes $m+1$, is demonstrated using the asymptotic queueing analysis under the heavy-tailed job requirements. We also discuss refinements of the comparison grouping mechanism that improve the accuracy of job classification at the expense of a small additional complexity.

Keywords: Comparison scheduling, scalability, fairness, adaptive thresholds, M/G/1 queue, processor sharing, shortest remaining processing time first, foreground background processor sharing, asymptotic analysis, heavy tails, medium size jobs

[*] This work is supported by NSF Grant 0615126.
Annals of Operations Research, in press.

Introduction

It has been widely recognized that heavy-tailed distributions are suitable for modeling job sizes in information service networks, e.g., see Jelenković and Momčilović (2003a,b) and the references therein. For heavy-tailed distributions, large jobs appear much more frequently than for the light-tailed ones, which imposes very different constraints in terms of optimizing the scheduling process as compared to the light-tailed scenarios. In particular, schedulers that may assign the server exclusively to a very large job, e.g., first come first serve (FIFO) discipline, can cause very large delays and, in general, suboptimal performance, as shown by Anantharam (1999).

Hence, most of the practical schedulers utilize either the processor sharing (PS) and foreground background processor sharing (FBPS) disciplines because of their inherent fairness, or the shortest remaining processing time first (SRPT) discipline because of its known optimality under quite general conditions. In particular, it was shown by Schrage (1968) that SRPT minimizes the number of customers in the G/G/1 queue over all work-conserving disciplines. For early references on these and other scheduling disciplines see Kleinrock (1976); Wolff (1989) and the references therein. Recently, the performance of these disciplines was revisited in the context of heavy tails; for a recent survey see Borst et al. (2003b). For practical applications of SRPT-based scheduling to improving Web server performance see Harchol-Balter et al. (2003); Rawat and Kshemkalyani (2003); also, for recent studies that are applying FBPS to reducing the latency of short TCP flows see Rai et al. (2004, 2005).

It is well known that the sojourn time distributions under PS, FBPS and SRPT scheduling disciplines are asymptotically equivalent for power law distributions (more precisely, regularly or intermediately regularly varying distributions). This was originally proved by Núñez-Queija (2000) and then later studied for regularly varying distributions in Theorems 2.2, 2.5 and 2.6 of Borst et al. (2003b); see also Theorem 2.1 of Jelenković and Momčilović (2003b) and Theorem 1 of Jelenković and Momčilović (2002). In other words, for large jobs, the waiting time does not depend on the choice of a specific scheduling discipline among PS, FBPS and SRPT.

In this paper, we introduce a new notion of conditional waiting time distribution which allows us to refine and distinguish the performance of PS/FBPS and SRPT schedulers for medium size jobs. Informally, our first main result, stated in Theorem 1.2, shows that even the relatively smaller jobs receive asymptotically the same residual capacity $1-\rho$ as the larger ones for SRPT discipline, while, for PS/FBPS schedulers, these smaller jobs share the residual capacity equally with the larger jobs in the system. Hence, it appears that SRPT provides much better and more uniform performance over a wide range of time scales. Furthermore, the performance improvement for conditionally smaller jobs is not achieved at the expense of larger jobs, i.e., SRPT is not only efficient but fair as well, which is in line with similar recent findings in the context of mean value analysis by Bansal and Harchol-Balter (2001); Wierman and Harchol-Balter (2003). To this end, we would like to point out that contrary to our findings, in the light-tailed context, it was shown by Ramanan and Stolyar (2001) that FIFO is optimal in terms of maximizing the decay rate of the waiting time distribution over all work conserving disciplines. For more recent results on the light-tailed asymptotic analysis see Nuyens and Zwart (2006) and the references therein.

Overall, using the SRPT scheduling is beneficial for a broad range of conditions and applications. However, as discussed in one of the very first papers on SRPT by Schrage and Miller (1966), this discipline may be quite difficult to implement. Clearly, its complicated preemptive nature requires keeping track of the remaining processing times for all jobs in the queue which may be prohibitive for systems with large job volumes, e.g., Web servers. In addition, Schrage and Miller (1966) show that the expected number of preemptions per job is proportional to the load of the system, which can be quite large. Hence, even as early as 1966, it was recognized by Schrage and Miller (1966) that one should try to approximate SRPT with less complex schedulers. The most apparent option, as suggested by Schrage and Miller (1966), is to design a threshold-based static priority approximation to SRPT. Basically, the idea is to select a fixed number of thresholds m and then group jobs into $m+1$ classes depending on which pair of thresholds a job size happens to fall between. Then, these classes are served according to the static priority discipline with higher priorities assigned to classes with smaller jobs. Since then, there has been a lot of work on threshold-based scheduling policies. For example, it was shown by Bansal and Gamarnik (2006) that even with a single threshold, one can obtain the performance comparable to SRPT up to a constant factor in terms of the mean sojourn time for M/M/1 queue as well as for M/G/1 queue with finite variance Pareto service distribution.

Although it is encouraging that one can achieve a provably very good approximation of M/G/1/SRPT queue even with a very small number of static thresholds (only one in the paper by Bansal and Gamarnik (2006)), these solutions are likely not to perform well in practice since the traffic characteristics are often nonstationary, highly correlated (long range dependent) and very bursty (e.g., batch arrivals, etc); see Park and Willinger (2000); Squillante et al. (1999). In order to overcome these difficulties, we propose a novel adaptive job classification (grouping) mechanism that is based on relative size comparison of a newly arriving job to the previous m arrivals; this scheduler is inspired by our conditional limit results. Specifically, if an arriving job is smaller than k and larger than $m-k$ of the previous m jobs, it is routed into class k. We also discuss refinements of the comparison grouping mechanism that improve the accuracy of the classification for both light-tailed and correlated job arrivals at the expense of a small (fixed) additional complexity in Subsection 2.1.1.

The good performance of our comparison classification mechanism is demonstrated using the asymptotic queueing analysis under heavy-tailed job sizes in Section 2.2. First, in Subsection 2.2.1 we study the queueing behavior of a class k process in isolation and show that the workload distribution decays faster for larger k. More precisely, for regularly varying (power law) service distribution, the tail of the workload distribution $\mathbb{P}[W^{(k)} > x]$ of a class k process, as implied by Theorem 2.1, is of the order of $x(\mathbb{P}[B > x])^{k+1}$, where B is the service requirement of a typical job before the comparison splitting. Hence, our comparison splitting procedure provides a proper ordering of jobs. Furthermore, in Subsection 2.2.2 we study the joint queueing behavior of all classes under the static priority (SP) discipline, with higher priorities assigned to classes with "smaller" jobs. Theorem 2.2 shows that the workload distribution of a class with a smaller index k (i.e., larger jobs) has the same queueing behavior as if it were served in isolation with the system capacity reduced by the mean arrival rates of the classes with smaller jobs. Roughly speaking, this is a similar behavior as seen in Theorem 1.2 for the SRPT discipline and, thus, the SP scheduling with our comparison splitting should provide a reasonable approximation to the SRPT discipline. Furthermore, in regard to the analysis, we would like to point out that the main technical difficulty is that the split processes are individually and mutually correlated. This statistical correlation makes other types of analyses, outside of the heavy-tailed context, possibly difficult.

In addition, we would like to point out that a preliminary version of this paper has appeared earlier in Jelenković et al. (2007) as part of the conference proceedings, which contains sketches of the proofs as well as the extensive simulation experiments. Those experiments demonstrated the good performance of our adaptive scheduler that, in particular, outperforms the static threshold policies when the arrival processes are statistically correlated and time varying. However, in contrast to the previous focus on simulations in Jelenković et al. (2007), this paper provides the rigorous details of the proofs.

The rest of this paper is structured as follows. In the next section, we introduce the new notion of conditional waiting time distribution that refines and differentiates the performance of PS/FBPS and SRPT schedulers for medium size jobs. Based on the conditional asymptotic result of the sojourn time distribution (stated in Theorem 1.2), we propose a novel comparison grouping scheme and its refined version in Section 2.1. To demonstrate its good performance, we conduct the asymptotic queueing analysis under heavy-tailed job sizes in Section 2.2. In the end, Section 3 summarizes our contributions.

1 Heavy-Tailed Limits for Medium Size Jobs with Popular Schedulers

1.1 Definitions and Preliminary Results

In this section we introduce the necessary notation and describe the existing and preliminary results. Let B_i and V_i denote the job size and the waiting time of the customer arriving at time T_i, respectively, where $\{B_i\}_{i>-\infty}$ are i.i.d. random variables. The arrival points $\{T_i\}_{i>-\infty}$ are assumed to be Poisson with rate λ and independent of job requirements $\{B_i\}_{i>-\infty}$. Hence, without loss of generality, in view of the PASTA property, we set $T_0 = 0$. The waiting time of a customer is defined as the amount of time between its arrival and departure, also referred to as sojourn time in the queuing literature. To present our main results, we need the following definitions.

Definition 1.1 *A nonnegative random variable X or its distribution function (d.f.) F is called intermediately regularly varying, $X \in \mathcal{IR}$, if*

$$\lim_{\eta \uparrow 1} \overline{\lim_{x\to\infty}} \frac{\mathbb{P}[X > \eta x]}{\mathbb{P}[X > x]} = 1.$$

Regularly varying distributions $\mathcal{R}_\alpha$ are the best-known examples from $\mathcal{IR}$.

Definition 1.2 *A nonnegative random variable X or its d.f. F is called regularly varying with index α, $X \in \mathcal{R}_\alpha$ ($F \in \mathcal{R}_\alpha$), if*

$$F(x) = 1 - \frac{l(x)}{x^\alpha}, \quad \alpha \geq 0,$$

where $l(x)$: $\mathbb{R}_+ \to \mathbb{R}_+$ is slowly varying, i.e., $\lim_{x\to\infty} l(\eta x)/l(x) = 1$, $\eta > 1$.

The preceding class includes the well-known power law distributions, e.g., $F(x) = 1 - 1/x^\alpha, x \geq 1, \alpha > 0$.

Let $\tilde{B}_i, 1 \leq i \leq m$ be the order statistics of $B_{-i}, 1 \leq i \leq m$ with the convention $\tilde{B}_0 = \infty$ and $\tilde{B}_{m+1} = 0$. To make the notation uniform, we assume that $\tilde{B}_0 = \infty$, $\tilde{B}_1 = 0$ for $m = 0$, and when it is necessary to emphasize the total number of random variables, we write explicitly $\tilde{B}_i^{(m)} \equiv \tilde{B}_i$.

Definition 1.3 *Let $\mathcal{A}_k^{(m)} \triangleq \{\tilde{B}_{k+1}^{(m)} \leq B_0 < \tilde{B}_k^{(m)}\}$ for $m \geq k \geq 0$.*

The asymptotic behavior of the sojourn time distribution for PS, FBPS and SRPT has been extensively studied under heavy-tails, e.g., see Zwart and Boxma (2000); Núñez-Queija (2000); Borst et al. (2003b); Jelenković and Momčilović (2002); Jelenković and Momčilović (2003b) and the references therein. We summarize these results for intermediately regularly varying distributions in the following theorem, which follows directly from our more general/refined result presented in Theorem 1.2 in the following section. In order to ease the notation we simply write $B \equiv B_0$ and $V \equiv V_0$.

For the rest of the paper, we assume that the system has reached stationarity. Also, we use H to denote a sufficiently large positive constant. The value of H is generally different in different places, for example, $H/2 = H$, $H^2 = H$, $H + 1 = H$, etc. Furthermore, we use the following standard notation. For any two real functions $a(t)$ and $b(t)$ and fixed $t_0 \in \mathbb{R} \cup \{\infty\}$ we will use $a(t) \sim b(t)$ as $t \to t_0$ to denote $\lim_{t \to t_0}[a(t)/b(t)] = 1$. Similarly, we say that $a(t) \gtrsim b(t)$ as $t \to t_0$ if $\liminf_{t \to t_0} a(t)/b(t) \geq 1$; $a(t) \lesssim b(t)$ has a complementary definition. In addition, we say that $a(t) = o(b(t))$ as $t \to t_0$ if $\lim_{t \to t_0} a(t)/b(t) = 0$. When $t_0 = \infty$, we often simply write $a(t) = o(b(t))$ without explicitly stating $t \to \infty$ in order to simplify the notation.

Theorem 1.1 *If $B \in \mathcal{IR}$ and $\mathbb{E}B^\alpha < \infty$ for some $\alpha > 1$, then, under the PS, FBPS or SRPT discipline, we have, as $x \to \infty$,*

$$\mathbb{P}[V > x] \sim \mathbb{P}[B > (1-\rho)x].$$

The preceding asymptotic insensitivity of the sojourn (waiting) time distribution on the scheduling discipline was first derived in Theorems 5.2.3, 5.2.4 and 5.2.5 of Núñez-Queija (2000) under somewhat more restrictive conditions; see also Theorems 2.2, 2.5 and 2.6 of Borst et al. (2003b). For PS, this result was proved in Theorem 2.1 of Jelenković and Momčilović (2003b) using a novel sample path approach that allows further extension of the result to moderately heavy distributions, e.g., lognormal, see Theorem 3.1 of Jelenković and Momčilović (2003b). Furthermore, as noted in Appendix B of Jelenković and Momčilović (2002), this sample path approach extends directly to SRPT and FBPS scheduling disciplines. Our proof of Theorem 1.2 in this paper relies directly on the arguments developed by Jelenković and Momčilović (2002); Jelenković and Momčilović (2003b).

1.2 Conditional Limits

The following theorem represents our first main result, which implies Theorem 1.1 by unconditioning on event $\mathcal{A}_k^{(m)}$, i.e., summing over all k, $0 \leq k \leq m$.

Theorem 1.2 *If $B \in \mathcal{IR}$ and $\mathbb{E}B^\alpha < \infty$ for some $\alpha > 1$, then, under either PS or FBPS discipline, we have for fixed k, as $x \to \infty$,*

$$\mathbb{P}\left[V > x, \mathcal{A}_k^{(m)}\right] \sim \mathbb{P}\left[B > \frac{(1-\rho)x}{(1+k)}, \mathcal{A}_k^{(m)}\right] \sim \frac{1}{k+1}\binom{m}{k}\mathbb{P}\left[B > \frac{(1-\rho)x}{k+1}\right]^{k+1}, \tag{1.1}$$

and under the SRPT discipline,

$$\mathbb{P}\left[V > x, \mathcal{A}_k^{(m)}\right] \sim \mathbb{P}\left[B > (1-\rho)x, \mathcal{A}_k^{(m)}\right] \sim \frac{1}{k+1}\binom{m}{k}\mathbb{P}\left[B > (1-\rho)x\right]^{k+1}. \tag{1.2}$$

Remark 1 These results can be easily extended to $GI/GI/1$ queue under the FBPS discipline, and possibly under the SRPT as well using the recent studies on SRPT by Nuyens et al. (2007). In order to provide a unified framework, we omit such possible extensions here and restrict ourselves to the $M/G/1$ framework. Furthermore, our focus in the second part of the paper is to exploit this idea of relative job comparisons to design adaptive and efficient approximation of SRPT, which we term *comparison scheduling.*

Remark 2 Note that on $\mathcal{A}_k^{(m)}$, the distribution of B has a much lighter tail of the order of $\mathbb{P}\left[B > x\right]^{k+1}$ and, thus, $\mathcal{A}_k^{(m)}$ partitions the probability space into jobs of decreasing sizes as k increases. Interestingly, the result shows that, for the SRPT discipline, even the relatively much smaller job receives the entire long-term residual capacity $1 - \rho$, while, for PS/FBPS, this smaller job shares equally the residual capacity with the k larger ones. Hence, SRPT outperforms PS/FBPS for medium size jobs and therefore provides much better and more uniform performance over a wide range of time scales, i.e., it appears that SRPT generates extra capacity. Informally, we believe that the explanation for this comes from the combined effect of the SRPT prioritization mechanism and the fact that jobs of "different" sizes occur on different time scales. Hence, the medium size jobs are basically not affected by the larger ones because of the higher priority assigned to them and the larger jobs are not impacted by the smaller ones due to the time scale separation.

In order to prove this theorem, we define the class of heavy-tailed distributions $\mathcal{L}$ that contains subexponential distributions and, in particular, the intermediately regularly varying class $\mathcal{IR}$, and establish the following two preliminary lemmas.

Definition 1.4 *A nonnegative random variable X or its d.f. F is called heavy-tailed $X \in \mathcal{L}$ (or $F \in \mathcal{L}$) if, for any fixed $y \in \mathbb{R}$,*

$$\lim_{x\to\infty} \frac{\mathbb{P}[X > x - y]}{\mathbb{P}[X > y]} = 1.$$

Lemma 1.1 *Let $\{X_i\}_{0\leqslant i\leqslant m}$ be i.i.d. random variables with $X_0 \in \mathcal{L}$ and, denote the order statistics of $X_1, X_2, \cdots, X_m$ by $\tilde{X}_1 \geqslant \tilde{X}_2 \geqslant \cdots \geqslant \tilde{X}_m$ with $\tilde{X}_0 = \infty$ and $\tilde{X}_{m+1} = 0$, then, for any $m \geq k \geq 0$, as $x \to \infty$, we have*

$$\begin{aligned}\mathbb{P}[X_0 > x, \tilde{X}_{k+1} \leq X_0 < \tilde{X}_k] &\sim \mathbb{P}[X_0 > x, X_0 < \tilde{X}_k] \\ &\sim \frac{1}{k+1}\binom{m}{k}(\mathbb{P}[X_0 > x])^{k+1}. \qquad (1.3)\end{aligned}$$

Remark 3 This result holds for all continuous distributions without the assumption $X_0 \in \mathcal{L}$. However, the assumption $X_0 \in \mathcal{L}$ is necessary in general since the result may not hold for light-tailed lattice valued distributions. Here, easy calculations show that the lemma does not hold for geometric distribution $\mathbb{P}[X_i = j] = p^j(1-p), j \geq 0$, where we obtain for $m = k = 1$ and positive integer $x \in \mathbb{N}$

$$\mathbb{P}[X_1 > X_0 > x] = \mathbb{E}\left[\mathbf{1}\{X_0 > x\}p^{X_0+1}\right] = \frac{p}{1+p}(\mathbb{P}[X_0 > x])^2.$$

Lemma 1.2 *If two arrival processes $\{(T_i, B_{1i})\}_{i>-\infty}$ and $\{(T_i, B_{2i})\}_{i>-\infty}$, satisfying $B_{1i} = 0$ for $i < 0$, $B_{10} = B_{20}$, and either $B_{1i} = B_{2i}$ or $B_{1i} = 0$ for $i > 0$, are served with SRPT discipline, then, the corresponding sojourn times V_1 and V_2 for the customer arriving at T_0 satisfy $V_2 \geq V_1$.*

The **proofs** of Lemma 1.1 and 1.2 are presented in the Appendix.

Proof of Theorem 1.2: Label the customer that arrives at time T_0, and define function $R_0(t) \equiv R_{B_0}(t)$ for $t \geq 0$ to be the amount of remaining work of the labeled customer at time t. Let L_m be the number of customers in the system just before time T_{-m}. For all the customers arriving between T_{-m} and T_0, define B^0_{-i} to be the remaining service time of $B_{-i}, 1 \leq i \leq m$ at time $t = 0$. For all the customers arriving before time T_{-m}, define $B_i^{(e)}(m), 1 \leq i \leq L_m$

to be the remaining service time at time T_{-m} and $B_i^{(e)}(0)$ the remaining service time at time $t=0$. Denote $x \wedge y \equiv \min(x,y)$.

1. Processor sharing discipline. Similarly as in Jelenković and Momčilović (2003b), we have the following min-plus identity

$$V_0 = B_0 + \sum_{i=1}^{m} B_{-i}^0 \wedge B_0 + \sum_{i=1}^{L_m} B_i^{(e)}(0) \wedge B_0 + \sum_{i=1}^{N(V_0)} B_i \wedge R_0(T_i). \tag{1.4}$$

First, we establish an *upper bound* for (1.1). Observing that the residual service B_{-i}^0 for customer $-i$ at time 0 is upper bounded by its original job size and using $B_i^{(e)}(0) \le B_i^{(e)}(m)$ as well as $R_0(T_i) \le B_0$, we derive on set $\mathcal{A}_k^{(m)}$

$$\begin{aligned} V_0 &\le B_0 + \sum_{i=1}^{m} B_{-i} \wedge B_0 + \sum_{i=1}^{L_m} B_i^{(e)}(m) \wedge B_0 + \sum_{i=1}^{N(V_0)} B_i \wedge B_0 \\ &\le (k+1)B_0 + (m-k)\tilde{B}_{k+1}^{(m)} + \sum_{i=1}^{L_m} B_i^{(e)}(m) \wedge B_0 + \sum_{i=1}^{N(V_0)} B_i \wedge B_0, \end{aligned}$$

where $\tilde{B}_{m+1}^{(m)} \equiv 0$ for $m=k$. Then, for $0<\delta<1-\rho$, we have

$$\begin{aligned} \mathbb{P}\left[V_0(1-\rho-\delta) > x, \mathcal{A}_k^{(m)}\right] &\le \mathbb{P}\left[(k+1)B_0 > (1-\delta)x, \mathcal{A}_k^{(m)}\right] + \mathbb{P}\left[(m-k)\tilde{B}_{k+1}^{(m)} > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\ &\quad + \mathbb{P}\left[W_{B\wedge B_0}^{\rho+\delta} > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] + \mathbb{P}\left[\sum_{i=1}^{L_m} B_i^{(e)}(m) \wedge B_0 > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\ &\triangleq I_1(x) + I_2(x) + I_3(x) + I_4(x), \end{aligned} \tag{1.5}$$

where $W_{B\wedge B_0}^{\rho+\delta}$ denotes the stationary workload in a queue with job sizes $\{B_i \wedge B_0\}_{i\ge 1}$ and service capacity $\rho+\delta$. Now, Lemma 1.1 implies

$$I_1(x) = \mathbb{P}\left[(k+1)B_0 > (1-\delta)x, \mathcal{A}_k^{(m)}\right] \sim \frac{1}{k+1}\binom{m}{k}\left(\mathbb{P}\left[B_0 > \frac{(1-\delta)x}{k+1}\right]\right)^{k+1}. \tag{1.6}$$

Then, denote the order statistics of $\{B_{-i}\}_{0\leqslant i\leqslant m}$ by $\{\tilde{B}_i^{(m+1)}\}_{0\leqslant i\leqslant m}$. For $k=m$, we have $I_2(x)=0$. And, for $0\leqslant k\leqslant m-1$, we obtain, from Lemma 1.1 and $B_0 \in \mathcal{IR}$,

$$\begin{aligned} I_2(x) = \mathbb{P}\left[(m-k)\tilde{B}_{k+1}^{(m)} > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] &\leqslant \mathbb{P}\left[\tilde{B}_{k+2}^{(m+1)} > \frac{\delta x}{3(m-k)}\right] \\ \sim \binom{m+1}{k+2}\left(\mathbb{P}\left[B_0 > \frac{\delta x}{3(m-k)}\right]\right)^{k+2} &= o(I_1(x)). \end{aligned} \tag{1.7}$$

Following the same technique that was developed by Jelenković and Momčilović (2003b), we have

$$\begin{aligned} I_3(x) &= \mathbb{P}\left[W_{B\wedge B_0}^{\rho+\delta} > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\ &\le \mathbb{P}\left[B_0 > \delta^2 x, \mathcal{A}_k^{(m)}\right]\mathbb{P}\left[W_B^{\rho+\delta} > \frac{\delta x}{3}\right] + \mathbb{P}\left[W_{B\wedge\delta^2 x}^{\rho+\delta} > \frac{\delta x}{3}\right], \end{aligned}$$

which, by Lemma 3.2 (i) in Jelenković and Momčilović (2003b), implies that for δ small enough,

$$I_3(x) = o\left(\mathbb{P}\left[B > \frac{x}{k+1}\right]^{k+1}\right) = o(I_1(x)). \tag{1.8}$$

Again, similarly as in Jelenković and Momčilović (2003b), for any integer n_0, we have

$$\begin{aligned}
I_4(x) &= \mathbb{P}\left[\sum_{i=1}^{L_m} B_i^{(e)}(m) \wedge B_0 > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\
&= \sum_{n=1}^{\infty}(1-\rho)\rho^n \mathbb{P}\left[\sum_{i=1}^{n} B_i^{(e)}(m) \wedge B_0 > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\
&\leqslant n_0 \mathbb{P}\left[\sum_{i=1}^{n_0} B_i^{(e)}(m) \wedge B_0 > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] + \sum_{n=n_0}^{\infty}(1-\rho)\rho^n \mathbb{P}\left[B_0 > \frac{\delta x}{3n}, \mathcal{A}_k^{(m)}\right] \\
&\triangleq I_{41} + I_{42}.
\end{aligned} \tag{1.9}$$

Here, it is easy to see that

$$\begin{aligned}
I_{41} &\leqslant n_0^2 \mathbb{P}\left[B_1^{(e)}(m) > \frac{\delta x}{3n_0}, \mathcal{A}_k^{(m)}\right] \mathbb{P}\left[B_0 > \frac{\delta x}{3n_0}, \mathcal{A}_k^{(m)}\right] \\
&= o\left(\mathbb{P}\left[B_0 > x, \mathcal{A}_k^{(m)}\right]\right).
\end{aligned} \tag{1.10}$$

Furthermore, since

$$s \triangleq \sup_{x \in [0,\infty)} \frac{\mathbb{P}\left[B_0 > x, \mathcal{A}_k^{(m)}\right]}{\mathbb{P}\left[B_0 > 2x, \mathcal{A}_k^{(m)}\right]} < \infty,$$

we obtain that, for any $\epsilon > 0, n \geq 1$, there exists $K_\epsilon > 0$ such that

$$\mathbb{P}\left[B_0 > \frac{\delta x}{3n}, \mathcal{A}_k^{(m)}\right] \leqslant s^{\lceil \log_2(n) \rceil} \mathbb{P}\left[B_0 > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \leqslant K_\epsilon (1+\epsilon)^n \mathbb{P}\left[B_0 > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right],$$

which, by choosing ϵ small enough with $\eta \triangleq \rho(1+\epsilon) < 1$, yields

$$\begin{aligned}
\sum_{n=n_0}^{\infty}(1-\rho)\rho^n \mathbb{P}\left[B_0 > \frac{\delta x}{3n}, \mathcal{A}_k^{(m)}\right] &\leqslant \sum_{n=n_0}^{\infty}(1-\rho)\rho^n K_\epsilon (1+\epsilon)^n \mathbb{P}\left[B_0 > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\
&\leqslant \frac{(1-\rho)K_\epsilon \eta^{n_0}}{1-\eta} \mathbb{P}\left[B_0 > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right].
\end{aligned} \tag{1.11}$$

By combining (1.9), (1.10) and (1.11), and then passing $n_0 \to \infty$, we obtain

$$I_4(x) = o\left(\mathbb{P}\left[B > \frac{x}{k+1}\right]^{k+1}\right) = o(I_1(x)),$$

which, in conjunction with (1.5), (1.6), (1.7), (1.8), and by passing $\delta \to 0$, yields

$$\mathbb{P}\left[V_0 > x, \mathcal{A}_k^{(m)}\right] \lesssim \mathbb{P}\left[B_0 > \frac{(1-\rho)x}{k+1}, \mathcal{A}_k^{(m)}\right]. \tag{1.12}$$

Next, we prove a *lower bound* for (1.1). Observe that within $\mathcal{A}_k^{(m)}$, we have

$$V_0 \geqslant B_0 + \sum_{i=1}^{m} B_{-i}^0 \wedge B_0 + \sum_{i=1}^{N(V_0)} B_i \wedge R_0(T_i) \geq (k+1)B_0 + mT_{-m} + \sum_{i=1}^{N(V_0)} B_i \wedge R_0(T_i), \quad (1.13)$$

where in the last inequality we applied $(x-y)\wedge z \geq x \wedge z - y$ for any $x, y, z \geqslant 0$; recall that $T_{-m} < 0$. Then, using the same arguments as in equation (3.11) in the proof of Theorem 2.1 in Jelenković and Momčilović (2003b), and the properties of $\mathcal{A}_k^{(m)}$, for $B_0 \in \mathcal{IR}$, we have

$$\mathbb{P}\left[V_0(1-\rho) > x, \mathcal{A}_k^{(m)}\right] \gtrsim \mathbb{P}\left[B_0 > \frac{x}{k+1}, \mathcal{A}_k^{(m)}\right]. \quad (1.14)$$

Combining (1.12) and (1.14) completes the proof of (1.1) for PS.

2. FBPS discipline. The proof is based on the sojourn time identity for FBPS

$$V_0 = B_0 + W_{B\wedge B_0}(T_0) + \sum_{i=1}^{N(V_0)} B_i \wedge B_0,$$

where $W_{B\wedge B_0}(T_n)$ denotes the stationary workload at T_n in a queue with Poisson arrival job sizes equal to $\{B_i \wedge B_0\}_{-\infty<i<n}$ and capacity 1; recall that $T_0 = 0$.

First, we establish an *upper bound*. Observe that within the set $\mathcal{A}_k^{(m)}$,

$$V_0 \leq (k+1)B_0 + (m-k)\tilde{B}_{k+1}^{(m)} + W_{B\wedge B_0}(T_{-m}) + \sum_{i=1}^{N(V_0)} B_i \wedge B_0,$$

which, for $0 < \delta < 1-\rho$, implies

$$\begin{aligned}\mathbb{P}\left[V_0(1-\rho-\delta) > x, \mathcal{A}_k^{(m)}\right] \leq{}& \mathbb{P}\left[(k+1)B_0 > (1-\delta)x, \mathcal{A}_k^{(m)}\right] + \mathbb{P}\left[(m-k)\tilde{B}_{k+1}^{(m)} > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\ &+ \mathbb{P}\left[W_{B\wedge B_0}(T_{-m}) > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] + \mathbb{P}\left[W_{B\wedge B_0}^{\rho+\delta} > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\ &\triangleq I_1(x) + I_2(x) + I_3(x) + I_4(x). \quad (1.15)\end{aligned}$$

Using the same arguments as in the proof of the upper bound for the PS case, we obtain

$$I_1(x) \sim \frac{1}{k+1}\binom{m}{k}\left(\mathbb{P}\left[B_0 > \frac{(1-\delta)x}{k+1}\right]\right)^{k+1}, \quad (1.16)$$

and similarly as in (1.7), (1.8), it follows that $I_2(x) = o(I_1(x)), I_3(x) = o(I_1(x)), I_4(x) = o(I_1(x))$. Therefore, by (1.15) and (1.16), we have

$$\mathbb{P}\left[V_0 > x, \mathcal{A}_k^{(m)}\right] \lesssim \mathbb{P}\left[B_0 > \frac{(1-\rho)x}{k+1}, \mathcal{A}_k^{(m)}\right]. \quad (1.17)$$

For a *lower bound*, within $\mathcal{A}_k^{(m)}$, we obtain

$$V_0 \geq (k+1)B_0 + T_{-m} + \sum_{i=1}^{N(V_0)} B_i \wedge B_0,$$

which is further lower bounded by the righthand side of (1.13). Combining (1.14) and (1.17) completes the proof of (1.1) for FBPS.

3. SRPT discipline. A similar sojourn time identity as in (1.4) can be derived for SRPT,

$$V_0 = B_0 + \sum_{i=1}^{L_m} B_i^{(e)}(0)\mathbf{1}\{B_i^{(e)}(0) \le B_0\} + \sum_{i=1}^{m} B_{-i}^0\mathbf{1}\{B_{-i}^0 \le B_0\} + \sum_{i=1}^{N(V_0)} B_i\mathbf{1}\{B_i < R_0(T_i)\},$$

where we use the convention that the jobs with earlier arrivals are served first in the case of equal remaining service times.

First, we prove a *lower bound*. For $l > 0$, define $B_{li} = 0$ for $i < 0$, and $B_{l0} = B_0$, $B_{li} = B_i\mathbf{1}(B_i \leqslant l)$ for $i > 0$. For the new queueing system with the arrival process $\{(T_i, B_{li})\}$, denote by $\{W_l(t)\}_{t\geqslant 0}$ the workload in the system without the labeled customer. Now, define the stopping time $T_{l0} \triangleq \inf\{t : R_0(t) \leqslant l\}$ and the corresponding residual capacity without the labeled customer $C(t) = \int_0^t \mathbf{1}(W_l(s) = 0)ds$. Clearly,

$$\mathbb{E}[C(t)] \sim (1-\rho_l)\,t \quad \text{as } t \to \infty, \tag{1.18}$$

where $\rho_l = \lambda\mathbb{E}[B\mathbf{1}(B \le l)] = \lim_{t\to\infty}\mathbb{P}[W_l(t) > 0]$. When $B_0 > l$, all the arrivals after time $T_0 = 0$ have shorter job requirements than the remaining service time of the labeled customer before time T_{l0}, and thus, the labeled customer can only receive service when there are no other customers present in the queue except itself. Therefore, conditional on $\{B_0 > l\}$, we have

$$C(T_{l0}) = B_0 - l. \tag{1.19}$$

Next, by the standard queueing stability result and (1.18), we have, for $\epsilon > 0$,

$$Z \triangleq \sup_{t\geq 0}\left(C(t) - (1-\rho_l+\epsilon)t\right) < \infty.$$

From (1.19), $V_{l0} \geqslant T_{l0}$ and the monotonicity of $C(t)$, we obtain, conditional on $\{B_0 > l\}$,

$$Z \geqslant C(V_{l0}) - (1-\rho_l+\epsilon)V_{l0} \geqslant B_0 - l - (1-\rho_l+\epsilon)V_{l0},$$

which, for large x, implies

$$\begin{aligned}\mathbb{P}\left[V_{l0} > x, \mathcal{A}_k^{(m)}\right] &\geqslant \mathbb{P}\left[B_0 > l, B_0 - l - Z > (1-\rho_l+\epsilon)x, \mathcal{A}_k^{(m)}\right]\\ &\geqslant \mathbb{P}\left[B_0 - l > (1+\epsilon)(1-\rho_l+\epsilon)x, \mathcal{A}_k^{(m)}\right] - \mathbb{P}\left[Z > \epsilon(1-\rho_l+\epsilon)x\right]. \end{aligned}\tag{1.20}$$

Furthermore, since the service requirements $\{B_{li}\}_{i\geqslant 1}$ are bounded by l, the busy period distribution of the corresponding workload $W_l(t)$ is exponentially bounded (e.g., see Nuyens and Zwart (2006); Palmowski and Rolski (2006)), implying that there exists $\delta > 0$, such that $\mathbb{P}[Z > x] = O(e^{-\delta x})$. This bound and (1.20), combined with Lemma 1.2 and $B \in \mathcal{IR}$, yield

$$\lim_{x\to\infty}\frac{\mathbb{P}\left[V_0 > x, \mathcal{A}_k^{(m)}\right]}{\mathbb{P}\left[B_0 > (1-\rho)x, \mathcal{A}_k^{(m)}\right]} \geqslant \lim_{x\to\infty}\frac{\mathbb{P}\left[B_0 > (1+\epsilon)(1-\rho_l+\epsilon)x, \mathcal{A}_k^{(m)}\right]}{\mathbb{P}\left[B_0 > (1-\rho)x, \mathcal{A}_k^{(m)}\right]}.$$

Passing $l \to \infty$, $\epsilon \to 0$ in the preceding inequality, we obtain the lower bound for SRPT.

For an *upper bound*, since the number of customers in system for SRPT at any time is not larger than the number of customers in system for any other rule applied on the same sequence of arrivals and service requirements, as shown by Schrage (1968), we use the stationary number of customers $L_m^{(PS)}$ at time T_{-m} in the corresponding PS queue to upper bound L_m. Furthermore, the workload W observed at time T_{-m} is an upper bound for the residual work R_i of a customer at time T_{-m}. Therefore,

$$V_0 \leq B_0 + \sum_{i=1}^{L_m^{(PS)}} W \wedge B_0 + \sum_{i=1}^{m} B_{-i}\mathbf{1}\{B_{-i} \leqslant B_0 - T_{-m}\} + \sum_{i=1}^{N(V_0)} B_i \wedge B_0,$$

which, for any $0 < \delta < 1 - \rho$, yields

$$\begin{aligned}
\mathbb{P}\left[V_0(1-\rho-\delta) > x, \mathcal{A}_k^{(m)}\right] \leqslant\ & \mathbb{P}\left[B_0 > (1-\delta)x, \mathcal{A}_k^{(m)}\right] \\
& + m\mathbb{P}\left[B_{-1}\mathbf{1}\{B_{-1} \leqslant B_0 - T_{-m}\} > \frac{\delta x}{3m}, \mathcal{A}_k^{(m)}\right] \\
& + \mathbb{P}\left[\sum_{i=1}^{L_m^{(PS)}} W \wedge B_0 > \frac{\delta x}{3m}, \mathcal{A}_k^{(m)}\right] + \mathbb{P}\left[W_{B\wedge B_0}^{\rho+\delta} > \frac{\delta x}{3}, \mathcal{A}_k^{(m)}\right] \\
& \triangleq I_1(x) + I_2(x) + I_3(x) + I_4(x). \qquad (1.21)
\end{aligned}$$

Similarly as in the proof of the upper bound for PS, we have

$$I_1(x) \sim \frac{1}{k+1}\binom{m}{k}\left(\mathbb{P}\left[B_0 > (1-\delta)x\right]\right)^{k+1} \qquad (1.22)$$

and

$$I_3(x) = o(I_1(x)), \quad I_4(x) = o(I_1(x)). \qquad (1.23)$$

The only difference, as compared to the PS case, is to evaluate $I_2(x)$. Noting that $\mathcal{A}_k^{(m)}$ is a subset of

$$\{B_0 \leqslant \tilde{B}_k\} = \bigcup_{1\leqslant i_1 < \cdots < i_k \leqslant m} \{B_{-i_1} \geqslant B_0, \cdots, B_{-i_k} \geqslant B_0\},$$

we obtain

$$\begin{aligned}
\frac{I_2(x)}{m} \leqslant\ & \mathbb{P}\left[B_{-1} > \frac{\delta x}{3m}, B_0 \leqslant \tilde{B}_k, B_{-1} < B_0\right] + \mathbb{P}\left[B_0 + \mid T_{-m} \mid \geqslant B_{-1} > \frac{\delta x}{3m}, B_0 \leqslant \tilde{B}_k, B_{-1} \geqslant B_0\right] \\
\triangleq\ & P_1 + P_2, \qquad (1.24)
\end{aligned}$$

where P_1 is derived by upper bounding the indicator function in $I_2(x)$ by 1. To estimate P_1, we use

$$\begin{aligned}
P_1 &\leqslant \binom{m-1}{k}\mathbb{P}\left[B_{-1} > \frac{\delta x}{3m}, B_0 > B_{-1}, \bigcap_{2\leqslant i \leqslant k+1} \{B_{-i} \geqslant B_0\}\right] \\
&\leqslant \binom{m-1}{k}\left(\mathbb{P}\left[B_{-1} > \frac{\delta x}{3m}\right]\right)^{k+2} = o\left(I_1(x)\right). \qquad (1.25)
\end{aligned}$$

Next, for $y \triangleq \delta x/(3m)$, it is easy to see

$$P_2 \leqslant \binom{m-1}{k-1}\mathbb{P}\left[B_0 + \mid T_{-m} \mid \geqslant B_{-1} > y, \bigcap_{1\leqslant i \leqslant k} \{B_{-i} \geqslant B_0\}\right],$$

where the preceding probability is further bounded by

$$\begin{aligned}
&\mathbb{P}\left[B_{-1} > y, B_0 \leqslant B_{-1} \leqslant B_0 + \sqrt{y}\right] \mathbb{P}\left[B_0 \geqslant y - \sqrt{y}\right]^{k-1} + \mathbb{P}[\mid T_{-m} \mid > \sqrt{y}] \\
&\leqslant (\mathbb{P}[B_{-1} \geqslant y, B_{-1} \leqslant B_0 + \sqrt{y}] - \mathbb{P}[B_{-1} \geqslant y, B_{-1} < B_0]) \\
&\quad \times \mathbb{P}\left[B_0 \geqslant y - \sqrt{y}\right]^{k-1} + me^{-\lambda\sqrt{y}/m}. \qquad (1.26)
\end{aligned}$$

Since $B_0, B_{-1} \in \mathcal{IR}$ and $\mathbb{P}[B_{-1} \geqslant y, B_{-1} \leqslant B_0 + \sqrt{y}] \lesssim \mathbb{P}\left[B_0 \geqslant y\right]^2 \sim \mathbb{P}[B_{-1} \geqslant y, B_{-1} < B_0]$, the right-hand side of inequality (1.26) is asymptotically equal to

$$o\left(\mathbb{P}\left[B_0 \geqslant y\right]^{k+1}\right) = o(I_1(x)),$$

which, in conjunction with (1.25) and (1.24), implies $I_2(x) = o(I_1(x))$. Finally, by replacing (1.22), (1.23) and the preceding estimation of $I_2(x)$ in (1.21), and then passing $\delta \to 0$, we finish the proof. □

2 Adaptive and Scalable Comparison Scheduling

Motivated by our conditional limits presented in Section 1, we propose a novel adaptive and scalable comparison scheduling scheme.

2.1 Comparison Splitting

In this section, we describe a new adaptive job classification mechanism that we term *comparison splitting*. The classification is based on relative size comparison of the arriving job to the previous m arrivals, $m \geq 1$. Specifically, if an arriving job is smaller than k and larger than $m-k$ of the previous m jobs, it is routed into class k, $0 \leqslant k \leqslant m$.

More formally, upon the arrival of job $i \geq 0$, we define $\tilde{B}_{i1} \geq \tilde{B}_{i2} \geq \cdots \tilde{B}_{im}$ to be the order statistics of $\{B_{i-m}, B_{i-m+1}, \cdots, B_{i-1}\}$ with $\tilde{B}_{i0} = \infty$ and $\tilde{B}_{i(m+1)} = 0$. Then, if $\tilde{B}_{i(k+1)} \leq B_i < \tilde{B}_{ik}$, the new arrival B_i is routed to class $k, 0 \leq k \leq m$ and the ith arrival in class k is denoted as $B_i^{(k)}$. In order to initiate the comparison splitting process, assume that $B_i, -m \leqslant i \leqslant -1$ are already known; otherwise, one can simply set $B_i \equiv 0, -m \leqslant i \leqslant -1$.

Here, we exemplify our splitting mechanism for $m = 3$ by dividing jobs into four classes S (small), M (medium), L (large) and XL (extra large) with the following rule,

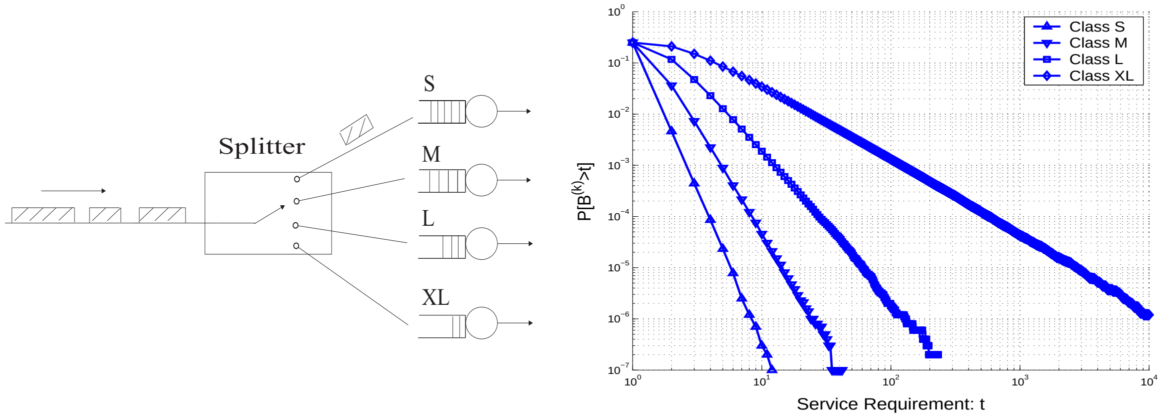
$$B_i \in \begin{cases} S & \text{if } B_i < \tilde{B}_{i3}, \\ M & \text{if } \tilde{B}_{i3} \leq B_i < \tilde{B}_{i2}, \\ L & \text{if } \tilde{B}_{i2} \leq B_i < \tilde{B}_{i1}, \\ XL & \text{if } \tilde{B}_{i1} \leq B_i; \end{cases}$$

this example is depicted in Figure 1 (A).

Now, we argue that our comparison splitting actually does order jobs into classes that contain smaller jobs for larger class indexes. Indeed, when $B \in \mathcal{L}$, Lemma 1.1 yields

$$\mathbb{P}\left[B_1^{(k)} > x\right] \sim \frac{1}{k+1}\binom{m}{k}\mathbb{P}[B > x]^{k+1} \quad \text{as } x \to \infty, \qquad (2.1)$$

which implies a decreasing distribution tail when k increases. Since the preceding expression is only an asymptotic result, it does not provide information on the possible ordering of the distributions $\mathbb{P}\left[B_1^{(k)} > x\right]$ for finite x. We address this question in the following example.



(A) Comparison splitter. (B) Job size distributions.

Figure 1: A comparison splitter with $m = 3$ and job size distributions for four different classes.

Example 1 In this example we simulate the performance of the comparison splitter for $m = 3$ (4 classes). Assume that the job sizes are distributed as power law $F(x) = 1 - 1/x^{\alpha}$ with $\alpha = 1.44$, which is the empirically measured file distribution by Jelenković and Momčilović (2003a); see Figure 1 on p. 577 therein. For a sample of 10^7 trials, we plot the simulated distributions of jobs for each class in Figure 1 (B). From the figure, it can be observed that the distributions $\mathbb{P}\left[B_1^{(k)} > x\right]$ are properly ordered for all values of x and k, not only for the asymptotic ones.

Based on the previous analysis and simulation example, we can see that our comparison splitter has the following advantages:

- it is adaptive since the comparing thresholds are defined by the preceding m arrivals;
- it is scalable because the system only needs to know the sizes of the previous m jobs;
- it provides accurate job classification as shown by equation (2.1) and Figure 1 (B).

Although our comparison splitter is very likely to provide a satisfactory ordering of distributions $\mathbb{P}\left[B_1^{(k)} > x\right]$, it may make errors on a sample path basis. Namely, it can occasionally classify smaller jobs into classes of smaller indexes and vice versa, and thus, possibly give a less accurate classification than a splitting mechanism that uses fixed thresholds. However, this possible small loss of accuracy is a fundamental tradeoff to gain the adaptability that is highly desirable in practice.

2.1.1 Refined Splitting

From the description of the comparison splitter, we can see that its adaptive thresholds are determined by the order statistics of the previous m arrivals. Thus, it is reasonable to expect that, at least for a stationary input, the accuracy of the classification will increase if we obtain these thresholds using a longer history (than the preceding m arrivals). However, the increase of history may reduce the adaptability and add to the complexity of the algorithm.

Here, we describe one such improved comparison splitter that is based on the order statistics of the preceding $ml, l \geq 1$ arrivals and parameterized by (m, l). Among other reasons, we continue to use the order statistics since the ordered list is easy to maintain dynamically. The splitter works as follows. At the time of arrival of a new job i, the algorithm maintains the job sizes of the previous ml arrivals, and orders them as $\tilde{B}_{i1} \geq \tilde{B}_{i2} \geq \cdots \geq \tilde{B}_{(i,ml)}$; when needed, we use the notation $\tilde{B}_{(i,j)} \equiv \tilde{B}_{ij}$ for improved clarity. We pick the subsequence $\tilde{B}_{(i,jl)}, 1 \leqslant j \leqslant m$ as the thresholds with $\tilde{B}_{(i,0)} = \infty, \tilde{B}_{(i,(m+1)l)} = 0$, and then, the new arrival is grouped into class k if its size lies in $[\tilde{B}_{(i,kl)}, \tilde{B}_{(i,(k-1)l)})$, $1 \leqslant k \leqslant m+1$.

In terms of engineering applications, this refined splitting algorithm is appealing because it can improve the accuracy for other types of arrivals, such as dependent processes and concentrated discrete distributions of job sizes. In order to measure how well the refined splitter classifies the input sequence, we compare the output of the refined splitter with a perfectly ordered input sequence. Denote the input sequence by $\{B_i\}_{1 \leqslant i \leqslant n}$, the output of the refined splitter by $\{O_i\}_{1 \leqslant i \leqslant n}$, and the increasing order of $\{B_i\}$ by $\{S_i\}_{1 \leqslant i \leqslant n}$. The output of the refined splitter $\{O_i\}$ is obtained by concatenating sequentially class $j+1$ after class j for all $1 \leqslant j \leqslant m-1$. Now, define the error rate to be

$$\eta(n) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{O_i \neq S_i\}.$$

Lemma 2.1 *For any fixed $0 < \epsilon < 1$, fixed m large enough, and an i.i.d. input sequence $\{B_i\}_{1 \leqslant i \leqslant n}$ taking finite number of values $\mathbb{P}[B_1 = b_j] = p_j, 1 \leqslant j \leqslant v$ with the splitter initialized by ml i.i.d. random copies of B_1 that are independent of $\{B_i\}_{1 \leqslant i \leqslant n}$, there exists $H_\epsilon, \xi_\epsilon > 0$, such that*

$$\mathbb{P}[\eta(n) > \epsilon] \leqslant H_\epsilon e^{-\xi_\epsilon \min(l,n)}.$$

The **proof** of Lemma 2.1 is presented in the appendix.

2.2 Queueing Analysis

In this section, we study the queueing performance of our comparison based scheduler assuming that jobs arrive according to a stationary renewal process $\{T_n\}$, $T_{-1} < 0 \leq T_0$ with finite mean $\mathbb{E}[T] < \infty$, where $T \overset{d}{=} T_1 - T_0$. The job sizes $\{B_n\}$ before the splitting are i.i.d and independent of $\{T_n\}$. To simplify the notation and analysis in this section, we say that the ith arrival to class k is equal to $B_i^{(k)} = B_i \mathbf{1}\{\tilde{B}_{i(k+1)} \leq B_i < \tilde{B}_{ik}\}$. This notation takes into account all the original arrival points even if $B_i \mathbf{1}\{\tilde{B}_{i(k+1)} \leq B_i < \tilde{B}_{ik}\} = 0$. The addition of zero size jobs in each class has no impact on queueing, but simplifies the exposition.

In Theorem 2.1, we characterize the workload asymptotics when each class is served in isolation. Then, in Theorem 2.2, we study the workload asymptotics of each individual class assuming that all the classes are served jointly according to a static priority discipline.

2.2.1 Queueing in Isolation

We first study the queueing characteristics of each class k when it is served in isolation with capacity c_k, $0 \leq k \leq m$. We use $W^{(k)}$ to denote the stationary workload of class k and define $B^{(k)} \overset{d}{=} B_1^{(k)}$.

Theorem 2.1 *If $\mathbb{P}[B > x] = l(x)/x^\alpha \in \mathcal{R}_\alpha$, $\alpha > 1$ and $\mathbb{E}[B^{(k)}] < c_k \mathbb{E}[T]$, then, as $x \to \infty$,*

$$\begin{aligned} \mathbb{P}\left[W^{(k)} > x\right] &\sim \frac{1}{c_k \mathbb{E}[T] - \mathbb{E}[B^{(k)}]} \int_x^\infty \mathbb{P}\left[B^{(k)} > u\right] du \\ &\sim \frac{1}{(k+1)(c_k \mathbb{E}[T] - \mathbb{E}[B^{(k)}])} \binom{m}{k} \frac{l(x)^{k+1}}{x^{\alpha k + \alpha - 1}}. \end{aligned}$$

Remark 4 Note that this theorem indicates that the workload distribution decays faster for larger k. To be more specific, the tail of the workload distribution for class k decays as $x(\mathbb{P}[B > x])^{k+1}$ and, thus, the jobs will have the waiting time distribution of the same order if served under FIFO. If, for example, each class were served according to PS/FBPS, one can expect that the waiting times will be of the same order as $(\mathbb{P}[B > x])^{k+1}$, as in our Theorem 1.2. However, this is much more difficult to prove because of the dependency in $\{B_n^{(k)}\}$.

Remark 5 Note that the result of Theorem 2.1 is of the same form as the one derived by Pakes (1975) for the $GI/GI/1$ queue. However, Pakes's result does not apply directly to our case since $\{B_n^{(k)}\}$ is m-dependent. For generalizations of Pakes's result to dependent processes see Jelenković and Lazar (1998); Asmussen et al. (1999). Note that, in principle, the approach from Asmussen et al. (1999) can be applied to prove our theorem. Instead, we present a direct proof that may be of independent interest.

In order to prove this theorem, we need the following definitions and lemmas. Define the partial sum of a stationary process $\{X_n\}_{n\in\mathbb{N}}$, where $X_n \in \mathcal{R}_\alpha$, as follows, $S_0 = 0$,

$$S_n = \sum_{i=1}^{n} X_i, \quad n \geq 1. \tag{2.2}$$

Definition 2.1 *For a stationary process $\{X_n\}_{n\in\mathbb{N}}$ and $m \in \mathbb{N}$, we say the process is m-dependent if X_n is independent of $\{X_i\}_{i<n-m}$ for all n.*

Lemma 2.2 *If we define*

$$S_n^{(m)} \triangleq \sum_{i=0}^{\lfloor \frac{n}{m} \rfloor} X_{im+1},$$

then

$$\mathbb{P}\left[\sup_{n\geq 0} S_n > x\right] \leqslant m\mathbb{P}\left[\sup_{n\geq 0} S_n^{(m)} > \frac{x}{m}\right].$$

Proof: Define

$$S_n^{(m,j)} = \sum_{i=0}^{\lfloor \frac{n}{m} \rfloor} X_{im+j},$$

where $1 \leqslant j \leqslant m$, and observe that $S_n \leq \sum_{j=1}^{m} S_n^{(m,j)}$. Therefore,

$$\begin{aligned}\mathbb{P}\left[\sup_{n\geq 0} S_n > x\right] &= \mathbb{P}\left[\sup_{n\geq 0} \sum_{j=1}^{m} S_n^{(m,j)} > x\right] \leqslant \mathbb{P}\left[\sum_{j=1}^{m} \sup_{n\geq 0} S_n^{(m,j)} > x\right] \\ &\leqslant \sum_{j=1}^{m} \mathbb{P}\left[\sup_{n\geq 0} S_n^{(m,j)} > \frac{x}{m}\right] \leqslant m\mathbb{P}\left[\sup_{n\geq 0} S_n^{(m)} > \frac{x}{m}\right],\end{aligned}$$

where the last equality follows from the stationarity of $\{X_n\}$. □

Lemma 2.3 *For a stationary m-dependent process $\{X_n\}_{n\in\mathbb{N}}$ with mean $\mathbb{E}X_1 = -\delta < 0$ and $X_1 \in \mathcal{R}_\alpha$, we have*

$$\mathbb{P}\left[\sup_{n\geqslant Hx} S_n > x\right] \leqslant \frac{1}{H^{\alpha-1}} O\left(\int_x^\infty \mathbb{P}[X_1 > u]du\right).$$

Proof: For simplicity of notation, in this section, we assume that $Hx \in \mathbb{N}$. Then, we define $M \triangleq \sup_{n \geqslant 0} S_n$ with $\mathbb{E}[X_n] = -\delta$, and note that

$$\sup_{n \geqslant Hx} S_n = S_{Hx} + \sup_{n \geqslant Hx} (S_n - S_{Hx}).$$

Since the process $\{X_n\}$ is stationary, we obtain

$$\sup_{n \geqslant Hx} (S_n - S_{Hx}) \overset{d}{=} M,$$

and therefore, $\mathbb{P}\left[\sup_{n \geqslant Hx} S_n > x\right]$ is upper bounded by

$$\begin{aligned}\mathbb{P}\left[S_{Hx} + \frac{\delta Hx}{2} + \sup_{n \geqslant Hx} (S_n - S_{Hx}) - \frac{\delta Hx}{2} > 0\right] &\leqslant \mathbb{P}\left[S_{Hx} + \frac{3\delta Hx}{4} > \frac{\delta Hx}{4}\right] + \mathbb{P}\left[M > \frac{\delta Hx}{2}\right] \\ &\triangleq I_1 + I_2.\end{aligned}$$

From the result of Pakes (1975) and Lemma 2.2, recalling that $X_1 \in \mathcal{R}_\alpha$, we have

$$I_2 \leqslant \frac{1}{H^{\alpha-1}} O\left(\int_x^\infty \mathbb{P}[X_1 > u] du\right). \tag{2.3}$$

Similarly, by defining $X_n^\delta = X_n + (3\delta)/4$ with the partial sum $S_n^\delta = \sum_1^n X_i^\delta$ and noting that $S_{Hx}^\delta \leq \sup_{n \geq 0} S_n^\delta$, we obtain

$$I_1 \leqslant \mathbb{P}\left[\sup_n S_n^\delta > \frac{\delta Hx}{4}\right] \leqslant \frac{1}{H^{\alpha-1}} O\left(\int_x^\infty \mathbb{P}[X_1 > u] du\right). \tag{2.4}$$

Combining (2.3) and (2.4) completes the proof. □

Proof of Theorem 2.1: By the classical result of Loynes (1962) (see also Chapter 2.2 of Baccelli and Bremaud (1994)), we have

$$W^{(k)} \overset{d}{=} \left(W^{(k)}(T_{-1}) + B_{-1}^{(k)} + c_k T_{-1}\right)^+,$$

where $W^{(k)}(T_{-1})$ is the stationary workload observed at the moment T_{-1}. Furthermore, $W^{(k)}(T_{-1}) \overset{d}{=} \sup_{n \geq 0} S_n$, with $S_n = \sum_{i=1}^n X_i, n \geq 1, S_0 = 0$ and $X_i \triangleq B_i^{(k)} - c_k(T_i - T_{i-1})$. Next, observe that for $x > 0$

$$\begin{aligned}\mathbb{P}[W^{(k)}(T_{-1}) > x] = \mathbb{P}\left[\sup_{n \geq 1} S_n > x\right] &\leq \mathbb{P}\left[\sup_{n \leqslant Hx} S_n > x\right] + \mathbb{P}\left[\sup_{n \geqslant Hx} S_n > x\right] \\ &\leq \mathbb{P}\left[\sup_{n \geq 1} \underline{S}_n^\epsilon > \delta x\right] + \mathbb{P}\left[\sup_{1 \leq n \leq Hx} \overline{S}_n^\epsilon > (1-\delta)x\right] + \mathbb{P}\left[\sup_{n \geqslant Hx} S_n > x\right] \\ &= I_1(x) + I_2(x) + I_3(x),\end{aligned} \tag{2.5}$$

where $\overline{X}_i^\epsilon = X_i \mathbf{1}\{X_i > \epsilon x\}$, $\underline{X}_i^\epsilon = X_i \mathbf{1}\{X_i \leqslant \epsilon x\}$, and $\overline{S}_n^\epsilon = \sum_{i=1}^n \left(\overline{X}_i^\epsilon + \mathbb{E}[X_i] + \delta\right)$, $\underline{S}_n^\epsilon = \sum_{i=1}^n \left(\underline{X}_i^\epsilon - \mathbb{E}[X_i] - \delta\right)$ are defined for some $\epsilon > 0, |\mathbb{E}[X_1]| > \delta > 0$.

First, let us prove an *upper bound* for (2.5). By Lemma 3.2(i) in Jelenković and Momčilović (2003b), for any $\beta > 0$, there exists $\epsilon > 0$ such that

$$I_1(x) = o(x^{-\beta}). \tag{2.6}$$

Furthermore, define $N_k = \sum_{i=1}^{Hx} \mathbf{1}\{\overline{X}_i^\epsilon > 0\}, 0 \le k \le m$; note that $\overline{X}_i^\epsilon$ depends on the class index k since $X_i = B_i^{(k)} - c_k(T_i - T_{i-1})$. To simplify the notation, we assume that Hx is an integer. Now, $\mathbb{P}[N_k \geqslant 2]$ is upper bounded by

$$\binom{Hx}{1}\mathbb{P}\left[B^{(k)} > \epsilon x\right]\binom{m-1}{1}\mathbb{P}\left[B > \epsilon x\right] + \binom{Hx}{2}\left(\mathbb{P}\left[B^{(k)} > \epsilon x\right]\right)^2 = o\left(x\left(\mathbb{P}[B > x]\right)^{k+1}\right).$$

In the preceding expression, the first term bounds the sum of probabilities $\mathbb{P}[\overline{X}_i^\epsilon > 0, \overline{X}_j^\epsilon > 0]$ for all indices $1 \le |i-j| \le m$ (note that in this case $\overline{X}_i^\epsilon$ and $\overline{X}_j^\epsilon$ are dependent); the second term provides a bound on the corresponding sum when $|i-j| > m$, using the fact that $\overline{X}_i^\epsilon$ and $\overline{X}_j^\epsilon$ are independent. Therefore,

$$\begin{aligned}
I_2(x) &\leqslant \mathbb{P}\left[\sup_{0\le n\le Hx} \overline{S}_n^\epsilon > (1-\delta)x, N_k = 1\right] + \mathbb{P}[N_k \geqslant 2] \\
&\le \sum_{n=1}^{Hx} \mathbb{P}\left[\overline{X}_i^\epsilon + n(\mathbb{E}[X_1] + \delta) > (1-\delta)x\right] + o\left(x\left(\mathbb{P}[B > x]\right)^{k+1}\right) \\
&\leqslant \int_0^\infty \mathbb{P}[X_1 > (1-\delta)x + u|\mathbb{E}[X_1] - \delta|]du + o\left(x\left(\mathbb{P}[B > x]\right)^{k+1}\right) \\
&\sim \frac{1}{|\mathbb{E}[X_1] - \delta|}\int_{(1-\delta)x}^\infty \mathbb{P}[X_1 > u]du. \qquad (2.7)
\end{aligned}$$

The estimate for $I_3(x)$ follows from Lemma 2.3. Using this estimate, (2.5), (2.6), (2.7) and passing $\delta, \epsilon \to 0$, $H \to \infty$, we obtain the upper bound.

Next, we prove the *lower bound* for (2.5)

$$\begin{aligned}
\mathbb{P}[W^{(k)}(T_{-1}) > x] &\ge \mathbb{P}\left[\sup_{1\le n\le Hx} S_n > x\right] \ge \mathbb{P}\left[\sup_{1\le n\le Hx} \overline{S}_n^\epsilon > x\right] \ge \mathbb{P}\left[\sup_{1\le n\le Hx} \overline{S}_n^\epsilon > x, N_k = 1\right] \\
&= \sum_{n=1}^{Hx} \mathbb{P}\left[\overline{X}_i^\epsilon + n(\mathbb{E}X_1 + \delta) > x\right] \ge \int_1^{Hx} \mathbb{P}\left[X_1 > x + u|\mathbb{E}X_1 - \delta|\right]du,
\end{aligned}$$

which by passing $x \to \infty$, using regular variation, and then passing $\delta \to 0$, results in

$$\mathbb{P}[W^{(k)}(T_{-1}) > x] \gtrsim \frac{1}{c_k\mathbb{E}[T] - \mathbb{E}[B^{(k)}]}\int_x^\infty \mathbb{P}\left[B^{(k)} > u\right]du. \qquad (2.8)$$

Finally, for any $0 < \epsilon < 1$, we have

$$\begin{aligned}
\mathbb{P}\left[W^{(k)} > x\right] &= \mathbb{P}\left[\left(W^{(k)}(T_{-1}) + B_{-1}^{(k)} + c_kT_{-1}\right)^+ > x\right] \\
&\leqslant \mathbb{P}\left[W^{(k)}(T_{-1}) > (1-\epsilon)x\right] + \mathbb{P}\left[B^{(k)} > \epsilon x\right],
\end{aligned}$$

which, by (2.8), and then passing $\epsilon \to 0$, yields

$$\mathbb{P}\left[W^{(k)} > x\right] \lesssim \mathbb{P}\left[W^{(k)}(T_{-1}) > x\right]. \qquad (2.9)$$

Also, since $W^{(k)}(T_{-1})$ is heavy-tailed and independent of T_{-1}, we obtain

$$\mathbb{P}\left[W^{(k)} > x\right] \geqslant \mathbb{P}\left[W^{(k)}(T_{-1}) + c_kT_{-1} > x\right] \sim \mathbb{P}\left[W^{(k)}(T_{-1}) > x\right]. \qquad (2.10)$$

Thus, (2.9) and (2.10) imply

$$\mathbb{P}\left[W^{(k)} > x\right] \sim \mathbb{P}\left[W^{(k)}(T_{-1}) > x\right], \tag{2.11}$$

which, combined with (2.8), completes the proof of the first asymptotics. The second asymptotic relationship of the theorem is implied by Lemma 1.1. □

2.2.2 Static Priority

In this subsection, we assume that there is only one server with capacity c and that the $m+1$ classes are served jointly with a preemptive static priority (SP) discipline between classes. Suppose that the priorities of the classes are assigned in a decreasing order of the class index k, $0 \leqslant k \leqslant m$, i.e., class k receives service only if classes $i, k+1 \leqslant i \leqslant m$ are empty. Denote by $W_0^{(k)}$ the stationary workload of class k observed at arrival point T_0. Let $\mu^{(k)} \triangleq \sum_{i=k}^{m} \mathbb{E}\left[B^{(i)}\right]$ and note that $\mu^{(0)} = \mathbb{E}[B]$.

Theorem 2.2 *If $\mathbb{P}[B > x] = l(x)/x^{\alpha} \in \mathcal{R}_{\alpha}$, $\alpha > 1$ and $\mathbb{E}[B] < c\mathbb{E}[T]$, then, as $x \to \infty$,*

$$\begin{aligned}\mathbb{P}\left[W_0^{(k)} > x\right] &\sim \frac{1}{c\mathbb{E}[T] - \mu^{(k)}} \int_x^{\infty} \mathbb{P}\left[B^{(k)} > u\right] du \\ &\sim \frac{1}{(k+1)(c\mathbb{E}[T] - \mu^{(k)})} \binom{m}{k} \frac{l(x)^{k+1}}{x^{\alpha k + \alpha - 1}}.\end{aligned}$$

Remark 6 This result shows that the distribution of the workload $W_0^{(k)}$ behaves asymptotically as if class k were served in isolation by a system with capacity reduced by the mean job sizes of classes with indices greater than k, which indicates a similar phenomenon as in Theorem 2.1. Thus, our SP scheduling with comparison splitter should approximate SRPT well.

Proof: Let $W^{(k)}(T_n)$ be the stationary workload of class k jobs at time T_n. First, we establish an *upper bound.* For $0 \leqslant k \leqslant m$, we group all the arrivals of classes $k, \cdots, m$ into a new class with the highest priority, while all the other classes remain the same. The workload of the new class is denoted as $\hat{W}^{(k)}(T_n)$, where $\hat{W}^{(k)}(T_n) \triangleq \sum_{i=k}^{m} W^{(i)}(T_n)$ and $\hat{W}_0^{(k)}$ represents a variable that is equal in distribution to $\hat{W}^{(k)}(T_n)$. Clearly,

$$W^{(k)}(T_n) \leqslant \hat{W}^{(k)}(T_n), \tag{2.12}$$

where the workload recursion for the new class satisfies

$$\hat{W}^{(k)}(T_{n+1}) = \left(\hat{W}^{(k)}(T_n) + \sum_{i=k}^{m} B_{n+1}^{(i)} - c(T_{n+1} - T_n)\right)^{+}.$$

Now, by Lemma 1.1, it is easy to see that, as $x \to \infty$,

$$\mathbb{P}\left[\sum_{i=k}^{m} B_{n+1}^{(i)} > x\right] \sim \mathbb{P}\left[B_{n+1}^{(k)} > x\right],$$

and, using the same argument as in the proof of the upper bound in Theorem 2.1, we obtain

$$\mathbb{P}\left[\hat{W}_0^{(k)} > x\right] \sim \frac{1}{c\mathbb{E}[T] - \mu^{(k)}} \int_x^{\infty} \mathbb{P}\left[B^{(k)} > u\right] du, \tag{2.13}$$

which, by (2.12), yields

$$\mathbb{P}\left[W_0^{(k)} > x\right] \lesssim \frac{1}{c\mathbb{E}[T] - \mu^{(k)}} \int_x^\infty \mathbb{P}\left[B^{(k)} > u\right] du. \tag{2.14}$$

Next, we prove a *lower bound.* For $\epsilon > 0$ and $k < m$, we have

$$\begin{aligned}\mathbb{P}\left[W_0^{(k)} > x\right] &\geqslant \mathbb{P}\left[W_0^{(k)} > x, \hat{W}_0^{(k+1)} \leqslant \epsilon x\right] \geqslant \mathbb{P}\left[\hat{W}_0^{(k)} > (1+\epsilon)x, \hat{W}_0^{(k+1)} \leqslant \epsilon x\right] \\ &\geqslant \mathbb{P}\left[\hat{W}_0^{(k)} > (1+\epsilon)x\right] - \mathbb{P}\left[\hat{W}_0^{(k+1)} > \epsilon x\right].\end{aligned}$$

Using the same argument as for (2.13) and passing $\epsilon \to 0$ in the preceding inequality imply

$$\mathbb{P}\left[W_0^{(k)} > x\right] \gtrsim \frac{1}{c\mathbb{E}[T] - \mu^{(k)}} \int_x^\infty \mathbb{P}\left[B^{(k)} > u\right] du.$$

The same asymptotic inequality can be easily shown to hold for $k = m$. This inequality, combined with (2.14), completes the proof of the first asymptotic relationship in Theorem 2.2. The second asymptotics follows directly from Lemma 1.1. □

3 Conclusion

We show in Theorem 1.2 that the medium size heavy-tailed jobs can have asymptotically much shorter sojourn times under SRPT than under PS/FBPS scheduling disciplines. Furthermore, the asymptotic performance of SRPT is uniformly good for the smaller as well as for the larger jobs, which implies that the performance gains of smaller jobs with SRPT, compared to PS/FBPS, are not achieved at the expense of larger jobs. Hence, in this asymptotic heavy-tailed context, SRPT is both efficient and fair, which complements similar findings obtained using the mean value analysis.

However, as early as in the paper by Schrage and Miller (1966), it was observed that SRPT may be difficult to implement because of its complicated preemptive nature that requires keeping track of the remaining processing times for all the jobs in the queue. Thus, it is natural to consider threshold-based static priority (SP) disciplines to approximate SRPT, as suggested originally by Schrage and Miller (1966), which was then followed by a considerable number of later studies. However, the main drawback of selecting static thresholds in practice is that the real world traffic is often nonstationary, highly correlated, bursty, etc.

Our second main contribution in this paper is the design of a scalable (low complexity) and adaptive comparison scheduling approximation to SRPT. The good performance of our comparison scheduler is demonstrated using our asymptotic queueing analysis under the heavy-tailed service requirements; additional verification of this scheduling algorithm was done by Jelenković et al. (2007) via simulations. We also discuss refinements of our mechanism that, at the expense of a small additional complexity, improve the accuracy of job classification for correlated arrivals and highly concentrated service distributions.

Finally, we would like to point out that, in addition to the static priority discipline analyzed in our paper, it may also be interesting to analyze the performance of our splitting mechanism for other disciplines, such as generalized processor sharing in Borst et al. (2003a), weighted fair queueing in Caprita et al. (2006), and hierarchical processor sharing.

Appendix

Proof of Lemma 1.1

Since the case $m = 0$ is immediate, we assume that $m \geq 1$. First, we show that the second asymptotics in (1.3) holds assuming that $\{X_i\}_{0 \leq i \leq m}$ are continuous. In this case, we have $\mathbb{P}[X_i = X_j] = 0, i \neq j$ and, thus

$$\mathbb{P}[X_0 > x, X_0 < \tilde{X}_k] = \binom{m}{k} \mathbb{P}\left[X_0 > x, X_0 \leq \min_{1 \leq i \leq k} X_i\right] = \frac{1}{k+1}\binom{m}{k}\mathbb{P}\left[\min_{0 \leq i \leq k} X_i > x\right]$$
$$= \frac{1}{k+1}\binom{m}{k}\mathbb{P}[X_0 > x]^{k+1}.$$

Next, the first asymptotics in (1.3) is implied by the preceding analysis and the following identity

$$\mathbb{P}[X_0 > x, \tilde{X}_{k+1} \leq X_0 < \tilde{X}_k] = \mathbb{P}[X_0 > x, X_0 < \tilde{X}_k] - \mathbb{P}[X_0 > x, X_0 < \tilde{X}_{k+1}].$$

If $\{X_i\}_{0 \leq i \leq m}$ are not continuous but in $\mathcal{L}$, (1.3) still holds asymptotically. This claim will follow from the preceding arguments if we show that for $X_i \in \mathcal{L}$, as $x \to \infty$,

$$\mathbb{P}[X_n > X_{n-1} \cdots > X_0 > x] \sim \mathbb{P}[X_n \geq X_{n-1} \cdots \geq X_0 > x]. \tag{3.1}$$

Since $X_i \in \mathcal{L}$, it is enough to prove the preceding relationship for $x \in \mathbb{N}$. Our proof starts with $n = 1$,

$$\mathbb{P}[X_0 > x, X_0 \leq X_1] = \mathbb{P}[X_0 > x, X_0 < X_1] + \mathbb{P}[X_0 > x, X_0 = X_1]. \tag{3.2}$$

Furthermore, for any $\epsilon > 0$ and x large,

$$\begin{aligned}\mathbb{P}[X_0 > x, X_0 = X_1] &= \sum_{y=x}^{\infty} \mathbb{P}[y < X_0 \leq y+1, X_0 = X_1, y < X_1 \leq y+1] \\ &\leq \sum_{y=x}^{\infty} (\mathbb{P}[y < X_0 \leq y+1])^2 \\ &= \sum_{y=x}^{\infty} \mathbb{P}[y < X_0 \leq y+1] \frac{\mathbb{P}[X_0 > y]}{\mathbb{P}[X_0 > y+1]} \mathbb{P}[X_0 > y+1] \\ &\quad - \sum_{y=x}^{\infty} \mathbb{P}[y < X_0 \leq y+1]\mathbb{P}[X_0 > y+1] \\ &\leq \epsilon \sum_{y=x}^{\infty} \mathbb{P}[y < X_0 \leq y+1]\mathbb{P}[X_0 > y+1] \qquad (3.3) \\ &\leqslant \epsilon(\mathbb{P}[X_0 > x])^2,\end{aligned}$$

where the last inequality is implied by the monotonicity of $\mathbb{P}[X_0 > y]$ and (3.3) follows from $X_0 \in \mathcal{L}$ since for any $\epsilon > 0$, we can choose x_0 such that for $y > x \geq x_0$,

$$\frac{\mathbb{P}[X_0 > y]}{\mathbb{P}[X_0 > y+1]} \leq 1 + \epsilon.$$

Combining (3.2), (3.3), using the fact that $\mathbb{P}[X_0 > x, X_0 < X_1]$ is of the same order as $(\mathbb{P}[X_0 > x])^2$, and passing $\epsilon \to 0$, yield the proof for $n = 1$. Now, for $n \geqslant 2$, we have

$$\begin{aligned}\mathbb{P}[X_n \geqslant X_{n-1} \cdots \geqslant X_0 > x] &= \mathbb{P}[X_n > X_{n-1} \geqslant \cdots \geqslant X_0 > x] + \mathbb{P}[X_n = X_{n-1} \geqslant \cdots \geqslant X_0 > x] \\ &\leqslant \mathbb{P}[X_n > X_{n-1} \geqslant \cdots \geqslant X_0 > x] + \mathbb{P}[X_n = X_{n-1} > x]\mathbb{P}[X_{n-2} \geqslant \cdots \geqslant X_0 > x] \\ &= \mathbb{P}[X_n > X_{n-1} \geqslant \cdots \geqslant X_0 > x] + o\left(\mathbb{P}[X_0 > x]^{n+1}\right),\end{aligned}$$

and by repeating the preceding procedure $n-1$ more times, we obtain

$$\mathbb{P}[X_n \geqslant X_{n-1} \cdots \geqslant X_0 > x] = \mathbb{P}[X_n > X_{n-1} > \cdots > X_0 > x] + o\left(\mathbb{P}[X_0 > x]^{n+1}\right).$$

Noting that $\mathbb{P}[X_n > X_{n-1} > \cdots > X_0 > x]$ is of the same order as $\mathbb{P}[X_0 > x]^{n+1}$ and $\mathbb{P}[X_n \geqslant X_{n-1} \cdots \geqslant X_0 > x] \geqslant \mathbb{P}[X_n > X_{n-1} > \cdots > X_0 > x]$, we finish the proof. □

Proof of Lemma 1.2

Let $R_{10}(t)$ and $R_{20}(t)$ be the remaining service times at time $t \geqslant 0$ for the labeled customer that arrives at T_0 under processes $\{(T_i, B_{1i})\}_{i > -\infty}$ and $\{(T_i, B_{2i})\}_{i > -\infty}$, respectively. By the same notion, we define $W_1(t)$ and $W_2(t)$ to be the workloads at time t in these two queues that need to be finished before the labeled customer can start receiving its service. In order to justify $V_1 \leqslant V_2$, it is enough to prove that $R_{10}(t) \leqslant R_{20}(t), t \geqslant 0$.

We use induction to prove the result and denote $\max(x, 0)$ by x^+. First, if $R_{10}(T_i) \leq R_{20}(T_i)$ and $W_1(T_i+) \leqslant W_2(T_i+)$, we have

$$\begin{aligned}W_1(t) = (W_1(T_i+) - (t - T_i))^+ &\leq (W_2(T_i+) - (t - T_i))^+ = W_2(t) \\ R_{10}(t) = R_{10}(T_i) - (t - T_i - W_1(T_i+))^+ &\leq R_{20}(T_i) - (t - T_i - W_2(T_i+))^+ = R_{20}(t) \qquad (3.4)\end{aligned}$$

for $T_i \leq t < T_{i+1}$. Note that $W_j(T_i+)$ and $W_j(T_i-)$ denote the right- and left-hand limits of $W_j(t)$ at T_i, respectively; i.e., the times right after and before the arrival at T_i. Hence, it is enough to prove that, all the customers arriving at T_i, $T_0 \leq T_i \leq V_1$, see $R_{10}(T_i) \leqslant R_{20}(T_i)$ and $W_1(T_i+) \leqslant W_2(T_i+)$ immediately after their arrival.

For the arrival at time T_0, the claim is obviously correct. Now, assuming that the result holds for $i = n$, we proceed to prove it for $i = n + 1$. Based on the hypothesis, (3.4) implies $R_{10}(T_{n+1}) \leqslant R_{20}(T_{n+1})$ and $W_1(T_{n+1}-) \leqslant W_2(T_{n+1}-)$ at the time immediately before T_{n+1}. Next, at time T_{n+1}, if $B_{1(n+1)} = 0 < B_{2(n+1)}$, then, we have

$$W_1(T_{n+1}+) = W_1(T_{n+1}-) \leq W_2(T_{n+1}-) + B_{2(n+1)}\mathbf{1}\left\{B_{2(n+1)} < R_{20}(T_{n+1})\right\} = W_2(T_{n+1}+),$$

since $W_1(T_{n+1}-) \leqslant W_2(T_{n+1}-)$.

The case $B_{1(n+1)} = B_{2(n+1)} = B_{n+1}$ results in the following three different scenarios:

1) If $B_{n+1} < R_{10}(T_{n+1})$, then

$$W_1(T_{n+1}+) = W_1(T_{n+1}-) + B_{n+1} \leqslant W_2(T_{n+1}-) + B_{n+1} = W_2(T_{n+1}+),$$

since $R_{10}(T_{n+1}) \leqslant R_{20}(T_{n+1})$ by induction hypothesis.

2) If $B_{n+1} > R_{20}(T_{n+1})$, then

$$W_1(T_{n+1}+) = W_1(T_{n+1}-) \leqslant W_2(T_{n+1}-) = W_2(T_{n+1}+).$$

3) If $R_{10}(T_{n+1}) \leqslant B_{n+1} \leqslant R_{20}(T_{n+1})$, then

$$W_1(T_{n+1}+) = W_1(T_{n+1}-) \leqslant W_2(T_{n+1}-) + B_{n+1} = W_2(T_{n+1}+).$$

Therefore, the result holds for $i = n+1$, which completes the induction, and implies that $V_2 \geqslant V_1$. □

Proof of Lemma 2.1

Without loss of generality we assume that $b_1 > \cdots > b_\nu$ and $\min\{p_k\}_{1\leqslant i\leqslant \nu} > 0$. Define $q_k \triangleq \sum_{i=1}^{k} p_i$, $1 \leqslant k \leqslant \nu$ with $q_0 = 0$ and choose $m > \min\{1/p_k\}_{1\leqslant k\leqslant \nu}$. When $B_i = b_k$, we say B_i is routed into the *right class* if B_i is either in class $\lfloor mq_{k-1} \rfloor$ or in class $\lceil mq_{k-1} - 1\rceil$ (note that if $mq_{k-1} \notin \mathbb{N}$, then $\lfloor mq_{k-1}\rfloor = \lceil mq_{k-1} - 1\rceil$). The condition $m > \min\{1/p_k\}_{1\leqslant k\leqslant \nu}$ guarantees that if $B_i \neq B_j$, then, the corresponding right classes for B_i and B_j are different since $mp_k > 1$ for all $1 \leqslant k \leqslant \nu$.

First, since both $\{O_i\}$ and $\{S_i\}$ are random, we construct a deterministic sequence $\{d_i\}_{1\leqslant i\leqslant n}$ for comparison purposes as follows: $d_i = b_k$, $\lfloor nq_{k-1}\rfloor + 1 \leqslant i \leqslant \lfloor nq_k \rfloor$. Then,

$$\begin{aligned}\mathbb{P}[\eta(n) > \epsilon] &\leqslant \mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}\{O_i \neq d_i\} > \frac{\epsilon}{2} n\right] + \mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}\{S_i \neq d_i\} > \frac{\epsilon}{2} n\right] \\ &\triangleq I_1 + I_2. \end{aligned} \tag{3.5}$$

For I_2, applying the union bound, we can easily prove that, for some $H, \xi > 0$,

$$I_2 \leqslant He^{-\xi n}. \tag{3.6}$$

Therefore, we only need to prove that $I_1 \leqslant He^{-\xi \min(l,n)}$, where H, ξ may be different from the ones chosen in (3.6).

Next, in order to evaluate I_1, we denote the event $\mathcal{E}_i = \{B_i \text{ is not in the right class}\}$ and prove that there exists $H, \xi > 0$, such that as $n \to \infty$,

$$\max_{1\leqslant i\leqslant \nu} \mathbb{P}[\mathcal{E}_i] \leqslant He^{-\xi l}. \tag{3.7}$$

To this end, if $\nu = 1$, it is obvious that $\mathbb{P}[\mathcal{E}_i] = 0$ for all i; if $\nu \geqslant 2$, noting that $\mathbb{P}[\mathcal{E}_i, B_i = b_1] = 0$, we have

$$\mathbb{P}[\mathcal{E}_i] = \sum_{k=1}^{\nu-1} \mathbb{P}[\mathcal{E}_i, B_i = b_{k+1}], \tag{3.8}$$

where $\mathbb{P}[\mathcal{E}_i, B_i = b_{k+1}]$ is upper bounded by

$$\begin{aligned} &\mathbb{P}\left[\left\{\tilde{B}_{(i,\lfloor mq_k + 1\rfloor l)} \leqslant b_{k+1} < \tilde{B}_{(i,\lceil mq_k - 1\rceil l)}\right\}^C\right] \\ &\quad\leqslant \mathbb{P}\left[b_{k+1} < \tilde{B}_{(i,\lfloor mq_k+1\rfloor l)}\right] + \mathbb{P}\left[b_{k+1} \geqslant \tilde{B}_{(i,\lceil mq_k - 1\rceil l)}\right] \\ &\quad= \mathbb{P}\left[\sum_{i=1}^{ml} \mathbf{1}\{B_i < b_{k+1}\} > \lfloor mq_k + 1\rfloor l\right] + \mathbb{P}\left[\sum_{i=1}^{ml} \mathbf{1}\{B_i < b_{k+1}\} \leqslant \lceil mq_k - 1\rceil l\right]. \end{aligned} \tag{3.9}$$

By noting that $\mathbb{E}[\mathbf{1}\{B_i < b_{k+1}\}] = q_k, 1 \leqslant k \leqslant \nu - 1$, and using the large deviation results with the condition $\lfloor mq_k + 1\rfloor > mq_k > \lceil mq_k - 1\rceil$, we obtain that for all $1 \leqslant k \leqslant \nu - 1$ and some

$H, \xi > 0$, the righthand side of (3.9) is further bounded by $He^{-\xi l}$. By substituting this upper bound for (3.9) into (3.8), we prove (3.7), and therefore, the total number of jobs

$$N_\epsilon \triangleq \sum_{i=1}^{n} \mathbf{1}\{\mathcal{E}_i\}$$

that are not in the right classes satisfies, for $0 < \delta < 1$ and some $H_\delta, \xi > 0$,

$$\mathbb{P}[N_\epsilon > \delta n] = \mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}\{\mathcal{E}_i\} > \delta n\right] \leqslant \frac{\mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}\{\mathcal{E}_i\}\right]}{\delta n} \leqslant H_\delta e^{-\xi l}. \tag{3.10}$$

Now, we continue with evaluating I_1. Since

$$I_1 \leqslant \sum_{k=1}^{\nu} \mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}\{O_i \neq d_i, O_i = b_k\} > \frac{\epsilon n}{2\nu}\right], \tag{3.11}$$

we only need to show that for each $1 \leqslant k \leqslant \nu$ and some $H_\epsilon, \xi_\epsilon > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}\{O_i \neq d_i, O_i = b_k\} > \frac{\epsilon n}{2\nu}\right] \leqslant H_\epsilon e^{-\xi_\epsilon \min(l,n)}.$$

To this end, we define $E_n^{(k)} \triangleq \sum_{i=1}^{n} \mathbf{1}\{O_i \neq d_i, O_i = b_k\}$ and denote by $N_k, 1 \leqslant k \leqslant \nu$ the total number of jobs of size b_k and by N_k^r the total number of jobs of size b_k that are routed into the right class with $N_0 = N_0^r = 0$. Obviously, by the definition of N_ϵ, we have $\left|\sum_{j=0}^{k}(N_j^r - N_j)\right| \leqslant N_\epsilon$ for $1 \leqslant k \leqslant \nu$. Now, we claim that, for $1 \leqslant k \leqslant \nu$,

$$E_n^{(k)} = \sum_{i=1}^{n} \mathbf{1}\{O_i \neq d_i, O_i = b_k\} \leqslant \left|\sum_{j=0}^{k-1} N_j^r - \lfloor nq_{k-1} + 1\rfloor\right| + \left|\sum_{j=0}^{k} N_j^r - \lfloor nq_k\rfloor\right| + 2N_\epsilon. \tag{3.12}$$

In order to prove (3.12), we define $\mathcal{R}_k \subset \{1, 2, \cdots, n\}$ to be the set of all the indices of the jobs in $\{O_i\}_{1 \leqslant i \leqslant \nu}$ that are routed to the right classes for job size b_k. Now, if there is no element i of $\mathcal{R}_k$ such that $O_i = b_k$, then the total number of jobs of size b_k in $\{O_i\}_{1 \leqslant i \leqslant \nu}$ is bounded by N_ϵ since none of the jobs of size b_k are in the right classes. Thus, in this case we obtain

$$E_n^{(k)} \leqslant \sum_{i=1}^{n} \mathbf{1}\{O_i = b_k\} \leqslant N_\epsilon.$$

Next, if $\mathcal{R}_k$ contains at least one index i such that $O_i = b_k$, we can always define $\tau_k = \min\{i \in \mathcal{R}_k : O_i = b_k\}$ and $\sigma_k = \max\{i \in \mathcal{R}_k : O_i = b_k\}$. Then, let $\mathcal{A} = \{i \in \mathcal{N} \mid \tau_k \leqslant i \leqslant \sigma_k\}$ and $\mathcal{B} = \{i \in \mathcal{N} \mid \lfloor nq_{k-1} + 1\rfloor \leqslant i \leqslant \lfloor nq_k\rfloor\}$. It is easy to see that all the indices in $\mathcal{A}$ but not in $\mathcal{B}$ are contributing to $E_n^{(k)}$ since $d_i \neq b_k$ for $i \in \mathcal{A}\backslash\mathcal{B}$, and therefore,

$$E_n^{(k)} \leqslant \mid \mathcal{A}\backslash\mathcal{B} \mid + N_\epsilon,$$

where N_ϵ contains all the errors $\mathbf{1}\{O_i \neq d_i, O_i = b_k\}$ for $i \notin \mathcal{R}_k$. Here, "\" represents set difference operation and $\mid \cdot \mid$ denotes the cardinality of a set. To compute the cardinality of the preceding set difference, we have the following four different scenarios.

- if $\tau_k \leqslant \lfloor nq_{k-1}+1 \rfloor$ and $\sigma_k \leqslant \lfloor nq_k \rfloor$, then $\mid \mathcal{A} \backslash \mathcal{B} \mid$ is upper bounded by $\lfloor nq_{k-1}+1 \rfloor - \tau_k$, which, by noting that $\sum_{j=0}^{k-1} N_j^r < \tau_k$, results in

$$E_n^{(k)} \leqslant \lfloor nq_{k-1}+1 \rfloor - \tau_k + N_\epsilon \leqslant \left| \sum_{j=0}^{k-1} N_j^r - \lfloor nq_{k-1}+1 \rfloor \right| + N_\epsilon;$$

- if $\tau_k \geqslant \lfloor nq_{k-1}+1 \rfloor$ and $\sigma_k \geqslant \lfloor nq_k \rfloor$, then $\mid \mathcal{A} \backslash \mathcal{B} \mid$ is upper bounded by $\sigma_k - \lfloor nq_k \rfloor$. By noting that $\sum_{j=0}^{k} N_j^r + N_\epsilon \geqslant \sigma_k$, we obtain

$$E_n^{(k)} \leqslant \sigma_k - \lfloor nq_k \rfloor + N_\epsilon \leqslant \sum_{j=0}^{k} N_j^r + N_\epsilon - \lfloor nq_k \rfloor + N_\epsilon;$$

- if $\lfloor nq_{k-1}+1 \rfloor \leqslant \tau_k \leqslant \sigma_k \leqslant \lfloor nq_k \rfloor$, then, $\mid \mathcal{A} \backslash \mathcal{B} \mid = 0$ and $E_n^{(k)}$ is upper bounded by the total number of jobs that are not in the right classes N_ϵ;

- if $\tau_k < \lfloor nq_{k-1}+1 \rfloor \leqslant \lfloor nq_k \rfloor < \sigma_k$, then, we obtain $\mid \mathcal{A} \backslash \mathcal{B} \mid \leqslant \lfloor nq_{k-1}+1 \rfloor - \tau_k + \sigma_k - \lfloor nq_k \rfloor$, which, by noting that $\sigma_k \leqslant \sum_{j=0}^{k} N_j^r + N_\epsilon$ and $\sum_{j=0}^{k-1} N_j^r < \tau_k$, yields

$$\begin{aligned} E_n^{(k)} &\leqslant \lfloor nq_{k-1}+1 \rfloor - \tau_k + \sigma_k - \lfloor nq_k \rfloor + N_\epsilon \\ &\leqslant \left| \sum_{j=0}^{k-1} N_j^r - \lfloor nq_{k-1}+1 \rfloor \right| + \sum_{j=0}^{k} N_j^r + N_\epsilon - \lfloor nq_k \rfloor + N_\epsilon. \end{aligned}$$

Therefore, by the above arguments, we prove the claim in (3.12).

Next, using (3.12), for $1 \leqslant k \leqslant \nu$, we derive

$$\begin{aligned} &\mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}\{O_i \neq d_i, O_i = b_k\} > \frac{\epsilon n}{2\nu} \right] \\ &\leqslant \mathbb{P}\left[\left| \sum_{j=0}^{k-1} N_j^r - \lfloor nq_{k-1}+1 \rfloor \right| + \left| \sum_{j=0}^{k} N_j^r - \lfloor nq_k \rfloor \right| + 2N_\epsilon > \frac{\epsilon n}{2\nu} \right] \\ &\leqslant \mathbb{P}\left[\left| \sum_{j=0}^{k-1} N_j - \lfloor nq_{k-1}+1 \rfloor \right| + \left| \sum_{j=0}^{k} N_j - \lfloor nq_k \rfloor \right| + 4N_\epsilon > \frac{\epsilon n}{2\nu} \right] \\ &\leqslant \mathbb{P}\left[\left| \sum_{j=0}^{k-1} N_j - \lfloor nq_{k-1}+1 \rfloor \right| > \frac{\epsilon n}{6\nu} \right] + \mathbb{P}\left[N_\epsilon > \frac{\epsilon n}{24\nu} \right] \\ &\quad + \mathbb{P}\left[\left| \sum_{j=0}^{k} N_j - \lfloor nq_k \rfloor \right| > \frac{\epsilon n}{6\nu} \right], \end{aligned}$$

which, by noting that $\mathbb{E}[N_j] = np_j$ for $1 \leqslant j \leqslant \nu$ and using Chernoff bound, (3.11) and (3.10), implies that $I_1 \leqslant He^{-\xi \min(l,n)}$ for some $H, \xi > 0$. Combining this bound, (3.5) and (3.6), we complete the proof. □

References

Anantharam, V. (1999). Scheduling strategies and long-range dependence. *Queueing Systems: Theory and Applications*, 33(1-3):73–89.

Asmussen, S., Schmidli, H., and Schmidt, V. (1999). Tail probabilities for non-standard risk and queueing processes with subexponential jumps. *Advances in Applied Probability*, 31(2):422–447.

Baccelli, F. and Bremaud, P. (1994). *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrence*. Springer Verlag.

Bansal, N. and Gamarnik, D. (2006). Handling load with less stress. *Queueing systems: Theory and Applications*, 54(1):45–54.

Bansal, N. and Harchol-Balter, M. (2001). Analysis of SRPT scheduling: investigating unfairness. In *Proceedings of ACM SIGMETRICS & Performance'01*, pages 279–290, Cambridge, MA.

Borst, S., Boxma, O., and Jelenković, P. (2003a). Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows. *Queueing Systems: Theory and Applications*, 43(4):273–306.

Borst, S. C., Boxma, O. J., Núñez-Queija, R., and Zwart, A. P. (2003b). The impact of the service discipline on delay asymptotics. *Performance Evaluation*, 54(2):175–206.

Caprita, B., Nieh, J., and Stein, C. (2006). Grouped distributed queues: distributed queue, proportional share multiprocessor scheduling. In *PODC'06: Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, pages 72–81, New York, NY, USA.

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

Jelenković, P. R., Kang, X., and Tan, J. (2007). Adaptive and scalable comparison scheduling. In *Proceedings of ACM SIGMETRICS'07*, volume 35, No.1, pages 215–226, San Diego, CA, USA.

Jelenković, P. R. and Lazar, A. A. (1998). Subexponential asymptotics of a Markov-modulated random walk with queueing applications. *Journal of Applied Probability*, 35(2):325–347.

Jelenković, P. R. and Momčilović, P. (2002). Resource sharing with subexponential distributions. In *Proceedings of IEEE INFOCOM'02*, volume 3, pages 1316–1325, New York, NY, USA.

Jelenković, P. R. and Momčilović, P. (2003a). Asymptotic loss probability in a finite buffer fluid queue with hetrogeneous heavy-tailed on-off processes. *Annals of Applied Probability*, 13(2):576–603.

Jelenković, P. R. and Momčilović, P. (2003b). Large deviation analysis of subexponential waiting times in a processor-sharing queue. *Mathematics of Operations Research*, 28(3):587–608.

Kleinrock, L. (1976). *Queueing Systems volume II: Computer Applications*. Wiley-Interscience.

Loynes, R. M. (1962). The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58:497–520.

Nuyens, M., Wierman, A., and Zwart, B. (2008). Preventing large sojourn times using SMART scheduling. *Operations Research*, 56(1):88–101.

Núñez-Queija, R. (2000). *Processor-Sharing Models for Integrated-Services Networks*. PhD thesis, Eindhoven University of Technology, the Netherlands.

Nuyens, M. and Zwart, B. (2006). A large-deviations analysis of the GI/GI/1 SRPT queue. *Queueing Systems: Theory and Applications*, 54(2):85–97.

Pakes, A. (1975). On the tails of waiting-time distributions. *Journal of Applied Probability*, 12:555–564.

Palmowski, Z. and Rolski, T. (2006). On the exact asymptotics of the busy period in GI/G/1 queues. *Advances in Applied Probability*, 38:792–803.

Park, K. and Willinger, W., editors (2000). *Self-similar Network Traffic and Performance Evaluation*. Wiley, New York.

Rai, I. A., Biersack, E. W., and Urvoy-Keller, G. (2005). Size-based scheduling to improve the performance of short TCP flows. *IEEE Network*, 19(1):12– 17.

Rai, I. A., Urvoy-Keller, G., Vernon, M. K., and Biersack, E. W. (2004). Performance analysis of LAS-based scheduling disciplines in a packet switched network. In *SIGMETRICS/Performance '04*, pages 106–117, New York, NY, USA.

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

Rawat, M. and Kshemkalyani, A. (2003). SWIFT: Scheduling in web servers for fast response time. In *Proceedings of the Second IEEE International Symposium on Network Computing and Applications*, page 15, Los Alamitos, CA, USA.

Schrage, L. E. (1968). A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16(3):687–690.

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

Squillante, M. S., Yao, D. D., and Zhang, L. (1999). Web traffic modeling and Web server performance analysis. *ACM SIGMETRICS Performance Evaluation Review*, 27(3):24–27.

Wierman, A. and Harchol-Balter, M. (2003). Classifying scheduling policies with respect to unfairness in an M/GI/1. In *Proceedings of ACM SIGMETRICS'03*, pages 238–249, San Diego, CA, USA.

Wolff, R. W. (1989). *Stochastic Modeling and Theory of Queues*. Prentice Hall.

Zwart, A. P. and Boxma, O. J. (2000). Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems*, 35(1-4):141–166.