Scheduling with Service-Time Information: The Power of Two Priority Classes

Yan Chen, Jing Dong

Columbia University, New York, NY 10027

Utilizing customers' service-time information, we study an easy-to-implement scheduling policy with two priority classes. By carefully designing the classes, the two-class priority rule achieves near-optimal performance. In particular, for a single-server queue, as the traffic intensity approaches 1, the policy achieves a scaling for the queue length processes that is similar to the shortest remaining processing time first policy. Our analysis quantifies how the tail of the service time distribution affects the benefits one can gain from service-time-based scheduling policies. When the service times are misspecified, we further quantify how imperfect observation of the service time affects the performance of the two-class priority rule through both theoretical and numerical analysis. Our results demonstrate the robustness of the two-class priority rule. Specifically, even with imperfect service-time information, the two-class priority rules can still achieve substantial performance improvement over the first come first served.

Key words: Priority queue, Scheduling, Heavy-traffic diffusion limit

1. Introduction

The development of statistical learning techniques and the growing availability of data facilitates us to gain more customer-side information. From the operations perspective, an important question to ask is how to utilize the customer-side information to achieve better system performance. In this paper, we look into one particular customer-side information – service times – and study how to use service-time information to do smarter scheduling for service systems.

When having perfect service-time information, it is well-known that the shortest remaining processing time first (SRPT) policy achieves superior performance. In particular, in a single-server queue, it has been established that SRPT is optimal with respect to minimizing the steady-state average sojourn time (Schrage and Miler 1966) and the average number of customers in the system at each point in time (Schrage 1968). This policy has been successfully implemented in scheduling jobs for machines (e.g., computers). In those settings, getting job-size information and administering non-first-come-first-served (FCFS) sequencing is relatively easy. Extending this policy to service systems faces two main obstacles: (1) managing a queue where the sequence of customers may change upon each arrival is hard, and (2) we may not have perfect job-size information. Motivated by the main insights from SRPT, in this paper, we propose a two-class priority rule that achieves comparable performance to SRPT. The two priority classes are defined by a single carefully-chosen threshold where customers whose service times are below the threshold get higher priority; that

is, it prioritizes smaller-size jobs. Within each priority class, customers are served as FCFS. Under this policy, we only need to maintain two FCFS queues, which greatly facilitate implementations in practice. As an added benefit of our scheduling policy and analysis framework, we can also provide some theoretical quantification of the effect of imperfect service times. We demonstrate that the policy is robust to service-time mis-specification.

Comparable performance to SRPT. We start by providing some intuition behind our development. In queueing systems, smart scheduling is especially important when the system is heavily loaded; that is, when the traffic intensity (server utilization rate), ρ, is close to 100% (see Section 2 for a precise definition of ρ). In a single-server queue, as ρ approaches 1, under FCFS, the queue scales as $1/(1-\rho)$, whereas under SRPT, the queue can scale much slower than $1/(1-\rho)$ when the service-time distribution has infinite support (Lin et al. 2011, Puha 2015). In other words, SRPT can provide order-of-magnitude performance improvement over FCFS. To see how such improvement is achieved, we notice that if we divide customers into different priority classes according to their remaining service times, i.e., the longer the service time, the lower the priority, then the SRPT policy can be viewed as the limit of a sequence of multi-class priority rules, where, in the limit, we send the number of classes to infinity (Schrage and Miler 1966). On the other hand, for multi-class queues, under certain conditions on the classes, the properly scaled queue length process only contains jobs in the lowest priority class (Reiman 1984). This result is formally known as state-space collapse in heavy-traffic asymptotic analysis. Under any work-conserving scheduling policy (the server is not idling when there are still customers waiting in the queue), the total workload scales as $1/(1-\rho)$. If all the "remaining" workload is from customers in the lowest priority class, then the queue size scales as $\mu_m/(1-\rho)$, where μ_m is the service rate (reciprocal of the average service time) of customers in the lowest priority class. The above discussion indicates that if the average service time of the lowest priority class goes to infinity as ρ approaches 1, then the queue size can scale more slowly than $1/(1-\rho)$.

We make two important observations from the above discussion. First, we only need to divide customers into two classes and focus on how to define the low priority class. Second, the definition of the lower priority class depends on the service time distribution and the traffic intensity of the system. In particular, as the traffic intensity ρ increases, we may need to shrink the size of the low priority class so that the average service time of jobs in the low priority class increases accordingly. These two observations are the rationale behind the development of our two-class priority rule. The main technical contribution of the paper is to provide a clear rule to define the low-priority class and to rigorously establish that the two-class priority rule with properly defined classes achieves comparable scaling for the queue size as SRPT under heavy traffic.

We highlight here the role of service time distribution in how much we can benefit from smart scheduling. For our two-class priority rule, the service time distribution determines the appropriate threshold to divide the two classes. This in turn determines the average service time of the low priority class, and thus the scaling of the queue as $\rho \to 1$. In particular, we show that the value of the threshold depends on the decay rate of the tail cumulative distribution function (tail cdf) of the service time distribution. The more slowly the tail cdf decays (the heavier the tail of the service time distribution), the more slowly the queue grows as $\rho \to 1$. This result indicates that systems in which the service time distributions have heavier tails tend to benefit more from the two-class priority rule. A similar phenomenon has been observed for the SRPT policy as well (Lin et al. 2011). Our analysis provides a clear explanation of how the service time distribution affects the performance gain from service-time-based priority rules.

Robustness to imperfect service time information. In real applications, the service-time requirements may not be exactly known to us, and are subject to various estimation errors. There are very limited works studying size-based scheduling with estimation errors, which has been listed as an open problem in Down (2019). There are several key challenges to study such problems: 1) Estimation errors arise in many different forms, depending on the modeling choice, the estimation error itself may affect the service time information. For example, if the service time is estimated using a linear regression model based on observable characteristics of the customer, then the distribution of the service time is a convolution of the variability in the observables and the estimation error. Thus, different estimation errors may require a very different set of analyses. 2) Process level analysis of system dynamics is challenging as we often have to keep track of either the age process or the remaining processing time process of all the jobs in the system. 3) Optimizing the scheduling policy would require updating the service-time estimation as the job spends more time in service. Our scheduling rule and analysis framework provide an analytically tractable way to investigate how imperfect service-time information affects system performance. In particular, for our two-class priority rule, we only need to study how the estimation error is affecting our ability to classify jobs into the two priority classes. Our asymptotic framework indicates that to achieve order-of-magnitude performance improvement over FCFS, we only require the classified high-priority jobs and low-priority jobs to satisfy certain scaling properties, which can be easy to verify.

We demonstrate the analytical tractability of our framework by studying models for three simple estimation error models. Our results generate several insights: 1) when the estimation errors have bounded support, we can achieve the same queue-length scaling as in the perfect information case; 2) when building classification models to assign customers to the two priority classes, it is more important to avoid wrongly classifying short jobs as low-priority jobs than wrongly classifying long jobs as high priority jobs; 3) when dealing with independent and identically distributed (iid)

measurement errors, if the tail of the error distribution is lighter than the tail of the service time distribution, we can still achieve $o(1/(1-\rho))$ scaling under the two-class priority rule with a properly chosen threshold. However, when the error distribution has a heavier tail, we may not be able to gain substantial performance improvement from the two-class priority rule.

We also complement our theoretical analysis with numerical experiments, in which we study the performance of the two-class priority rule with predicted service times. We consider linear regression models, nonlinear regression models, and generalized linear models. In all cases, the two-class priority rule with a properly chosen threshold achieves significant performance improvement over FCFS. Most noticeably, even when the prediction model is only able to explain 3% of the variability in the actual service times, the two-class priority rule utilizing the predicted service times is still able to reduce the steady-state average queue length by more than 50% over FCFS in our simulated example. This is because even with highly inaccurate prediction models, we still have a very low probability of wrongly classifying short jobs as low-priority jobs.

We next provide a brief review of the literature. The goal is to put our work in the right context.

1.1. Related Literature

The literature on scheduling using service-time (job-size) information has a long history. Schrage and Miler (1966) are among the first to study the SRPT policy and various other scheduling policies for $M/GI/1$ queues. Because these scheduling policies require perfect job-size information, we see most applications in scheduling of jobs for machines (Harchol-Balter et al. 2003). In addition to SRPT, popular scheduling policies include processor sharing (PS), shortest job first. Recently, Scully et al. (2018) develop a unified framework to analyze these scheduling policies. Relatively few works look at the imperfect information setting (Dell'Amico et al. 2014, Down 2019). Most noticeably, the foreground-background policy does not require any prior job-size information and performs well when the service time distribution has a decreasing failure rate (Nuyens and Wierman 2006). More generally, policies based on the Gittins index is known to minimize mean response time under various information structures (Aalto et al. 2009, Scully et al. 2020). However, the priority structure can be computationally intractable. Emadi et al. (2019) study the benefit of smart scheduling with only predicted service-time information using data from a call center. Our work contributes to this line of work by proposing a simple two-class priority rule that achieves superior performance and is robust to service-time misspecification.

Our result builds on heavy-traffic asymptotic analysis. Process level asymptotic analysis of job-size-based scheduling policies is challenging, as it often requires keeping track of the remaining processing time of each job in the system. Gromoll (2004) establishes the diffusion limit for PS queues. Gromoll et al. (2011), Puha (2015), Banerjee et al. (2020) develop the diffusion limit for

processes related to SRPT queues. Similar to these works, we establish the diffusion limit of the queue length processes under the two-class priority rule. Our scaling is similar to that established in Puha (2015), Banerjee et al. (2020) for SRPT. The heavy-traffic limit of the steady-state average response time under SRPT is studied in Lin et al. (2011). A key observation from this line of work is that the performance of the SRPT policy depends heavily on the tail property of the service time distribution. Our work provides a clear explanation for this observation.

Our work is also related to the line of research on scheduling/prioritizing policies for single-server queues with multiple classes of customers. Cox and Smith (1961) are among the earliest to analyze the optimality of an important index policy – the $c\mu$-rule, where c is the per unit time holding cost per customer and μ is the service rate. Under the $c\mu$-rule, we give priority to the class with a larger value of $c\mu$. Later, Van Mieghem (1995) and Mandelbaum and Stolyar (2004) extend the $c\mu$-rule to more general settings using heavy-traffic asymptotic analysis. In our model, we can think of everyone having the same linear holding cost. Thus, the $c\mu$-rule reduces to one where we prioritize the class with a shorter average service time. The main insight from this line of work is that under the diffusion scaling, the queue only contains jobs in the lowest priority class. We derive a similar state-space-collapse result. The main difference between our work and the existing literature is that our class division is not fixed. The lower priority class is shrinking as the traffic intensity increases. The diffusion scaling we apply is also smaller than the one in the $c\mu$-rule literature. Recent works also study the $c\mu$-rule with imperfect information (Argon and Ziya 2009, Sun et al. 2018). Our work also contributes to this line of literature by providing an analytical framework to study the effect of estimation errors.

Also note that the optimality of scheduling policies with simple structures, such as SRPT or the $c\mu$-rule, has only been established for single-server queues. In the asymptotic sense, we can extend the optimality results to multiple-server queues under the conventional heavy-traffic regime where the number of servers is held fixed as the traffic intensity approaches 1. For example, Grosof et al. (2018) recently establish the optimality of SRPT for multi-server queues in the conventional heavy-traffic regime. In many-server heavy-traffic regime where the number of servers is sent to infinity with the arrival rate, the asymptotic optimality of $c\mu$-type of scheduling policies may no longer hold (see, e.g., Harrison and Zeevi (2004), Kim et al. (2018)).

1.2. Notations

Throughout the paper, we refer to waiting time as the total amount of time that a customer spends in the system. This time includes both the time spent in the queue and the time spent being served, and is also referred to as the sojourn time or the flow time in the literature. We refer to the number of customers in the queue or the queue size as the number of customers in the system. This queue

length includes both customers waiting to be served and customers in service (being served). We use the words customer and job interchangeably.

We define $\eta(t) := t$ and $\zeta(t) := 0$ for all $t \geq 0$. We denote $D[0, \infty)$ as the space of functions from $[0, \infty)$ to $\mathbb{R}$ that are right continuous with left limits, and is endowed with Skorohod J_1 topology. Letting $f(n)$ and $g(n)$ be two nonnegative functions, we write $f(n) \sim g(n)$ if there exist $0 < c < C < \infty$ and $n_0 \in (0, \infty)$, such that for any $n > n_0$, $cg(n) < f(n) < Cg(n)$. We write $f(n) = O(g(n))$ if there exists positive constants $C < \infty$ and n_0, such that for any $n \geq n_0$, $f(n) \leq Cg(n)$. We write $f(n) = o(g(n))$ if for any $\epsilon > 0$, there exists $n_0(\epsilon) \in (0, \infty)$, such that for any $n > n_0(\epsilon)$, $f(n) \leq \epsilon g(n)$.

2. Two-class priority rule

Consider a single-server queue with a general renewal arrival process and general independent and identically distributed (iid) service times, i.e., a $GI/GI/1$ queue. We denote λ as its arrival rate and μ as its service rate. Then, the traffic intensity ρ is defined as

$$\rho := \lambda/\mu.$$

We also denote F as the cumulative distribution function (cdf) of the service time, and $\bar{F}$ as its tail cdf, i.e, $\bar{F} = 1 - F$. The two priority classes are defined by the threshold

$$K(\rho, F) = \bar{F}^{-1}\left((1 - \rho)^{(1-\delta)}\right) \text{ for some fixed } 0 < \delta < 1. \tag{1}$$

Customers whose service times are smaller than or equal to $K(\rho, F)$ are in the high priority class, Class 1, whereas customers whose service times are larger than $K(\rho, F)$ are in the low priority class, Class 2. δ is a parameter that balance the load between the high priority class and the low priority class. We suggest setting δ small, for example, between 0.01 and 0.1, as indicated by our analysis in Section 3.

The threshold, $K(\rho, F)$, depends on both the service time distribution and the traffic intensity of the system. It is designed such that $K(\rho, F)$ increases as ρ increases. Note that if the service time distribution has infinite support, $K(\rho, F) \to \infty$ as $\rho \to 1$. However, $K(\rho, F)$ can not grow too fast as ρ approaches 1. In particular, it needs to be properly chosen to control the workload of Class 1 customers (this notion will be made precise in Section 3).

In Table 1, we compare the performance of the two-class priority rule with other benchmark policies in $M/M/1$ queues with different traffic intensities. The benchmark policies include FCFS, shortest-job-first (SJF), where we prioritize the job with a smaller size but do not allow preemption, and SRPT, where we prioritize the job with a shorter remaining processing time in a preemptive way. We consider two implementations of the two-class priority rule, one allows preemption (two-class P) and the other does not (two-class NP). We observe that compared with FCFS, the two-class

priority rule is able to reduce the average queue length by about a half or more across different traffic intensities. We also note that SRPT performs the best, as has been established in the literature. However, the gaps between the two-class priority rules and SRPT are small. This finding suggests that even though we can add more priority classes to improve the system performance, the marginal gains from doing so would be small.

Table 1   **Steady-state average queue length for** $M/M/1$ **queues under different scheduling policies (**$\mu = 1$**,** $\lambda = \rho$**,** $K = \bar{F}((1-\rho)^{(1-0.05)}) = -(1-0.05)\log(1-\rho)$**).**

ρ	0.8	0.85	0.9	0.95
FCFS	4.00	5.67	9.00	19.00
Two-Class NP	2.68	3.43	4.76	8.19
Two-Class P	2.29	2.94	4.13	7.30
SJF	2.31	2.86	3.78	5.95
SRPT	1.88	2.36	3.20	5.26

In actual service systems, we may not have perfect service-time information. In the next example, we consider the case in which the service times take the form

$$v = \exp\big(\beta^T X + \epsilon\big)$$

where X is a vector of observable customer characteristics, β is the vector of coefficients for the features in X, $\epsilon \sim N(0, \sigma_e^2)$ and is independent of X. With this service time model, $\mathbb{E}[\log(v)|X] = \beta^T X$. Assume that we have access to $\beta^T X$. Thus, we can use $\exp(\beta^T X)$ to predict v. Let $\hat{F}$ denote the cdf of the predicted service time, i.e., $\beta^T X$. Customers whose predicted service times are smaller than or equal to $K(\rho, \hat{F})$ are assigned to Class 1, and others are assgined to Class 2.

In Table 2, we compare the performance of our two-class priority rule (NP) with FCFS and SJF using the predicted service times. Set $\beta^T X = 0.1 + 0.1X_1 + 0.4X_2 + 0.4X_3$, where $X_1 \sim N(1,1)$, $X_2 \sim$Exp(1), and $X_3 \sim$Uniform$[0,1]$. We also set $\sigma_e = 0.5$. In this case, the prediction model is able to explain 61% of the variability in the service times. We keep the service time distribution fixed and vary the arrival rate to achieve different traffic intensities. We observe that even with prediction errors, the two-class priority rules can still achieve significant queue length reduction from FCFS. In particular, when $\rho = 0.95$, the two-class priority rules leads to a more than 50% queue length reduction over FCFS. SJF performs slightly better than our proposed policy, suggesting there is limited value in adding more priority classes. These observations demonstrate the robustness of the job-size based scheduling rules to estimation errors. We provide detailed analysis about how the estimation errors affect the performance of the two-class priority rule in Sections 6 and 7.

Table 2 Steady-state average queue length for $M/GI/1$ queues under different scheduling policies using predicted service times.

ρ	0.8	0.85	0.9	0.95
$K(\rho, \hat{F})$	3.6	4	4.7	6.2
FCFS	4.54	6.60	10.04	21.21
Two-Class NP	3.65	4.42	6.82	10.89
SJF	2.99	3.87	5.02	8.46

In what follows, we start by analyzing the two-class priority rule with perfect service-time information.

3. Asymptotic performance analysis

To quantify the superior performance of the two-class priority rule and to derive insights on how to choose the appropriate class-division threshold $K(\rho, F)$, we take a heavy-traffic asymptotic approach. We consider a sequence of systems where the traffic intensity converges to 1 in an appropriate manner, and we study how the queue length process scales along the sequence. More specifically, consider a sequence of $GI/GI/1$ queues indexed by n, starting from empty at time 0. The service time distribution is fixed with mean $1/\mu = 1$; that is, we set a unit of time as the average service time. The arrival rate for the n-th system is $\lambda^n = 1 - \beta/\sqrt{n}$ for some $\beta > 0$. Under these parameter specifications, the traffic intensity for the sequence of systems approaches 1 at rate $1/\sqrt{n}$, i.e.,

$$\rho^n := \lambda^n/\mu = 1 - \beta/\sqrt{n}.$$

Let $A^n = \{A^n(t) : t \geq 0\}$ denote the arrival process of the n-th system. Let $\tau^n(k)$ denote the interarrival time between the k-th arriving customer and the $(k-1)$-th customer. We assume $\tau^n(k)$'s are iid and have the same distribution as τ^∞/λ_n, where τ^∞ is a random variable with $\mathbb{E}[\tau^\infty] = 1$. We denote G as the cdf of τ^∞. We also write $v^n(k)$ as the service time of the k-th arriving customer in system n. Because the service time distribution does not change for different scales of systems, we use v to denote a generic random variable following the service time distribution F.

We impose the following assumption on v.

ASSUMPTION 1. *The service time has a continuous distribution with probability density function f. There exists $C > 0$, such that $f(x) > 0$ for any $x > C$. There exists $\delta > 0$, such that $\mathbb{E}[v^{2+\delta}] < \infty$.*

The first part of Assumption 1 essentially requires that the service time distribution has infinite support (unbounded). Even under SRPT, when the service time distribution has bounded support, the queue scales as $1/(1-\rho)$ (Lin et al. 2011). In this case, no order-of-magnitude improvement arises from "smart" scheduling. The moment condition in Assumption 1 is standard to establish

the diffusion limit for the queue length processes (see, e.g., Whitt (2002)). Under Assumption 1, we denote $\sigma_s^2 := \mathrm{Var}(v)$.

We also impose the following light-tail assumption on τ^∞.

ASSUMPTION 2. *There exits $\epsilon > 0$ such that $\mathbb{E}[\exp(\theta\tau^\infty)] < \infty$ for $\theta \in (-\epsilon, \epsilon)$, i.e., τ^∞ has a finite moment generating function in a neighborhood of the origin.*

Under Assumption 2, τ^∞ has finite moments of all order. We denote $\sigma_a^2 := \mathrm{Var}(\tau^\infty)$. We also write $\psi_a(\theta) := \log \mathbb{E}[\exp(\theta\tau^\infty)]$ and $\bar{\theta} := \inf\{\theta : \psi_a(\theta) = \infty\}$. Lastly, we impose the following technical assumption on τ^∞.

ASSUMPTION 3. *ψ_a is differentiable everywhere in $(-\infty, \bar{\theta})$ and $\lim_{\theta \uparrow \bar{\theta}} \psi_a(\theta) = \psi_a(\bar{\theta})$. If $\psi_a(\bar{\theta}) < \infty$, $\lim_{\theta \uparrow \bar{\theta}} \psi_a'(\theta) = \infty$.*

We note that Assumptions 2 and 3 are satisfied by most commonly used interarrival time distributions. Examples include phase-type distributions (e.g., Erlang, hyperexponential) and Weibull distributions.

3.1. The critical threshold value

For the n-th system, we define K^n as the threshold value for the two classes; that is, customers with service time less than or equal to K^n are classified into Class 1; others are classified into Class 2. For simplicity of exposition, we assume class 1 customers have preemptive priority over class 2 customers.

We start by introducing a few more notations. Denote $\lambda_1^n = \lambda^n F(K^n)$ and $\lambda_2^n = \lambda^n \bar{F}(K^n)$ as the respective arrival rates of Class 1 and Class 2 customers in the n-th system. We also write $v_i^n(k)$ as the service time of the k-th Class i arrival in system n. Then,

$$\mu_1^n = (\mathbb{E}[v_1^n(k)])^{-1} = \left(\frac{1}{F(K^n)}\int_0^{K^n} xf(x)dx\right)^{-1} \text{ and } \mu_2^n = (\mathbb{E}[v_2^n(k)])^{-1} = \left(\frac{1}{\bar{F}(K^n)}\int_{K^n}^\infty xf(x)dx\right)^{-1}$$

denote the service rates of the two classes respectively. Let $\rho_i^n = \lambda_i^n/\mu_i^n$ denote the traffic intensity for Class i, $i = 1, 2$.

We next introduce the general idea underlying our development. From the existing results for priority queues (Reiman 1984), we make the following two important observations:

1. The Class 1 queue scales as $(1 - \rho_1^n)^{-1}$. Note that if Class 1 customers have preemptive priority over Class 2 customers, the Class 1 queue evolves as if there are only Class 1 customers in the system.

2. The total workload process scales as $(1 - \rho^n)^{-1} = O(\sqrt{n})$. Indeed, this scaling for the workload process holds for any work-conserving scheduling policies.

Note that if the "limiting" queue only contains Class 2 customers, the queue length process scales as $\mu_2^n \sqrt{n}$. To achieve a smaller scaling than $\sqrt{n}$ for the queue length process, we need $\mu_2^n \to 0$ as $n \to \infty$. However, we need to send μ_2^n to zero with delicacy. In particular, we need to make sure that under the scaling $\mu_2^n \sqrt{n}$, we see no Class 1 customers in the queue asymptotically, i.e., $\mu_2^n \sqrt{n}(1 - \rho_1^n) \to \infty$ as $n \to \infty$.

Above all, the key is to choose the threshold K^n in an appropriate way. Our choice of K^n satisfies the following assumption.

ASSUMPTION 4. $K^n \to \infty$, and there exists $\delta \in (0, 1/2)$, such that $n^{1/2-\delta}\bar{F}(K^n) \to \infty$ as $n \to \infty$.

We next provide some comments about Assumption 4. Let $\gamma^n := K^n \mu_2^n$. Under Assumption 1, $2/3 \leq \gamma^n \leq 1$. Then, Assumption 4 implies that $\mu_2^n \to 0$ as $n \to \infty$. This helps ensure that the class 2 queue scales slower than $\sqrt{n}$. We also note

$$\rho_1^n = \lambda_1^n/\mu_1^n = \left(1 - \frac{\beta}{\sqrt{n}}\right) \int_0^{K^n} x f(x) dx = 1 - \frac{\bar{F}(K^n)}{\mu_2^n} + O(1/\sqrt{n}).$$

Thus, under Assumption 4, there exists $\delta \in (0, 1/2)$, such that

$$\frac{n^{1/2-\delta}}{K^n}(1 - \rho_1^n) \to \infty \text{ as } n \to \infty, \tag{2}$$

or equivalently,

$$n^{1/2-\delta}\mu_2^n(1 - \rho_1^n) \to \infty \text{ as } n \to \infty.$$

This helps ensure that the class 1 queue scales even slower than the class 2 queue.

Lastly, recall that in (1), we suggest setting $K^n = \bar{F}^{-1}((1 - \rho_n)^{(1-\delta)})$ for some $0 < \delta < 1$. This choice of K^n satisfies Assumption 4 when the service time distribution satisfies Assumption 1.

3.2. Performance of the two-class priority rule

In this section, we rigorously quantify the performance of the two-class priority rule with K^n satisfying Assumption 4. In particular, we analyze how the queue length process scales as $n \to \infty$. Denote $Q_i^n = \{Q_i^n(t) : t \geq 0\}$ as the queue length process for Class i, $i = 1, 2$, in system n, and $Q^n = Q_1^n + Q_2^n$ as the total queue length. The following theorem is the main result of this paper. It shows that under the two-class priority rule, the queue length process scales as $\sqrt{n}/K^n$. As $K^n \to \infty$ as $n \to \infty$ under assumption 4, the queue length process scales slower than $\sqrt{n}$.

Let $\gamma := \lim_{n \to \infty} K^n \mu_2^n$. Define $\hat{Q} = \{\hat{Q}(t) : t \geq 0\}$ as a reflected Brownian motion with drift coefficient $-\gamma\beta$ and diffusion coefficient $\gamma\sqrt{\sigma_a^2 + \sigma_s^2}$.

THEOREM 1. *For the preemptive two-class priority rule, under Assumptions 1 – 4,*

$$\frac{K^n}{\sqrt{n}}Q_1^n(nt) \Rightarrow \zeta(t) \text{ and } \frac{K^n}{\sqrt{n}}Q_2^n(nt) \Rightarrow \hat{Q}(t) \text{ in } D[0,\infty) \text{ as } n \to \infty,$$

which implies

$$\frac{K^n}{\sqrt{n}}Q^n(nt) \Rightarrow \hat{Q}(t) \ \text{in} \ D[0,\infty) \ \text{as} \ n \to \infty.$$

Alternatively, we have

$$\frac{1}{\sqrt{n}\mu_2^n}Q^n(nt) \Rightarrow \text{RBM}(-\beta, \sigma_a^2 + \sigma_s^2) \ \text{in} \ D[0,\infty) \ \text{as} \ n \to \infty,$$

where $\text{RBM}(-\beta, \sigma_a^2 + \sigma_s^2)$ denotes a eflected Brownian motion with drift coefficient $-\beta$ and diffusion coefficient $\sqrt{\sigma_a^2 + \sigma_s^2}$.

Denote $\hat{Q}(\infty)$ as the stationary distribution of $\hat{Q}$. Then, $\hat{Q}(\infty)$ follows an exponential distribution with rate $2\beta/(\gamma(\sigma_a^2 + \sigma_s^2))$ (Harrison and Williams 1987).

We also establish the interchange-of-limits and uniform integrability result for the queue length processes of $M/GI/1$ queues. Let $Q^n(\infty)$ denote the stationary distribution of $Q^n(t)$.

THEOREM 2. *For the preemptive two-class priority rule, under Assumptions 1 and 4, if the arrival process is Poisson, then for any $x \geq 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{K^n}{\sqrt{n}}Q^n(\infty) \leq x\right) = 1 - \exp\left(-\frac{2\beta}{\gamma(1+\sigma_s^2)}x\right)$$

and

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{K^n}{\sqrt{n}}Q^n(\infty)\right] = \frac{\gamma(1+\sigma_s^2)}{2\beta}.$$

Equivalently, we have,

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}\mu_2^n}\mathbb{E}[Q^n(\infty)] = \frac{1+\sigma_s^2}{2\beta}.$$

The proof of Theorem 1 is delayed to Section 4, where we also present several other results about the asymptotic behavior of the workload processes and the virtual waiting time processes for the two priority classes. The proof of Theorem 2 can be found in Appendix B.

3.3. The effect of the service time distribution

In this section, we study the effect of service time distribution on the performance of the two-class priority rule. We also compare the queue-length scaling under the two-class priority to that under SRPT. A very appealing aspect of Theorems 1 & 2 is that they explicitly characterize the impact of service time distribution on the benefit of smart scheduling through K^n. In particular, Q^n scales as $\sqrt{n}/K^n$ and K^n is determined by the service time distribution. In general, the more slowly the tail of the service time distribution decays, the larger K^n tends to be, which in turn leads to a smaller scaling of $\mathbb{E}[Q^n(\infty)]$. We demonstrate this through two specific classes of service time distributions. These two classes of distributions include a lot of commonly used service time distributions and are widely studied in the SRPT literature (see, for example, Lin et al. (2011), Puha (2015)).

To mark the explicit dependence of threshold K on service time distribution F and the traffic intensity ρ, we use the notation $K(\rho, F)$ introduced in (1).

Pareto service time distributions Consider the family of Pareto distributions: $\bar{F}_\alpha(x) = (m_\alpha/x)^\alpha$ with parameters $\alpha > 1$ and $m_\alpha = (\alpha - 1)/\alpha$. In this case,

$$K(\rho, F_\alpha) = m_\alpha(1 - \rho)^{-(1-\delta)/\alpha}.$$

We note that the smaller the value of the parameter α, the more slowly $\bar{F}_\alpha$ decays, and the larger the value of $K(\rho, F)$ tends to be.

Weibull service time distribution Consider the family of Weibull distributions: $\bar{F}_\alpha(x) = \exp(-(x/\nu_\alpha)^\alpha)$ with parameters $\alpha > 1$ and $v_\alpha = 1/\Gamma(1 + 1/\alpha)$, where Γ is the gamma function. In this case

$$K(\rho, F_\alpha) = \nu_\alpha \left(-(1 - \delta)\log(1 - \rho)\right)^{1/\alpha}.$$

We again note that the smaller the value of the parameter α, the more slowly $\bar{F}_\alpha$ decays, and the larger the value of $K(\rho, F)$ is.

We next compare the two-class priority rule with SRPT. For Pareto service time distributions, Lin et al. (2011) establish that under SRPT,

$$n^{-\frac{\alpha-2}{2(\alpha-1)}} \mathbb{E}[Q^n(\infty)] \to \frac{\pi/(1-\alpha)}{2\sin(\pi/(1-\alpha))} \frac{(1+\sigma_s^2)}{2m_\alpha\beta} \text{ as } n \to \infty.$$

For our two-class priority rule, if we set $K^n = m_\alpha n^{\frac{1}{\alpha}(1/2-\delta)}$ for some $\delta > 0$, we have

$$n^{-\frac{\alpha-1-2\delta}{2\alpha}} \mathbb{E}[Q^n(\infty)] \to \frac{(1+\sigma^2)}{2m_\alpha\beta} \text{ as } n \to \infty.$$

In this case, even though δ can be set arbitrarily small, $\frac{\alpha-1}{2\alpha}$ is larger than $\frac{\alpha-2}{2(\alpha-1)}$, i.e., the scaling of the average queue length is larger under the two-class priority rule than that under SRPT. However, the difference in scaling decreases as α increases.

For Weibull service time distributions, Lin et al. (2011) establish that under SRPT

$$\frac{\log(\sqrt{n})^{1/\alpha}}{\sqrt{n}} \mathbb{E}[Q^n(\infty)] \to \frac{1+\sigma_s^2}{2\nu_\alpha\beta} \text{ as } n \to \infty.$$

For our two-class priority rule, if we set $K^n = \nu_\alpha(\log(n^{1/2-\delta}))^{1/\alpha}$ for some $\delta > 0$, we have

$$\frac{\log(n^{1/2-\delta})^{1/\alpha}}{\sqrt{n}} \mathbb{E}[Q^n(\infty)] \to \frac{1+\sigma_s^2}{2\nu_\alpha\beta} \text{ as } n \to \infty.$$

In this case, the two scheduling policies achieve the same scaling.

4. Proof of Theorem 1

In this section, we provide the proof of Theorems 1. The proof is divided into two steps. We first establish state-space collapse results for the workload process and the queue length process for Class 1 (Proposition 1). We then establish proper limits for the virtual waiting time process and the queue length process for Class 2 (Proposition 2). The detailed proofs of the propositions are deferred to Appendix A. Note that these intermediate results are of independent interests as they highlight the difference in performance of the two priority classes.

Let $A_i^n = \{A_i^n(t) : t \geq 0\}$ denote the arrival process of Class i customers. Note that by our class-division rule, A_i^n's are still renewal processes, but they may not be independent of each other. We also define

$$V^n(t) = \sum_{k=1}^{A^n(t)} v^n(k) - t,$$

$$U^n(t) = V^n(t) - \inf_{0 \leq s \leq t} V^n(s) \wedge 0,$$

$$V_1^n(t) = \sum_{k=1}^{A_1^n(t)} v_1^n(k) - t,$$

$$U_1^n(t) = V_1^n(t) - \inf_{0 \leq s \leq t} V_1^n(s) \wedge 0,$$

$$U_2^n(t) = U^n(t) - U_1^n(t).$$

Note that U^n is the total unfinished workload process, and U_i^n is the unfinished workload process of Class i jobs.

Under any work-conserving service policy and Assumptions 1 and 2, it has been established that (Reiman 1984)

$$\frac{1}{\sqrt{n}} U^n(nt) \Rightarrow \text{RBM}(-\beta, \sigma_a^2 + \sigma_s^2) \text{ in } D[0, \infty) \text{ as } n \to \infty. \tag{3}$$

As Class 1 customers have preemptive priority over Class 2 customers, $U_1^n(t)$ is also the virtual waiting time process for Class 1. The following proposition establishes state-space collapse for $K^n U_1^n(nt)/\sqrt{n}$ and $K^n Q_1^n(nt)/\sqrt{n}$

PROPOSITION 1. *Under Assumptions 1 – 4,*

$$\frac{K^n}{\sqrt{n}} U_1^n(nt) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty,$$

and

$$\frac{K^n}{\sqrt{n}} Q_1^n(nt) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty.$$

Proposition 1 implies the virtual waiting time and the queue length of Class 1 scale as $o(\sqrt{n}/K^n)$, or equivalently, $o(\sqrt{n}\mu_2^n)$.

Define

$$\hat{U}^n(t) = \frac{U^n(nt)}{\sqrt{n}} \text{ and } \hat{U}_2^n(t) = \frac{U_2^n(nt)}{\sqrt{n}}.$$

From Proposition 1, we have

$$\hat{U}^n - \hat{U}_2^n \Rightarrow \zeta \text{ in } D[0,\infty) \text{ as } n \to \infty,$$

and thus,

$$\hat{U}_2^n \Rightarrow \mathrm{RBM}(-\beta, (\sigma_a^2 + \sigma_s^2)) \text{ in } D[0,\infty) \text{ as } n \to \infty.$$

For Class 2 customers, as it has lower priority, the unfinished workload, $U^n(t)$, is smaller than its virtual waiting time at t. To characterize the waiting time of Class 2, we define

$$B^n(t, U) := \inf\{s \geq 0 : V_1^n(t+s) - V_1^n(t) + U \leq 0\}.$$

Then, $B^n(t, U^n(t))$ is the virtual waiting time process for Class 2. We also define

$$\hat{B}^n(t) := B^n(nt, U^n(nt))/\sqrt{n}.$$

PROPOSITION 2. *Under Assumptions 1 – 4,*

$$\rho_2^n \hat{B}^n(t) - \hat{U}^n(t) \Rightarrow \zeta(t) \text{ in } D[0,\infty) \text{ as } n \to \infty,$$

and

$$\frac{K^n}{\sqrt{n}} Q_2^n(nt) \Rightarrow \hat{Q}(t) \text{ in } D[0,\infty) \text{ as } n \to \infty.$$

Proposition 2 indicates the virtual waiting time process for Class 2 scales as

$$\sqrt{n}/\rho_2^n \sim \sqrt{n}/(K^n \bar{F}(K^n)).$$

Note that under FCFS, the virtual waiting time process for Class 2 scales as

$$\sqrt{n} + 1/\mu_2^n \sim \sqrt{n}.$$

Note that under Assumption 4, $K^n \bar{F}(K^n) \to 0$ as $n \to \infty$. Thus the virtual waiting time for Class 2 under the two-class priority rule has a larger scaling than under FCFS. The difference in scaling suggests the overall improvement in average queue length and average waiting time under the two-class priority rule is at the expense of having Class 2 customers waiting for a longer time. We refer to Wierman and Harchol-Balter (2003) for more discussions about fairness under different scheduling policies for the $M/GI/1$ queue.

Combining Proposition 1 and Proposition 2, we have proved Theorem 1.

5. Numerical experiments for the two-class priority rule

In this section, we provide some numerical experiments to illustrate the pre-limit performance of the two-class priority rule. In particular, we study the impact of the service time distribution on system performance. We also look into the class-dependent performance of the two-class priority rule.

We first note that a wide range of $K(\rho, F)$'s satisfies Assumption 4. Thus, we start by conducting some sensitivity analysis for the effect of different values of $K(\rho, F)$ on the performance of the two-class priority rule. Figure 1 plots the steady-state average queue length for different values of the threshold under the two-class priority rule with and without preemption. Note the expected queue length does vary with different values of $K(\rho, F)$. However, the magnitude of variation is very small for the range of values of $K(\rho, F)$ plotted. This finding indicates the two-class priority rule is relatively insensitive to the choice of the threshold within a reasonable range.

Figure 1 The steady-state average queue length for $M/M/1$ queue under the two-class priority rule (P) with different values of the threshold.
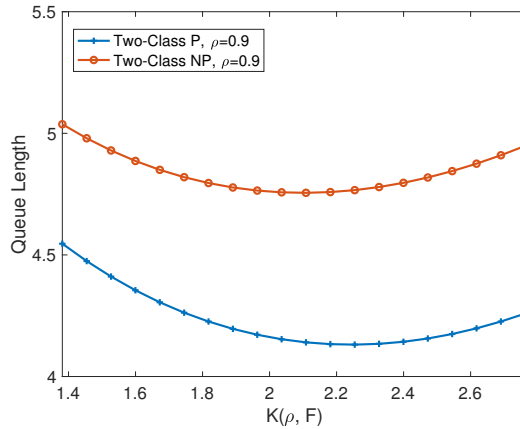
In Table 3, we compare the steady-state average queue length for $M/GI/1$ queues with the same unit service rate but different service time distributions. We consider the class of Pareto distribution with parameter α, where we set $m_\alpha = (\alpha - 1)/\alpha$ and the tail cdf $\bar{F}(x) = (m_\alpha/x)^\alpha$. Note that the larger the value of α, the faster $\bar{F}$ decays, and thus, the lighter the tail of the service time distribution is.

Table 3 Steady-state average queue length for $M/GI/1$ queues with Pareto(m_α, α) service time distributions for different values of α ($\mu = 1$, $\lambda = \rho$, $K(\rho, F) = \bar{F}((1-\rho)^{(1-0.05)})$).

	FCFS			Two-class (P)			SRPT		
ρ	0.8	0.9	0.99	0.8	0.9	0.99	0.8	0.9	0.99
$\alpha = 2.5$	3.68	8.19	89.20	2.53	4.43	21.60	1.88	3.05	10.54
$\alpha = 5$	2.51	5.22	53.26	2.32	4.34	30.20	2.11	3.83	23.03
$\alpha = 7.5$	2.44	5.05	51.18	2.38	4.57	35.91	2.20	4.16	31.64
$\alpha = 10$	2.42	5.00	50.61	2.40	4.70	39.16	2.25	4.35	34.81

We observe that when comparing the two-class priority rule to FCFS, the two class-priority rule always achieves a smaller queue. The queue length reduction under the two-class priority rule is becoming larger when the traffic intensity increases. For example for Pareto service distribution with $\alpha = 2.5$, when $\rho = 0.8$, the two-class priority reduces the average queue length by 32%; when $\rho = 0.9$, the reduction is 46%; in the extreme case when $\rho = 0.99$, the reduction is as large as 76%. These observations are consistent with our asymptotic analysis in Theorems 1 & 2. Specifically, the queue length process scales slower as ρ increases under the two-class priority rule than that under FCFS, and this scaling difference is more apparent when ρ is closer to 1.

We next take a closer look at the effect of the service time distributions. When comparing the two-class priority rule with FCFS, we note that we gain more reduction in queue length for heavier tail service time distributions, i.e., when α is smaller. For example, for $\rho = 0.9$, when $\alpha = 2.5$, we achieve a queue length reduction of 46% using the two-class priority rule; when $\alpha = 5$, the reduction is 17%; when $\alpha = 10$, the reduction is only 6%.

We also oberve that under FCFS, as the tail of the service time distribution becomes lighter, i.e., α increases, the average queue length decreases. This finding is expected, because for $M/GI/1$ queue with unit service rate, the steady-state expected queue length under FCFS takes the form

$$\rho + \frac{(1 + \sigma_s^2)\rho^2}{2(1 - \rho)},$$

and σ_s^2 decreases as α increases. Under the two-class priority rule, Theorem 2 suggests that

$$\mathbb{E}[Q(\infty)] \approx \frac{\gamma(1 + \sigma_s^2)}{2(1 - \rho)K(\rho, F_\alpha)} \text{ where } K(\rho, F_\alpha) = m_\alpha(1 - \rho)^{-(1-\delta)/\alpha}.$$

Note that when α increases, both σ_s^2 and $K(\rho, F_\alpha)$ decreases. As σ_s^2 does change with ρ, the effect of $K(\rho, F_\alpha)$ will dominate when ρ is close enough to 1. However, when ρ is small, whether the smaller σ_s^2 or the smaller K^n will dominate is unclear. From Table 3, we observe that when $\rho = 0.8$, the average queue length is first decreasing and then increasing as α increases. In this regime, the effect of σ_s^2 plays a role. However, when $\rho = 0.99$, the average queue length increases as α increases. In this regime, the effect of $K(\rho, F_\alpha)$ dominates.

Lastly, when comparing SRPT with the two-class priority rule, we note that when α is small, SRPT can achieve substantially shorter average queue length than the two-class priority rule. However, the expected queue lengths under the two scheduling policies are getting closer as α increases. For example, for $\rho = 0.9$, when $\alpha = 5$, the optimality gap of the two-class priority rule is 13%; when $\alpha = 10$, the gap is 8%. This observation is consistent with our analysis in Section 3.3.

We next demonstrate the class-dependent performance under the two-class priority rule. Figure 2 compares the steady-state average queue length for Class 1 and Class 2 under the two-class priority

rule for an $M/M/1$ queue. We also plot the average total queue length. We observe that because Class 1 is prioritized, it has a much shorter queue than Class 2. Table 4 compares the steady-state average waiting times for Class 1, Class 2, and all customers together (All) under the two-class priority rule and FCFS. We observe that compared with FCFS, despite the huge improvement in average waiting times for Class 1 customers, Class 2 customers do incur a significant increase in waiting times under the two-class priority rule. These observations are consistent with our asymptotic analysis in Propositions 1 & 2.

Figure 2   **Steady-state average queue lengths of $M/M/1$ queues with different traffic intensities under the two-class priority rule (P).**
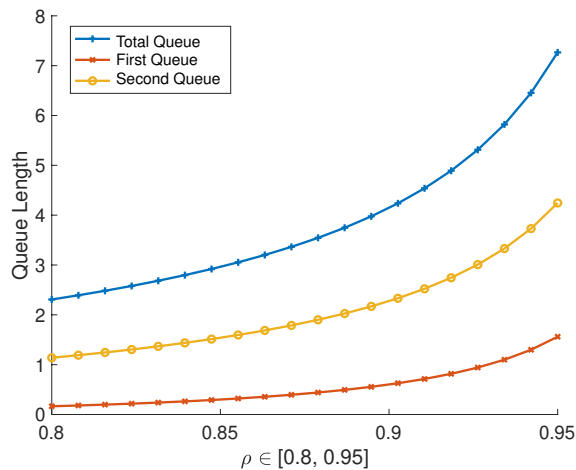
Table 4   **Steady-state average waiting time for $M/M/1$ queues under the two-class priority rule (P) and FCFS $\left(\mu=1,\ \lambda=\rho,\ K(\rho)=\bar{F}((1-\rho)^{(1-0.05)})\right)$.**

	Two-Class (P)			FCFS		
	Class 1	Class 2	All	Class 1	Class 2	All
$\rho=0.8$	0.83	10.19	2.86	4.58	6.53	5.00
$\rho=0.85$	1.07	15.61	3.46	6.31	8.47	6.67
$\rho=0.9$	1.52	28.98	4.59	9.72	12.19	10.00
$\rho=0.95$	2.79	89.34	7.68	19.82	22.85	20.00

6. Imperfect service-time information

As an added benefit to our two-class priority rule, we can provide some theoretical quantification for the effect of imperfect service-time information on system performance. Employing the same heavy-traffic asymptotic mode of analysis as in Section 3, our goal is to be able to gain "order-of-magnitude" performance improvement using the two-class priority rule over FCFS. When we do not have perfect job-size information, what we classified as Class 1 jobs may be different from jobs whose actual service times are less than or equal to K^n. For $i=1,2$, let $\hat{\lambda}_i^n$, $\hat{\mu}_i^n$, and $\hat{\rho}_i^n$ denote

the arrival rate, service rate, and traffic intensity of the classified Class i jobs respectively. The key insights from our perfect information analysis are that as long as the following two conditions hold, the queue length under the two-class priority rule still scales slower than $(1-\rho_n)^{-1}$.

C1) The average service time of the classified Class 2 jobs goes to infinity as the traffic intensity approaches 1. In particular, if $\hat{\mu}_2^n \to 0$ as $n \to \infty$, the Class 2 queue still scales slower than $(1-\rho_n)^{-1}$.

C2) The queue of the classified Class 1 jobs under a proper scaling (determined by the size of the Class 2 queue) diminishes as the traffic intensity approaches 1. In particular, if $n^{1/2-\delta}\hat{\mu}_2^n(1-\hat{\rho}_1^n) \to \infty$ as $n \to \infty$ for some $0 < \delta < 1/2$, the Class 1 queue still scales slower than the Class 2 queue. These two conditions provide quite some flexibility to accommodate estimation errors for the service times. In our subsequent development, we study how different estimation errors would affect the above two conditions (C1 and C2).

We denote $v^n(k)$ as the actual size of the k-th job in the n-th system and $\hat{v}^n(k)$ as its estimated size. We also write $\varepsilon^n(k) = v^n(k) - \hat{v}^n(k)$ as the estimation error. As the service time does not scale with n, we denote v as a generic service time, $\hat{v}$ as a generic estimated service time, and ε as a generic estimation error. We refer to the real Class 1 jobs as jobs whose actual service times are no larger than the threshold K^n, i.e, $v^n(k) \leq K^n$, and similarly for the real Class 2 jobs. We refer to the classified Class 1 jobs as jobs who are assigned to Class 1 based on some estimation, and similarly for the classified Class 2 jobs.

We consider three specific forms of estimation errors. The first one has bounded estimation errors. We employ a worst-case analysis to demonstrate the robustness of the two-class priority rule. The second one is a simple classification model, which allows us to investigate the impact of two different types of classification errors: wrongly classifying a Class 1 job as Class 2 versus wrongly classifying a Class 2 job as Class 1. Our third model assumes iid measurement errors, which allows us to investigate the interplay between the tail of the measurement error distribution and the tail of the service time distribution. Even though these are simplified forms of estimation errors, they provide important insights into the effect of estimation errors on system performance under the two-class priority rule. In Section 7, we complement the theoretical analysis in this section with numerical experiments for problems with more general predicted service times and in more realistic settings.

6.1. Bounded estimation errors

Consider the case where there exits $M \in (0, \infty)$ such that $\varepsilon^n(k) \in [-M, M]$ with probability one for all $k \in \mathbb{N}$. We do not impose further assumptions on the estimation errors. For example, $\varepsilon^n(k)$ can have non-zero mean and can depend on $v^n(k)$ and/or $\hat{v}^n(k)$. For the two-class priority rule, we assign the jobs whose predicted service time is smaller than or equal to K^n, i.e., $\hat{v}^n(k) \leq K^n$, to Class 1; and the others are assigned to Class 2.

With bounded estimation errors, we can take a worst case analysis approach. In particular, the worst-case scenario that is "against" our two-class priority rule is the one where we wrongly classify all jobs whose actual service times are in $[K^n - M, K^n]$ as Class 2, and all the jobs whose actual service times are in $(K^n, K^n + M]$ as Class 1. (This notion will be made precise in the proof of Proposition 3.) Even in this worst-case scenario, we are still able to show that $1/\hat{\mu}_2^n \sim K^n$ and $n^{1/2-\delta}\hat{\mu}_2^n(1 - \rho_2^n) \to \infty$ as $n \to \infty$ for some $0 < \delta < 1/2$. Thus, we achieve the same scaling for the queueing length process as in the perfect information setting:

PROPOSITION 3. *For the preemptive two-class priority rule with estimated service times, under Assumptions 1 – 4, suppose there exits a constant $M \in (0, \infty)$ such that $\varepsilon \in [-M, M]$ with probability one. Then, $1/\hat{\mu}_2^n \sim K^n$ and*

$$\frac{1}{\sqrt{n}\hat{\mu}_2^n}Q^n(nt) \Rightarrow RBM(-\beta, \sigma_a^2 + \sigma_s^2) \ in \ D[0, \infty) \ as \ n \to \infty.$$

The proof of Proposition 3 and all subsequent results in this section can be found in Appendix C.

6.2. A special classification model

In this section, we assume that there exists a classification model that classifies incoming jobs into the two priority classes. The classification errors depend on the underlying classification model. For example, if there is a prediction model for service times, e.g, the model considered in Section 6.1, and we classify jobs whose predicted service times are less than or equal to K^n as Class 1, then the probability of wrongly classifying a Class 2 job as Class 1 depends on the actual service time of the job and the accuracy of the prediction model. In this case, jobs whose actual service times are closer to K^n may be more likely to be wrongly classified.

In this section, we consider a simplified setting where for the n-th system, each real Class 1 job has an equal probability of being wrongly classified as a Class 2 job. We denote this probability as p_{12}^n. Likewise, each real Class 2 job has an equal probability of being wrongly classified as a Class 1 job. We denote this probability as p_{21}^n. The goal is to highlight the difference between the two types of classification errors, i.e, p_{12}^n versus p_{21}^n.

Let $p_{11}^n = 1 - p_{12}^n$ and $p_{22}^n = 1 - p_{21}^n$. Then, we have

$$\hat{\lambda}_i^n = p_{1i}^n \lambda^n F(K^n) + p_{2i}^n \lambda^n \bar{F}(K^n),$$

$$\hat{\mu}_i^n = \left(\frac{p_{1i}^n \lambda^n F(K^n)}{\hat{\lambda}_i^n} \frac{\int_0^{K^n} xf(x)dx}{F(K^n)} + \frac{p_{2i}^n \lambda^n \bar{F}(K^n)}{\hat{\lambda}_i^n} \frac{\int_{K^n}^{\infty} xf(x)dx}{\bar{F}(K^n)} \right)^{-1}.$$

PROPOSITION 4. *For the preemptive two-class priority rule with a classification model where each real Class i job has equal probability of being classified as Class j, $i, j = 1, 2$, assume Assumptions 1 – 4 hold.*

i) If $p_{12}^n = O(\bar{F}(K^n))$ and $p_{21}^n \le a$ for some $0 \le a < 1$, then $1/\hat{\mu}_2^n \sim K^n$ and

$$\frac{1}{\sqrt{n}\hat{\mu}_2^n} Q^n(nt) \Rightarrow RBM(-\beta, \sigma_a^2 + \sigma_s^2) \text{ in } D[0,\infty) \text{ as } n \to \infty.$$

ii) If $p_{12}^n/\bar{F}(K^n) \to \infty$ as $n \to \infty$ and $p_{21}^n \le a$ for some $0 \le a < 1$, then $\lim_{n\to\infty} \hat{\mu}_2^n = 1$.

Proposition 4 indicates that it is more important to avoid wrongly classifying real Class 1 customers as Class 2 than the other way around. In particular, as long as p_{12}^n, decays to zero sufficiently fast and the probability of the second type of error, p_{21}^n, is bounded away from 1, we can achieve the same queue-length scaling as in the perfect information case. On the other hand, if p_{12}^n decays to 0 too slowly, the queue may scale the same as that under FCFS (Case ii in Proposition 4).

6.3. Measurement errors

To gain more insights into the interplay between the service time distribution and the distribution of estimation errors, in this section, we consider a specific form of estimated service times that may arise due to measurement errors. We assume $\varepsilon^n(k)$'s are iid with finite mean and variance, and $\varepsilon^n(k)$ is independent of $v^n(k)$. Under the two class priority rule, if the estimated job size is smaller than or equal to K^n, we assign the job to Class 1; otherwise, the job is assigned to Class 2.

In this case, without loss of generality, we write

$$\hat{v}^n(k) = v^n(k) + \varepsilon^n(k)$$

instead of $\hat{v}^n(k) = v^n(k) - \varepsilon^n(k)$. We assume ε is a continuous random variable with probability density function ϕ. We also write Φ as the cdf of ε and $\bar{\Phi} := 1 - \Phi$ as its tail cdf. Then,

$$\hat{\lambda}_1^n = \lambda^n \mathbb{P}(v + \varepsilon < K^n) = \left(1 - \frac{\beta}{\sqrt{n}}\right) \int_0^\infty f(t)\Phi(K^n - t)dt;$$

$$\hat{\lambda}_2^n = \lambda^n \mathbb{P}(v + \varepsilon \ge K^n) = \left(1 - \frac{\beta}{\sqrt{n}}\right) \int_0^\infty f(t)\bar{\Phi}(K^n - t)dt;$$

$$1/\hat{\mu}_1^n = \mathbb{E}[v|v + \varepsilon < K^n] = \frac{\int_0^\infty tf(t)\Phi(K^n - t)dt}{\int_0^\infty f(t)\Phi(K^n - t)dt};$$

$$1/\hat{\mu}_2^n = \mathbb{E}[v|v + \varepsilon \ge K^n] = \frac{\int_0^\infty tf(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt}.$$

We first note that

$$1 - \hat{\rho}_1^n = 1 - \left(1 - \frac{\beta}{\sqrt{n}}\right) \int_0^\infty tf(t)\Phi(K^n - t)dt$$

$$= \int_0^\infty tf(t)\bar{\Phi}(K^n - t)dt + O(1/\sqrt{n}).$$

In addition, note that for $\delta > 0$ chosen according to Assumption 4,

$$
\begin{aligned}
n^{1/2-\delta}\hat{\mu}_2^n(1-\hat{\rho}_1^n) &= n^{1/2-\delta}\int_0^\infty f(t)\bar{\Phi}(K^n-t)dt + o(1) \\
&\geq n^{1/2-\delta}\int_{K^n}^\infty f(t)\bar{\Phi}(K^n-t)dt + o(1) \\
&\geq \bar{\Phi}(0)n^{1/2-\delta}\bar{F}(K^n) \to \infty \text{ as } n \to \infty.
\end{aligned}
$$

This indicates that $n^{1/2-\delta}\mu_2^n(1-\rho_n) \to \infty$ as $n \to \infty$. In particular, even with estimation errors, the Class 1 queue still scales slower than the Class 2 queue. Thus, our main focus in subsequent analysis is to decide whether $\hat{\mu}_2^n \to 0$ as $n \to \infty$, i.e., to see whether the Class 2 queue scales slower than $(1-\rho^n)^{-1}$.

The following assumption put restrictions on the tail of the distribution of ε. In particular, it requires that ε has a lighter tail than v.

ASSUMPTION 5. *There exists $c > 0$ such that for any $0 < \theta < c$, $\mathbb{E}[\exp(-\theta\varepsilon)] < \infty$. For any $M > 0$, $\lim_{t\to\infty} \bar{\Phi}(t-M)/\bar{F}(t) = 0$.*

PROPOSITION 5. *For the preemptive two-class priority rule using estimated service times with iid measurement errors, under Assumptions 1 - 5, $\hat{\mu}_2^n \to 0$ and*

$$
\frac{1}{\sqrt{n}\hat{\mu}_2^n}Q^n(nt) \Rightarrow RBM(-\beta, \sigma_a^2 + \sigma_s^2) \text{ in } D[0,\infty) \text{ as } n \to \infty.
$$

If we further assume that for any $a > 0$, $\lim_{t\to\infty} \bar{\Phi}(at)/\bar{F}(t) = 0$, then $1/\hat{\mu}_2^n \sim K^n$.

Proposition 5 indicates that if the estimation error distribution has a lighter tail than the service time distribution, we can still achieve $o(\sqrt{n})$ scaling for the queue length process in heavy traffic. From the proof of Proposition 5, we note that the exact order of the scaling may depend on "how much lighter" the estimation error distribution is in comparison to the service time distribution.

Given Theorem 5, a natural follow-up question is what would happen if the estimation error distribution has a heavier tail than the service time distribution. We next study a special case of this:

LEMMA 1. *Suppose the distribution of the estimation error is regularly varying. In particular, there exists a function $\xi : (0,\infty) \to (0,\infty)$ such that for any $a > 0$, $\lim_{x\to\infty} \bar{\Phi}(ax)/\bar{\Phi}(x) = \xi(a)$. Under Assumptions 1 - 4, if*

$$
\lim_{x\to\infty} \frac{\bar{\Phi}(x)}{x\bar{F}(x)} = c < \infty,
$$

then $\lim_{n\to\infty} \hat{\mu}_2^n > 0$.

Lemma 1 requires the distribution of ε to be regularly varying, For example, the error distribution can have a Pareto tail, $\bar{\Phi}(x) = \left(\frac{m}{x+m\alpha/(\alpha-1)}\right)^\alpha$. It also requires that the error distribution has a

heavier tail than the service time distribution. For example, if the service time distribution is Pareto with parameter α_s, $\alpha_s > 3$, and the error distribution has a Pareto tail with parameter $\alpha_e \in (1, \alpha_s - 1]$, the conditions of Lemma 1 are satisfied. In this case, the two-class priority rule gives rise to the same queue-length scaling as FCFS.

7. Numerical experiments for predicted service times

In this section, we consider several classical prediction models for the service times, and use simulation to evaluate the performance of the two-class priority rule when we only have access to predicted service times. Focusing on more realistic service system settings, we study the non-preemptive version of the two-class priority rule (2NP), and choose FCFS and SJF with predicted service times as two benchmark policies. All the reported steady-state average queue length is estimated based on the average queue length over $10^7/\mu$ units of time (long-time average). To facilitate comparison, common random numbers are used when simulating the system under different scheduling policies.

EXAMPLE 1 (A LINEAR REGRESSION MODEL). Consider the setting where we know certain features of the customers and the service time model takes the form

$$v_i = \beta^T X_i + \epsilon_i,$$

where X_i denotes the vector of observable features of the i-th customer, β is the vector of coefficients, and ϵ_i is the residual that is assumed to have mean zero and is uncorrelated with X_i. Note that in this case, $\mathbb{E}[v_i|X_i] = \beta^T X_i$.

In Table 5, we consider the linear model: $\mathbb{E}[v_i|X_i] = 0.1 + 0.1X_{1i} + 0.4X_{2i} + 0.4X_{3i}$, where $X_{1i} \sim N(1,1)$, $X_{2i} \sim$ Exponential(1), and $X_{3i} \sim$ Uniform$[0,2]$, and ϵ_i's are iid $N(0, \sigma_e^2)$ and are independent of X_i's. Note that in this case, ϵ_i and v_i are dependent. Thus, this model is different from the measurement error model studied in Section 6.3. When generating service times, we set

$$v_i = (\mathbb{E}[v_i|X_i] + \epsilon_i) \vee 0.001 \tag{4}$$

to ensure that the service times are positive. The value of σ_e varies form 0 to 0.5. Due to (4), setting σ_e too large would require a non-negligible amount of truncations which makes the service time model quite different from the linear regression model. Thus, σ_e is capped at 0.5. When running 2NP and SJF, we use $\beta^T X_i$ as the predicted service time. The traffic intensity of the system is kept at 95% and the threshold of the two-class priority rule is set to be the 95-th quantile of the predicted service time distribution, which is equal to 1.8 in this case.

We observe from Table 5 that as σ_e increases from 0 to 0.5, the queue length reduction we gain from 2NP over FCFS only decreases slightly, i.e., from 36% when $\sigma_e = 0$ to 35% when $\sigma_e = 0.5$.

This suggests that 2NP is quite robust to estimation errors. We also observe that the two-class priority rule achieves comparable but worse performance than SJF.

Table 5 Steady-state average queue length of $M/G/1$ queues with predicted service times $\beta^T X_i$. $\lambda = 0.95\mu$. For the two class priority rule, we set the threshold $K = 1.8$.

σ_e	0	0.25	0.5
2NP	7.67	7.92	8.31
SJF	6.39	6.50	6.96
FCFS	12.00	12.35	12.79

EXAMPLE 2 (A LOG-LINEAR REGRESSION MODEL). Consider the following service time model

$$v_i = \exp\big(\beta^T X_i + \epsilon_i\big),$$

where X_i denotes the vector of observable features for the i-th customer, β is the vector of coefficients, and ϵ_i has mean zero and is uncorrelated with X_i. In this case, $\mathbb{E}[\log(v_i)|X_i] = \beta^T X_i$.

In Table 6, we set $\mathbb{E}[\log(v_i)|X_i] = 0.1 + 0.1X_{1i} + 0.4X_{2i} + 0.4X_{3i}$, where $X_{1i} \sim N(1,1)$, $X_{2i} \sim$ Exponential(1), and $X_{3i} \sim$ Uniform[0, 2], and ϵ_i's are iid $N(0, \sigma_e^2)$ and are independent of X_i's. The value of σ_e is varied from 0 to 1.5. When running 2NP and SJF, we use $\exp(\beta^T X_i)$ as the predicted service time. Note that in this case, the predicted service time is biased, i.e., it does not have the same expectation as the actual service time, and the estimation errors conditional on the observable features, i.e, $v_i - \exp(\beta^T X_i)$ conditional on X_i, are not identically distributed. The traffic intensity of the system is kept at 95% and the threshold of the two-class priority rule is set to be the 95% quantile of the predicted service time distribution, which is equal to 6.2 in this case.

We first observe that across all levels of estimation errors tested in Table 6, 2NP leads to a more than 50% reduction in the steady-state average queue length when comparing to FCFS. In the extreme case when $\sigma_e = 1.5$, the predicted service times explain only 3% of the variability in the actual service times. Even in this case, 2NP achieves a 51% reduction in average queue length. We also note that under 2NP, the proportion reduction is not monotonically decreasing as σ_e increases. This is because in this example, as σ_e increases, the coefficient of variation of the service time distribution increases significantly. For example, when $\sigma_e = 0$, the coefficient of variation is only 0.92; when $\sigma_e = 1.5$, the coefficient of variation is 4.50. When the service times are more heterogeneous, we may gain more from prioritizing shorter jobs. Lastly, we observe that SJF achieves an even smaller average queue length than 2NP. For example, when $\sigma_e = 0.5$, 2NP leads to a 49% reduction while SJF leads to a 60% reduction in average queue length over FCFS. When $\sigma_e = 1$, 2NP leads to a 55% reduction while SJF leads to a 68% reduction. This suggests that we can gain further performance improvement by introducing more priority classes, even with

estimation errors. Note that SJF can be viewed as the limit of a properly defined sequence of multi-class priority rules as the number of priority classes goes to infinity (Schrage and Miler 1966).

Table 6 Steady-state average queue length of $M/G/1$ queues with predicted service times $\exp(\beta^T X_i)$. $\lambda = 0.95\mu$. **For the two class priority rule, we set the threshold** $K = 6.5$.

σ_e	0	0.25	0.5	0.75	1	1.25	1.5
2NP	8.36	9.20	10.89	14.32	20.76	34.70	67.58
SJF	6.52	6.74	8.46	10.09	14.75	24.06	47.18
FCFS	16.82	18.47	21.21	30.67	46.00	74.52	137.54

We next take a closer look at why even with very noisy predictions, utilizing the predicted service times to do smart scheduling can still achieve significant performance improvement. Table 7 reports the R^2, which measures the proportion of variability in the actual service times that can be explained by the prediction model, p_{12}, which is the probability of wrongly classifying a Class 1 job into Class 2, and p_{21}, which is the probability of wrongly classifying a Class 2 job into Class 1. We observe that even though R^2 decreases substantially and p_{21} increases substantially as σ_e increases, p_{12} only increases by a very small amount and is still very close to zero. In particular, p_{12} increases from 0.011 when $\sigma_e = 0.25$ to 0.033 when $\sigma_e = 1.5$. Based on our analysis in Section 6.2, p_{12} plays a more important role in determining the performance of the two-class priority rule. In particular, Proposition 4 indicates as long as p_{12} is sufficiently small and $p_{21} < 1$, we can achieve similar performance as in the perfect information case.

Table 7 Estimation errors for Example 2

σ_e	0.25	0.5	0.75	1	1.25	1.5
R^2	87%	61%	33%	20%	8%	3%
p_{12}	0.011	0.017	0.022	0.026	0.029	0.033
p_{21}	0.324	0.630	0.778	0.837	0.868	0.886

EXAMPLE 3 (A GAMMA MODEL). Consider a simplified supermarket checkout setting where we get to observe the number of items a customer is purchasing, which we denoted by L_i for the i-th customer. Given L_i, we assume the actual service time v_i has a Gamma distribution with parameters L_i and θ, where θ is average processing time of each item, which we normalize to be one unit of time.

In Table 8, we assume L_i's are iid Poisson random variables with rate 10. When running 2NP and SJF, we use L_i as the predicted service time. The traffic intensity ρ varies from 80% to 95%. For a particular value of ρ, the threshold K is set to be the $100\rho\%$ quantile of L_i.

We observe from Table 8 that as ρ increases, we gain more performance improvement from 2NP over FCFS. Specifically, when $\rho = 0.8$, the reduction in steady-state average queue length under

2NP is only 13%. When $\rho = 0.95$, the reduction is 22%. In addition, we observe again that SJF achieves a slighter better performance than 2NP. However, SJF may be harder to implement in practice as the sequence of waiting customers may change upon each new arrival.

Table 8   **Steady-state average queue length of** $M/G/1$ **queues with predicted service times** L_i**.** $\lambda = 0.95/10$**.**

ρ	0.8	0.85	9	0.95
K	12	13	14	15
2NP	2.36	3.20	4.79	9.15
SJF	2.27	3.00	4.32	8.08
FCFS	2.71	3.69	5.72	11.72

8.   Concluding remarks

In this paper, we propose a simple two-class priority rule that achieves comparable performance to SRPT. We characterize the process-level diffusion limit for the queue length process in heavy-traffic. The diffusion scaling is nonstandard and depends on the service time distribution. We also demonstrate the robustness of the two-class priority rule to service-time misspecification through both theoretical analyses and numerical experiments.

Our theoretical analysis of service-time misspecification focuses on simplified models of estimation errors to gain analytical tractability. Our numerical experiments consider more general regression models for service times and demonstrate the good performance of the two-class priority rule. Theoretical analysis of the two-class priority rule with more general predicted service times would be an interesting future research direction. We also observe in our numerical experiments that SJF with predicted service times, in general, achieves better performance than the two-class priority rule. This suggests that in some cases, it might be beneficial to add more priority classes. Theoretical analysis of the performance of SJF with predicted service times would be another interesting future research direction. Lastly, it would be interesting to extend the analysis to multi-server queues (Some numerical experiments for multi-server queues is provided in Appendix E.2). When having multiple servers, we may be able to strike a better balance between system efficiency (average waiting time) and fairness among jobs of different sizes.

Appendix A: Proofs of the results in Section 4

We start by defining a few more notations. Let

$$\hat{U}_1^n(t) = \frac{K^n}{\sqrt{n}} U_1^n(nt) \text{ and } \hat{Q}_i^n(t) = \frac{K^n}{\sqrt{n}} Q_i^n(nt) \text{ for } i = 1, 2.$$

A.1. Proof of of Proposition 1

Let W_k^n denote the waiting time of the k-th Class 1 arrival in queue in the n-th system. Then based on Lindley's recursion for the single-server queue, we have for any $T > 0$ (Blanchet et al. 2018),

$$\sup_{0 \le t \le T} U_1^n(t) \le \sup_{1 \le k \le A_1^n(T)} (W_k^n + v_1^n(k)).$$

Thus, for any fixed constant $a > 0$, we have

$$\mathbb{P}\left(\sup_{0 \le t \le T} \hat{U}_1^n(t) > \epsilon\right) \le \mathbb{P}\left(\max_{1 \le k \le A_1^n(nT)} \frac{K^n}{\sqrt{n}}(W_k^n + v_1^n(k)) > \epsilon\right)$$

$$\le \underbrace{\mathbb{P}\left(\max_{1 \le k \le (1+a)nT} \frac{K^n}{\sqrt{n}}(W_k^n + v_1^n(k)) > \epsilon\right)}_{\text{(I)}} + \underbrace{\mathbb{P}(A_1^n(nT) > (1+a)nT)}_{\text{(II)}}.$$

We next analyze (I) and (II) one by one.

For (I), define $\Delta_k^n := v_1^n(k-1) - \tau^n(k)$, and $S_0^n = 0$ and $S_k^n = S_{k-1}^n + \Delta_k^n$. Then,

$$W_{k+1}^n = \max\{W_k^n + \Delta_{k+1}^n, 0\} = \max_{0 \le i \le k+1}\{S_{k+1}^n - S_i^n\} \overset{D}{=} \max_{0 \le i \le k+1} S_i^n.$$

Let $M^n = \sup_{k \ge 1} S_k^n$. Then,

$$\mathbb{P}\left(\max_{1 \le k \le N} W_k^n > a\right) \le \sum_{k=1}^N \mathbb{P}(W_k^n > a) \le N\mathbb{P}(M^n > a).$$

This further indicates that

$$\mathbb{P}\left(\max_{1 \le k \le (1+a)nT} \frac{K^n}{\sqrt{n}}(W_k^n + v_1^n(k)) > \epsilon\right) \le \underbrace{\mathbb{P}\left(\max_{1 \le k \le (1+a)nT} \frac{K^n}{\sqrt{n}} v_1^n(k) > \epsilon/2\right)}_{\text{(i)}} + \underbrace{(1+a)nT\mathbb{P}\left(\frac{K^n}{\sqrt{n}} M^n > \epsilon/2\right)}_{\text{(ii)}}.$$

For (i), because $v_1^n(k) \le K^n$ and $(K^n)^2/\sqrt{n} \to 0$ as $n \to \infty$ under Assumptions 1 and 4, we have (i) converges to 0 as $n \to \infty$.

For (ii), because

$$\mathbb{E}[\Delta_k^n] = \frac{1}{F(K^n)} \int_0^{K^n} x f(x) dx - \frac{1}{\lambda^n F(K^n)} = \frac{1}{\lambda^n F(K^n)}(\rho_1^n - 1) < 0,$$

M^n is the all time maximum of a random walk with negative. We define

$$\psi_n(\theta) := \log \mathbb{E}[\exp(\theta \Delta_k^n)].$$

As $v_1^n(k-1) \le K^n$, there exists $\epsilon > 0$, such that $\psi_n(\theta) < \infty$ for any $\theta > -\epsilon$. As $\psi_n'(0) = \mathbb{E}[\Delta_k^n] < 0$ and $\psi_n''(0) = \text{Var}(\Delta_k^n) > 0$, there exists $\theta_n > 0$ such that $\psi_n(\theta_n) = 0$. By Cramér-Lundberg bound (Asmussen 2008),

$$\mathbb{P}(M^n > c) \le \exp(-\theta_n c).$$

From Corollary 3 in Glynn and Whitt (1994b), we also have

$$\theta_n = 2\frac{1}{\text{Var}(\Delta_k^n)\lambda^n F(K^n)}(1 - \rho_1^n) + o(\mathbb{E}[\Delta_k^n]).$$

Because $\text{Var}(\Delta_k^n) \to \sigma_a^2 + \sigma_s^2$ and $\lambda^n F(K^n) \to 1$ as $n \to \infty$,

$$\lim_{n \to \infty} \theta_n/(1-\rho_1^n) = 2/(\sigma_a^2 + \sigma_s^2).$$

This implies that $\theta_n \sim 1 - \rho_1^n$. Then,

$$\text{(ii)} \leq (1+a)nT \exp\left(-\theta_n \frac{\sqrt{n}\epsilon}{K^n 2}\right) \to 0 \text{ as } n \to \infty.$$

For (II), let $A_0(t)$ denote the renewal process with interarrival times distributed as τ^∞. Under Assumptions 2 and 3, by Theorem 1 in Glynn and Whitt (1994a), we have

$$\lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}[\exp(\theta A_0(t))] = -\psi_a^{-1}(-\theta).$$

Then by Gartner-Ellis theorem, there exists $a > 0$, such that

$$\frac{1}{t} \log \mathbb{P}(A_0(t) > (1+a)t) = -I(a), \text{ where } I(a) = \sup_\theta\{\theta a + \psi_a^{-1}(-\theta)\} > 0. \tag{5}$$

In particular, the probability $\mathbb{P}(A_0(t) > (1+a)t)$ decays exponentially fast in t. This implies

$$\mathbb{P}(A_1^n(nT) > (1+a)nT) \leq \mathbb{P}(A_0(nT) > (1+a)nT) \to 0 \text{ as } n \to \infty.$$

Above all, if $a > 0$ is chosen according to (5), we have established that both (I) and (II) converge to zero as $\epsilon \to \infty$. This concludes the proof of state-space collapse for $K^n U_1^n(nt)/\sqrt{n}$.

We next establish the state-space collapse for $K^n Q_1^n(nt)/\sqrt{n}$. It follows similar lines of arguments as the proof of Theorem 4 in Section 3.2 of Reiman (1984). However, there are several extra technical difficulties we need to address along the way. We denote $a_1^n(t)$ as the arrival time of the Class 1 customer that is in service at time t. If there is no Class 1 customer in service, we set $a_1^n(t) = t$. We also define $\bar{a}_1^n(t) = a_1^n(nt)/n$.

As

$$A_1^n(t) - A_1^n(a_1^n(t)) \leq Q_1^n(t) \leq A_1^n(t) - A_1^n(a_1^n(t)) + 1,$$

$$\frac{K^n}{\sqrt{n}}\left(A_1^n(nt) - A_1^n(n\bar{a}_1^n(t))\right) \leq \hat{Q}_1^n(t) \leq \frac{K^n}{\sqrt{n}}\left(A_1^n(nt) - A_1^n(n\bar{a}_1^n(t))\right) + \frac{K^n}{\sqrt{n}}. \tag{6}$$

We also note that as

$$U_1^n(a_1^n(t)) \leq t - a_1^n(t) \leq U_1^n(a_1^n(t)) + v_1^n(A_1^n(a_1^n(t))),$$

$$\hat{U}_1^n(\bar{a}_1^n(t)) \leq K^n\sqrt{n}(t - \bar{a}_1^n(t)) \leq \hat{U}_1^n(\bar{a}_1^n(t)) + \frac{K^n}{\sqrt{n}}v_1^n(A_1^n(a_1^n(nt))). \tag{7}$$

As $v_1^n(k) \leq K^n$,

$$\frac{K^n}{\sqrt{n}}v_1^n(A_1^n(a_1^n(nt))) \leq \frac{(K^n)^2}{\sqrt{n}} \to 0 \text{ as } n \to \infty,$$

where the convergence holds under Assumptions 1 and 4. Then, from (7),

$$\sup_{0 \leq t \leq T}\left|K^n\sqrt{n}(t - \bar{a}_1^n(t)) - \hat{U}_1^n(\bar{a}_1^n(t))\right| \Rightarrow 0 \text{ as } n \to \infty. \tag{8}$$

From (6), we have

$$\sup_{0 \leq t \leq T}\hat{Q}_1^n(t) \leq \underbrace{\sup_{0 \leq t \leq T}\frac{K^n}{\sqrt{n}}|A_1^n(nt) - A_1^n(n\bar{a}_1^n(t))|}_{\text{(A)}} + \frac{K^n}{\sqrt{n}}.$$

We first note under Assumptions 1 and 4, $K^n/\sqrt{n} \to 0$ as $n \to 0$.

Next, for (A), set $\delta_n = \frac{\epsilon}{4(1+a)} \frac{\sqrt{n}}{K^n}$, where a is chosen according to (5). Then,

$$\mathbb{P}\left(\sup_{0 \le t \le T} \frac{K^n}{\sqrt{n}} |A_1^n(nt) - A_1^n(n\bar{a}_1^n(t))| > \epsilon \right)$$

$$= \mathbb{P}\left(\sup_{0 \le t \le T} \frac{K^n}{\sqrt{n}} |A_1^n(nt) - A_1^n(n\bar{a}_1^n(t))| > \epsilon;\ \sup_{0 \le t \le T} |nt - n\bar{a}_1^n(t)| > \delta_n \right)$$

$$+ \mathbb{P}\left(\sup_{0 \le t \le T} \frac{K^n}{\sqrt{n}} |A_1^n(nt) - A_1^n(n\bar{a}_1^n(t))| > \epsilon;\ \sup_{0 \le t \le T} |nt - n\bar{a}_1^n(t)| \le \delta_n \right)$$

$$\le \underbrace{\mathbb{P}\left(\sup_{0 \le t \le T} \frac{K^n}{\sqrt{n}} |nt - n\bar{a}_1^n(t)| > \frac{\epsilon}{4(1+a)} \right)}_{\text{(a)}} + \underbrace{\mathbb{P}\left(\sup_{0 \le t \le nT} \frac{K^n}{\sqrt{n}} |A_1^n(t + \delta_n) - A_1^n(t)| > \frac{\epsilon}{2} \right)}_{\text{(b)}}$$

For (a), from (8), we have

$$\frac{K_n}{\sqrt{n}} (nt - n\bar{a}_1^n(t)) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty.$$

Thus, (a) converges to zero as $n \to \infty$. For (b),

$$\mathbb{P}\left(\sup_{0 \le t \le nT} \frac{K^n}{\sqrt{n}} A_1^n(t + \delta_n) - A_1^n(t) > \frac{\epsilon}{2} \right)$$

$$\le \mathbb{P}\left(\sup_{i=0,\ldots,\lceil \frac{nT}{\delta_n} \rceil} A_1^n(i\delta_n + 2\delta_n) - A_1^n(i\delta_n) > \frac{\epsilon}{2} \frac{\sqrt{n}}{K^n} \right)$$

$$\le \left\lceil \frac{nT}{\delta_n} \right\rceil \mathbb{P}\left(A_1^n(2\delta_n) > \frac{\epsilon}{2} \frac{\sqrt{n}}{K^n} \right)$$

$$\le \left\lceil \frac{nT}{\delta_n} \right\rceil \mathbb{P}(A_0(2\delta_n) > (1+a)2\delta_n)$$

$$\le \left\lceil \frac{nT}{\delta_n} \right\rceil \exp(-\delta_n I(a) + \epsilon) \text{ from (5) for } n \text{ large enough}$$

$$\to 0 \text{ as } n \to \infty.$$

This completes the proof of the second part of Proposition 1. □

A.2. Proof of of Proposition 2

Define

$$\hat{V}_1^n(t) = \frac{1}{\sqrt{n}} \left(V_1^n(nt) + nt(1 - \rho_1^n) \right).$$

Then,

$$\hat{B}^n(t) := \inf\{s \ge 0 : \hat{V}_1^n(t + s/\sqrt{n}) - \hat{V}_1^n(t) + \hat{U}^n(t) \le (1 - \rho_1^n)s\}.$$

Let $\tau_1^n(k)$ be the interarrival time between the k-th and $(k-1)$-th arrival in A_1^n. Under Assumptions 1 – 4, we have $\mathbb{E}[\tau_1^n(k)] \to 1$ and $\text{Var}(\tau_1^n(k)) \to \sigma_a^2$ as $n \to \infty$. In addition, $\sup_{n \ge 1} \mathbb{E}[\tau_1^n(k)^2] \le \mathbb{E}[(\tau_1^\infty)^3] < \infty$. Then,

$$\hat{V}_1^n \Rightarrow \text{BM}(0, \sigma_a^2 + \sigma_s^2) \text{ in } D[0, \infty) \text{ as } n \to \infty,$$

where $\text{BM}(\mu, \sigma^2)$ denote a Brownian motion with drift coefficient μ and diffusion coefficient σ (Reiman 1984). The convergence implies that for any $T > 0$,

$$\mathbb{P}\left(\sup_{0 \le t \le T} \sup_{0 \le s \le n^{1/2 - \delta}/K^n} |\hat{V}_1^n(t + s/\sqrt{n}) - \hat{V}_1(t)| \le \epsilon \right) \to 1 \text{ as } n \to \infty, \tag{9}$$

where δ is chosen according to (2), i.e., $(1-\rho_1^n)n^{1/2-\delta}/K^n \to \infty$ as $n \to \infty$.

We also note assuming

$$\sup_{0 \le t \le T} \sup_{0 \le s \le n^{1/2-\delta}/K^n} |\hat{V}_1^n(t + s/\sqrt{n}) - \hat{V}_1(t)| \le \epsilon,$$

if $\sup_{0 \le t \le T} \hat{B}^n(t) \le n^{1/2-\delta}/K^n$,

$$\sup_{0 \le t \le T} |(1-\rho_1^n)\hat{B}^n(t) - \hat{U}^n(t)| \le \epsilon;$$

and if $\hat{B}^n(t) > n^{1/2-\delta}/K^n$ for some $t > 0$,

$$\hat{U}^n(t) > (1-\rho_1^n)n^{1/2-\delta}/K^n - \epsilon.$$

Because $\hat{U}^n \Rightarrow \mathrm{RBM}(-\beta, 1+\sigma^2)$ as $n \to \infty$, then, by (2),

$$\mathbb{P}\left(\sup_{0 \le t \le T} \hat{U}^n(t) > (1-\rho_1^n)n^{1/2-\delta}/K^n - \epsilon\right) \to 0 \text{ as } n \to \infty.$$

Thus, the convergence in (9) implies

$$\mathbb{P}\left(\sup_{0 \le t \le T} |(1-\rho_1^n)\hat{B}^n(t) - \hat{U}^n(t)| \le \epsilon\right) \to 1 \text{ as } n \to \infty.$$

In addition, because $\rho_2^n/(1-\rho_1^n) \to 1$ as $n \to \infty$, we have

$$\rho_2^n \hat{B}^n(t) - \hat{U}^n(t) \Rightarrow \zeta(t) \text{ in } D[0, \infty) \text{ as } n \to \infty.$$

This concludes the proof of the first part of Proposition 2.

We now prove the second part. Let $a_2^n(t)$ denote the time of the oldest class 2 customer in the system at time t. We also define $\bar{a}_2^n(t) = a_2^n(nt)/n$. Then,

$$A_2^n(t) - A_2^n(a_2^n(t)) \le Q_2^n(t) \le A_2^n(t) - A_2^n(a_2^n(t)) + 1 \tag{10}$$

and

$$B^n(a_2^n(t), U_n(a_2^n(t)-)) \le t - a_2^n(t) \le B^n(a_2^n(t), U_n(a_2^n(t))). \tag{11}$$

From (11), we have

$$\rho_2^n \hat{B}^n(\bar{a}_2^n(t)-) \le \rho_2^n \sqrt{n}(t - \bar{a}_2^n(t)) \le \rho_2^n \hat{B}^n(\bar{a}_2^n(t)).$$

From the first part of the proposition, $\rho_2^n \hat{B}_n(t) \Rightarrow \mathrm{RBM}(-\beta, 1+\sigma^2)$. Therefore,

$$\sup_{0 \le t \le T} \left|\rho_2^n \hat{B}^n(\bar{a}_2^n(t)) - \rho_2^n \sqrt{n}(t - \bar{a}_2^n(t))\right| \Rightarrow 0 \text{ as } n \to \infty. \tag{12}$$

Because $\rho_2^n \sqrt{n} \to \infty$ as $n \to \infty$, we also have $\bar{a}_2^n \Rightarrow \eta$ in $D[0, \infty)$ as $n \to \infty$.

From (10), we have

$$\frac{K^n}{\sqrt{n}}\left(A_2^n(nt) - A_2^n(n\bar{a}_2^n(t))\right) \le \hat{Q}_2^n(t) \le \underbrace{\frac{K^n}{\sqrt{n}}\left(A_2^n(nt) - A_2^n(n\bar{a}_2^n(t))\right)}_{(I)} + \frac{K^n}{\sqrt{n}}.$$

Under Assumptions 1 and 4, $K^n/\sqrt{n} \to 0$ as $n \to \infty$. For (I), we first note that

$$\frac{K^n}{\sqrt{n}}\left(A_2^n(nt) - A_2^n(n\bar{a}_2^n(t))\right)$$

$$= \underbrace{K^n\sqrt{\lambda_2^n}\frac{1}{\sqrt{\lambda_2^n n}}(A_2^n(nt) - \lambda_2^n nt)}_{(i)} - \underbrace{K^n\sqrt{\lambda_2^n}\frac{1}{\sqrt{\lambda_2^n n}}(A_2^n(n\bar{a}_2^n(t)) - \lambda_2^n n\bar{a}_2^n(t))}_{(ii)} + \underbrace{\gamma^n\sqrt{n}\rho_2^n(t - \bar{a}_2^n(t))}_{(iii)}.$$

We next analyze $(A_2^n(nt) - \lambda_2^n nt)/\sqrt{\lambda_2^n n}$. Let $\tau_2^n(k)$ denote the interarrival time between the k-th and the $(k-1)$-th arrival in A_2^n. Then, $\mathbb{E}[\tau_2^n(k)] = 1/\lambda_2^n$ and $\text{Var}(\tau_2^n(k)) = (1 - \bar{F}(K^n))/(\lambda_2^n)^2 + \sigma_a^2/(\lambda^n \lambda_2^n)$. We also define $\tilde{A}_2^n$ as a renewal process in which the interarrival time between the k-th and $(k-1)$-th arrival is $\lambda_2^n \tau_2^n(k)$. Then $\tilde{A}_2^n(\lambda_2^n t) = \tilde{A}_2^n(t)$. For $\tilde{A}_2^n$, the interarrival time has mean $\mathbb{E}[\lambda_2^n \tau_2^n(k)] = 1$ and variance

$$\text{Var}(\lambda_2^n \tau_2^n(k)) = (\lambda_2^n)^2 \frac{1 - \bar{F}(K^n)}{(\lambda_2^n)^2} + (\lambda_2^n)^2 \frac{\sigma_a^2}{\lambda^n \lambda_2^n} \to 1 \text{ as } n \to \infty.$$

In addition,

$$\sup_{n \geq 1} \mathbb{E}[(\lambda_2^n \tau_2^n(k))^3] \leq \sup_{n \geq 1} \left(\bar{F}(K^n)^2 - 6\bar{F}(K^n) + 6 \right) \mathbb{E}[(\tau^\infty)^2] < \infty.$$

Because $\lambda_2^n n \to \infty$ as $n \to \infty$,

$$\frac{A_2^n(nt) - \lambda_2^n nt}{\sqrt{\lambda_2^n n}} = \frac{\tilde{A}_2^n(\lambda_2^n nt) - \lambda_2^n nt}{\sqrt{\lambda_2^n n}} \Rightarrow \text{BM}(0,1) \text{ in } D[0,\infty) \text{ as } n \to \infty. \tag{13}$$

Now, for (i), under Assumption 1 and 4, $K^n \sqrt{\lambda_2^n} \to 0$ as $n \to \infty$. Then, from (13), we have (i)$\Rightarrow \zeta$ in $D[0,\infty)$ as $n \to \infty$. Similarly, for (ii), because $\bar{a}_2 \to \eta$ and $K^n \sqrt{\lambda_2^n} \to 0$, (ii)$\Rightarrow \zeta$ in $D[0,\infty)$ as $n \to \infty$. Lastly, because $\gamma^n \to \gamma$ as $n \to \infty$, from the first part of Proposition 2 and (12), we have (iii)$\Rightarrow \gamma \text{RBM}(-\beta, \sigma_a^2 + \sigma_s^2)$ in $D[0,\infty)$ as $n \to \infty$.

Putting (i), (ii), and (iii) together, we have (I)$\Rightarrow \gamma \text{RBM}(-\beta, \sigma_a^2 + \sigma_s^2)$ in $D[0,\infty)$ as $n \to \infty$. This further implies that

$$\hat{Q}_2^n(t) \Rightarrow \hat{Q}(t) \text{ in } D[0,\infty) \text{ as } n \to \infty.$$

$\square$

Appendix B: Proof of Theorem 2

From Theorem 1, we have for any $x \geq 0$,

$$\lim_{t \to \infty} \lim_{n \to \infty} \mathbb{P}\left(\frac{K^n}{\sqrt{n}} Q^n(nt) \leq x \right) = 1 - \exp\left(-\frac{2\beta}{\gamma(1 + \sigma_s^2)} x \right). \tag{14}$$

For the first result in Theorem 2, we essentially want to show that we can change the limit on the right-hand side of (14).

For the n-th system, as $\rho^n < 1$ and the scheduling policy is non-idling, the stationary distribution of $Q^n(t)$ is well defined. In particular, under the two-class preemptive priority rule, the steady-state queue length satisfies (Schrage and Miler 1966)

$$\mathbb{E}[Q^n(\infty)] = \rho_1^n + \frac{\rho_2^n}{1 - \rho_1^n} + \frac{\lambda_1^{n,2} \mathbb{E}[(v_1^n)^2]}{2(1 - \rho_1^n)} + \frac{\lambda^n \lambda_2^n \mathbb{E}[v^2]}{2(1 - \rho_1^n)(1 - \rho^n)}.$$

As $\frac{\sqrt{n}}{K^n}(1 - \rho_1^n) \to \infty$, $\sqrt{n}(1 - \rho^n) \to \beta$, and $K^n \mu_2^n \to \gamma$ as $n \to \infty$, we have

$$\frac{K^n}{\sqrt{n}} \mathbb{E}[Q^n(\infty)] \to \frac{\gamma(1 + \sigma_s^2)}{2\beta} \text{ as } n \to \infty. \tag{15}$$

This establishes the second convergence result in Theorem 2.

The convergence in (15) also implies $\frac{K^n}{\sqrt{n}} Q^n(\infty)$ is tight. Then, the interchange of limit result then follows from Gamarnik and Zeevi (2006). $\square$

Appendix C: Proof of the results in Section 6

C.1. Proof of Proposition 3

As $\varepsilon \in [-M, M]$ with probability one, we have the following bounds for $\hat{\mu}_2^n$ and $\hat{\rho}_1^n$:

$$\frac{1}{\hat{\mu}_2^n} = \mathbb{E}[v|\hat{v} > K^n] = \mathbb{E}[v|\hat{v} > K^n, v > K^n - M] \geq K^n - M,$$

$$\frac{1}{\hat{\mu}_2^n} = \mathbb{E}[v|\hat{v} > K^n] = \frac{\mathbb{E}[v1_{\hat{v} > K^n, K^n - M < v < K^n + M}] + \mathbb{E}[v1_{v > K^n + M}]}{\mathbb{P}(\hat{v} > K^n, K^n - M < v < K^n + M) + \mathbb{P}(v > K^n + M)}$$

$$\leq \mathbb{E}[v|\hat{v} > K^n, K^n - M < v < K^n + M] + \mathbb{E}[v|v > K^n + M]$$

$$\leq (K^n + M) + \frac{3}{2}(K^n + M) \text{ by Assumption 1,}$$

and

$$1 - \hat{\rho}_1^n = 1 - \lambda^n \mathbb{E}[v1_{\hat{v} \leq K^n}] = 1 - \mathbb{E}[v1_{\hat{v} \leq K^n}] + \frac{\beta}{\sqrt{n}}\mathbb{E}[v1_{\hat{v} \leq K^n}]$$

$$= \mathbb{E}[v1_{\hat{v} > K^n}] + \frac{\beta}{\sqrt{n}}\mathbb{E}[v1_{\hat{v} \leq K^n}]$$

$$\geq (\hat{\mu}_2^n)^{-1}\mathbb{P}(\hat{v} > K^n) \geq (\hat{\mu}_2^n)^{-1}\bar{F}(K^n + M)$$

As $M \in (0, \infty)$ is fixed, $1/\hat{\mu}_2^n \sim K^n$ and for $0 < \delta < 1/2$ chosen according to Assumption 4

$$n^{1/2 - \delta}\hat{\mu}_2^n(1 - \hat{\rho}_1^n) \geq n^{1/2 - \delta}\bar{F}(K^n + M) \to \infty \text{ as } n \to \infty.$$

We have thus checked conditions C1 and C2. The rest of the process level convergence proof follows exactly the same lines of argument as the proof of Theorem 1. We thus omit it here. $\square$

C.2. Proof of Proposition 4.

Let $\mathcal{C}_i^n$ denote the event that the job is classified as Class i, $i = 1, 2$, in system n. We first note that

$$1 - \hat{\rho}_1^n = 1 - \lambda^n \mathbb{E}[v1_{\mathcal{C}_1^n}] = 1 - \mathbb{E}[v1_{\mathcal{C}_1^n}] + \frac{\beta}{\sqrt{n}}\mathbb{E}[v1_{\mathcal{C}_1^n}]$$

$$= \mathbb{E}[v1_{\mathcal{C}_2}] + \frac{\beta}{\sqrt{n}}\mathbb{E}[v1_{\mathcal{C}_1}]$$

$$\geq (\hat{\mu}_2^n)^{-1}\mathbb{P}(\mathcal{C}_2^n) = (\hat{\mu}_2^n)^{-1}\left(p_{12}^n F(K^n) + p_{22}^n \bar{F}(K^n)\right)$$

Thus,

$$n^{1/2 - \delta}\hat{\mu}_2^n(1 - \hat{\rho}_1^n) \geq n^{1/2 - \delta}\left(p_{12}^n F(K^n) + p_{22}^n \bar{F}(K^n)\right) \to \infty \text{ as } n \to \infty,$$

where the convergence follows because $n^{1/2 - \delta}F(K^n) \to \infty$, $n^{1/2 - \delta}\bar{F}(K^n) \to \infty$, and $\lim_{n \to \infty} p_{12}^n + p_{22}^n > 0$ under the assumption that $p_{21}^n \leq a$ for some $0 \leq a < 1$. We have thus verified condition C2.

We next check condition C1, i.e., whether $\hat{\mu}_2^n$ converges to zero. When $p_{12}^n \leq C\bar{F}(K^n)$ for some $C > 0$ and $p_{21}^n \leq a$,

$$\frac{1}{\hat{\mu}_2^n} = \frac{p_{12}^n \int_0^{K^n} xf(x)dx + p_{22}^n \int_{K^n}^\infty xf(x)dx}{p_{12}^n F(K^n) + p_{22}^n \bar{F}(K^n)} \leq \frac{C\bar{F}(K^n) + \int_{K^n}^\infty xf(x)dx}{(1-a)\bar{F}(K^n)} \sim K^n$$

and

$$\frac{1}{\hat{\mu}_2^n} \geq \frac{(1-a)\int_{K^n}^\infty xf(x)dx}{C\bar{F}(K^n) + \bar{F}(K^n)} \sim K^n.$$

Therefore, $1/\hat{\mu}_2^n \sim K^n$. The rest of the process level convergence proof follows from the proof of Theorem 1.

On the other hand, when $p_{12}^n/\bar{F}(K^n) \to \infty$ as $n \to \infty$ and $p_{21}^n \leq a$,

$$\frac{1}{\hat{\mu}_2^n} = \frac{p_{12}^n \int_0^{K^n} xf(x)dx + p_{22}^n \int_{K^n}^\infty xf(x)dx}{p_{12}^n F(K^n) + p_{22}^n \bar{F}(K^n)} \to 1 \text{ as } n \to \infty.$$

In this case, the Class 2 queue scales as $\sqrt{n}$. $\square$

C.3. Proof of Proposition 5

For any fixed $M > 0$, we first note

$$\frac{1}{\hat{\mu}_2^n} = \frac{\int_0^\infty t f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt} \geq \frac{M\int_M^\infty f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt}$$
$$= M\left(1 - \frac{\int_0^M f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt}\right).$$

We also note

$$\frac{\int_0^M f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt} \leq \frac{\bar{\Phi}(K^n - M)\int_0^M f(t)dt}{\bar{\Phi}(0)\int_{K^n}^\infty f(t)dt} \leq \frac{\bar{\Phi}(K^n - M)}{\bar{\Phi}(0)\bar{F}(K^n)} \to 0,$$

as $n \to \infty$ under Assumption 5. Thus, for any $M > 0$, there exists $N > 0$ such that for $n > N$, $1/\hat{\mu}_2^n > M/2$. As M can be set arbitrarily large, we have $1/\hat{\mu}_2^n \to \infty$ as $n \to \infty$.

We next analyze the case where for any fixed $a > 0$, $\lim_{t\to\infty} \bar{\Phi}(at)/\bar{F}(t) = 0$. Fix $a \in (0, 1)$. Then,

$$\frac{1}{\hat{\mu}_2^n} = \frac{\int_0^\infty t f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt} \geq \frac{(1-a)K^n \int_{(1-a)K^n}^\infty f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt}$$
$$= (1-a)K^n\left(1 - \frac{\int_0^{(1-a)K^n} f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt}\right)$$
$$\geq (1-a)K^n\left(1 - \frac{\bar{\Phi}(aK^n)}{\bar{\Phi}(0)\bar{F}(K^n)}\right).$$

As $\lim_{t\to\infty} \bar{\Phi}(at)/\bar{F}(t) = 0$, for n large enough, $1/\hat{\mu}_2^n > (1-a)K^n/2$.

On the other hand,

$$\frac{1}{\hat{\mu}_2^n} = \frac{\int_0^\infty t f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt} = \frac{\int_0^{K^n} t f(t)\bar{\Phi}(K^n - t)dt + \int_{K^n}^\infty t f(t)\bar{\Phi}(K^n - t)dt}{\int_0^{K^n} f(t)\bar{\Phi}(K^n - t)dt + \int_{K^n}^\infty f(t)\bar{\Phi}(K^n - t)dt}$$
$$\leq \frac{K^n \int_0^{K^n} f(t)\bar{\Phi}(K^n - t)dt}{\int_0^{K^n} f(t)\bar{\Phi}(K^n - t)dt} + \frac{\int_{K^n}^\infty t f(t)dt}{\bar{\Phi}(0)\int_{K^n}^\infty f(t)dt}$$
$$= \left(1 + \frac{1}{\bar{\Phi}(0)\gamma^n}\right)K^n \leq \left(1 + \frac{3}{2\bar{\Phi}(0)}\right)K^n \text{ by Assumption 1.}$$

Thus, $1/\hat{\mu}_2^n \sim K^n$.

The rest of the proof follows exactly the same lines of argument as the proof of Theorem 1. □

C.4. Proof of Lemma 1

Note that

$$\hat{\mu}_2^n = \frac{\int_0^\infty f(t)\bar{\Phi}(K^n - t)dt}{\int_0^\infty t f(t)\bar{\Phi}(K^n - t)dt}$$
$$= \frac{\int_0^{K^n/2} f(t)\bar{\Phi}(K^n - t)dt + \int_{K^n/2}^\infty f(t)\bar{\Phi}(K^n - t)dt}{\int_0^{K^n/2} t f(t)\bar{\Phi}(K^n - t)dt + \int_{K^n/2}^\infty t f(t)\bar{\Phi}(K^n - t)dt}$$
$$\geq \frac{\bar{\Phi}(K^n)F(K^n/2) + \bar{\Phi}(K^n/2)\bar{F}(K^n/2)}{\bar{\Phi}(K^n/2) + \int_{K^n/2}^\infty t f(t)dt}$$
$$\geq \frac{\bar{\Phi}(K^n)F(K^n/2) + \bar{\Phi}(K^n/2)\bar{F}(K^n/2)}{\bar{\Phi}(K^n/2) + K^n\bar{F}(K^n/2)}$$
$$= \frac{\frac{\bar{\Phi}(K^n)}{\bar{\Phi}(K^n/2)}F(K^n/2) + \bar{F}(K^n/2)}{1 + \frac{K^n\bar{F}(K^n/2)}{\bar{\Phi}(K^n/2)}} \to \frac{1}{\xi(1/2)(1 + 2c)} \text{ as } n \to \infty.$$

As $(\xi(1/2)(1 + 2c))^{-1} > 0$, this completes the proof of Lemma 1. □

Appendix D: Exact analysis of different scheduling rules

For $M/GI/1$ queues, we have closed-form expression for the steady-state average queue length under five different scheduling rules (Schrage and Miler 1966): 1) two-class priority rule without preemption (NP); 2) two-class priority rule with preemption (P); 3) two-class priority rule with semi-preemption, where the priority of the job depends on its remaining size, i.e., if the remaining size of a job is smaller than K^n, then it is classified as a Class 1 job, otherwise, it is classified as a Class 2 job; 4) SJF; and 5) SRPT. We use $Q^n(\infty; i)$ to denote the stationary number of customers in the system under scheduling policy i, $i = 1, 2, \ldots, 5$. Assume $\mu = 1$. Under Two-Class NP,

$$\mathbb{E}[Q^n(\infty; 1)] = \rho^n + \frac{\lambda^n \lambda_1^n \mathbb{E}[v^2]}{2(1 - \rho_1^n)} + \frac{\lambda^n \lambda_2^n \mathbb{E}[v^2]}{2(1 - \rho_1^n)(1 - \rho^n)}.$$

Under Two-Class P,

$$\mathbb{E}[Q^n(\infty; 2)] = \rho_1^n + \frac{\rho_2^n}{1 - \rho_1^n} + \frac{(\lambda_1^n)^2 \mathbb{E}[(v_1^n)^2]}{2(1 - \rho_1^n)} + \frac{\lambda^n \lambda_2^n \mathbb{E}[v^2]}{2(1 - \rho_1^n)(1 - \rho^n)}.$$

Under Two-Class semi-preemptive,

$$\mathbb{E}[Q^n(\infty; 3)] = \rho_1^n + \frac{\rho_2^n}{1 - \rho_1^n} - \frac{\lambda_2^n \rho_1^n K^n}{1 - \rho_1^n} + \frac{\lambda_1^n \lambda_2^n (K^n)^2}{2(1 - \rho_1^n)} + \frac{(\lambda_1^n)^2 \mathbb{E}[(v_1^n)^2]}{2(1 - \rho_1^n)} + \frac{\lambda^n \lambda_2^n \mathbb{E}[v^2]}{2(1 - \rho_1^n)(1 - \rho^n)}.$$

Under SJF,

$$\mathbb{E}[Q^n(\infty; 4)] = \rho^n + \frac{(\lambda^n)^2 \mathbb{E}[v^2]}{2} \int_0^\infty \frac{1}{\left(1 - \lambda^n \int_0^x y \, dF(y)\right)^2} \, dF(x).$$

Under SRPT,

$$\mathbb{E}[Q^n(\infty; 5)] = \lambda^n \int_0^\infty \int_0^x \frac{1}{1 - \lambda^n \int_0^y u \, dF(u)} \, dy \, dF(x) + \frac{(\lambda^n)^2}{2} \int_0^\infty \frac{\int_0^x y^2 \, dF(y) + \bar{F}(x) x^2}{\left(1 - \lambda^n \int_0^x y \, dF(y)\right)^2} \, dF(x).$$

For the first three scheduling policies, $i = 1, 2, 3$, under Assumption 4, simple calculation reveals that

$$\frac{K^n}{\sqrt{n}} \mathbb{E}[Q^n(\infty; i)] \to \frac{\gamma(1 + \sigma^2)}{2\beta} \quad \text{as } n \to \infty,$$

or equivalently,

$$\frac{1}{\sqrt{n} \mu_2^n} \mathbb{E}[Q^n(\infty; i)] \to \frac{(1 + \sigma^2)}{2\beta} \quad \text{as } n \to \infty.$$

These limits agree with what we derived in Theorem 2, and suggests that preemption does not have an effect in the limit. In particular, the two-class priority rules achieve the same asymptotic performance under different preemption assumptions.

When comparing the two-class priority rule to SJF and SRPT, we note that

$$\mathbb{E}[Q^n(\infty; 1)] \geq \mathbb{E}[Q^n(\infty; 4)] \geq \mathbb{E}[Q^n(\infty; 5)].$$

To quantify the actual differences between the three policies, we need to restrict ourselves to specific forms of service time distributions (see, e.g., Lin et al. (2011)).

For multi-server queues, very few exact formulas are known. We refer to Wang et al. (2015) for exact analysis of two-class $M/M/c$ queues where one class has preemptive priority over the other.

Appendix E: Additional Numerical Experiments

In this section, we provide some additional numerical experiments for the performance of the two-class priority rule. In particular, we compare the two-class priority rule to another benchmark policy: foreground-background (FB) policy, which does not require any prior job-size information. We also study the performance of the two-class priority rule in multi-server queues.

E.1. Comparison to the foreground-background policy

Under FB, priority is given to the job that has received the least amount of service. If there are multiple such jobs, then they are served in a processor sharing manner (Nuyens and Wierman 2006). FB has several slightly different implementations. We choose this one as it has been shown to be optimal under certain conditions (Nuyens and Wierman 2006). Note that process-sharing is in general not applicable in service operations applications, and is thus not considered in rest of this paper.

FB is known to perform well for service time distributions with decreasing failure rates (heavy-tailed service time distributions). However, when the service time distribution has an increasing failure rate, it can perform worse than FCFS. In Table 9, we compare the performance of our two-class priority rule using predicted service time to the FB policy for $M/GI/1$ queues. We test two different service time distributions: one has a decreasing failure rate (Pareto($m_{2.5}, 2.5$)) and the other has an increasing failure rate (Weibull($\nu_{1.5}, 1.5$)). We observe that when the service time distribution has a decreasing failure rate, FB performs much better than FCFS but worse than our two-class priority rule. This observation is well-expected because the two-class priority rule utilizes more job-size information than FB. When the service time distribution has an increasing failure rate, FB actually performs worse than FCFS, while our two-class priority rule still performs very well, much better than FCFS. This observation suggests that the two-class priority rule is more robust to different service time distributions. We also emphasize that when predicted service time information is available, the two-class priority rule is much easier to implement in service systems than FB.

Table 9 Steady-state average queue length for an $M/GI/1$ queue under different scheduling policies. For the two-class priority rules (P and NP), we use predicted service times where $\hat{v} = v + \epsilon$ and ϵ's are iid $N(0, 0.3^2)$. We set the threshold $K(\rho, F) = \bar{F}((1-\rho)^{-(1-0.05)})$. For the Pareto distribution, $\bar{F}(x) = (m_{2.5}/x)^{2.5} 1\{x \geq m_{2.5}\}$. For the Weibull distribution, $\bar{F}(x) = \exp\left(-(x/\nu_{1.5})^{1.5}\right) 1\{x \geq 0\}$.

	Pareto				Weibull			
ρ	0.8	0.85	0.9	0.95	0.8	0.85	0.9	0.95
FCFS	3.68	5.19	8.19	17.20	3.14	4.37	6.82	14.14
Two-Class NP	2.87	3.72	5.18	8.64	2.42	3.14	4.46	8.01
Two-Class P	2.53	3.23	4.43	7.29	2.29	2.98	4.26	7.71
FB	3.22	4.27	6.12	10.43	4.64	6.77	11.20	25.30

E.2. Multi-server queues

As discussed in Section 1.1, even though we only prove the scaling results for single-server queues, we expect the same results to hold for multi-server queues under the conventional heavy-traffic scaling. In this section,

we analyze the performance of the two-class priority rule with nonpreemption in multi-server queues. We choose FCFS and SJF as two benchmark policies.

In Table 10, we compare the steady-state average queue length for $M/M/c$ queues under different scheduling policies. We set $\mu = 1$ and vary the traffic intensity, $\rho = \lambda/(c\mu)$, by varying the value of the arrival rate λ. We observe that for large values of c, when ρ is small, there is almost no performance improvement from "smart" scheduling. The reason is that these systems require almost no waiting. In particular, the system operates like an $M/M/\infty$ queue, in which all customers enter service immediately upon arrival. Thus, no scheduling decision needs to be made. However, when ρ is large enough, we still gain significant performance improvement from job-size-based scheduling rules. In addition, our two-class priority rule achieves almost the same performance as SJF, but is much easier to administer in practice.

Table 10   **Steady-state average queue length for** $M/M/c$ **queues under different scheduling policies**
$$\left(\mu = 1, \lambda = \rho c,\ K(\rho, F) = \bar{F}((1-\rho)^{(1-0.05)})\right)\textbf{.}$$

	$c = 10$			$c = 100$		
ρ	0.80	0.90	0.99	0.80	0.90	0.99
FCFS	9.64	15.02	105.31	80.0	92.0	186.4
Two-Class NP	8.96	11.87	37.74	80.0	90.9	124.5
SJF	8.77	11.14	26.72	80.0	90.7	114.5

References

Aalto, Samuli, Urtzi Ayesta, Rhonda Righter. 2009. On the gittins index in the m/g/1 queue. *Queueing Systems* **63**(1-4) 437.

Argon, N., S. Ziya. 2009. Priority assignment under imperfect information on customer type identities. *Manufacturing and Service Operations Management* **11**(4) 674–693.

Asmussen, Søren. 2008. *Applied probability and queues*, vol. 51. Springer Science & Business Media.

Banerjee, Sayan, Amarjit Budhiraja, Amber L Puha. 2020. Heavy traffic scaling limits for shortest remaining processing time queues with heavy tailed processing time distributions. *arXiv preprint arXiv:2003.03655* .

Blanchet, J., J. Dong, Y. Pei. 2018. Perfect sampling of GI/GI/c queues. *Queueing Systems* **90** 1–33.

Cox, D.R., W. L. Smith. 1961. *Queues*. Methuen, London.

Dell'Amico, M., D. Carra, M. Pastorelli, P. Michiardi. 2014. Revisiting size-based scheduling with estimated job sizes. *IEEE 22nd International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems* .

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

Emadi, S., R. Ibrahim, S. Kesavan. 2019. Can "very noisy" information go a long way? an exploratory analysis of personalized scheduling in service systems. Working paper.

Gamarnik, D., A. Zeevi. 2006. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Advances in Applied Probability* **16**(1) 56–90.

Glynn, P. W., W. Whitt. 1994a. Large deviation behavior of counting processes and their inverses. *Queueing Systems* **17** 107–128.

Glynn, P. W., W. Whitt. 1994b. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability* 131–156.

Gromoll, H. C., L. Kruk, A. L. Puha. 2011. Diffusion limits for shortest remaining processing time queues. *Stochastic System* **1** 1–16.

Gromoll, H.C. 2004. Diffusion approximation for a processor sharing queue in heavy traffic. *Annals of Applied Probability* **14**(2) 555–611.

Grosof, I., Z. Scully, M. Harchol-Balter. 2018. SRPT for multi-server systems. *Performance Evaluation* **127-128** 154–175.

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

Harrison, J. M., R. J. Williams. 1987. Multidimensional reflected brownian motions having exponential stationary distributions. *The annuals of probability* **15**(1) 115–137.

Harrison, J.M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the halfin–whitt heavy traffic regime. *Operations Research* **52** 242–257.

Kim, J., R. Randhawa, A. Ward. 2018. Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. *Manufacturing and Service Operations Management* **20**(2) 285–301.

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

Mandelbaum, A., A. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* **52**(6) 836–855.

Nuyens, M., A. Wierman. 2006. The foreground-background queue: A survey. *Performance Evaluation* 89–148.

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

Reiman, M.I. 1984. Some diffusion approximations with state space collapses. *Modeling and Performance Evaluation Methodology* **60** 207–240.

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

Schrage, L.E., L.W. Miler. 1966. The queue $M/G/1$ with the shortest remaining processing time discipline. *Operations Research* **14**(4).

Scully, Z., I. Grosof, M. Harchol-Balter. 2020. The gittins policy is nearly optimal in the M/G/k under extremely general conditions. *POMACS*.

Scully, Z., M. Harchol-Balter, A. Scheller-Wolf. 2018. SOAP: One clean analysis of all age-based scheduling policies. *Proceedings of the ACM on Measurement and Modeling of Computer Systems*, vol. 2. 16:1 – 16:30.

Sun, Z., N.T. Argon, S. Ziya. 2018. Priority scheduling of jobs with unknown types. Working Paper.

Van Mieghem, J. 1995. Dynamic scheduling with convex delay costs: The generalized cμ-rule. *The Annuals of Applied Probability* **5**(3) 908–833.

Wang, J., O. Baron, A. Scheller-Wolf. 2015. M/M/c queue with two priority classes. *Operations Research* **63**(3) 733–749.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.

Wierman, A., M. Harchol-Balter. 2003. Classifying scheduling policies with respect to unfairness in an M/GI/1. *ACM SIGMETRICS international conference on measurement and modeling of computer systems*. New York, NY.