Submitted to
manuscript (Please, provide the manuscript number!)

Shortest-Job-First Scheduling in Many-Server Queues with Impatient Customers and Noisy Service-Time Estimates

Jing Dong
Columbia University, 3022 Broadway, New York, NY 10027, jing.dong@gsb.columbia.edu

Rouba Ibrahim
University College London, 1 Canada Square, London E14 5AB, rouba.ibrahim@ucl.ac.uk

Size-based scheduling has been extensively studied, for more than 50 years, yet almost exclusively in single-server queues with infinitely patient jobs and perfectly known service times. Much less is known about its performance in multiserver queues, particularly under noisy service-time information. In this paper, we present the first theoretical analysis of the non-preemptive Shortest-Job-First (SJF) policy in multiserver queues with abandonment and noisy service-time estimates. In particular, we consider the $GI/GI/s+GI$ queue and demonstrate that, in the many-sever overloaded regime, performance in the noisy SJF queue is equivalent, asymptotically in steady state, to a non-preemptive two-class priority queue where customers with short predicted service times (below a threshold) are served without wait, and customers with long predicted service times (above a threshold) eventually abandon without service. We prove that the noisy SJF discipline asymptotically maximizes the system throughput, among all non-preemptive scheduling disciplines that use the noisy service-time information.

Key words: SJF, impatience, noisy service times, multiserver queues.

1. Introduction

In this paper, we study how to schedule customers based on noisy service-time estimates. In particular, we focus on analyzing system performance under the non-preemptive noisy **s**hortest-**j**ob-**f**irst (SJF) policy, which prioritizes the customer with the smallest service-time *estimate*, in multiserver

queues with customer abandonment. Size-based scheduling, e.g., SJF, has been extensively studied for over 50 years, yet almost exclusively in single-server queues and assuming perfectly known service times. In contrast, the performance analysis of SJF with estimation errors remains largely an open problem, even in the relatively simple single-server queueing model (Down 2019, Scully et al. 2021). We are, to the best of our knowledge, the first to derive theoretical analysis for SJF with noisy service-time information in a multiserver setting with impatient customers. Throughout this paper, we use SJF to denote the noisy SJF policy which relies on noisy point estimates of the service times, rather than assume full knowledge of the service times.

1.1. Size-Based Scheduling in Service Systems

Our focus in this paper is on service systems. In service systems, such as healthcare facilities or contact centers, service requests are typically processed by multiple agents working in parallel. Furthermore, queued customers who are waiting to be served by an available agent do not wait indefinitely for service. In other words, customers have finite patience times, and they abandon the queue if they have to wait for too long. Most importantly, the times needed to process customer service requests are typically not perfectly known before entry to service. For example, in the context of a call center, predictions of future call durations are notoriously imprecise (Ibrahim et al. 2016). Lastly, we assume that customers in service cannot be preempted, as per common practice.

For a realistic queueing-model representation of a service system, we consider here a multiserver queueing system with customer abandonment where only noisy information is available, ex-ante, about the service times, e.g., in the form of service-time predictions. It seems natural to exploit service-time information, if it is available, when making customer scheduling decisions. For example, it is well known that, with perfectly known service times, scheduling the shortest job first is typically effective in reducing the overall system congestion. However, there are relatively few papers which study system performance when scheduling is based on noisy service-time estimates, even in simple single-server queueing settings without abandonment; e.g., see Wierman and Nuyens (2008) and Scully et al. (2020, 2021). Since size-based scheduling holds great promise, in theory, there is a need to investigate how much of its superior performance extends to more general queueing models, e.g., models that are practically relevant to the design of service systems. That is the aim of this paper.

1.2. Size-Based Scheduling: A Long History

Size-based scheduling policies, such as non-preemptive SJF or preemptive **s**hortest-**r**emaining-**p**rocessing-**t**ime (SRPT), have received much attention in the literature due to their attractive properties, e.g., in minimizing the mean response time in the system. Indeed, a substantial body of literature demonstrates the superior performance of the SRPT policy in the $M/G/1$ queue (Schrage and Miller 1966, Lin et al. 2011). Moreover, it has been shown that, in the $GI/GI/1$ queue, SJF and SRPT yield asymptotically the same performance at fluid scale (Gromoll and Keutel 2012). Despite decades of research, the theoretical analysis of SRPT and SJF remains (almost) exclusively confined to single-server queues with infinitely patient customers and perfectly known service times. We believe that this can be attributed to two main reasons. First, much of the interest in the performance analysis of size-based scheduling stems, traditionally, from applications in the analysis and design of computer systems. In such systems, unlike in service systems, certain human-customer-related traits, such as having a finite patience, are not particularly relevant. Second, as we explain below, analyzing size-based scheduling in more general queueing models involves substantial technical challenges that are hard to address.

In broad terms, analyzing size-based scheduling policies is complicated because one must keep track of the service time of each customer in the queue, leading to a curse of dimensionality issue. While asymptotic analysis, e.g., under heavy traffic, generally allows for simpler descriptions of the system, it involves studying suitably scaled measure-valued system-state descriptors, which is technically challenging (Banerjee et al. 2020). In moving from a single server to multiple servers, the main technical challenge in analyzing SRPT and SJF is that multiserver queues are not order-conserving so that classical arguments in the single-server setting (e.g., the tagged job method) do not readily extend (see Section 4.2 in Grosof et al. (2018)). Allowing for customer abandonment complicates the analysis even further. Indeed, scheduling decisions in systems with abandonment is well-known to be difficult, because the optimal scheduling policy can be state-dependent and varies for different patience-time distributions (Puha and Ward 2019). Dong and Ibrahim (2021)

studies the asymptotic performance of SRPT in the $M/GI/s+GI$ queue, in steady state, with perfect service-time information. To overcome the analytical challenges, that paper relies on a coupling between the SRPT queueing model and an analytically tractable loss queue. With noisy service-time information, as we consider here, such coupling proofs do not readily extend because it is difficult to ensure a strict ordering of the sample paths across coupled systems. So, there is a need to develop alternative proof techniques.

1.3. This Paper's Contributions

Our theoretical results take steps towards filling some important gaps in the literature. First, the performance analysis of the SRPT or the SJF policies with estimation errors remains largely an open problem, even in the relatively simple single-server queueing models (Down 2019, Scully et al. 2021). Second, with the exceptions of Grosof et al. (2018), which considers infinitely patient customers, and Dong and Ibrahim (2021), which allows for finite patience times, there are no known theoretical results about the performance of the SRPT or the SJF policies in multiserver queues, even when service times are perfectly known in advance. In this paper, we are the first to derive theoretical results on the performance of non-preemptive SJF with noisy service time information in the $GI/GI/s+GI$ queueing model, in steady state. We assume that we have a noisy point estimate of the service time of each customer. We focus here on a many-server asymptotic mode of analysis and the overloaded regime, which is also known as the efficiency-driven regime. This regime is appropriate because queueing times are negligible, in large systems with abandonment, under moderate or light load, i.e., the critically-loaded or underloaded regimes (Garnett et al. 2002). In the many-server overloaded regime, a non-negligible proportion of customers abandon the queue. Thus, carefully designing the scheduling policy to optimize the throughput is crucial in this setting.

Here is our main theoretical contribution. Under a mild assumption on the monotonicity of the mean actual service time conditional on its (random) prediction, e.g., when predictions are based on a regression model, we prove a state-space collapse result in the many-server overloaded limit. In

particular, we demonstrate that system performance, in steady state, under SJF is asymptotically equivalent to performance in a non-preemptive two-class priority queue. In the limit, customers with small *predicted* service times (below a threshold) are served immediately, and customers with large predicted service times (above a threshold) are kept waiting until they abandon. The key to establishing this result is to show that the SJF policy arises as the limit of a sequence of finite multi-class priority policies, as the number priority classes increases to infinity. Since multi-class priority systems are analytically tractable, we can leverage results from the extant literature, in particular Atar et al. (2014), to derive performance measures in the SJF system. In particular, we show that, in the many-server overloaded limit, increasing the number of priority classes beyond two classes has an asymptotically negligible effect. This is a practically important result because implementing the SJF policy can be challenging in practice given that it involves dynamically rank ordering the service requirements of all customers in queue. In contrast, implementing a two-class priority scheduling is much simpler: It relies on a static classification of customers into two classes only, based on a (noisy) point estimate of the customer's service time. Importantly, we demonstrate that the two-class priority policy asymptotically maximizes the system throughput, among all non-preemptive policies that only utilize the service-time predictions. To wit, this includes blind scheduling policies, such as first-come-first-served and last-come-first-served, which do not exploit the service-time information at all. Thus, the non-preemptive SJF policy asymptotically maximizes system throughput as well.

Finally, we establish a monotonicity property on the asymptotic system throughput under a bivariate stochastic-order relation on the random pairs of actual and predicted service times. As a corollary, we show that, in the practically relevant case of lognormally-distributed service times, the higher the correlation between the actual and predicted service times, the higher the asymptotic throughput under SJF.

We supplement our theoretical analysis with a detailed numerical study where we explore, e.g., the accuracy of the two-class approximation as a function of the noise in the service-time prediction.

We find that the noisier the service-time predictions, the better the two-class approximation. Recalling that service-time predictions are typically very noisy in practice lends further support to the usefulness of our theoretical analysis.

The rest of this paper is organized as follows. In Section 2, we review the literature. In Section 3, we describe our model. In Section 4, we state and prove our main theorem by drawing the connection between the SJF queueing system and the two-class priority queue, through state-space collapse. In Section 5, we relate the accuracy of the service-time prediction to the throughput. In Section 6, we describe results from a numerical study. In Section 7, we draw conclusions. We relegate the proofs of standard results and additional numerical results to the appendix.

2. Literature Review

Size-based policies, such as SJF and SRPT, have been extensively studied for over 50 years, yet almost exclusively in single-server queues with infinitely patient jobs and perfectly know service times. For example, Schrage (1968) and Schrage and Miller (1966) demonstrate optimality properties of SRPT in the $M/G/1$ system. There is a notable stream of works that studies SRPT under heavy traffic (Down et al. (2009), Gromoll et al. (2011), Puha et al. (2015)). Scully et al. (2018) develops a unified framework to analyze several age-based scheduling policies. This framework enables the study of scheduling policies with non-monotone age-based index rules, which specify the order in which jobs are scheduled. Size-based scheduling with noisy service time information is rarely considered in the literature, even in the single-server setting. Notable exceptions are Wierman and Nuyens (2008), Mitzenmacher (2021), Scully et al. (2021), and Scully et al. (2020). All of these papers study single-server queues without customer abandonment, focus on the objective of minimizing mean response time or mean holding cost, and consider a specific form of estimation errors, e.g., bounded error sizes. For example, Wierman and Nuyens (2008) consider additive or multiplicative errors, where the estimate of a job of size s is within $[s-\sigma, s+\sigma]$ or $[s(1-\sigma), s(1-\sigma)]$. Scully et al. (2021) considers more general multiplicative errors, where the estimate of a job of size s is within $[\beta s, \alpha s]$ for $\alpha \geq \beta > 0$. Mitzenmacher (2021) considers a single classifier that predicts

whether the job is above a given threshold. We study multiserver queues with abandonment, focus on objective of maximizing system throughput, and consider a new quantification of estimation errors through a mild assumption on the conditional mean of the actual service time. This new error quantification covers the important case where job-size predictions are based on regression models. We measure the accuracy of the prediction by the positive quadratic dependence order, which is related to the correlation between actual and predicted service times when service times are lognormal.

With multiple servers, we know that SRPT is not necessarily optimal; e.g., see Leonardi and Raz (2007). An important reference is Grosof et al. (2018), which studies the performance of the SRPT policy in a multiserver queueing system with Poisson arrivals, general service times, and no abandonment, i.e., the $M/G/k$ system. In this setting, the SRPT policy is shown to achieve an asymptotically optimal mean sojourn time in the conventional heavy-traffic regime. Scully et al. (2020) extends this result and demonstrates that the Gittins policy is optimal, in heavy traffic, in the $M/G/k$ system. The Gittins policy adapts to any amount of available information about service times (including, e.g., having a noisy point estimate of the service time), and is equivalent to SRPT when the service times are fully known, but the index can be hard to compute in practice. To the best of our knowledge, Dong and Ibrahim (2021) is the first to derive theoretical results about the performance of SRPT in multiserver queues with abandonment. The results of that paper were incomplete in two main dimensions: (1) they are based on the unrealistic assumption that service times are perfectly known, and (2) they allow for preemptions which may be practically infeasible (this simplifies the analysis and enables the coupling proof in that paper). Here, we go beyond those two limiting assumptions. We also use a completely different proof technique based on fluid limits for many-server systems with a finite number of priority classes.

Overall, given those gaps in the literature, there is a need to investigate the extent to which the superior performance of SJF and SRPT continues to hold in multiserver queues where patience times are finite, and where service times may or may not be perfectly known.

3. Modelling Framework

In this section, we set the stage for our subsequent theoretical development by describing our modeling framework and defining our many-server asymptotic mode of analysis.

3.1. Model Description

We consider the $GI/GI/s+GI$ queueing system in steady state, i.e., we assume that the arrival process is a renewal process with continuously distributed interarrival times with mean $1/\lambda$, service times are independent and identically distributed (i.i.d.) continuous random variables with a cumulative distribution function (cdf) G, a probability density function (pdf) g, and mean $1/\mu$, and times to abandon are i.i.d. continuous random variables with a cdf F, a pdf g, and mean $1/\theta$. Let $\zeta_G(x) \coloneqq g(x)/(1-G(x))$ and $\zeta_F(x) \coloneqq f(x)/(1-F(x))$ denote the hazard rate functions for the service time and patience time, respectively. We define $M_G \coloneqq \sup\{x : G(x) < 1\}$ and $M_F \coloneqq \sup\{x : F(x) < 1\}$, and note that we allow M_G and M_F to equal infinity. As in Assumption 4.2 in Atar et al. (2014), whose fluid-limit results for finite-priority systems we exploit in our analysis, we make the following assumptions on the system primitives:

ASSUMPTION 1. *For the patience-time distribution, we have i)* $\sup_{x\in[0,M_F)} \zeta_F(x) < \infty$ *is bounded and ii)* $f(x) > 0$ *for* $x \in [0, M_F)$*. For the service-time distribution, we have* $\zeta_G(x)$ *is either bounded or lower semi-continuous on* $(0, M_G)$*.*

We consider the non-preemptive SJF policy. Specifically, a customer who, upon arrival, finds an empty server goes to service immediately. If all servers are busy at the arrival epoch, then the new arriving customer must join the queue. When a server becomes available, the customer in queue with the smallest predicted service time (point estimate) will be the next to begin service. Service preemptions are not allowed. Customers have finite patience times, generated at the arrival epoch of the customer. We assume that the arrival, service, and abandonment processes are mutually independent. We define the traffic intensity $\rho \coloneqq \lambda/s\mu$. Because abandonment is allowed in the system, it is not necessary to assume $\rho < 1$ for the system to reach a steady state.

3.2. Many-Server Overloaded Regime

We consider a sequence of $GI/GI/s_\lambda + GI$ queues, indexed by the arrival rate λ. We fix the traffic intensity in system λ to $\rho_\lambda = \lambda/(s_\lambda \mu) \equiv \rho > 1$, i.e., we consider an overloaded setting. We hold the service-time and patience-time distributions fixed, independently of λ, and let λ and s_λ increase without bound. Let U denote the cdf of the baseline interarrival time with mean 1. We assume the intarrival time of the λ-th system has cdf $U^\lambda(x) = U(\lambda x)$, i.e., the intarrival times are scaled down by λ. Let $A^\lambda(t)$ denote the number of arrivals in the λ-th system by time t and let $A(t)$ denote the baseline renewal arrival process. Under the scaling introduced above, we have

$$\frac{A^\lambda(\cdot)}{\lambda} = \frac{A(\lambda \cdot)}{\lambda} \to \eta(\cdot) \text{ as } \lambda \to \infty \text{ almost surely uniformly over compact time intervals,}$$

where $\eta(t) \coloneqq t$. For simplicity of exposition, we assume that the system starts empty at time 0.

Let S denote a generic service time and $\hat{S}$ denote the corresponding generic service-time prediction. We assume $\hat{S}$ is a continuously random variable and let h denote its pdf. In addition, let S_i and $\hat{S}_i$ denote the service time and predicted service time of the i-th arriving customer. We omit reference to the customer index when the index itself does not matter. We define

$$\mu(y) \coloneqq \mathbb{E}[S|\hat{S} = y, \text{Serv}], \tag{1}$$

to be the conditional mean actual service time conditional on the corresponding $\hat{S} = y$ and on the customer being served. We note that for any non-anticipative and non-preemptive scheduling policy where the scheduling decision is based on service time prediction only, we can write $\mu(y) = \mathbb{E}[S|\hat{S} = y]$ since conditional on the predicted service time $\hat{S}$, the actual service time is independent of whether the customer is served. Define the threshold, τ, satisfying

$$\lambda \cdot \mathbb{P}(\hat{S} \le \tau) \cdot \mathbb{E}[S|\hat{S} \le \tau] = \lambda \mathbb{E}[S\mathbf{1}(\hat{S} \le \tau)] = \lambda \int_0^\tau \mu(y) h(y) dy = s_\lambda, \tag{2}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. That is, we choose τ such that the total workload of customers with predicted service times smaller than or equal to τ matches the service capacity of the system. Our main result is a state-space collapse approximation of the system: We show

that the non-preemptive SJF queue is asymptotically equivalent to a two-class priority queue. The high-priority class, defined as jobs whose predicted service times are smaller than or equal to the threshold τ, has non-preemptive priority over the low-priority class, defined as jobs whose predicted service times are larger than τ. We emphasize that τ, as defined in (2), does not depend on λ since $s_\lambda/\lambda = 1/(\rho\mu)$ is held fixed under our scaling.

3.3. Positive Regression Dependence

Throughout this paper, we make the following assumptions about the distribution and accuracy of the service time estimate $\hat{S}$. Recall that $h(x)$ denotes the pdf of $\hat{S}$, and let $H(x)$ denote its cdf. We also define $M_H \coloneqq \sup\{x : H(x) < 1\}$.

ASSUMPTION 2. *The conditional mean actual service time $\mathbb{E}[S|\hat{S}=y]$ is increasing in y. For the service time estimate, we have $h(x) > 0$ for $x \in [0, M_H)$ and $\sum_{x\in[0,M_H)} h(x) < \infty$.*

To see when $\mathbb{E}[S|\hat{S}=y]$ is increasing in y, we recall the concept of positive regression dependence of the actual service time, S, on its predicted value, $\hat{S}$ (Lehmann 1966). Equivalently, we say that S is stochastically increasing in $\hat{S}$ (Barlow and Proschan 1975).

DEFINITION 1. S is said to be stochastically increasing in $\hat{S}$ or, equivalently, S is positive regression dependent on $\hat{S}$, if $\mathbb{P}(S > y|\hat{S}=x)$ is increasing in x for all y, where $x, y \geq 0$.

It immediately follows from this definition that $\mu(y)$ in (1) increases in y under positive regression dependence (this is a sufficient but not necessary condition). Positive regression dependence is a mild assumption on the random pairs $(S, \hat{S})$. For example, it is satisfied when service-time predictions are obtained based on a regression model, which is quite common in practice.

Example. If, for the i^{th} customer, $S_i = \hat{S}_i + \epsilon_i$ where the prediction $\hat{S}_i$ and the noise term ϵ_i are independent random variables, then S_i is stochastically increasing in $\hat{S}_i$. To show this, we note that

$$\mathbb{P}(S_i > y|\hat{S}_i = x) = \mathbb{P}(\hat{S}_i + \epsilon_i > y|\hat{S}_i = x) = \mathbb{P}(\epsilon_i > y - x|\hat{S}_i = x) = \mathbb{P}(\epsilon_i > y - x),$$

which is clearly increasing in x for a fixed value of y.

4. Many-Server Limits under SJF Scheduling

In this section, we study steady-state performance of the overloaded $GI/GI/s+GI$ queue under the non-preemptive SJF policy with noisy service-time estimates, in the many-server asymptotic limit. Specifically, we consider the asymptotic regime defined in Section 3.2. Our key theoretical contribution is to demonstrate that SJF asymptotically maximizes system throughput, among all non-preemptive scheduling policies that exploit the noisy service-time information. We also show that the steady-state performance under SJF is asymptotically equivalent to the performance in a two-class priority queue where customers with predicted service times below the threshold in (2) are given priority over customers whose predicted service times are above that threshold.

4.1. Throughput Maximization

Let $X^\lambda(t)$ and $Q^\lambda(t)$ denote the number of customers in the system and in queue at time t, in the λ-th system, respectively. Recall that $A^\lambda(t)$ denotes the number of arrivals by time t. We further denote $D^\lambda(t)$ as the number of departures from service by time t. We note that $D^\lambda(t)$ depends on the scheduling policy in the system.

The structure of our reasoning in this section is as follows. We begin by considering a properly defined two-class priority rule. We demonstrate (Lemmas 1 and 2) that this policy asymptotically maximizes the throughput in the system, among all non-preemptive policies that exploit the noisy service-time information. Then, we prove that the SJF policy can be defined as the limit of an appropriately chosen sequence of finite-class priority rules, as the number of classes increases to infinity. We also demonstrate that these finite-class priority rules asymptotically maximize the throughput in the system as well (Lemma 3). Thus, we conclude our main result that SJF asymptotically maximizes the throughput (Theorem 1).

4.1.1. Two-class priority policy. The first step in our analysis is to study the asymptotic throughput in a system operating under the following two-class priority scheduling rule, which we denote by π_0. Under π_0, all customers with $\hat{S} \le \tau$, for τ in (2), are given high non-preemptive priority, and the remaining customers, i.e., the ones with $\hat{S} > \tau$, are given low priority. In Lemma 1, we derive the asymptotic throughput under π_0.

LEMMA 1. *Under Assumption 1, for the sequence of systems under π_0, we have*

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_0}(t)]=\lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_0}(t)]=\int_0^\tau h(y)dy.$$

Proof. From Theorem 4.4 in Atar et al. (2014), we have

$$\lim_{\lambda\to\infty}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_0}(t)]=\lim_{\lambda\to\infty}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_0}(t)]=\delta$$

for some $\delta\geq 0$ which is characterized by the unique invariant state of the unique fluid-model equations. There remains to characterize the value of δ. We do so next by characterizing the corresponding invariant state of the fluid limit. Note that, under the non-preemptive two-class priority rule π_0, the higher priority class, class 1, has $\rho_1=1$. Let δ_1 denote the departure rate of class 1 fluid in the invariant state, and δ_2 denote the departure rate of class 2 fluid. From Theorem 3.3 in Atar et al. (2014), we have that

$$\delta_1=\int_0^\tau h(y)dy \quad\text{and}\quad \delta_2=0.$$

Then, $\delta=\delta_1+\delta_2=\int_0^\tau h(y)dy$. ∎

For any scheduling policy π which exploits the noisy service-time information, we define

$$\bar{\text{Th}}^{\pi}:=\limsup_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi}(t)]$$

as the asymptotic throughput in the system. In Lemma 2, we demonstrate that the throughput under any non-preemptive scheduling policy which exploits the noisy service-time information is upper bounded by the throughput under π_0, i.e., π_0 maximizes the asymptotic throughput.

LEMMA 2. *Under Assumption 2, for any non-preemptive non-anticipative scheduling policy π using the noisy service-time information,*

$$\bar{Th}^{\pi}\leq\int_0^\tau h(y)dy.$$

Proof. Let $\{t_n\}_{n\geq 1}$ denote the subsequence for which

$$\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)]=\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi}(t)].$$

Note that the above limit exist, because $D^{\lambda,\pi}(t)\leq A^{\lambda}(t)$. Let $\hat{\gamma}^{\lambda}(y)$ denote the long-run average probability of getting service when the customer's predicted service time $\hat{S}=y$ along the subsequence $\{t_n\}_{n\geq 1}$. Then,

$$\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)]=\lambda\int_0^{\infty}\hat{\gamma}^{\lambda}(y)h(y)dy.$$

We also note that

$$\left(\lambda\int_0^{\infty}\hat{\gamma}^{\lambda}(y)h(y)dy\right)\left(\frac{\int_0^{\infty}\mu(y)\hat{\gamma}^{\lambda}(y)h(y)dy}{\int_0^{\infty}\hat{\gamma}^{\lambda}(y)h(y)dy}\right)\leq s^{\lambda}\text{ and }\hat{\gamma}^{\lambda}(y)\in[0,1]. \tag{3}$$

Based on (3), $\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)]$ is upper bounded by λ multiplied by the objective of the following optimization problem

$$\begin{aligned}&\max_{\gamma}\int_0^{\infty}\gamma(y)h(y)dy\\&\text{s.t.}\int_0^{\infty}\mu(y)\gamma(y)h(y)dy\leq\frac{s^{\lambda}}{\lambda}=\int_0^{\tau}\mu(y)h(y)dy\\&\gamma(y)\in[0,1]\end{aligned} \tag{4}$$

From the first constraint in (4), since $\mu(y)$ is increasing in y, we have

$$\int_0^{\tau}\mu(y)\gamma(y)h(y)dy+\mu(\tau)\int_{\tau}^{\infty}\gamma(y)h(y)dy\leq\int_0^{\infty}\mu(y)\gamma(y)h(y)dy\leq\int_0^{\tau}\mu(y)h(y)dy.$$

This further implies that

$$\mu(\tau)\int_{\tau}^{\infty}\gamma(y)h(y)dy\leq\int_0^{\tau}\mu(y)(1-\gamma(y))h(y)dy\leq\mu(\tau)\int_0^{\tau}(1-\gamma(y))h(y)dy.$$

Thus,

$$\int_{\tau}^{\infty}\gamma(y)h(y)dy\leq\int_0^{\tau}(1-\gamma(y))h(y)dy.$$

Rearranging the above inequality, we have

$$\int_0^{\infty}\gamma(y)h(y)dy\leq\int_0^{\tau}h(y)dy.$$

This implies that

$$\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)]\le\lambda\int_0^\tau h(y)dy.$$

Thus, $\bar{\mathrm{Th}}^{\pi}\le\int_0^\tau h(y)dy$. ∎

4.1.2. Finite-class priority policy. We now formulate the SJF policy as the limit of an appropriately defined sequence of finite-class priority rules, as the number of classes increases without bound. We consider the following sequence of class division rules based on the predicted service times, $\hat{S}$. At the 0-the level, we divide the customers into two classes: $\hat{S}\in[0,\tau]$ and $\hat{S}\in(\tau,\infty)$, where τ is defined in (2). At the n-th level, $n\ge1$, we divide the customers into $2^{2n}+1$ classes:

$$\left[0,\frac{1}{2^n}\tau\right],\left(\frac{1}{2^n}\tau,\frac{2}{2^n}\tau\right],\left(\frac{2}{2^n}\tau,\frac{3}{2^n}\tau\right],\ldots,\left(\frac{2^{2n}-1}{2^n}\tau,\frac{2^{2n}}{2^n}\tau\right],\left(\frac{2^{2n}}{2^n}\tau,\infty\right).$$

Let π_n denote the priority rule induced by the n-th level segmentation of priority classes. Let π_{SJF} denote the SJF policy. In Lemma 3, we prove that the limit of the sequence π_n exists and is unique, and that it must coincide with π_{SJF}.

LEMMA 3. *Under assumption 2, for the sequence of priority rules, $(\pi_n)_{n\ge0}$, we have $\lim_{n\to\infty}\pi_n$ exists and is uniquely defined to be π_{SJF}, i.e., the non-preemptive SJF policy.*

Proof. We map the set of priority classes of π_n to the interval $[0,1]$ by identifying each class $i\in\{1,2,\ldots,2^{2n}+1\}$ with the number

$$H\left(\frac{i-1}{2^n}\tau\right)\in\left\{0,H\left(\frac{1}{2^n}\tau\right),\ldots,H\left(\frac{2^{2n}}{2^n}\tau\right)\right\}\subset[0,1].$$

Let $\bar{C}\coloneqq\sum_{0<x<M_H}h(x)<\infty$. Since $0\le H\left(\frac{i}{2^n}\tau\right)-H\left(\frac{i-1}{2^n}\tau\right)=\int_{(i-1)\tau/2^n}^{i\tau/2^n}h(x)dx\le\bar{C}2^{-n}$ and $\bar{C}<\infty$,

$$\max_{0\le i\le2^{2n}+1}\left\{H\left(\frac{i}{2^n}\tau\right)-H\left(\frac{i-1}{2^n}\tau\right)\right\}\to0\text{ as }n\to\infty.$$

In addition, since $\lim_{n\to\infty}H(2^n\tau)\to1$, the limit is the continuous interval $[0,1]$. In the limit, priority classes are indexed by $[0,1]$, and for each $x\in[0,1)$, any class within $[0,x]$ has non-preemptive

priority over every class within $(x,1]$. In particular, for any two jobs, $\hat{S}_1$ and $\hat{S}_2$ with $0<\hat{S}_1<\hat{S}_2<\infty$, since $H(\hat{S}_1)<H(\hat{S}_2)$, $\hat{S}_1$ has a higher priority over $\hat{S}_2$, which is equivalent to the SJF rule. ∎

In the following lemma, we derive an expression for the asymptotic throughput under π_n.

LEMMA 4. *Under Assumption 1, for the sequence of $GI/GI/s_\lambda+GI$ systems under π_n, we have*

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_n}(t)]=\lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_n}(t)]=\int_0^\tau h(y)dy.$$

Proof. For the system operating under π_n, let $L^n=2^n$. From Theorem 4.4 in Atar et al. (2014), we have

$$\lim_{\lambda\to\infty}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_n}(t)]=\lim_{\lambda\to\infty}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_n}(t)]=\delta,$$

for some $\delta\geq 0$, which is characterized by the unique invariant state of the unique fluid-model equations. We next characterize the value of δ by characterizing the corresponding invariant state of the fluid limit. For the non-preemptive $(2^{2n}+1)$-class priority rule π_n, let ρ_i denote the traffic intensity of class i and δ_i denote the departure rate of class i fluid in the invariant state, $i=1,2,\ldots,2^{2n}+1$. Note that $\sum_{i=1}^{L^n}\rho_i=1$. Based on Theorem 3.3 in Atar et al. (2014), we have that for $i\leq L^n$,

$$\delta_i=\int_{((i-1)/2^n)\tau}^{(i/2^n)\tau}h(y)dy;$$

for $i>L^n$, $\delta_i=0$. Thus,

$$\delta=\sum_{i=1}^{2^{2n}+1}\delta_i=\sum_{i=1}^{2^n}\int_{((i-1)/2^n)\tau}^{(i/2^n)\tau}h(y)dy=\int_0^\tau h(y)dy.$$

∎

Combining Lemmas 4 and 2 implies that π_n's also maximize the asymptotic throughput among all non-preemptive scheduling policies that use the noisy service-time information.

4.1.3. Asymptotic throughput under SJF. We are now ready to study the asymptotic throughput under SJF, and prove that SJF indeed maximizes this throughput.

THEOREM 1. *Under Assumptions 1 and 2, π_{SJF} maximizes the asymptotic throughput. In particular,*

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{SJF}}(t)]=\lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{SJF}}(t)]=\int_0^\tau h(y)dy.$$

Proof. From Lemma 4, we have

$$\lim_{n\to\infty}\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_n}(t)]=\lim_{n\to\infty}\lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_n}(t)]=\int_0^\tau h(y)dy.$$

Since $D^{\lambda,\pi_n}(t)\le A^\lambda(t)$, by the Dominated Convergence Theorem and Lemma 3, we have

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{\mathrm{SJF}}}(t)]=\lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{\mathrm{SJF}}}(t)]=\int_0^\tau h(y)dy.$$

By Lemma 2, π_{SJF} maximizes the asymptotic throughput. ∎

4.2. Asymptotic Steady-State Performance

In this section, we explore the steady-state asymptotic performance in the system under SJF. The proof of Theorem 2 is similar to the proof of Theorem 1 in Dong and Ibrahim (2021); we relegate it to the appendices. Let $\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}}}$ denote the event that a "tagged" customer arriving to a random system state drawn from the system's steady-state distribution is served, and $W^\lambda_{\pi_{\mathrm{SJF}}}$ denote the customer's waiting time.

THEOREM 2. *Under Assumptions 1 and 2, for the sequence of $GI/GI/s_\lambda+GI$ queues under SJF, we have*

(a) $\lim_{\lambda\to\infty}\mathbb{P}(Serv^\lambda_{\pi_{SJF}}|\hat{S}\le\tau)=1$ *and* $\lim_{\lambda\to\infty}\mathbb{P}(Serv^\lambda_{\pi_{SJF}}|\hat{S}>\tau)=0$.

(b) $\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{SJF}}|\hat{S}\le\tau]=0$ *and* $\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{SJF}}|\hat{S}>\tau]=1/\theta$.

(c) $\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{SJF}}]=\frac{1}{\theta}\int_\tau^\infty h(y)dy$.

Theorem 2 demonstrates a state-space collapse result: In steady state, performance in the SJF queueing system is asymptotically indistinguishable from performance in a two-class priority queueing system where customers with short predicted service times, below τ, have non-preemptive

priority over customers with long predicted service times, above τ. This result has practical implications. Based on our asymptotic results, we conclude that, in large congested systems, there is demonstrably little advantage of implementing the SJF policy relative to a much coarser two-class priority policy. This is significant because the SJF policy is not easy to implement in practice, as it requires keeping track of the predicted service times of everyone in queue. In contrast, the coarse two-class priority rule is much simpler as it only requires classifying an incoming customer as having long or short predicted service time.

5. A Comparison of Prediction Models

In this section, our aim is to deepen our understanding into how an improvement in prediction accuracy for the service times translates into improvement in system performance when scheduling according to SJF. In particular, we establish a monotonicity property on the asymptotic throughput in the system under a bivariate stochastic order on the random pairs of actual and predicted service times (Theorem 3) . As a corollary, we show that, in the practically relevant case of lognormally-distributed service times, the higher the correlation between the actual and predicted service times, the higher the asymptotic throughput under non-preemptive SJF.

5.1. PQD Dependence Order

We let $\mathcal{F}(f_X, f_Y)$ denote the set of all bivariate distributions with the same marginal densities f_X and f_Y. Positive Quadrant Dependence (PQD) is a bivariate stochastic order that is defined as follows; see Chapter 9 in Shaked and Shanthikumar (2007).

Definition 2. (PQD order) Let (X_1, Y_1) and (X_2, Y_2) have joint complementary cumulative distribution functions (ccdf) $\bar{G}_1$ and $\bar{G}_2$ and the same univariate marginals, i.e., both in $\mathcal{F}(f_X, f_Y)$. Then,

$$(X_1, Y_1) \leq_{PQD} (X_2, Y_2) \quad \text{if, and only if,} \quad \bar{G}_1(x, y) \leq \bar{G}_2(x, y) \quad \text{for all } (x, y).$$

Theorem 3. *Let $(S, \hat{S}_1), (S, \hat{S}_2) \in \mathcal{F}(f, h)$. The following holds:*

$$\textit{If} \quad (S, \hat{S}_1) \leq_{PQD} (S, \hat{S}_2) \quad \textit{then} \quad \bar{Th}^{SJF}(S, \hat{S}_1) \leq \bar{Th}^{SJF}(S, \hat{S}_2).$$

Proof. Let $(S, \hat{S})$ denote a generic actual and predicted service-time pair. Define $\tilde{\mu}(y) \coloneqq \mathbb{E}[S|\hat{S}=y]$. Let $f(x|y)$ denote the conditional density of S given $\hat{S}=y$, and $F(x|y)$ and $\bar{F}(x|y)$ denote the corresponding cdf and ccdf, respectively. Recall that $h(y)$ is the pdf of $\hat{S}$ and $H(y)$ is its cdf. We also denote $\bar{H}(y)=1-H(y)$. Lastly, we denote $G(x,y)=\mathbb{P}(S\le x,\hat{S}\le y)$ as the joint cdf of S and $\hat{S}$ and $\bar{G}(x,y)=\mathbb{P}(S>x,\hat{S}>y)$. Note that

$$\tilde{\mu}(y)=\int_0^\infty xf(x|y)dx=\int_0^\infty \bar{F}(x|y)dx.$$

For a fixed threshold τ, the workload of the higher priority ($\hat{S}\le\tau$) and lower priority ($\hat{S}>\tau$) classes can be written as follows, where the exchange of integral is due to Tonelli's theorem:

$$\begin{aligned}\lambda\int_0^\tau \tilde{\mu}(y)h(y)dy&=\lambda\int_0^\tau\left(\int_0^\infty \bar{F}(x|y)dx\right)h(y)dy\\&=\lambda\int_0^\infty\left(\int_0^\tau \bar{F}(x|y)h(y)dy\right)dx=\lambda\int_0^\infty \mathbb{P}(S>x,\hat{S}\le\tau)dx\end{aligned}$$

and

$$\lambda\int_\tau^\infty \tilde{\mu}(y)h(y)dy=\lambda\int_0^\infty\left(\int_\tau^\infty \bar{F}(x|y)h(y)dy\right)dx=\lambda\int_0^\infty \mathbb{P}(S>x,\hat{S}>\tau)dx=\lambda\int_0^\infty \bar{G}(x,\tau)dx.$$

Next, we consider two different service-time predictions $\hat{S}_1$ and $\hat{S}_2$. We assume $\hat{S}_1$ and $\hat{S}_2$ have the same marginal distribution, h, and $\bar{G}_1(x,y)\le\bar{G}_2(x,y)$ for all (x,y), i.e., $(S,\hat{S}_1)\le_{PQD}(S,\hat{S}_2)$. Then, for any fixed τ,

$$\int_0^\infty \bar{G}_1(x,\tau)dx\le\int_0^\infty \bar{G}_2(x,\tau)dx.$$

We let τ_1 and τ_2 be the thresholds corresponding to the $\hat{S}_1$ and $\hat{S}_2$ service-time predictions, i.e., obtained using (2). Then, we note from (2) that

$$\lambda\mathbb{E}[S\mathbf{1}(\hat{S}\le\tau)]=\lambda\int_0^\infty \mathbb{P}(S\ge x,\hat{S}\le\tau)=\lambda\int_0^\infty\Big(\mathbb{P}(S\ge x)-\mathbb{P}(S\ge x,\hat{S}\ge\tau)\Big)dx=s_\lambda.$$

This leads to:

$$\int_0^\infty \bar{G}_1(x,\tau_1)dx=\int_0^\infty \bar{G}_2(x,\tau_2)dx=\mathbb{E}[S]-\frac{s_\lambda}{\lambda},$$

which must mean that $\tau_1<\tau_2$. Next, since $\hat{S}_1$ and $\hat{S}_2$ have the same marginal distribution h, we must have that $H(\tau_1)<H(\tau_2)$, which implies that $\bar{\mathrm{Th}}^{SJF}(S,\hat{S}_1)\le\bar{\mathrm{Th}}^{SJF}(S,\hat{S}_2)$. ∎

5.2. Correlation for Lognormal Service Times

We now consider lognormal service times, which arise a lot in practice. Let $r[X,Y]$ denote the correlation between random variables X and Y. We let Z, $\hat{Z}_1$, and $\hat{Z}_2$ denote normally-distributed random variables. We assume $\hat{Z}_1$ and $\hat{Z}_2$ have identical marginal distribution, but they can have different correlations with Z. Then, $(Z,\hat{Z}_1)$ and $(Z,\hat{Z}_2)$ each follow a bivariate normal distribution. This guarantees that, for $j=1,2$, if $r[Z,\hat{Z}_j]>0$, then $(S,\hat{S}_j)$ satisfies positive regression dependence as defined in Definition 1. We consider two sets of service-time predictions: $(S,\hat{S}_1)\stackrel{d}{=}(e^Z,e^{\hat{Z}_1})$ and $(S,\hat{S}_2)\stackrel{d}{=}(e^Z,e^{\hat{Z}_2})$.

LEMMA 5. *For $(S,\hat{S}_1)$ and $(S,\hat{S}_2)$ as defined above,*

$$r[Z,\hat{Z}_1]\le r[Z,\hat{Z}_2] \quad \textit{if, and only if,} \quad (S,\hat{S}_1)\le_{PQD}(S,\hat{S}_2).$$

Proof. Assume that $(S,\hat{S}_1)\le_{PQD}(S,\hat{S}_2)$. Note that the PQD order is preserved under monotonically increasing transformations, so that $(Z,\hat{Z}_1)\le_{PQD}(Z,\hat{Z}_2)$. By Lemma 1 in Wu et al. (2018), it follows that $r[Z,\hat{Z}_1]\le r[Z,\hat{Z}_1]$. For the converse, assume that $r[Z,\hat{Z}_1]\le r[Z,\hat{Z}_2]$. Then, by Lemma 3 of Wu et al. (2019) , it holds that $(Z,\hat{Z}_1)\le_{PQD}(Z,\hat{Z}_2)$. Thus, $(S,\hat{S}_1)\le_{PQD}(S,\hat{S}_2)$. ∎

Combining Theorem 3 and Lemma 5 implies that, with lognormal service times, there is an easy way to check which of several sets of service-time predictions leads to a higher asymptotic throughput under SJF. In particular, provided that these predictions have the same marginal distributions, one would only have to compute correlations with the actual service times: The higher the correlation, the higher the throughput. We also note that the assumption on having the same marginal distribution for alternative service-time predictions may be restrictive. Thus, we consider in Section 6 alternative service-time prediction models where this assumption does not hold, as a robustness check, and we reach consistent conclusions there.

6. Numerical Study

In this section, we describe results from a simulation study where we: (i) study the accuracy of the two-class priority approximation (Section 6.2); (ii) quantify the impact of increasing the number

of classes in the system (Section 6.3); and (iii) investigate the importance of selecting the right threshold in the two-class priority approximation (Section 6.4). We summarize our main results here, and relegate tables with detailed simulation results to the appendices.

6.1. Description of the Experiments

We simulate the $M/GI/s+M$ queueing system. For the arrival process, we consider a Poisson arrival process. We focus on overloaded systems where the arrival rate exceeds the total service rate. In particular, we consider values of the traffic intensity $\rho = 1.4$ and $\rho = 1.8$. For the number of servers, we consider values ranging from $s = 20$ to $s = 1000$. For each set of simulation results, we report point estimates of performance measures which are based on averaging across 10 independent simulation replications of length 1,000,000 arrivals each. For each point estimate, we calculate 95% confidence intervals, but we do not report these in the tables because we found them to be very narrow: The half widths are consistently below 0.05% of the corresponding point estimates.

We consider lognormally-distributed (actual) service times where we fix, without loss of generality, the mean service time to be equal to 1. Let S_i denote the actual service time of customer i. We let Z_i be a normally-distributed random variable with mean $-\ln(2)/2$ and variance $\ln(2)$. This makes $S_i = e^{Z_i}$ lognormally distributed with mean 1 and variance 1. Let α be a scalar such that $0 \leq \alpha \leq 1$, and we let $\hat{Z}_i(\alpha)$ and $\epsilon_i(\alpha)$ be such that

$$Z_i = \hat{Z}_i(\alpha) + \epsilon_i(\alpha), \tag{5}$$

where $\hat{Z}_i(\alpha)$ is normally distributed with mean $-\ln(2)/2$ and variance $\alpha \ln(2)$ and, independently of $\hat{Z}_i(\alpha)$, $\epsilon_i(\alpha)$ is normally distributed with mean 0 and variance $(1-\alpha)\ln(2)$. We define the service-time prediction $\hat{S}_i(\alpha) \coloneqq e^{\hat{Z}_i(\alpha)}$. We note that S_i and $\hat{S}_i$ defined in this manner satisfy positive regression dependence as defined in Definition 1. We emphasize that both the marginal distribution of $\hat{S}_i(\alpha)$ and the joint distribution of $(S_i, \hat{S}_i(\alpha))$ depend on α, i.e., they are not fixed, unlike the model in Section 5.2. We assume so deliberately because we would like to test the robustness of our results beyond the model in Section 5.2. Specifically, we vary α to alter the correlation between

Z_i and $\hat{Z}_i(\alpha)$, i.e., between the actual and predicted service times: Smaller values of α correspond to noisier predictions. We consider values of α ranging from $\alpha = 0.001$ ($r[Z_i, \hat{Z}_i(\alpha)] = 0.032$ and $r[S_i, \hat{S}_i(\alpha)] = 0.028$) to $\alpha = 0.98$ ($r[Z_i, \hat{Z}_i(\alpha)] = 0.99$ and $r[S_i, \hat{S}_i(\alpha)] = 0.99$).

In the appendices (Table 13, Table 14, and Figure 2), we also consider a model for $(S_i, \hat{S}_i)$ which is consistent with our description in Section 5.2. In particular, we fix the marginal distributions of S_i and $\hat{S}_i$ to be lognormally distributed with mean 1 and variance 1. We vary the correlation between Z_i and $\hat{Z}_i$ and consider values ranging from 0.005 to 0.99. The conclusions that we reach based on our simulation study are consistent under both service-time prediction models.

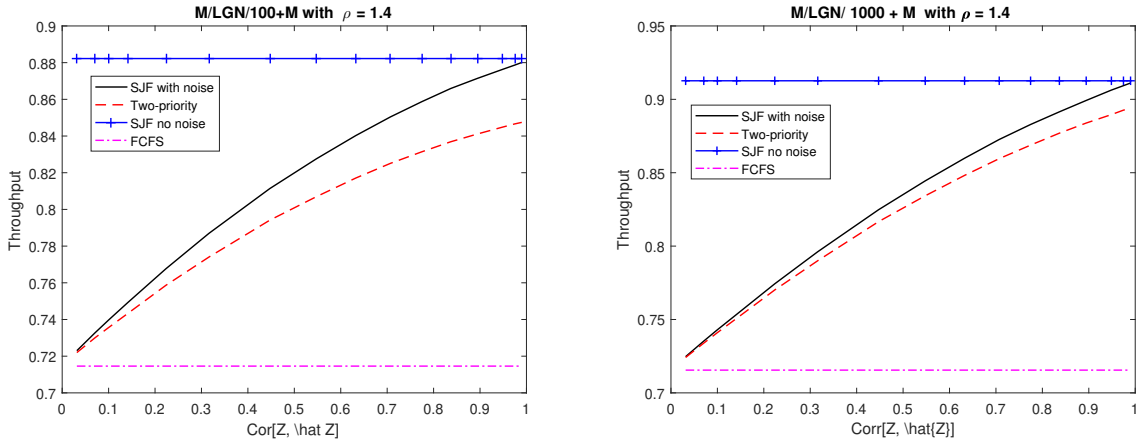

Figure 1 **Long-run throughput, in steady state, in the $M/LGN/n+M$ model.**

6.2. Accuracy of the Two-Priority Approximation

In this section, we study the accuracy of the approximations based on the limits in Theorem 2. We consider the $M/LGN/s+M$ system where we fix $\rho = 1.4$ and vary the number of servers $s = 20, 50, 70, 100, 500, 1000$. The limits in Theorem 2 hold as the number of servers increases without bound. Thus, we expect the accuracy of the corresponding approximations to be superior in large systems. Nevertheless, we deliberately consider small values of s too in order to validate the usefulness of those approximations in relatively small systems. We are also interested in quantifying the dependence of that accuracy on the degree of noise in the service-time prediction. The numerical results of this section are presented in Tables 1-6 of the appendices.

6.2.1. Size of the system. As expected, Tables 1-6 show that the accuracy of the approximations improves as the number of servers increases. For one example, when $\alpha = 0.3$, the relative errors in the throughput range from 2% for $s = 20$ to 1% for $s = 1000$. For another example, when $\alpha = 0.98$, those relative errors range from 5% for $s = 20$ to 2% for $s = 1000$. It is worth noting that the quality of the two-class priority approximations is reasonable for a relatively small number of servers (e.g., $s = 20$). This is important because it implies that our approximations are useful to describe system dynamics even when the system is not unrealistically large.

6.2.2. Noise in the service-time prediction. We now turn to investigating the impact of noise in the service-time predictions. While it is to be expected that noisier service-time predictions would lead to worse performance in the system under SJF, e.g., smaller throughput, the extent of this degradation in performance is not clear and is worthwhile investigating.

In Figure 1, we plot point estimates of the throughput in the $M/LGN/100 + M$ and $M/LGN/1000 + M$ models, as a function of the correlation between Z and $\hat{Z}(\alpha)$, for $\rho = 1.4$. We consider values of the correlation ranging from 0.03 ($\alpha = 0.001$) to 0.99 ($\alpha = 0.98$). We also include in the plots, as benchmarks, curves corresponding to the SJF policy assuming perfect knowledge of the service times and the first-come-first-served (FCFS) policy. We note that the throughput for SJF with no noise (top curve in the plot) and FCFS (bottom curve in the plot) are constant as a function of the correlation as they do not depend on the predicted service time. For the SJF and two-class priority rules with noisy service time information, when the service-time prediction is very noisy (low values of α), we expect performance in the system to be close to FCFS performance. In contrast, when the prediction is very accurate (high values of α), we expect the performance to be close to performance under SJF with perfect knowledge of service times. This is confirmed by the plots in Figure 1. Overall, the deterioration in throughput, as α decreases, is (loosely) upper bounded by the difference in throughput between FCFS and SJF with no noise, which is roughly equal to 20% for both $s = 100$ and $s = 1000$.

Interestingly, it is also apparent in Figure 1 that the quality of the two-class priority approximation *degrades* as the correlation increases. For both $s = 100$ and $s = 1000$, the relative errors

between the throughput in the SJF system with noise and the two-class priority system fall within a reasonable range, with the performance of the approximation degrading as the number of servers decreases. Indeed, for example, Table 4 shows that, for $s = 100$, the relative error in the throughput ranges from 0.14% for $\alpha = 0.001$ (first row) to 3.7% for $\alpha = 0.98$ (last row). This is practically meaningful because service-time predictions in service systems are usually not very accurate, which is when the two-class priority approximation is most accurate.

6.3. Increasing the Number of Classes

In Theorem 2, we derived limits of expected steady-state performance measures in the SJF system, and demonstrated that performance in this system is asymptotically equivalent to performance in a two-class priority queueing system. The SJF system itself can be thought of as an infinite-class priority queue where each customer defines their own priority class, based on their predicted service times. The asymptotic results in Theorem 2 hold as the size of the system increases without bound. For very large systems, our main result indicates that increasing the number of classes beyond two should have an almost negligible impact on performance. However, there remains to study the impact of increasing the number of classes on performance when the number of servers in the system is relatively small. We carry out that investigation in this section.

We consider here three different ways of increasing the number of classes, and we report point estimates of the probability of abandonment in each case. We report these results in Tables 7, 8, and 9. We fix $s = 100$ and let $\rho = 1.4$. We continue to consider a range of correlation values between Z and $\hat{Z}(\alpha)$ ranging from 0.03 to 0.99, i.e., α ranges from 0.001 to 0.98. In Table 7, we calculate the threshold τ as in (2). For the system with two classes, we let the high-priority class correspond to $\hat{S}(\alpha) \leq \tau$ and the low-priority class correspond to $\hat{S}(\alpha) > \tau$. To increase the number of classes, we make the high-priority class more granular by splitting it into equal intervals. For example, for a system with three classes, we consider three intervals for $\hat{S}$ with decreasing priority: $(0, \tau/2]$, $(\tau/2, \tau]$, and (τ, ∞). For a system with four classes, we consider four intervals for $\hat{S}$ with decreasing priority: $(0, \tau/3]$, $(\tau/3, 2\tau/3]$, $(2\tau/3, \tau]$ and (τ, ∞). In Table 8, we proceed in

a similar fashion, however we make the class $(0.8\tau, \tau]$ more granular instead. For example, for a system with three classes, we consider the following intervals for $\hat{S}$: $(0, 0.8\tau]$, $(0.8\tau, \tau]$ and (τ, ∞). For another example, for a system with four classes, we consider the following intervals for $\hat{S}$: $(0, 0.8\tau]$, $(0.8\tau, 0.9\tau]$, $(0.9\tau, \tau]$ and (τ, ∞). Finally, in Table 9, we proceed similarly by splitting the class $(\tau, 1.2\tau]$ in a similar fashion instead.

As can be seen from Table 7, we generally do not see a significant improvement in performance in going beyond two classes. This is especially true when the noise in the prediction is substantial (first rows of the table). This is because, as noted earlier, the two-class approximation is especially accurate where there is significant noise in the service-time prediction. In contrast, when service-time predictions are very accurate (last rows of the table), we do observe some performance improvement (e.g., in throughput) when we move to a more granular split of classes, especially in going from two to three classes. For example, with $\alpha = 0.98$ (last row), the probability of abandonment decreases from 15.3% to 13.8% in going from two to three classes. Beyond three classes, we consistently see only a small reduction in the probability of abandonment. We can make similar observations based on Tables 8 and 9 as well, which we do not discuss separately for brevity.

6.4. Choosing the Right Threshold in Two-Class Priority Rule

In Tables 10 - 12, we investigate the effect of choosing the right threshold, τ, on performance in the system. In particular, our aim is to quantify the performance improvement which results from selecting a threshold that accounts for the noise in the service-time prediction. To do so, we consider two systems. In the first system, we implement a two-class priority scheduling policy where customers with service-time prediction $\hat{S} \leq \tau_1$ are given non-preemptive priority over customers with $\hat{S} > \tau_1$. We calculate τ_1 by solving (2). In the second system, we consider a two-class priority system with a threshold τ_2 instead. We let τ_2 be the solution of the equation:

$$\lambda \mathbb{P}(S \leq \tau_2)\mathbb{E}[S|S \leq \tau_2] = s;$$

that is, we assume that τ_2 is calculated by assuming out any noise in the service-time prediction.

We consider $s = 100, 1000$ and $\rho = 1.4$ (Tables 10 and 11) and $s = 1000$ and $\rho = 1.8$ (Table 12). We observe that when service-time predictions are extremely noisy (small values of α), there is negligible advantage from implementing the correct threshold, i.e., based on (2). The reason is that performance in the system is close to performance under FCFS in this case since the service-time information is so noisy that it does not offer a significant advantage over a blind service policy which does not exploit the service-time information at all. On the other hand, when service-time predictions are extremely accurate (large values of α), there is also negligible advantage from implementing the correct threshold. In this case, the reason is that there is almost perfect knowledge of the service times, so the two thresholds, τ_1 and τ_2, are very close to each other. In contrast, we see significant improvement from implementing the correct threshold for moderate values of α. For example, for $s = 100$ and $\rho = 1.4$, Table 10 shows that the probability of abandonment reduces by almost 40% when implementing the correct threshold for $\alpha = 0.4$. We make consistent observations for all parameter values tested.

7. Conclusions

In this paper, we presented the first theoretical analysis of performance in an SJF queueing system with multiple servers, impatient customers, and noisy service-time predictions. We considered an overloaded regime and carried out a many-server asymptotic mode of analysis. We proved state-space collapse, and showed that steady-state performance measures converge to their counterparts in a non-preemptive two-class priority system where customers with short predicted service times (below a threshold) have priority over customers with long predicted service times (above a threshold).

We can glean managerial insights based on our theoretical results. Our main theorem implies that a service discipline which splits customers into just two priority classes can yield as good performance as SJF. This is practically important because implementing SJF can be quite challenging in practice, since it involves keeping track of the predicted service requirements and rank ordering every customer in the queue. The accuracy of this approximation is superior in large and congested

systems, and performs reasonably well in small systems too, as was substantiated in our numerical study. Thus, a manager may achieve the desired superior performance by implementing a coarse customer classification instead.

We also observed that the two-class priority approximation is especially accurate when there is a considerable amount of noise in the service-time prediction. This is practically important as this is typically the case in e.g., service systems such as hospitals and call centers. We leave providing theoretical evidence to substantiate this numerical observation to future research.

Relatedly, we investigated the impact of increasing the number of classes in a system with a finite number of servers. We found that having multiple classes (more than two) leads to a marginal improvement in performance when service times are very noisy, as two classes suffice in this case. In contrast, when service-time predictions are accurate, we can have a moderate gain in going from two to three customer classes, but the improvement in performance is slow beyond three classes. Thus, a manager is recommended to use three classes when service-time predictions are generally accurate, and two classes otherwise.

In terms of how to split customers into two classes, we investigated the importance of choosing the correct threshold. We found that there is not much advantage to accounting for noise in the prediction (in calculating the threshold) when the service-time prediction is either very noisy or very accurate. In these extreme cases, a manager may implement the two-class priority policy based on a threshold calculated using the actual service time distribution. However, when the service-time prediction is moderately noisy, the manager would greatly benefit from selecting the correct threshold, i.e., incorporating the noise, in the two-class priority system.

Appendix A: Proof of Theorem 2.

Proof. We begin by proving part (a). We note that the maximizer in problem (4) of Lemma 2 is given by $\gamma^*(y) = 1(y \le \tau)$. By Theorem 1,

$$\lim_{\lambda\to\infty} \mathbb{P}(\text{Serv}^{\lambda}_{\pi_{\text{SJF}}}) = \lim_{\lambda\to\infty} \frac{1}{\lambda} \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{\text{SJF}}}(t)] = \int_0^{\tau} h(y)dy = \int_0^{\infty} \gamma^*(y)h(y)dy = \mathbb{P}(\hat{S} \le \tau).$$

Thus, it must be that $\lim_{\lambda\to\infty} \mathbb{P}(\text{Serv}^{\lambda}_{\pi_{\text{SJF}}} | \hat{S} \le \tau) = 1$ and $\lim_{\lambda\to\infty} \mathbb{P}(\text{Serv}^{\lambda}_{\pi_{\text{SJF}}} | \hat{S} > \tau) = 0$.

We now prove part (b). Let T_c denote the patience time of a typical customer c, and V_c^λ and W_c^λ be the offered wait (if the customer has infinite patience) and the actual waiting time experienced by that customer, respectively, i.e., $W_c^\lambda = \min\{T_c, V_c^\lambda\}$. Then, for customers with $\hat{S} \leq \tau$:

$$\mathbb{P}(T_c > V_c^\lambda | \hat{S}_c \leq \tau) = \mathbb{P}(\text{Serv}_c^\lambda | \hat{S}_c \leq \tau) \to 1 \text{ as } \lambda \to \infty.$$

For a patience time distribution with $g(0) > 0$, e.g., exponential patience time, we have $(V_c^\lambda | \hat{S}_c \leq \tau) \Rightarrow 0$ as $\lambda \to \infty$. Because $0 \leq W_c^\lambda \leq V_c^\lambda$, $(W_c^\lambda | \hat{S}_c \leq \tau) \Rightarrow 0$. Next, as $W_c^\lambda \leq T_c$ and $\mathbb{E}[T_c] = 1/\theta < \infty$, by dominated convergence theorem,

$$\mathbb{E}[W_c^\lambda | \hat{S}_c \leq \tau] \to 0 \text{ as } \lambda \to \infty.$$

For customers with $\hat{S} > \tau$, by part (a),

$$\mathbb{P}(T_c \leq V_c^\lambda | \hat{S}_c > \tau) = \mathbb{P}(\text{Aband}_c^\lambda | \hat{S}_c > \tau) \to 1 \text{ as } \lambda \to \infty.$$

This implies that $(W_c^\lambda | \hat{S}_c > \tau) \Rightarrow T_c$. Because $W_c^\lambda \leq T_c$ and $\mathbb{E}[T_c] < \infty$, by dominated convergence theorem,

$$\mathbb{E}[W_c^\lambda | \hat{S}_c > \tau] \to \mathbb{E}[T_c] = 1/\theta \text{ as } \lambda \to \infty.$$

Lastly, part (c) follows from the fact that $\mathbb{E}[W_c^\lambda] = \mathbb{E}[W_c^\lambda | \hat{S}_c \leq \tau]\mathbb{P}(\hat{S}_c \leq \tau) + \mathbb{E}[W_c^\lambda | \hat{S}_c > \tau]\mathbb{P}(\hat{S}_c > \tau)$. ∎

Appendix B: Supporting Tables and Figures

In this appendix, we present tables and figures with numerical results that provide further support to the numerical study in Section 6. In particular, in Tables 1-6, we provide support to Section 6.2 by summarizing performance measures that quantify the accuracy of the two-priority approximation in the $M/LGN/s+M$ model with varying number of servers, s, and fixing $\rho = 1.4$. We let s range from $s = 20$ to $s = 1000$. In Tables 8 and 9, we provide additional support to Section 6.3 by presenting point estimates of the probability of abandonment under two alternative ways of increasing the number of classes in the $M/LGN/100+M$ model with $\rho = 1.4$. In Tables 10 - 12, we provide support to Section 6.4 by exploring the effect of selecting the wrong threshold in the two-priority classification. All the tables previously mentioned consider the service-time model of Section 6 of the paper. Finally, in Tables 13 and 14 and Figure 2, we present results corresponding to the service-time model of Section 5.2 in the paper.

			SJF			Two class		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.1278	7.8799	0.2820	0.1653	7.9186	0.2834
0.005	0.0636	0.0707	0.1274	7.6873	0.2753	0.1640	7.7899	0.2792
0.01	0.0877	0.1000	0.1275	7.5972	0.2714	0.1635	7.6878	0.2757
0.02	0.1151	0.1414	0.1269	7.4058	0.2643	0.1626	7.5570	0.2707
0.05	0.1899	0.2236	0.1263	7.0310	0.2510	0.1610	7.3198	0.2621
0.1	0.2688	0.3162	0.1246	6.6656	0.2383	0.1596	7.0796	0.2534
0.2	0.3894	0.4472	0.1227	6.1890	0.2214	0.1579	6.7627	0.2413
0.3	0.4767	0.5477	0.1213	5.8858	0.2099	0.1572	6.5568	0.2341
0.4	0.5691	0.6325	0.1192	5.6276	0.2006	0.1568	6.4022	0.2288
0.5	0.6413	0.7071	0.1175	5.3845	0.1921	0.1562	6.2689	0.2236
0.6	0.7154	0.7746	0.1165	5.2308	0.1865	0.1559	6.1534	0.2195
0.7	0.7878	0.8367	0.1147	5.0559	0.1807	0.1552	6.0390	0.2154
0.8	0.8602	0.8944	0.1132	4.9112	0.1753	0.1556	5.9817	0.2131
0.9	0.9333	0.9487	0.1123	4.8034	0.1713	0.1555	5.9217	0.2110
0.95	0.9641	0.9747	0.1116	4.7389	0.1694	0.1558	5.8850	0.2102
0.98	0.9866	0.9899	0.1112	4.7026	0.1677	0.1559	5.8922	0.2102

Table 1 **Accuracy of the approximations in Theorem 2 in the $M/LGN/20+M$ system with $\rho = 1.4$.**

			SJF			Two class		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0936	19.4651	0.2783	0.1420	19.5063	0.2791
0.005	0.0636	0.0707	0.0940	18.8228	0.2694	0.1401	18.9993	0.2722
0.01	0.0877	0.1000	0.0943	18.3953	0.2629	0.1392	18.6868	0.2676
0.02	0.1151	0.1414	0.0944	17.7968	0.2542	0.1379	18.2297	0.2610
0.05	0.1899	0.2236	0.0945	16.6173	0.2375	0.1356	17.3616	0.2484
0.1	0.2688	0.3162	0.0947	15.4222	0.2205	0.1335	16.4520	0.2352
0.2	0.3894	0.4472	0.0943	13.9603	0.1993	0.1308	15.3454	0.2191
0.3	0.4767	0.5477	0.0936	12.9847	0.1850	0.1294	14.6476	0.2089
0.4	0.5691	0.6325	0.0925	12.2118	0.1742	0.1282	14.0640	0.2006
0.5	0.6413	0.7071	0.0915	11.5702	0.1651	0.1275	13.6136	0.1940
0.6	0.7154	0.7746	0.0905	11.0284	0.1576	0.1271	13.2339	0.1885
0.7	0.7878	0.8367	0.0894	10.5843	0.1510	0.1266	12.9298	0.1842
0.8	0.8602	0.8944	0.0885	10.2153	0.1456	0.1269	12.7138	0.1809
0.9	0.9333	0.9487	0.0876	9.8765	0.1409	0.1267	12.4845	0.1779
0.95	0.9641	0.9747	0.0870	9.7176	0.1388	0.1266	12.3774	0.1768
0.98	0.9866	0.9899	0.0864	9.6256	0.1374	0.1264	12.3194	0.1757

Table 2 **Accuracy of the approximations in Theorem 2 in the $M/LGN/50+M$ system with $\rho = 1.4$.**

			SJF			Two class		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0807	27.1550	0.2773	0.1323	27.2593	0.2783
0.005	0.0636	0.0707	0.0811	26.2476	0.2681	0.1303	26.4921	0.2710
0.01	0.0877	0.1000	0.0815	25.6401	0.2614	0.1292	26.0129	0.2659
0.02	0.1151	0.1414	0.0815	24.7235	0.2522	0.1280	25.3404	0.2587
0.05	0.1899	0.2236	0.0821	23.0294	0.2348	0.1255	24.0218	0.2451
0.1	0.2688	0.3162	0.0826	21.2272	0.2164	0.1233	22.6268	0.2308
0.2	0.3894	0.4472	0.0825	18.9647	0.1934	0.1201	20.8776	0.2130
0.3	0.4767	0.5477	0.0822	17.4954	0.1783	0.1181	19.7510	0.2014
0.4	0.5691	0.6325	0.0819	16.3368	0.1666	0.1170	18.8728	0.1923
0.5	0.6413	0.7071	0.0815	15.3890	0.1570	0.1163	18.1527	0.1853
0.6	0.7154	0.7746	0.0808	14.6181	0.1490	0.1158	17.5697	0.1791
0.7	0.7878	0.8367	0.0804	13.9946	0.1425	0.1152	17.1118	0.1742
0.8	0.8602	0.8944	0.0795	13.4497	0.1371	0.1154	16.7522	0.1704
0.9	0.9333	0.9487	0.0787	12.9479	0.1320	0.1155	16.4372	0.1672
0.95	0.9641	0.9747	0.0778	12.6766	0.1297	0.1156	16.2680	0.1658
0.98	0.9866	0.9899	0.0777	12.5777	0.1284	0.1155	16.1964	0.1649

Table 3 **Accuracy of the approximations in Theorem 2 in the $M/LGN/70+M$ system with $\rho = 1.4$.**

			SJF			Two class		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0682	38.7296	0.2770	0.1212	38.8721	0.2780
0.005	0.0636	0.0707	0.0684	37.3598	0.2673	0.1194	37.6858	0.2700
0.01	0.0877	0.1000	0.0687	36.4471	0.2603	0.1181	36.9477	0.2644
0.02	0.1151	0.1414	0.0690	35.0869	0.2507	0.1167	35.9017	0.2567
0.05	0.1899	0.2236	0.0696	32.4679	0.2323	0.1140	33.7302	0.2413
0.1	0.2688	0.3162	0.0701	29.7250	0.2128	0.1110	31.5432	0.2260
0.2	0.3894	0.4472	0.0709	26.3390	0.1881	0.1078	28.7060	0.2059
0.3	0.4767	0.5477	0.0709	23.9961	0.1714	0.1061	26.9172	0.1930
0.4	0.5691	0.6325	0.0705	22.1714	0.1586	0.1048	25.5386	0.1829
0.5	0.6413	0.7071	0.0704	20.7567	0.1485	0.1040	24.4331	0.1749
0.6	0.7154	0.7746	0.0705	19.6973	0.1407	0.1035	23.5301	0.1686
0.7	0.7878	0.8367	0.0704	18.8018	0.1340	0.1030	22.7404	0.1631
0.8	0.8602	0.8944	0.0700	18.0100	0.128	0.1028	22.1475	0.1589
0.9	0.9333	0.9487	0.0698	17.3162	0.1236	0.1028	21.6270	0.1551
0.95	0.9641	0.9747	0.0695	16.9740	0.1211	0.1027	21.4076	0.1534
0.98	0.9866	0.9899	0.0692	16.7739	0.1198	0.1032	21.3123	0.1526

Table 4 **Accuracy of the approximations in Theorem 2 in the $M/LGN/100+M$ system with $\rho = 1.4$.**

			SJF			Two class		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0313	193.0422	0.2760	0.0760	193.5315	0.2766
0.005	0.0636	0.0707	0.0314	185.5035	0.2654	0.0743	186.7599	0.2673
0.01	0.0877	0.1000	0.0313	180.6657	0.2580	0.0733	182.0929	0.2603
0.02	0.1151	0.1414	0.0314	173.4318	0.2474	0.0721	175.6152	0.2509
0.05	0.1899	0.2236	0.0316	158.5673	0.2268	0.0693	162.1856	0.2322
0.1	0.2688	0.3162	0.0320	143.2425	0.2051	0.0676	148.9357	0.2133
0.2	0.3894	0.4472	0.0321	123.8862	0.1770	0.0639	131.2422	0.1880
0.3	0.4767	0.5477	0.0323	110.3833	0.1576	0.0618	119.6017	0.1711
0.4	0.5691	0.6325	0.0327	99.7599	0.1425	0.0611	110.6577	0.1580
0.5	0.6413	0.7071	0.0331	91.2479	0.1302	0.0600	103.1668	0.1474
0.6	0.7154	0.7746	0.0335	84.4268	0.1204	0.0588	96.4971	0.1385
0.7	0.7878	0.8367	0.0337	78.7296	0.1120	0.0583	91.2719	0.1312
0.8	0.8602	0.8944	0.0341	73.5779	0.1048	0.0584	87.1753	0.1251
0.9	0.9333	0.9487	0.0344	68.8307	0.0983	0.0584	83.6712	0.1199
0.95	0.9641	0.9747	0.0347	66.9124	0.0954	0.0586	82.2796	0.1177
0.98	0.9866	0.9899	0.0346	65.6831	0.0937	0.0589	81.2640	0.1163

Table 5 **Accuracy of the approximations in Theorem 2 in the $M/LGN/500+M$ system with $\rho = 1.4$.**

			SJF			Two class		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0221	385.1895	0.2753	0.0605	385.9085	0.2759
0.005	0.0636	0.0707	0.0222	369.7601	0.2647	0.0591	371.5782	0.2661
0.01	0.0877	0.1000	0.0222	359.7233	0.2570	0.0578	362.3138	0.2589
0.02	0.1151	0.1414	0.0222	345.4297	0.2465	0.0561	348.9148	0.2492
0.05	0.1899	0.2236	0.0228	315.7488	0.2257	0.0550	321.1919	0.2299
0.1	0.2688	0.3162	0.0228	284.5682	0.2038	0.0526	293.4705	0.2099
0.2	0.3894	0.4472	0.0226	244.5212	0.1752	0.0501	256.1455	0.1833
0.3	0.4767	0.5477	0.0231	217.6770	0.1555	0.0479	231.6201	0.1657
0.4	0.5691	0.6325	0.0233	196.1077	0.1401	0.0471	212.4476	0.1518
0.5	0.6413	0.7071	0.0236	178.0930	0.1273	0.0460	196.7703	0.1404
0.6	0.7154	0.7746	0.0239	164.0412	0.1171	0.0457	183.0561	0.1312
0.7	0.7878	0.8367	0.0241	152.3888	0.1085	0.0448	171.2986	0.1230
0.8	0.8602	0.8944	0.0244	141.5950	0.1008	0.0444	161.9419	0.1161
0.9	0.9333	0.9487	0.0245	131.6309	0.0937	0.0446	154.1588	0.1104
0.95	0.9641	0.9747	0.024	127.2698	0.0907	0.0441	150.1600	0.1074
0.98	0.9866	0.9899	0.0249	124.7826	0.0889	0.0446	148.4281	0.1062

Table 6 **Accuracy of the approximations in Theorem 2 in the $M/LGN/1000+M$ system with $\rho = 1.4$.**

α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	SJF	2 classes	3 classes	5 classes	10 classes
0.001	0.0282	0.0316	0.2770	0.2780	0.2780	0.2780	0.2780
0.005	0.0636	0.0707	0.2673	0.2700	0.2700	0.2700	0.2695
0.01	0.0877	0.1000	0.2603	0.2644	0.2644	0.2643	0.2632
0.02	0.1151	0.1414	0.2507	0.2567	0.2567	0.2562	0.2540
0.05	0.1899	0.2236	0.2323	0.2413	0.2523	0.2343	0.2326
0.1	0.2688	0.3162	0.2128	0.2260	0.2264	0.2147	0.2135
0.2	0.3894	0.4472	0.1881	0.2059	0.1971	0.1911	0.1903
0.3	0.4767	0.5477	0.1714	0.1930	0.1799	0.1761	0.1757
0.4	0.5691	0.6325	0.1586	0.1829	0.1677	0.1651	0.1650
0.5	0.6413	0.7071	0.1485	0.1749	0.1591	0.1572	0.1570
0.6	0.7154	0.7746	0.1407	0.1686	0.1528	0.1513	0.1512
0.7	0.7878	0.8367	0.1340	0.1631	0.1478	0.1466	0.1465
0.8	0.8602	0.8944	0.1284	0.1589	0.1436	0.1426	0.1426
0.9	0.9333	0.9487	0.1236	0.1551	0.1405	0.1395	0.1394
0.95	0.9641	0.9747	0.1211	0.1534	0.1387	0.1382	0.1381
0.98	0.9866	0.9899	0.1198	0.1526	0.1382	0.1374	0.1374

Table 7 $M/LGN/100 + M$ **with** $\rho = 1.4$ **where we divide the high class (**$< \tau$**) into equally-sized classes.**

α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	SJF	2 classes	3 classes	5 classes	10 classes
0.001	0.0282	0.0316	0.2770	0.2780	0.2780	0.2780	0.2776
0.005	0.0636	0.0707	0.2673	0.2700	0.2700	0.2691	0.2684
0.01	0.0877	0.1000	0.2603	0.2644	0.2642	0.2624	0.2621
0.02	0.1151	0.1414	0.2507	0.2567	0.2557	0.2534	0.2529
0.05	0.1899	0.2236	0.2323	0.2413	0.2375	0.2360	0.2357
0.1	0.2688	0.3162	0.2128	0.2260	0.2193	0.2180	0.2178
0.2	0.3894	0.4472	0.1881	0.2059	0.1960	0.1952	0.1951
0.3	0.4767	0.5477	0.1714	0.1930	0.1811	0.1805	0.1803
0.4	0.5691	0.6325	0.1586	0.1829	0.1699	0.1694	0.1695
0.5	0.6413	0.7071	0.1485	0.1749	0.1615	0.1612	0.1612
0.6	0.7154	0.7746	0.1407	0.1686	0.1554	0.1552	0.1551
0.7	0.7878	0.8367	0.1340	0.1631	0.1503	0.1500	0.1500
0.8	0.8602	0.8944	0.1284	0.1589	0.1459	0.1457	0.1456
0.9	0.9333	0.9487	0.1236	0.1551	0.1425	0.1424	0.1424
0.95	0.9641	0.9747	0.1211	0.1534	0.1410	0.1409	0.1409
0.98	0.9866	0.9899	0.1198	0.1526	0.1402	0.1401	0.1401

Table 8 **Increasing the number of classes: $M/LGN/100 + M$ with $\rho = 1.4$ where we make the $(0.8\tau, \tau)$ class more granular with equally-sized classes.**

α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	SJF	2 classes	3 classes	5 classes	10 classes
0.001	0.0282	0.0316	0.2770	0.2780	0.2780	0.2780	0.2777
0.005	0.0636	0.0707	0.2673	0.2700	0.2700	0.2697	0.2691
0.01	0.0877	0.1000	0.2603	0.2644	0.2642	0.2634	0.2629
0.02	0.1151	0.1414	0.2507	0.2567	0.2563	0.2549	0.2545
0.05	0.1899	0.2236	0.2323	0.2413	0.2396	0.2382	0.2382
0.1	0.2688	0.3162	0.2128	0.2260	0.2229	0.2217	0.2216
0.2	0.3894	0.4472	0.1881	0.2059	0.2017	0.2012	0.2010
0.3	0.4767	0.5477	0.1714	0.1930	0.1887	0.1881	0.1880
0.4	0.5691	0.6325	0.1586	0.1829	0.1785	0.1782	0.1782
0.5	0.6413	0.7071	0.1485	0.1749	0.1708	0.1705	0.1705
0.6	0.7154	0.7746	0.1407	0.1686	0.1651	0.1650	0.1649
0.7	0.7878	0.8367	0.1340	0.1631	0.1606	0.1605	0.1604
0.8	0.8602	0.8944	0.1284	0.1589	0.1565	0.1564	0.1565
0.9	0.9333	0.9487	0.1236	0.1551	0.1534	0.1533	0.1534
0.95	0.9641	0.9747	0.1211	0.1534	0.1520	0.1519	0.1519
0.98	0.9866	0.9899	0.1198	0.1526	0.1513	0.1513	0.1512

Table 9 **Increasing the number of classes: $M/LGN/100+M$ with $\rho=1.4$ where we make the $(\tau, 1.2\tau)$ class more granular with equally-sized classes.**

			Two class right threshold			Two class wrong threshold		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.1212	38.8721	0.2780	0.2361	39.7960	0.2851
0.005	0.0636	0.0707	0.1194	37.6858	0.2700	0.2361	39.7820	0.2850
0.01	0.0877	0.1000	0.1181	36.9477	0.2644	0.2366	39.8898	0.2857
0.02	0.1151	0.1414	0.1167	35.9017	0.2567	0.2367	39.9327	0.2858
0.05	0.1899	0.2236	0.1140	33.7302	0.2413	0.2370	40.0037	0.2860
0.1	0.2688	0.3162	0.1110	31.5432	0.2260	0.2372	40.0213	0.2860
0.2	0.3894	0.4472	0.1078	28.7060	0.2059	0.2362	39.9446	0.2848
0.3	0.4767	0.5477	0.1061	26.9172	0.1930	0.2281	38.9449	0.2772
0.4	0.5691	0.6325	0.1048	25.5386	0.1829	0.2120	36.8546	0.2622
0.5	0.6413	0.7071	0.1040	24.4331	0.1749	0.1896	33.9810	0.2420
0.6	0.7154	0.7746	0.1035	23.5301	0.1686	0.1639	30.6543	0.2182
0.7	0.7878	0.8367	0.1030	22.7404	0.1631	0.1401	27.4672	0.1956
0.8	0.8602	0.8944	0.1028	22.1475	0.1589	0.1225	24.9204	0.1773
0.9	0.9333	0.9487	0.1028	21.6270	0.1551	0.1111	22.9721	0.1637
0.95	0.9641	0.9747	0.1027	21.4076	0.1534	0.1076	22.2185	0.1585
0.98	0.9866	0.9899	0.1032	21.3123	0.1526	0.1057	21.8478	0.1557

Table 10 **Effect of choosing the wrong threshold in the $M/LGN/100+M$ system with $\rho = 1.4$.**

			Two class right threshold			Two class wrong threshold		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0605	385.9085	0.2759	0.2386	397.2644	0.2845
0.005	0.0636	0.0707	0.0591	371.5782	0.2661	0.2384	396.8136	0.2844
0.01	0.0877	0.1000	0.0578	362.3138	0.2589	0.2389	397.9682	0.2849
0.02	0.1151	0.1414	0.0561	348.9148	0.2492	0.2391	398.3045	0.2850
0.05	0.1899	0.2236	0.0550	321.1919	0.2299	0.2392	398.5709	0.2850
0.1	0.2688	0.3162	0.0526	293.4705	0.2099	0.2395	398.9648	0.2850
0.2	0.3894	0.4472	0.0501	256.1455	0.1833	0.2384	398.2404	0.2841
0.3	0.4767	0.5477	0.0479	231.6201	0.1657	0.2307	388.8506	0.2766
0.4	0.5691	0.6325	0.0471	212.4476	0.1518	0.2142	367.6738	0.2615
0.5	0.6413	0.7071	0.0460	196.7703	0.1404	0.1910	338.1445	0.2405
0.6	0.7154	0.7746	0.0457	183.0561	0.1312	0.1617	301.7280	0.2149
0.7	0.7878	0.8367	0.0448	171.2986	0.1230	0.1283	261.3477	0.1862
0.8	0.8602	0.8944	0.0444	161.9419	0.1161	0.0911	217.5845	0.1550
0.9	0.9333	0.9487	0.0446	154.1588	0.1104	0.0540	172.5530	0.1230
0.95	0.9641	0.9747	0.0441	150.1600	0.1074	0.0465	157.9092	0.1127
0.98	0.9866	0.9899	0.0446	148.4281	0.1062	0.0453	152.7485	0.1090

Table 11 **Effect of choosing the wrong threshold in the $M/LGN/1000+M$ system with $\rho = 1.4$.**

			Two class right threshold			Two class wrong threshold		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0554	779.1574	0.4330	0.3244	795.0558	0.4432
0.005	0.0636	0.0707	0.0550	755.7675	0.4205	0.3244	794.9826	0.4432
0.01	0.0877	0.1000	0.0537	741.2371	0.4118	0.3250	796.9677	0.4438
0.02	0.1151	0.1414	0.0524	719.1483	0.3995	0.3254	798.0134	0.4440
0.05	0.1899	0.2236	0.0516	674.0468	0.3745	0.3255	798.4892	0.4439
0.1	0.2688	0.3162	0.0493	624.7043	0.3477	0.3247	797.0073	0.4429
0.2	0.3894	0.4472	0.0480	559.9949	0.3112	0.3123	777.9405	0.4319
0.3	0.4767	0.5477	0.0467	512.1316	0.2846	0.2887	740.5255	0.4102
0.4	0.5691	0.6325	0.0456	474.2025	0.2632	0.2588	692.0790	0.3830
0.5	0.6413	0.7071	0.0448	442.1810	0.2455	0.2245	636.0262	0.3519
0.6	0.7154	0.7746	0.0438	414.8870	0.2303	0.1862	574.2843	0.3182
0.7	0.7878	0.8367	0.0432	392.1642	0.2173	0.1452	510.4045	0.2830
0.8	0.8602	0.8944	0.0433	370.1393	0.2055	0.1007	443.6124	0.2461
0.9	0.9333	0.9487	0.0428	350.5303	0.1944	0.0535	373.5181	0.2076
0.95	0.9641	0.9747	0.0426	341.1217	0.1893	0.0398	345.1128	0.1923
0.98	0.9866	0.9899	0.0424	336.0697	0.1864	0.0404	336.4412	0.1872

Table 12 **Effect of choosing the wrong threshold in the $M/LGN/1000+M$ system with $\rho = 1.8$.**

		SJF			Two class		
$r[S,_i \hat{S}_i]$	$r[Z,_i \hat{Z}_i]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.0025	0.005	0.0220	395.3176	0.2832	0.0612	395.5416	0.2831
0.0071	0.01	0.0220	393.6909	0.2819	0.0609	394.4173	0.2821
0.0366	0.05	0.0222	379.7860	0.2711	0.0592	381.0778	0.2719
0.0711	0.1	0.0222	361.5732	0.2578	0.0576	363.7636	0.2595
0.1501	0.2	0.0223	324.0653	0.2308	0.0531	329.4904	0.2348
0.2312	0.3	0.0226	287.2849	0.2046	0.0489	295.5902	0.2106
0.3196	0.4	0.0230	251.8599	0.1793	0.0455	264.6596	0.1883
0.4161	0.5	0.0232	217.6817	0.1548	0.0439	235.7478	0.1677
0.5137	0.6	0.0234	185.7213	0.1321	0.0452	209.6101	0.1493
0.6233	0.7	0.0237	157.6626	0.1125	0.0467	186.6144	0.1332
0.7402	0.8	0.0243	138.1429	0.0984	0.0465	168.6685	0.1203
0.8662	0.9	0.0247	126.5476	0.0902	0.0445	155.0831	0.1104
0.9318	0.95	0.0251	123.9812	0.0885	0.0444	151.4413	0.1079
0.9724	0.98	0.0251	123.2799	0.0881	0.0442	149.5975	0.1066
0.9862	0.99	0.0254	123.2587	0.0882	0.0449	149.7034	0.1069

Table 13 **Accuracy of the approximation in the $M/LGN/1000+M$ system with $\rho=1.4$ under service-time model of Section 5.2.**

		SJF			Two class		
$r[S,_i \hat{S}_i]$	$r[Z_i, \hat{Z}_i]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.0025	0.005	0.0171	792.5415	0.4417	0.0551	792.9762	0.4417
0.0071	0.01	0.0170	790.2275	0.4402	0.0548	790.5204	0.4401
0.0366	0.05	0.0169	768.3630	0.4272	0.0539	770.1773	0.4278
0.0711	0.1	0.0173	739.7211	0.4108	0.0531	742.8040	0.4123
0.1501	0.2	0.0177	681.5804	0.3782	0.0505	686.7148	0.3810
0.2312	0.3	0.0181	620.6161	0.3441	0.0479	629.1391	0.3487
0.3196	0.4	0.0185	558.1607	0.3097	0.0452	570.1010	0.3162
0.4161	0.5	0.0189	496.6495	0.2751	0.0427	512.6401	0.2839
0.5137	0.6	0.0192	435.2636	0.2415	0.0419	455.2422	0.2525
0.6233	0.7	0.0195	380.2486	0.2109	0.0409	404.3327	0.2241
0.7402	0.8	0.0200	337.9368	0.1875	0.0407	364.2818	0.2020
0.8662	0.9	0.0203	312.3646	0.1735	0.0405	339.4761	0.1885
0.9318	0.95	0.0209	306.9997	0.1705	0.0413	334.3451	0.1857
0.9724	0.98	0.0210	305.1819	0.1697	0.0418	332.6789	0.1848
0.9862	0.99	0.0210	304.5051	0.1697	0.0426	332.5934	0.1852

Table 14 **Accuracy of the approximation in the $M/LGN/1000+M$ system with $\rho=1.8$ under service-time model of Section 5.2 .**

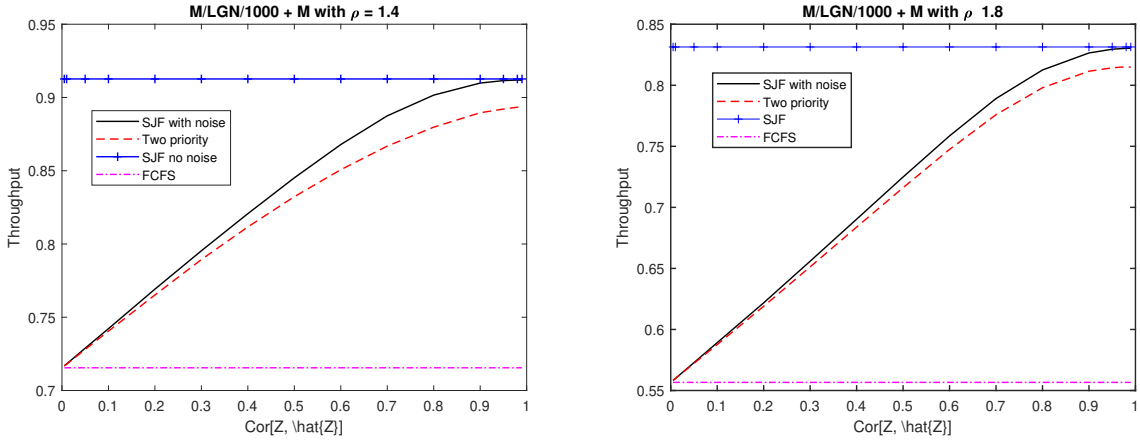

Figure 2 **Long-run throughput, in steady state, in the $M/LGN/1000+M$ model under the service-time model of Section 5.2.**

References

Atar, R., H. Kaspi, and N. Shimkin (2014). Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research 39*(3), 672–696.

Banerjee, S., A. Budhiraja, and A. L. Puha (2020). Heavy traffic scaling limits for shortest remaining processing time queues with heavy tailed processing time distributions.

Barlow, R. E. and F. Proschan (1975). Statistical theory of reliability and life testing: probability models. Technical report, Florida State Univ Tallahassee.

Dong, J. and R. Ibrahim (2021). Srpt scheduling discipline in many-server queues with impatient customers. *Management Science 67*(12), 7708–7718.

Down, D. G. (2019). Open problem—size-based scheduling with estimation errors. *Stochastic Systems 9*(3), 295–296.

Down, D. G., H. C. Gromoll, and A. L. Puha (2009). Fluid limits for shortest remaining processing time queues. *Mathematics of Operations Research 34*(4), 880–911.

Garnett, O., A. Mandelbaum, and M. Reiman (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management 4*(3), 208–227.

Gromoll, H. C. and M. Keutel (2012). Invariance of fluid limits for the shortest remaining processing time and shortest job first policies. *Queueing Systems 70*(2), 145–164.

Gromoll, H. C., Ł. Kruk, and A. L. Puha (2011). Diffusion limits for shortest remaining processing time queues. *Stochastic Systems 1*(1), 1–16.

Grosof, I., Z. Scully, and M. Harchol-Balter (2018). Srpt for multiserver systems. *Performance Evaluation 127*, 154–175.

Ibrahim, R., H. Ye, P. L'Ecuyer, and H. Shen (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting 32*(3), 865–874.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics 37*(5), 1137–1153.

Leonardi, S. and D. Raz (2007). Approximating total flow time on parallel machines. *Journal of Computer and System Sciences 73*(6), 875–891.

Lin, M., A. Wierman, and B. Zwart (2011). Heavy-traffic analysis of mean response time under shortest remaining processing time. *Performance Evaluation 68*(10), 955–966.

Mitzenmacher, M. (2021). Queues with small advice. In *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA21)*, pp. 1–12. SIAM.

Puha, A. L. et al. (2015). Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. *The Annals of Applied Probability 25*(6), 3381–3404.

Puha, A. L. and A. R. Ward (2019). Scheduling an overloaded multiclass many-server queue with impatient customers. In *Operations Research & Management Science in the Age of Analytics*, pp. 189–217. INFORMS.

Schrage, L. (1968). Letter to the editor-a proof of the optimality of the shortest remaining processing time discipline. *Operations Research 16*(3), 687–690.

Schrage, L. E. and L. W. Miller (1966). The queue m/g/1 with the shortest remaining processing time discipline. *Operations Research 14*(4), 670–684.

Scully, Z., I. Grosof, and M. Harchol-Balter (2020). The gittins policy is nearly optimal in the m/g/k under extremely general conditions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 4*(3), 1–29.

Scully, Z., I. Grosof, and M. Mitzenmacher (2021). Uniform bounds for scheduling with job size estimates. *arXiv preprint arXiv:2110.00633*.

Scully, Z., M. Harchol-Balter, and A. Scheller-Wolf (2018). Soap: One clean analysis of all age-based scheduling policies. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 2*(1), 1–30.

Shaked, M. and J. G. Shanthikumar (2007). *Stochastic orders*. Springer Science & Business Media.

Wierman, A. and M. Nuyens (2008). Scheduling despite inexact job-size information. In *Proceedings of the 2008 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pp. 25–36.

Wu, C., A. Bassamboo, and O. Perry (2019). Service system with dependent service and patience times. *Management Science 65*(3), 1151–1172.