Queueing Syst (2016) 84:145–202
DOI 10.1007/s11134-016-9493-y

Comparisons of ticket and standard queues

Otis B. Jennings[2] · **Jamol Pender**[1]

Received: 27 May 2015 / Revised: 8 April 2016 / Published online: 22 July 2016

Abstract Upon arrival to a ticket queue, a customer is offered a slip of paper with a number on it—indicating the order of arrival to the system—and is told the number of the customer currently in service. The arriving customer then chooses whether to take the slip or balk, a decision based on the perceived queue length and associated waiting time. Even after taking a ticket, a customer may abandon the queue, an event that will be unobservable until the abandoning customer would have begun service. In contrast, a standard queue has a physical waiting area so that abandonment is apparent immediately when it takes place and balking is based on the actual queue length at the time of arrival. We prove heavy traffic limit theorems for the generalized ticket and standard queueing processes, discovering that the processes converge together to the same limit, a regulated Ornstein–Uhlenbeck process. One conclusion is that for a highly utilized service system with a relatively patient customer population, the ticket and standard queue performances are asymptotically indistinguishable on the scale typically uncovered under heavy traffic approaches. Next, we heuristically estimate several performance metrics of the ticket queue, some of which are of a sensitivity typically undetectable under diffusion scaling. The estimates are tested using simulation and are shown to be quite accurate under a general collection of parameter settings.

✉ Jamol Pender
jjp274@cornell.edu

Otis B. Jennings
otisbjennings@gmail.com

[1] School of Operations Research and Information Engineering, Cornell University, Ithaca, USA

[2] Dallas, TX, USA

Keywords Ticket Queues · Unobservable Queues · Abandonment · Heavy Traffic · Limit Theorems · Diffusion Approximations

Mathematics Subject Classification 60K25 · 90B22 · 68M20 · 60F17

1 Introduction

In many service settings, newly arriving customers are given information about the number of individuals preceding them in line, even when this line is virtual. There are several ways in which this information may be passed on. For example, a customer visiting either a delicatessen or the department of motorized vehicles (DMV) is offered a ticket with a number on it and, via some physical display, is informed of the current customer being serviced. In restaurants, dinner parties may either be told about the estimated wait or told how many similarly configured dinner parties are ahead of them. Often these two forms of information are roughly interchangeable: given the service rate, knowledge of the queue length yields an estimate of the delay, and vice versa. Being informed about the delay in service provision, customers then choose whether to join the queue or to balk.

The fact that a customer initially accepts the estimated delay and joins the queue does not guarantee that the customer will wait around until service can begin. Customers may renege on their initial decision and abandon the queue. In environments where customers are physically waiting in line for their service—such as at a bank or grocery store—abandonment is immediately apparent to service providers and other customers alike. However, neither the delicatessen service personnel, fellow ticket holders, nor potential ticket holders are aware when someone has chosen to abandon their ticket. This event is not discovered until the ticket's number is called and no one responds. In general, the number of outstanding tickets may be larger than the number of customers actually waiting for service.

A *ticket queue* refers to the setup typically employed in delicatessens, but can be thought of more generally as a mechanism for tracking the number of *potential* customers yet to be served and for maintenance of a first-come-first-served protocol. By potential, it is meant that these customers have been triaged, joined the queue, and have committed in principle to be served, yet may ultimately abandon before service can actually begin. Some of the customers may renege on this implicit commitment and leave; reneging customers typically will not inform the system manager of their decision to forgo their place in line. As a result, what is generally perceived as the queue length—the number of potential customers to be served—will in fact be an upper bound.

A visual comparison of abandonment in the two queueing types is depicted in Fig. 1. The left side of the figure shows a progression of three events as they are experienced in the standard queue. The first image has a queue with four customers, numbered 1 through 4 and with customer #1 currently in service. In the next frame, customer #3 abandons the queue, the system immediately detects the defection and customer #4 replaces customer #3 in line. Then customer #1 completes service. Finally customer #4 reaches the front of the queue and begins service in the 4th frame after the service completion for customer #2.

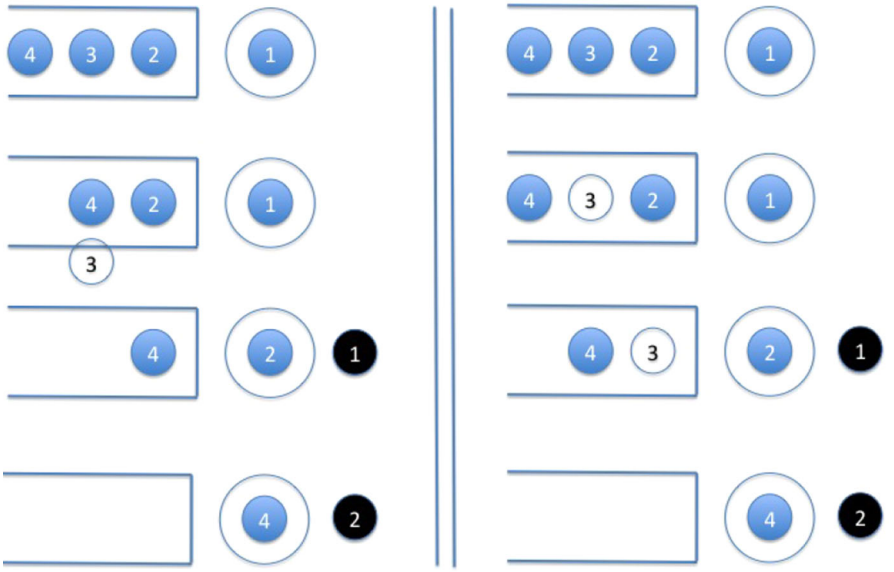

Fig. 1 Queueing dynamics of the ticket and standard queues when abandonment occurs

The right side of the image depicts the same sequence of events but as they are experienced in a ticket queue. Notice that when customer #3 abandons his ticket, the system is unaware. Thus, the perceived queue length (including the customer in service) is four. After customer #1 completes service, all customers advance and customer #2 begins service. After customer #2's service is complete, an attempt is made to handle customer #3, and it is determined immediately that customer #3 has left. So customer #4 begins service at this time. What we see here is that in two of the four frames, the ticket queue's perceived length is larger than that of the standard queue.

Several questions emerge: What are the dynamics of the collection of outstanding tickets? What is the actual number of customers remaining, i.e., those who have yet to renege? What fraction of customers renege? What fraction of customers balk? Does the difference in implementation of ticket and standard queues lead to a marked difference in performance between the two customer-organizing methods?

Some heuristics seeking to address these questions were provided by [22]. Our paper revisits the ticket queue to provide new heuristics and intuition that is complementary to that of [22]. In particular, informed by a heavy traffic limit theorem, we conclude that for highly utilized systems with relatively patient customers, ticket queues and conventional queues are indistinguishable. One of the conclusions of [22] is that for heavily loaded systems with relatively *impatient* customers, the ticket queue experiences a higher percentage of balking. In addition to these divergent insights, what sets this paper apart from [22] is that it provides estimates of the distribution of the queue length process, which can then be used to estimate other stochastic elements. More details about these estimates are provided below. Moreover, general assumptions about the abandonment distribution and balking distributions are employed; random variables in [22] are all exponentially distributed.

The paper appeals to heavy traffic limit theory in the pursuit of its approximations and comparisons of the ticket and standard queues. In this regard, it is similar in approach to [19], which studies a generalized diffusion-scaled single-server queue with abandonment *or* balking. Our paper concerns a realistic situation where there is both balking *and* abandonment. Moreover, technical complications arise as we base balking on the queue that customers perceive, whereas abandonment is based on the delays that customers experience. Ultimately, what one obtains under diffusion scaling is a regulated (at zero) Ornstein–Uhlenbeck (ROU) process whose constant drift is related to the difference between the arrival and service rates and whose restorative drift involves the derivatives of the abandonment and balking distributions, both evaluated at zero. The diffusion limit of the critically loaded ticket queue is identical, despite the fact that for the ticket queue, abandonment of a ticket is not matched with a shortening of the queue until the ticket of the abandoning customer reaches the front of the queue. Further, given the same primitive random variable elements, the diffusion-scaled ticket and standard queueing processes converge together to the same diffusion limit, a stronger notion of process similarity. In other words, not only are the dynamics of the ticket and standard queues asymptotically similar, the processes are asymptotically coupled.

As mentioned above, we estimate the distribution of the queueing processes employing heuristic interpretations of the limit theorem. Additionally, we estimate performance metrics such as abandonment and balking probabilities and the expected number of abandoned tickets in circulation. Each of these additional quantities is typically lost under conventional heavy traffic limit theory, but are manifest on a smaller—i.e., more sensitive—scaling. The approach for estimating the expected number of abandoned tickets in circulation provides the additional insight needed to compare the subtle differences between the standard and ticket queues.

We should also mention that the differences between standard and ticket queues also highlight the different ways that delay information can be communicated to the customer. This communication between the service and its customers is important because customers will make their decision to wait or to leave the queue based on the information that they receive from the manager of the system. For example, see the following papers on research on queues with delay announcements and estimation [1,2,8,9,15,20,21].

As for the assumptions used throughout, the paper also fits within the growing literature on queues with generally distributed abandonment distributions, specifically those that are not exponentially distributed. Some recent examples of such papers include [3,7,11–13,18,19,23]. The first two of these are in a multiserver setting, whereas the others are in the single-server regime. The last two papers use measure-valued processes to model system dynamics.

The remainder of the paper proceeds as follows. Both the ticket queueing model and standard queueing model (with abandonment and balking) are presented in the next section. The main result, Theorem 3.1, is presented in Sect. 3. Section 4 contains the heavy traffic-inspired approximations for performance measures and some interpretations. The proofs of the main results are provided in Sect. 5. Extensive numerical results are presented in Sect. 6. Concluding remarks and extensions follow.

1.1 Notation

We conclude this introduction with notational conventions. Let $\mathcal{R}$ denote the set of all reals, $\mathcal{R}_+$ the set of non-negative reals, and $\mathcal{N}$ the set of strictly positive integers. For a Polish space $\mathcal{S}$, let $\mathbf{D}(\mathcal{R}_+, \mathcal{S})$ denote the space of right continuous with left limits functions from $\mathcal{R}_+$ into $\mathcal{S}$. The Polish spaces we consider here are $\mathcal{R}_+$ and $\mathcal{R}_+^2$.

For ease of exposition, quantities that are related to the ticket queue will be appended with the subscript T and those associated with the standard queue will have the subscript S. We use the subscript α as a place holder for either ticket ($\alpha = T$) or standard ($\alpha = S$) queues. The standard hazard rate function $h : \mathcal{R} \mapsto \mathcal{R}_+$ is the ratio of the density of the standard normal distribution to the tail of the standard normal distribution: $h(x) = \phi(x)/(1-\Phi(x))$ for every $x \in \mathcal{R}$. For enumeration purposes, we use the letters i, j, and k to represent non-negative integers. The letters q, r, s, and t are used to represent time. Typically, the symbols δ, ϵ, η are used to represent small positive real numbers. In contrast, the letters K and L are used to represent large quantities, predominately as upper bounds.

2 The model basics

In this section we provide the primitive random variables for the queueing processes, describe the construction of the ticket and standard queues, and discuss the intricacies of simulating the ticket queues.

2.1 Random variables

We construct the ticket and standard queueing processes using the same collection of random variables. Each customer *that arrives after time zero* has an interarrival time, (potential) service time, initial time tolerance, and (potential) abandonment time. For the ith customer, these times are captured in the quadruple (u_i, v_i, b_i, d_i). The letters 'b' and 'd' denote balking and deadline, respectively. The mutually independent sequences $\{u_i, i \geq 1\}, \{v_i, i \geq 1\}, \{b_i, i \geq 1\}, \{d_i, i \geq 1\}$, which are all defined on the same probability space $(\Omega, \mathcal{F}, \mathcal{P})$, are each i.i.d. The exogenous arrival rate of jobs is λ and the service rate is μ. The sequences $\{u_i, i \geq 1\}$ and $\{v_i, i \geq 1\}$ have unitary means, and the interarrival and service times of the ith customer are u_i/λ and v_i/μ, respectively. The arrival time of the ith job occurs at time

$$t_i = (1/\lambda) \sum_{j=1}^{i} u_j.$$

The unitary interarrival and service times have variances σ_a^2 and σ_s^2, respectively. The quantities b_i and d_i represent the random variables associated with balking and reneging, respectively. Let F_b and F_d denote their respective cumulative distribution functions. We assume the these functions both vanish at zero. Moreover, their derivatives exist at zero and the sum of these derivatives is strictly positive. We define the sum as

$$\theta \equiv F_b'(0) + F_d'(0). \tag{2.1}$$

2.2 Initial conditions

Consider jobs that are present at time zero. These jobs do not require arrival times. Neither do they require balking random variables. Hence, for these initial jobs, we provide only two sequences of random variables: unitized potential service times, $\{\hat{v}_i, i \geq 1\}$, and *residual* deadline quantities, $\{\hat{d}_i, i \geq 1\}$. Let $Q(0)$ denote the number of initial jobs. If initial job $i \leq Q(0)$ has not begun service before time $\hat{d}_i$, this job will abandon. If this initial job has not been abandoned, then its service time will be $\hat{v}_i/\mu$. The initial potential service times are i.i.d. and have the same distribution as the potential service times of the jobs arriving after time zero. The residual deadlines do not necessarily have the same distribution. Let $\hat{F}_{d,i}$ be the cumulative distribution of $\hat{d}_i$. We impose the following uniform restriction on their distributions near zero: There exist an $\hat{f} > 0$ and an $h_0 > 0$ such that

$$\sup_i \frac{\hat{F}_{d,i}(h)}{h} \leq \hat{f}, \quad \forall h \leq h_0.$$

The workload at time 0, denoted $W(0)$, is the amount of effort required to process jobs present at time zero. Because of deadlines running out before service has begun, some of the $Q(0)$ jobs present at time zero will not be served. Therefore, it is not as simple as adding up the service times of the first $Q(0)$ jobs. Instead, let $\hat{w}_i$ denote the cumulative amount of server effort required among the first i jobs in-queue at time zero. The ith job will be served if and only if it is sufficiently patient, or if $\hat{d}_i > \hat{w}_{i-1}$, where $\hat{w}_0 = 0$. We can define the $\hat{w}_i$'s recursively as follows:

$$\hat{w}_i = \sum_{j=1}^{i} \frac{\hat{v}_j}{\mu} \cdot 1(\hat{d}_j > \hat{w}_{j-1}) = \hat{w}_{i-1} + \frac{\hat{v}_i}{\mu} \cdot 1(\hat{d}_i > \hat{w}_{i-1}), \quad i \geq 1.$$

It follows that $W(0) = \hat{w}_{Q(0)}$.

2.3 The queueing processes

Let the process $Q_T = \{Q_T(t), t \geq 0\}$ track the dynamics of the ticket queue and $Q_S = \{Q_S(t), t \geq 0\}$ track that of the standard queue. The difference between the ticket and standard queue is the timing of when the abandonment is accounted for. Otherwise the system dynamics are identical. For example, customers arrive to the ticket and standard queue in the same sequence, with the same balking and reneging tendencies, and with the same service requirements. The common arrival process is $A = \{A(t), t \geq 0\}$.

All other processes are indexed by the abandonment protocol $\alpha \in \{S, T\}$. In particular, $B_\alpha = \{B_\alpha(t), t \geq 0\}$ is the balking process that tracks as a function of time the number of arriving customers who leave immediately upon arrival. The reneging process $R_\alpha = \{R_\alpha(t), t \geq 0\}$ tracks the number of jobs *that arrive after time zero* who have abandoned, and whose abandonment has been detected in the system. Recall

that in the standard queue an abandoned customer is immediately detected once they leave, whereas for the ticket queue, abandonment is only apparent at the moment at which service for that customer would have begun. The process $S_\alpha = \{S_\alpha(t), t \geq 0\}$ tracks the number of service completions *of jobs that arrived after time zero*, as a function of how much effort the server has expended. Relatedly, the busy time process $T_\alpha = \{T_\alpha(t), t \geq 0\}$ reports how much time has been spent processing jobs as a function of time t, including initial jobs. The idle time process $I_\alpha = \{I_\alpha(t), t \geq 0\}$ is complementary to T_α: $I_\alpha(t) = t - T_\alpha(t)$ for each $t \geq 0$. Lastly, the workload process $W_\alpha = \{W_\alpha(t), t \geq 0\}$ reports as function of time the amount of effort required by the server to process those customers currently in-queue that will not abandon.

The system must clear out all initial jobs in-queue before it can start processing jobs that arrive after time zero. Let $\hat{Q}_\alpha = \{\hat{Q}_\alpha(t), t \geq 0\}$ track the number of remaining initial jobs at time t. Some of these jobs may abandon. Let $\hat{R}_\alpha = \{\hat{R}_\alpha(t), t \geq 0\}$ track, as a function of time, the number of jobs *who arrive before time zero*, who have abandoned, and whose abandonment has been detected in the system. The standard and ticket queues start with the same collection of jobs at time zero. Hence, among these initial jobs, the same subset of jobs abandon both the standard and ticket queues. What is different about the abandonment processes is the timing of when the abandonment is detected. What is identical between the ticket and standard queues is the initial job service completion process: $\hat{S} = \{\hat{S}(t), t \geq 0\}$, which is defined in the next section.

Equations governing the ticket and standard queue are below. For each $t \geq 0$ and $\alpha \in \{S, T\}$,

$$Q_\alpha(t) = \hat{Q}_\alpha(t) + A(t) - B_\alpha(t) - R_\alpha(t) - S_\alpha((T_\alpha(t) - W(0))^+), \tag{2.2}$$

$$\hat{Q}_\alpha(t) = Q(0) - \hat{R}_\alpha(t) - \hat{S}(t), \tag{2.3}$$

$$A(t) = \sup\left\{ j \geq 0 : \sum_{i=1}^{j} u_i/\lambda \leq t \right\}, \tag{2.4}$$

$$B_\alpha(t) = \sum_{i=1}^{A(t)} 1(b_i \leq Q_\alpha(t_i-)/\mu), \tag{2.5}$$

$$R_T(t) = \sum_{i=1}^{A(t)} 1(b_i > Q_T(t_i-)/\mu) \cdot 1(d_i \leq W_T(t_i-)) \cdot 1(W_T(t_i-) \leq t - t_i), \tag{2.6}$$

$$\hat{R}_T(t) = \sum_{i=1}^{Q(0)} 1(\hat{d}_i \leq \hat{w}_{i-1}) \cdot 1(\hat{w}_{i-1} \leq t), \tag{2.7}$$

$$R_S(t) = \sum_{i=1}^{A(t)} 1(b_i > Q_S(t_i-)/\mu) \cdot 1(d_i \leq \min(W_S(t_i-), t - t_i)), \tag{2.8}$$

$$\hat{R}_S(t) = \sum_{i=1}^{Q(0)} 1(\hat{d}_i \leq \min(\hat{w}_{i-1}, t)), \tag{2.9}$$

$$T_\alpha(t) = \int_0^t 1(Q_\alpha(s) > 0)\mathrm{d}s, \tag{2.10}$$

$$I_\alpha(t) = t - T_\alpha(t), \tag{2.11}$$

and

$$W_\alpha(t) = W(0) - T_\alpha(t) + \sum_{i=1}^{A(t)} (v_i/\mu) \cdot 1(b_i > Q_\alpha(t_i-)/\mu) \cdot 1(d_i > W_\alpha(t_i-)). \tag{2.12}$$

Interpreting (2.2) and (2.3), the queue length process consists of the remaining initial jobs and may increase with each arrival, provided the corresponding customer does not balk. In addition to the departure of the initial jobs, the queue length process decreases whenever there is an abandonment or a service completion among the jobs that arrive after time zero. Initially, service is allocated entirely to the initial jobs and remains so until $W(0)$, when those jobs have departed entirely. At this point, service allocation is given entirely to jobs arriving after time zero. The initial jobs experience only abandonment and service completion. The remaining job process is a decreasing function, hits zero, and remains there.

As for the balking process and (2.5), an arriving customer joins if their initial delay threshold b_i is sufficiently large. When a customer arrives to the system, it is first triaged and given the opportunity to join the queue. Under the ticket queue implementation, joining the queue involves the acceptance of the offered numbered ticket. In addition to knowing which ticket it will be given, the customer is told the number of the ticket holder currently in service. For the standard queue, joining the queue involves standing in a physical line. In both cases, the decision of whether or not to join the queue is based on the customer's expectation of the delay until service and her tolerance for such a delay. Customers convert the queue information into an expectation of delay until service. We assume the conversion is naive and the same for both ticket and standard queueing environments: Given the queue length, the customer estimates the delay by dividing the queue length by the service rate μ, a quantity assumed to be known by all customers. That is, if the ith customer arrives at time t, the customer joins the queue if $b_i > Q_\alpha(t-)/\mu$; otherwise the customer balks. The sequence of balking tolerances has common distribution function F_b. Naturally, the probability that a customer arriving at time t balks is $F_b(Q_\alpha(t-)/\mu)$.

Consider (2.6) and (2.8). A customer who joins the queue is not guaranteed to stick around for service. If the delay that customer i experiences in the queue reaches d_i then that customer will abandon from the queue. The time that a customer arriving at time t would have to wait is captured by $W_\alpha(t-)$. An analogous workload process is the main object of study in [12]. The distribution of the abandonment time random variables d_i is denoted F_d; the 'd' stands for customer *deadline*. It follows that a non-balking customer arriving at time t will abandon the queue with probability $F_d(W_\alpha(t-))$. What sets the ticket queue apart from the standard one is the time at which the process Q_α reflects the abandonment of a job. Hence the need to express $R_T(t)$ and $R_S(t)$ separately in (2.6) and (2.8), respectively. Consider (2.6) and how it captures customer abandonment.

Not only must the tolerance d_i be smaller than the delay that the customer must endure before service, the system does not know that the customer has abandoned until that delay has expired. Equations (2.7) and (2.9) are the analogous formulations for reneging customers who are present at time zero.

The workload process is also referred to as the virtual waiting time process because it tracks, as a function of time, the amount of time a sufficiently patient, non-balking customer would have to wait before receiving service. The virtual waiting time process increases by the service time whenever a job arrives to the system that will eventually receive service. The process decreases at rate one whenever it is greater than zero, or equivalently, whenever the queue length is non-zero. The cumulative amount that the workload has decreased by time t is precisely equal to the total busy time $T_\alpha(t)$.

2.4 State space descriptors and simulation of the ticket queue

Simulating the ticket queue is more complicated than simulating the standard queue. We describe below the intricacies of simulating the ticket queue under both Markovian and non-Markovian assumptions. We use simulation later to assess the accuracy of our approximations and heuristics.

Under Markovian assumptions, the state space of the ticket queue can be captured by a vector of zeros and ones. The length of the vector corresponds to the number of outstanding tickets, including the ticket of the customer currently in service. (For convenience, assume that the first element of the vector is the leftmost element.) If the vector has a non-zero length, the first element corresponds to the customer in service and by convention is a one. The other elements correspond to the other unresolved tickets. Further, the order of these elements reflects the relative order of the corresponding customers' arrivals and, under our first-come-first-served assumptions, the order in which resolution will take place. Ones in the vector represent customers who have not abandoned the queue. Some of these customers may abandon before resolution takes place. When a customer abandons, the corresponding element turns into a zero. When service of a sufficiently patient customer takes place, the state vector shifts to the left because at least the first element of the vector must be removed. Either the second element is a zero or it is a one. In the latter case, the entire state vector shifts by one element. If the second element is a zero, this represents an abandoned ticket. Starting with this zero in the second element, all contiguous zeros will be removed from the state descriptor. The assumption here is that resolution of abandoned tickets is instantaneous. When a job arrives to the system and chooses to join the queue, a one is appended to the end of the state vector. If the customer balks, no change in the state takes place.

The system transition is governed by exponential clocks for each unresolved ticket that has yet to be abandoned and is not being processed, one clock for the job in service, and one for the next arriving job. If the clock associated with an unresolved ticket not in service expires, then the associated customer abandons and the element in the state descriptor changes from a 1 to a 0. If there is an arrival, then another random variable is generated and compared to the weighted queue length to determine whether the customer balks; if balking occurs the state does not change. If the clock associated

with the job in service expires, service completion ensues and the state changes as described above.

Alternatively, one could have one exponential clock for all unresolved tickets that are not in service and have not been abandoned. The rate of this clock is equal to the number of such tickets multiplied by the abandonment rate of a single individual. If this clock is the one that expires next, then the actual abandoning customer is found by randomly choosing between the non-abandoned waiting customers with equal probability. When there is a change in system state, this clock must be recalculated because the number of unresolved tickets will have changed as well.

Under general assumptions on the random variables, the state space must contain the residual interarrival time of the next customer to arrive, the residual service time of the customer at the front of the queue, and for each customer yet to reach the front of the queue, the residual abandonment time. There is an alternative state space formulation, if one is content with only knowing which of the customers in-queue will *eventually* be served. For this alternative, one must track the virtual waiting time process, W_T, which yields as a function of time the amount of time that a customer must wait until service begins, and the eventual service time of jobs that will be served. These service times can be kept in a vector similar to the vector of zeros and ones above. From the time of their arrival, jobs that will have abandoned before reaching the front of the queue have a zero in their corresponding element of the vector. The virtual waiting time process jumps at the time of an arrival by the service time of the corresponding customer only if this customer actually joins the queue and is sufficiently patient; see, for instance, [19] or [12]. What is lost in this formulation is the timing of the individual jobs' abandonment times. Gained is the freedom from having to track residual abandonment times for each job in-queue.

3 Comparing the queue processes

In this section we provide the main result, a heavy traffic limit theorem that serves as the theoretical underpinning of the heuristics forwarded in the subsequent section.

To facilitate comparing the ticket and standard queue, we appeal to heavy traffic limit theory. To this end, we consider a sequence of systems, indexed by n. The arrival and service rates of the nth system are λ^n and μ^n. Equations (2.2)–(2.10) have straightforward analogs with the λ and μ replaced by λ^n and μ^n, respectively. For each $t \geq 0$, and $\alpha \in \{S, T\}$,

$$Q^n_\alpha(t) = \hat{Q}^n_\alpha(t) + A^n(t) - B^n_\alpha(t) - R^n_\alpha(t) - S^n_\alpha(T^n_\alpha(t) - W^n(0))^+, \tag{3.1}$$

$$\hat{Q}^n_\alpha(t) = Q^n(0) - \hat{R}^n_\alpha(t) - \hat{S}^n(t), \tag{3.2}$$

$$A^n(t) = \sup\left\{ j \geq 0 : \sum_{i=1}^{j} u_i/\lambda^n \leq t \right\}, \tag{3.3}$$

$$B^n_\alpha(t) = \sum_{i=1}^{A^n(t)} 1(b_i \leq Q^n_\alpha(t^n_i-)/\mu^n), \tag{3.4}$$

$$R_T^n(t) = \sum_{i=1}^{A^n(t)} 1(b_i > Q_T^n(t_i^n-)/\mu^n) \cdot 1(d_i \le W_T^n(t_i-)) \cdot 1(W_T^n(t_i^n-) \le t - t_i), \tag{3.5}$$

$$\hat{R}_T^n(t) = \sum_{i=1}^{Q^n(0)} 1(\hat{d}_i \le \hat{w}_{i-1}^n) \cdot 1(\hat{w}_{i-1}^n \le t), \tag{3.6}$$

$$R_S^n(t) = \sum_{i=1}^{A^n(t)} 1(b_i > Q_S^n(t_i^n-)/\mu^n) \cdot 1(d_i \le \min(W_S^n(t_i^n-), t - t_i^n)), \tag{3.7}$$

$$\hat{R}_S^n(t) = \sum_{i=1}^{Q^n(0)} 1(\hat{d}_i \le \min(\hat{w}_{i-1}^n, t)), \tag{3.8}$$

$$T_\alpha^n(t) = \int_0^t 1(Q_\alpha^n(s) > 0)\mathrm{d}s, \tag{3.9}$$

$$\begin{aligned} W_\alpha^n(t) = {} & W^n(0) - T_\alpha^n(t) \\ & + \sum_{i=1}^{A^n(t)} (v_i/\mu^n) \cdot 1(b_i > Q_\alpha^n(t_i^n-)/\mu^n) \cdot 1(d_i > W_\alpha^n(t_i^n-)), \end{aligned} \tag{3.10}$$

and

$$I_\alpha^n(t) = t - T_\alpha^n(t), \tag{3.11}$$

where

$$t_i^n = (1/\lambda^n) \sum_{j=1}^{i} u_j,$$

$$W^n(0) = \hat{w}_{Q^n(0)}^n, \quad \hat{w}_0^n = 0, \qquad \text{and} \quad \hat{w}_i^n = \sum_{j=1}^{i} \frac{\hat{v}_j}{\mu^n} \cdot 1(\hat{d}_j > \hat{w}_{j-1}^n), \quad i \ge 1.$$

The associated scaled processes are $Q_\alpha^n = \{Q_\alpha^n(t), t \ge 0\}$, $\hat{Q}_\alpha^n = \{\hat{Q}_\alpha^n(t), t \ge 0\}$, $A^n = \{A^n(t), t \ge 0\}$, $B_\alpha^n = \{B_\alpha^n(t), t \ge 0\}$, $R_\alpha^n = \{R_\alpha^n(t), t \ge 0\}$, $\hat{R}_\alpha^n = \{\hat{R}_\alpha^n(t), t \ge 0\}$, $S_\alpha^n = \{S_\alpha^n(t), t \ge 0\}$, $\hat{S}_\alpha^n = \{\hat{S}_\alpha^n(t), t \ge 0\}$, $T_\alpha^n = \{T_\alpha^n(t), t \ge 0\}$, $I_\alpha^n = \{I_\alpha^n(t), t \ge 0\}$, and $W_\alpha^n = \{W_\alpha^n(t), t \ge 0\}$.

We envision the arrival and service rates each being order n and differing by a quantity that is order $\sqrt{n}$. So, in the absence of abandonment and balking, one would expect the queue length to be order $\sqrt{n}$ and for the workload process to be order $1/\sqrt{n}$. In fact, this intuition is true in the presence of both balking and abandonment. Hence, we define for each $\alpha \in \{S, T\}$ the diffusion-scaled queue length process $\tilde{Q}_\alpha^n = \{\tilde{Q}_\alpha^n(t), t \ge 0\}$ and the inflated workload process $\tilde{W}_\alpha^n = \{\tilde{W}_\alpha^n(t), t \ge 0\}$, where

for each $t \geq 0$,

$$\tilde{Q}^n_\alpha(t) = \frac{Q^n_\alpha(t)}{\sqrt{n}} \quad \text{and} \quad \tilde{W}^n_\alpha(t) = \sqrt{n} W^n_\alpha(t).$$

Notice that we do not scale time as the arrival rates and service rates are already proportional to n. Also notice that the balking and abandonment times do not change with n. The reasoning is that demand may change and service speed must adjust accordingly, however, individuals will still have the same desires for and assessment of service quality.

We introduce the processes $\epsilon^n_\alpha = \{\epsilon^n_\alpha(t), t \geq 0\}$ for each $\alpha \in \{S, T\}$, where for each $t \geq 0$,

$$\epsilon^n_\alpha(t) = \hat{R}^n_\alpha(t) + \left(R^n_\alpha(t) - \sum_{i=1}^{A^n(t)} 1(d_i \leq Q^n_\alpha(T^n_i -)/\mu^n) \right). \tag{3.12}$$

The idea is to replace the reneging process R^n_α with a process that ignores whether the job has balked when considering whether it will renege. In reality, a balking customer leaves and, as a result, the question of whether balking customers would renege is a moot one. The introduction of this process also eliminates the concern of when the reneging customer causes a decrease in the queue length. Here we assume that the customer never actually enters the queue. One further subtlety is that the workload is replaced by the weighted queue length—the quantity used to determine balking—so that one needs only to track the process Q^n_α rather than the joint process (Q^n_α, W^n_α). Lastly, a benefit of this formulation is that it allows for the ticket and standard queues to be handled simultaneously. Ultimately, we will show that the process ϵ^n_α is negligible under diffusion scaling, which partially argues why the diffusion-scaled ticket and standard queues converge together to the same limit. The process $\epsilon^n_\alpha(\cdot)$ also eliminates the reneging of the initial jobs altogether; the reneging of initial jobs is shown to be negligible in Proposition 5.13.

The process S^n_α tracks the number of customers (who arrive after time 0) served as a function of total effort dedicated to customers. As not all customers receive service, the service times that determine S^n_α are a subset of $\{v_i, i \geq 1\}$, the collection of potential service times of arriving jobs. This subset differs for the ticket queue and the standard queue. The index of the ith job whose service time contributes to S^n_α—that is, who is actually served—is

$$j^n_\alpha(i) = \inf \left\{ k \geq 1 : \sum_{\ell=1}^{k} 1(b_i > Q^n_\alpha(t^n_\ell -)/\mu^n) \cdot 1(d_i > W^n_\alpha(t^n_\ell -) \geq i \right\}, \quad i \geq 1,$$
$$\alpha \in \{S, T\}.$$

The sequence of service times that are actually used is denoted $\{v^n_\alpha(i), i \geq 1\}$, where $v^n_\alpha(i) = v^n_{j^n_\alpha(i)}$. Because the service time of job $j^n_\alpha(i)$ is independent of all random variables that dictate whether this service time is used, the filtered sequence $\{v^n_\alpha(i), i \geq 1\}$ is i.i.d. and has the same distribution as the original (unfiltered) collection of service times. Therefore, any property of the unfiltered service times—such as weak laws of

large numbers or invariance principles—holds for the filtered sequence. Finally, we can write $S^n_\alpha(t)$ for each $t \geq 0$ as

$$S^n_\alpha(t) = \sup\left\{k \geq 0 : (1/\mu^n)\sum_{i=1}^{k} v^n_\alpha(i) \leq t\right\}, \quad \alpha \in \{S, T\}. \tag{3.13}$$

The processes $\hat{S} = \{\hat{S}(t), t \geq 0\}$ and $\hat{S}^n = \{\hat{S}^n(t), t \geq 0\}$ are also defined analogously.

Now we can write the diffusion-scaled queue length processes for each $t \geq 0$ as

$$\begin{aligned}\tilde{Q}^n_\alpha(t) = {} & \tilde{Q}^n(0) + \tilde{A}^n(t) - \tilde{M}^n_{b,\alpha}(\bar{A}^n(t)) - \tilde{M}^n_{d,\alpha}(\bar{A}^n(t)) - \tilde{\epsilon}^n_\alpha(t) - \tilde{\delta}^n_\alpha(t) \\ & - \tilde{S}^n_\alpha((T^n_\alpha(t) - W^n(0))^+) \\ & - \theta\int_0^t \tilde{Q}^n_\alpha(s)\mathrm{d}s + \frac{(\lambda^n - \mu^n)}{\sqrt{n}}t + \tilde{Y}^n_\alpha(t), \quad \alpha \in \{S, T\},\end{aligned} \tag{3.14}$$

where, for each $\alpha \in \{S, T\}$, $\tilde{Q}^n_\alpha(0) = (1/\sqrt{n})Q^n(0)$ is the scaled initial queue length and, for each $t \geq 0$,

$$\tilde{A}^n(t) = (1/\sqrt{n})\left(A^n(t) - \lambda^n t\right), \tag{3.15}$$

$$\bar{A}^n(t) = (1/n)A^n(t), \tag{3.16}$$

$$\tilde{M}^n_{b,\alpha}(t) = (1/\sqrt{n})\sum_{i=1}^{\lfloor nt\rfloor}\left(1(b_i \leq Q^n_\alpha(t^n_i-)/\mu^n) - F_b(\sqrt{n}\tilde{Q}^n_\alpha(t^n_i-)/\mu^n)\right), \tag{3.17}$$

$$\tilde{M}^n_{d,\alpha}(t) = (1/\sqrt{n})\sum_{i=1}^{\lfloor nt\rfloor}\left(1(d_i \leq Q^n_\alpha(t^n_i-)/\mu^n) - F_d(\sqrt{n}\tilde{Q}^n_\alpha(t^n_i-)/\mu^n)\right), \tag{3.18}$$

$$\tilde{S}^n_\alpha(t) = (1/\sqrt{n})\left(S^n_\alpha(t) - \mu^n t\right), \tag{3.19}$$

$$\tilde{\epsilon}^n_\alpha = (1/\sqrt{n})\epsilon^n_\alpha(t), \tag{3.20}$$

$$\begin{aligned}\tilde{\delta}^n_\alpha(t) = {} & \frac{1}{\sqrt{n}}\sum_{i=1}^{A^n(t)}\left(F_b(\sqrt{n}\tilde{Q}^n_\alpha(t^n_i-)/\mu^n) + F_d(\sqrt{n}\tilde{Q}^n_\alpha(t^n_i-)/\mu^n)\right) \\ & - \theta\int_0^t \tilde{Q}^n_\alpha(s)\mathrm{d}s + \frac{1}{\sqrt{n}}\left(\hat{S}^n(t) - \mu^n\min(t, W^n(0))\right),\end{aligned} \tag{3.21}$$

and

$$\tilde{Y}^n_\alpha(t) = \left(\frac{\mu^n}{n}\right)\tilde{I}^n_\alpha(t) = \left(\frac{\mu^n}{\sqrt{n}}\right)I^n_\alpha(t). \tag{3.22}$$

We refer to $\tilde{A}^n = \{\tilde{A}^n(t), t \geq 0\}$ as the diffusion-scaled arrival process and to $\bar{A}^n = \{\bar{A}^n(t), t \geq 0\}$ as its fluid-scaled analog. The reader may notice that the process $\tilde{\delta}^n_\alpha = \{\tilde{\delta}^n_\alpha(t), t \geq 0\}$ has what looks like an instantaneous drift that is proportionate to the value of the scaled queue length process. The remaining processes are centered and

diffusion-scaled versions of their original analogs: $\tilde{M}^n_{b,\alpha} = \{\tilde{M}^n_{b,\alpha}(t), t \geq 0\}$, $\tilde{M}^n_{d,\alpha} = \{\tilde{M}^n_{d,\alpha}(t), t \geq 0\}$, $\tilde{S}^n_\alpha = \{\tilde{S}^n_\alpha(t), t \geq 0\}$, $\tilde{\epsilon}^n_\alpha = \{\tilde{\epsilon}^n_\alpha(t), t \geq 0\}$, and $\tilde{Y}^n_\alpha = \{Y^n_\alpha(t), t \geq 0\}$.

3.1 A heavy traffic limit theorem

In order to prove a heavy traffic limit theorem, we assume for our sequence of systems indexed by n that arrival and service rates are order n quantities and are asymptotically identical; that is, as $n \to \infty$,

$$\lambda^n/n \to \mu \quad \text{and} \quad \mu^n/n \to \mu. \tag{3.23}$$

Further, the difference between the two should be an order $\sqrt{n}$ quantity such that as we take the limit $n \to \infty$,

$$(\lambda^n - \mu^n)/\sqrt{n} = \beta^n \to \beta \in (-\infty, \infty). \tag{3.24}$$

One can refer to (3.23) as the heavy traffic condition; the expression implies that

$$\rho^n = \lambda^n/\mu^n \to 1, \tag{3.25}$$

as $n \to \infty$. We assume that the random variables associated with balking and abandonment are unaffected by the change in the index n. Define

$$\sigma \equiv \mu\sqrt{\sigma_a^2 + \sigma_s^2} \tag{3.26}$$

as the standard deviation associated with the arrival and service times. Lastly, define $B = \{B(t), t \geq 0\}$ as a Brownian motion with no drift and an infinitesimal variance of 1.

The framework developed in [18] justifies the alternative representation of (3.14):

$$(\tilde{Q}^n_\alpha, \tilde{Y}^n_\alpha) = (\Phi_\theta, \Psi_\theta)(\tilde{Q}^n(0) + \tilde{X}^n_\alpha), \tag{3.27}$$

where $(\Phi_\theta, \Psi_\theta) : \mathbf{D}(\mathcal{R}_+, \mathcal{R}) \mapsto \mathbf{D}(\mathcal{R}_+, \mathcal{R}^+_+)$ is a Lipschitz continuous map, $\tilde{X}^n_\alpha = \{\tilde{X}^n_\alpha(t), t \geq 0\}$, and for each $t \geq 0$ and $\alpha \in \{S, T\}$,

$$\begin{aligned}\tilde{X}^n_\alpha(t) = \tilde{A}^n(t) - \tilde{M}^n_{b,\alpha}(\bar{A}^n(t)) - \tilde{M}^n_{d,\alpha}(\bar{A}^n(t)) - \tilde{\epsilon}^n_\alpha(t) - \tilde{\delta}^n_\alpha(t) - \tilde{S}^n_\alpha(T^n_\alpha(t)) \\ + \frac{(\lambda^n - \mu^n)}{\sqrt{n}}t.\end{aligned} \tag{3.28}$$

The elements of $\tilde{X}^n$ are those that either will converge to Brownian motions or that are asymptotically negligible. The limiting stochastic process is the following:

$$\tilde{X} = \beta e + \sigma B. \tag{3.29}$$

We now present our main result for the diffusion-scaled queue length and workload processes.

Theorem 3.1 *If*

$$(\tilde{Q}^n(0), \tilde{W}^n(0)) \Rightarrow (\tilde{Q}_0, \tilde{Q}_0/\mu), \quad as\ n \to \infty, \tag{3.30}$$

then

$$((\tilde{Q}^n_S, \tilde{W}^n_S, \tilde{Y}^n_S), (\tilde{Q}^n_T, \tilde{W}^n_T, \tilde{Y}^n_T)) \Rightarrow ((\tilde{Q}, \tilde{Q}/\mu, \tilde{Y}), (\tilde{Q}, \tilde{Q}/\mu, \tilde{Y})), \quad as\ n \to \infty, \tag{3.31}$$

where $\tilde{Q}(0)$ *is equal in distribution to* $\tilde{Q}_0$, $\tilde{Q} = \Phi_\theta(\tilde{Q}(0) + \tilde{X})$, $\tilde{Y} = \Psi_\theta(\tilde{Q}(0) + \tilde{X})$, *and together* $\tilde{Q}$ *and* $\tilde{Y}$ *obey the following stochastic differential equation:*

$$d\tilde{Q}(t) = -\theta(\beta/\theta - \tilde{Q}(t))dt + \sigma dB(t) + d\tilde{Y}(t). \tag{3.32}$$

Remark 3.2 The process $\tilde{Q}$ is referred to as an ROU process. The steady-state distribution of $\tilde{Q}$ is a truncated (at zero) normal variable,

$$\tilde{Q}(\infty) = \text{Normal}\left(\frac{\beta}{\theta}, \frac{\sigma^2}{2\theta}, 0, \infty\right), \tag{3.33}$$

whose mean is

$$E[\tilde{Q}(\infty)] = \frac{\beta}{\theta} + \frac{\sigma}{\sqrt{2\theta}} h\left(-\frac{\beta}{\sigma\sqrt{\theta/2}}\right), \tag{3.34}$$

where the hazard function $h(\cdot)$ is defined as the ratio of the density and the tail of the standard normal distribution:

$$h(x) = \frac{\varphi(x)}{1 - \Phi(x)}, \quad \text{for all } x \in \mathcal{R}. \tag{3.35}$$

3.2 Preliminaries

We conclude this section with several results that are well-known in the heavy traffic literature. As such, we do not provide proofs. The lemmas are all similar in substance to those in Lemma 3.1 of [10].

Lemma 3.3 (Bounded total arrivals) *For any* $t \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(A^n(t) > 2\mu n t\right) = 0.$$

Lemma 3.4 (Bounded maximum service time) *For any* $\epsilon, K, t \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{i\leq Knt} v_i^n > \epsilon/\sqrt{n} \right) = \lim_{n\to\infty} \mathbb{P}\left(\sup_{i\leq Knt} \frac{\hat{v}_i}{\mu^n} > \epsilon/\sqrt{n} \right) = 0.$$

Lemma 3.5 (Functional law of large numbers for initial service times) *For any* $\epsilon, b > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{j,k\leq b\sqrt{n}} \sqrt{n} \left| \sum_{i=j+1}^{k} \frac{\hat{v}_i}{\mu^n} - \frac{(k-j)}{\mu n} \right| > \epsilon \right) = 0.$$

Lemma 3.3 places an upper bound on the arrival process. This bound allows us to replace the number of arrivals in an interval with a deterministic upper bound. Likewise, Lemma 3.4 gives a uniform upper bound on service times that are of order n in quantity. Used in conjunction with Lemma 3.3, Lemma 3.4 places an upper bound on all service times during any finite interval of time. Lemma 3.5 places a bound on the amount by which the service times of initial jobs can differ from their expected value.

The following four results pertain to the arrival of jobs and the arrival of potential work. The first result states that jobs arrive in a linear fashion.

Lemma 3.6 (Uniformly bounded fluid arrivals) *For any* ϵ *and* $t > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{s\leq t} \left| \bar{A}^n(s) - \mu s \right| > \epsilon \right) = 0.$$

The second result, based on the heavy traffic condition, is a functional law of large numbers and states that, asymptotically, potential work arrives at rate 1, uniformly over compact intervals.

Lemma 3.7 (Uniformly bounded fluid potential workload) *For any* ϵ *and* $t > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{s\leq t} \left| \sum_{i=1}^{A^n(s)} v_i^n - s \right| > \epsilon \right) = 0.$$

The above lemma is a key component for demonstrating that server idleness is asymptotically negligible; see Proposition 5.9. The lemma is also used for the proof of tightness of our sequence of scaled queueing processes; see Proposition 5.15. However, we also need another version of the above lemma, but for short time intervals:

Lemma 3.8 (Net potential workload tightness) *For any* ϵ *and* $t > 0$, *there exists a* $\delta > 0$ *such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{u<v\leq t, v-u<\delta} \left| \sum_{i=A^n(u)+1}^{A^n(v)} v_i^n - (v-u) \right| > \frac{\epsilon}{\sqrt{n}} \right) = 0.$$

For intuition as to why this lemma holds, consider the required server effort that all arrivals would contribute to the workload process if none of the customers abandoned or balked. Then center this process at each time t by t itself, the potential amount of time that the server could have been working had the server never idled. If this centered process is then scaled up by $\sqrt{n}$ and the limit is taken, the result is a Brownian motion. It follows that the sequence of processes is tight and this fact is used in the proof of Proposition 5.15.

4 Approximations and interpretations

Now we use the limits of the previous section to approximate several performance metrics. Throughout this section, assume that we have a queueing system with arrival rate λ, service rate μ, balking distribution F_b, abandonment distribution F_d, standard deviation of interarrival times σ_a, and standard deviation of service times σ_s. To draw connections between the formal limiting procedure with the original queueing system, we make the following notational conventions:

$$\mu^n = n, \quad \beta = \frac{\lambda - \mu}{\sqrt{\mu}}, \qquad \lambda^n = n + \beta\sqrt{n}$$

and

$$\hat{\sigma} = \sqrt{(\sigma_a \cdot \lambda)^2 + (\sigma_s \cdot \mu)^2}.$$

Notice that defining any two of λ, μ, and β uniquely defines the third.

4.1 Distribution of the ticket queue in steady state

Noting the scaling $Q^n = \sqrt{n}\tilde{Q}^n \approx \sqrt{\mu}\tilde{Q}$, we approximate the steady-state distribution of our queueing process using the steady-state distribution of the ROU process:

$$Q \approx \text{Normal}\left(\frac{\lambda - \mu}{\theta}, \frac{\mu\hat{\sigma}^2}{2\theta}, 0, \infty\right), \tag{4.1}$$

that is, the ticket queue distribution is approximated by a normal distribution with mean $(\lambda - \mu)/\theta$, variance $\mu\hat{\sigma}^2/(2\theta)$, and truncated to lie within $[0, \infty)$. Note the substitution $\beta \approx (\lambda - \mu)/\sqrt{\mu}$.

4.2 The expected ticket queue length

We can also approximate the expected queue length as

$$E[Q] \approx \frac{\lambda - \mu}{\theta} + \hat{\sigma}\sqrt{\frac{\mu}{2\theta}}h\left(\frac{(1-\rho)}{\hat{\sigma}}\sqrt{\frac{2\mu}{\theta}}\right), \tag{4.2}$$

where $\rho = \lambda/\mu$. If one is also interested in approximations for higher-order cumulant moments, see, for example, [14,16,17].

4.3 The fraction of abandonment

There are three approximations forwarded for the abandonment probability. To simplify notation we let $g(0) = F_d'(0)$ and $f(0) = F_b'(0)$. The first approximation takes the rate of abandonment from the queue and divides by the total arrival rate:

$$\alpha_1 \approx \frac{g(0)E[Q]}{\lambda} \approx \frac{\rho-1}{\rho}\frac{g(0)}{\theta} + \frac{\hat{\sigma}\, g(0)}{\rho\sqrt{2\theta\mu}} h\left(\frac{(1-\rho)}{\hat{\sigma}}\sqrt{\frac{2\mu}{\theta}}\right). \tag{4.3}$$

The second approach starts with computing the cumulative distribution function evaluated at the expected delay:

$$\begin{aligned}\alpha_2 &\approx E\left[G\left(\frac{Q}{\mu}\right)\right] = E\left[G\left(\frac{\tilde{Q}^n}{\sqrt{\mu}}\right)\right] \approx g(0)\frac{E[\tilde{Q}^n]}{\sqrt{\mu}} = g(0)\frac{E[Q]}{\mu} \\ &\approx (\rho-1)\frac{g(0)}{\theta} + \frac{\hat{\sigma}\, g(0)}{\sqrt{2\theta\mu}} h\left(\frac{(1-\rho)}{\hat{\sigma}}\sqrt{\frac{2\mu}{\theta}}\right).\end{aligned} \tag{4.4}$$

The last approach is the simplification of the first two under the assumption that $\rho = 1$:

$$\alpha_3 \approx \frac{\sigma g(0)}{2\sqrt{\pi\theta\mu}}. \tag{4.5}$$

4.4 The fraction of balking customers

The balking probabilities are similar to the abandonment ones and, as such, have three versions:

$$\gamma_1 \approx \frac{\rho-1}{\rho}\frac{f(0)}{\theta} + \frac{\hat{\sigma}\, f(0)}{\rho\sqrt{2\theta\mu}} h\left(\frac{(1-\rho)}{\hat{\sigma}}\sqrt{\frac{2\mu}{\theta}}\right), \tag{4.6}$$

$$\gamma_2 \approx (\rho-1)\frac{f(0)}{\theta} + \frac{\hat{\sigma}\, f(0)}{\sqrt{2\theta\mu}} h\left(\frac{(1-\rho)}{\hat{\sigma}}\sqrt{\frac{2\mu}{\theta}}\right), \tag{4.7}$$

and when $\rho = 1$, we have that

$$\gamma_3 \approx \frac{\sigma f(0)}{2\sqrt{\pi\theta\mu}}. \tag{4.8}$$

4.5 The expected number of unresolved abandoned tickets

Given the number of unresolved tickets, a fixed fraction of these are expected to be abandoned:

$$\mathbb{E}[X(t)] \approx \frac{1}{2} G\left(\frac{Q(t)}{\mu}\right) Q(t) \approx \frac{g(0)Q(t)^2}{2\mu} \approx \frac{g(0)}{2}\tilde{Q}^2 \approx \frac{g(0)}{2} E[\tilde{Q}(\infty)]^2. \tag{4.9}$$

4.6 Interpretation

So why should we believe that there is very little difference between the ticket queue and the standard queue in steady state? In the absence of balking, one would assume that the number of customers in the standard queue would be smaller than that in the ticket queue, as the former rids itself of customers who will not add to the server workload. One mechanism that reduces this difference is that the ticket queue, albeit longer, ultimately sees less work than its queue length would suggest. Hence it must be resolving its ticket queue faster than the standard queue is processing its customers. Moreover, the concentration of abandoned tickets is typically greater among the tickets close to the front of the queue, as these tickets have been in circulation the longest. But the closer the ticket is to the front, the sooner it gets resolved. The more abandoned tickets, the faster the server resolves such tickets. Hence, the ticket queue tends to drive itself back toward the standard queue status the farther away it deviates from it.

Adding in the balking customers further lessens the difference between the two queueing scenarios. If the ticket queue is longer than the standard queue then the former has more customers balking at the front end. To conclude, as the distance grows between the length of the ticket and standard queues, so do the forces that force the coupling of the two queueing models. This notion is formally expressed in Theorem 5.8.

5 Proof of the main results

The results that follow lead up to the proof of the main result at the conclusion of this section. Some proofs are delayed until the Appendix.

5.1 Asymptotic boundedness

We argue first that the scaled queue length processes and the workload processes are asymptotically bounded.

Lemma 5.1 *Under (3.30), we have that for any $t, \eta > 0$ there exists a $K = K(\eta) > 0$ such that for each $\alpha \in \{S, T\}$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{Q}_\alpha^n(s) > K\right) < \eta \tag{5.1}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}_{\alpha}^{n}(s) > K \right) < \eta. \tag{5.2}$$

Proof The queue length processes for the ticket and standard queues can both be bounded pathwise by a third queueing process that contains neither balking nor abandonment. It is standard that this third scaled queue length process converges to a reflected Brownian motion. It also follows that this third scaled queue length process exhibits the boundedness expressed in (5.1); for example, see Lemma 3.4 of [10]. And because this third process bounds the ticket and standard queueing processes for every time t, the result in (5.1) follows. The same arguments hold for the workload processes in (5.2) and this concludes the proof. □

Lemma 5.1 emphasizes the orders of magnitude of the queueing and workload processes. This lemma will be used frequently in conjunction with the balking and abandonment distributions to place bounds on abandonment and balking frequencies.

5.2 Abandonment and balking frequencies

Next, we cover several properties of the accumulation of balking and abandonment events among the arriving jobs. The following lemmas, which besides Lemma 5.3 are provided without proof, use the fact that the derivatives of the balking and abandonment distributions exist at zero; see (2.1). The first lemma is used throughout this section and follows from a straightforward application of Taylor's expansion.

Lemma 5.2 *For any $K > 0$,*

$$\frac{F_b(K/\sqrt{n})}{K/\sqrt{n}} + \frac{F_d(K/\sqrt{n})}{K/\sqrt{n}} < 2\theta$$

for sufficiently large n.

The second lemma is similar.

Lemma 5.3 *For any $\delta, K > 0$,*

$$\sup_{s\in[0,K]} \left(\frac{F_b((s+\delta)/\sqrt{n}) - F_b(s/\sqrt{n})}{\delta/\sqrt{n}} + \frac{F_d((s+\delta)/\sqrt{n}) - F_d(s/\sqrt{n})}{\delta/\sqrt{n}} \right) < 2\theta$$

for sufficiently large n.

The next result shows that one can choose a sufficiently small δ such that, uniformly over all subintervals of $[0, t]$ of size δ, the total number of jobs that arrive in any subinterval that either abandon or balk is arbitrarily small.

The proof of this and subsequent results can be found in the Appendix.

Proposition 5.4 *For any $\epsilon, \eta, t > 0$ and $K > 0$, there exists a $\delta > 0$ such that*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\le t} \sum_{i=A^n(s)+1}^{A^n(s+\delta)} \left(1(b_i \le K/\sqrt{n}) + 1(d_i \le K/\sqrt{n})\right) > \epsilon\sqrt{n} \right) < \eta. \quad (5.3)$$

Likewise, for a sufficiently small time interval, the amount of potential workload contribution associated with balking or abandoning jobs arriving during the interval is smaller than order $1/\sqrt{n}$. This result is a key element in the proof of tightness of our scaled queue length and workload processes; see Proposition 5.15.

Proposition 5.5 *For any $\eta, t > 0$ and $K > 0$, there exists a δ such that*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\le t} \sum_{i=A^n(s)+1}^{A^n(s+\delta)} v_i^n \cdot \left(1(b_i \le K/\sqrt{n}) + 1(d_i \le K/\sqrt{n})\right) > \frac{\epsilon}{\sqrt{n}} \right) < \eta. \quad (5.4)$$

So far our propositions have been able to replace the queueing and workload processes with upper bounds early in the proofs. For the following result, where we show that the centered and scaled balking and approximate abandonment processes converge to zero, such substitutions cannot be made immediately.

Proposition 5.6 (Centered balking and reneging processes are negligible) *Under the assumptions of Theorem 3.1, for each $\alpha \in \{S, T\}$ and any $\varepsilon, \eta, t > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{M}_{b,\alpha}^n(s) \right| > \varepsilon \right) < \eta \quad (5.5)$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{M}_{d,\alpha}^n(s) \right| > \varepsilon \right) < \eta. \quad (5.6)$$

The implications here are that the balking and reneging random variables can be replaced with their respective distribution functions.

5.3 Coupled processes

An interpretation of Lemma 5.1 is that the queue length is order $\sqrt{n}$. In a model with no balking or abandonment, this fact would be sufficient to draw a linear relationship between the queue length and the workload of the form $Q/\mu \approx W$. In the presence of balking or abandonment, this relationship is justified in [19]. The key is that the number of jobs in-queue who do not contribute to the workload is negligible with respect to $\sqrt{n}$. We have the same result here.

Proposition 5.7 (State space collapse) *Under the conditions of* (3.30)*, we have that for any* $t, \varepsilon, \eta > 0$ *and each* $\alpha \in \{S, T\}$,

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{Q}^n_\alpha(s) - \mu \tilde{W}^n_\alpha(s) \right| > \varepsilon \right) < \eta.$$

Next, we establish that the sequences $\{\tilde{Q}^n_T, n \geq 1\}$ and $\{\tilde{Q}^n_S, n \geq 1\}$ converge to the same limit, if anything at all, as do $\{\tilde{W}^n_T, n \geq 1\}$ and $\{\tilde{W}^n_S, n \geq 1\}$.

Theorem 5.8 (Asymptotic coupling) *Under the assumptions of Theorem 3.1, for any* $\varepsilon, \eta, t > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{Q}^n_S(s) - \tilde{Q}^n_T(s) \right| > \varepsilon \right) < \eta \tag{5.7}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{W}^n_S(s) - \tilde{W}^n_T(s) \right| > \varepsilon \right) < \eta. \tag{5.8}$$

Proof We first show that (5.8) holds. Then (5.7) follows by applying the triangle inequality twice and Propositions 5.7 for both $\alpha = S$ and $\alpha = T$.

Fix $\varepsilon, \eta, t > 0$. Removing the absolute value signs yields

$$\begin{aligned} &\mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{W}^n_S(s) - \tilde{W}^n_T(s) \right| > \varepsilon \right) \\ &\quad \leq \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}^n_S(s) - \tilde{W}^n_T(s) > \varepsilon \right) + \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}^n_T(s) - \tilde{W}^n_S(s) > \varepsilon \right). \end{aligned} \tag{5.9}$$

We will show that

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}^n_T(s) - \tilde{W}^n_S(s) > \varepsilon \right) < \eta/2, \tag{5.10}$$

and then

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}^n_S(s) - \tilde{W}^n_T(s) > \varepsilon \right) < \eta/2 \tag{5.11}$$

will follow by symmetry.

There are two steps to demonstrating that (5.10) holds. First we restrict the amount by which the gap between $\tilde{W}_T$ and $\tilde{W}_S$ can grow at any instant. Then we show that

once the gap is of a certain size—one that is smaller than ϵ—the gap necessarily must shrink. To this end, introduce the following notation for any $s_0 \leq t$:

$$\tau^n(s_0) = \inf\{s \in [s_0, t] : \tilde{W}_T^n(s) - \tilde{W}_S^n(s) \geq \varepsilon\}$$

and

$$\gamma^n(s_0) = \inf\{s \in [s_0, t] : \tilde{W}_T^n(s) - \tilde{W}_S^n(s) \leq \varepsilon/2\},$$

where either stopping time is equal to t if the corresponding infimum is taken over an empty set. Suppose $\tilde{W}_T^n$ is greater than $\tilde{W}_S^n$. The depletion of the former is always as fast as the that of the latter since the servers work at the same rate. Therefore, the gap between $\tilde{W}_T^n$ and $\tilde{W}_S^n$ can only increase due to jumps in $\tilde{W}_T^n$. If jumps in $\tilde{W}_T^n$ are all strictly less than $\varepsilon/4$, then before the gap can exceed ε, it must first assume some value in $(3\varepsilon/4, \varepsilon)$. Then, once in this interval, the process must hit ε before falling below $\varepsilon/2$. Otherwise, the gap must again assume a value in $(3\varepsilon/4, \varepsilon)$ before it reaches ε. Hence, we have that

$$\begin{aligned}\mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}_T^n(s) - \tilde{W}_S^n(s) > \varepsilon\right) \leq\ & \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}_T^n(s) - \tilde{W}_T^n(s-) \geq \varepsilon/4\right) \\ & + \mathbb{P}\left(\exists s_0 \leq t \text{ s.t. } \tilde{W}_T^n(s_0)\right. \\ & \left. - \tilde{W}_S^n(s_0) \in \left(\frac{3\varepsilon}{4}, \varepsilon\right) \quad \text{and} \quad \tau^n(s_0) < \gamma^n(s_0)\right).\end{aligned} \tag{5.12}$$

For the first term on the right-hand side, a jump in the ticket queue workload must be due to a large service time associated with an arriving job. By Lemmas 3.3 and 3.4,

$$\begin{aligned}\mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}_T^n(s) - \tilde{W}_T^n(s-) \geq \varepsilon/4\right) \leq\ & \mathbb{P}\left(A^n(t) > 2\mu n t\right) \\ & + \mathbb{P}\left(\sup_{i \leq 2\mu n t} v^n(i) \geq \varepsilon/4\right) < \frac{\eta}{2}.\end{aligned} \tag{5.13}$$

As for the second term on the right-hand side, $\tilde{W}_T^n$ is greater than $\tilde{W}_S^n$ throughout the interval $[s_0, \min(\tau^n(s_0), \gamma^n(s_0)]$. It follows then that any new job that arrives to both queues during that interval and eventually abandons the standard queue must also abandon the ticket queue. So for the gap between the processes to increase during this interval, it must be due to jobs that balk at the standard queue but not at the ticket queue; this is possible only if the ticket queue is smaller than the standard queue during this interval. It follows that

$$\mathbb{P}\left(\exists s_0 \leq t \text{ s.t. } \tilde{W}_T^n(s_0) - \tilde{W}_S^n(s_0) \in \left(\frac{3\varepsilon}{4}, \epsilon\right) \quad \text{and} \quad \tau^n(s_0) < \gamma^n(s_0)\right)$$

$$\leq \mathbb{P}\left(\exists s \leq t \text{ s.t. } \tilde{W}_T^n(s) - \tilde{W}_S^n(s) \geq \frac{\varepsilon}{2} \quad \text{and} \quad \tilde{Q}_T^n(s) - \tilde{Q}_S^n(s) < 0\right)$$
$$\leq \mathbb{P}\left(\sup_{s\in[0,t]} \left|\frac{\tilde{Q}_T^n(s)}{\mu} - \tilde{W}_T^n(s)\right| > \frac{\varepsilon}{4}\right) + \mathbb{P}\left(\sup_{s\in[0,t]} \left|\frac{\tilde{Q}_S^n(s)}{\mu} - \tilde{W}_S^n(s)\right| > \frac{\varepsilon}{4}\right),$$

where the second inequality is a consequence of the triangle inequality. Applying Proposition 5.7 twice yields

$$\mathbb{P}\left(\exists s_0 \leq t \text{ s.t. } \tilde{W}_T^n(s_0) - \tilde{W}_S^n(s_0) \in \left(\frac{3\varepsilon}{4}, \epsilon\right) \quad \text{and} \quad \tau^n(s_0) < \gamma^n(s_0)\right) < \frac{\eta}{4}. \tag{5.14}$$

Hence, (5.10) follows from (5.12)–(5.14); (5.8) follows from (5.9)–(5.11); and (5.7) follows from Proposition 5.7, (5.8), and the triangle inequality. □

Theorem 5.8 allows us to focus all of our efforts in proving that one of these sequences—say $\{\tilde{Q}_T^n, n \geq 1\}$—converges because the other sequence is brought along with it.

The service allocation process T^n converges to the identity function.

Proposition 5.9 (Convergence of the allocation process) *Under the assumptions of Theorem 3.1, for each $\alpha \in \{S, T\}$ and any $\varepsilon, \eta, t > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left|T_\alpha^n(s) - s\right| > \varepsilon\right) < \eta.$$

A further implication of this result is that the sequence of idle time processes converges to zero.

5.4 Simplifying the reneging process

Reneging can happen only if the associated job actually joins the queue. This complication makes for involved expressions for the reneging processes in (3.5) and (3.7). Furthermore, notice that the expressions include both the queue length process as well as the workload process. The process $\tilde{\epsilon}_\alpha^n$ allows one to replace the workload process with the queue length process, to ignore whether or not the jobs have balked when considering whether they will abandon, and to ignore the timing of when reneging is detected by the system. The following proposition justifies this approximation.

For each $n \geq 1$ and α, we define the processes $R_\alpha^{0,n} = \{R_\alpha^{0,n}(t), t \geq 0\}$, $R_\alpha^{1,n} = \{R_\alpha^{1,n}(t), t \geq 0\}$, and $R_\alpha^{2,n} = \{R_\alpha^{2,n}(t), t \geq 0\}$, where

$$R_\alpha^{0,n}(t) = \sum_{i=1}^{A^n(t)} 1(d_i \leq Q_\alpha^n(t_i^n-)/\mu^n), \tag{5.15}$$

$$R_\alpha^{1,n}(t) = \sum_{i=1}^{A^n(t)} 1(d_i \le W_\alpha^n(t_i^n-)) \tag{5.16}$$

and

$$R_\alpha^{2,n}(t) = \sum_{i=1}^{A^n(t)} 1(b_i > Q_\alpha^n(t_i^n-)/\mu^n) \cdot 1(d_i \le W_\alpha^n(t_i^n-)). \tag{5.17}$$

Working backwards, notice that $R_\alpha^{2,n}$ is the fictitious abandonment process where the abandonment takes place upon arrival. Next, $R_\alpha^{1,n}$ is the fictitious abandonment process whereby abandonment happens upon arrival and jobs may abandon even if they have already balked. Finally, $R_\alpha^{0,n}$ is the process by which abandonment happens upon arrival, is independent of the balking process, and is based on the weighted queue length upon arrival instead of the workload upon arrival. The diffusion-scaled analogs have the form $\tilde{R}_\alpha^{k,n} = \{\tilde{R}_\alpha^{k,n}(t), t \ge 0\}$, where $\tilde{R}_\alpha^{k,n}(t) = (1/\sqrt{n})R_\alpha^{k,n}(t)$ for each $k = 0, 1, 2$ and $\alpha \in \{S, T\}$. Ultimately, we would like to replace R_α^n with $R_\alpha^{0,n}$; see Proposition 5.14. The following three propositions arrive at that conclusion progressively.

The first result verifies that abandonment may be treated as taking place upon arrival.

Proposition 5.10 *For any ϵ, η and $t > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left(\tilde{R}_\alpha^{2,n}(s) - \tilde{R}_\alpha^n(s) \right) > \epsilon \right) < \eta.$$

An immediate consequence of the next result is that, effectively, no customer could have abandoned if their balking random variable was small enough to cause it to balk as well. That is, the number of jobs that are candidates for both balking and abandonment is asymptotically negligible.

Proposition 5.11 *For any ϵ, η and $t > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left(\tilde{R}_\alpha^{1,n}(s) - \tilde{R}_\alpha^{2,n}(s) \right) > \epsilon \right) < \eta.$$

Now, we show that abandonment can be based on the weighted queue length as opposed to the workload.

Proposition 5.12 *For any ϵ, η and $t > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{R}_\alpha^{0,n}(s) - \tilde{R}_\alpha^{1,n}(s) \right| > \epsilon \right) < \eta.$$

Proposition 5.13 (Reneging among initial jobs is negligible) *For every $\eta > 0$ and each $\alpha \in \{S, T\}$, there exists an $L > 0$ such that, under the assumptions of Theorem 3.1,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\hat{R}^n_\alpha(t) > L\right) < \eta.$$

Proposition 5.14 (Reneging effectively ignores balking and takes place upon arrival) *For each $\alpha \in \{S, T\}$, and under the assumptions of Theorem 3.1,*

$$\tilde{\epsilon}^n_\alpha \to 0$$

in probability as $n \to \infty$.

Proof Notice that, for any $t \geq 0$ and $\alpha \in \{S, T\}$,

$$\tilde{\epsilon}^n_\alpha(t) = \frac{\hat{R}^n_\alpha(t)}{\sqrt{n}} + \left(\tilde{R}^n_\alpha(t) - \tilde{R}^{2,n}_\alpha(t)\right) + \left(\tilde{R}^{2,n}_\alpha(t) - \tilde{R}^{1,n}_\alpha(t)\right) + \left(\tilde{R}^{1,n}_\alpha(t) - \tilde{R}^{0,n}_\alpha(t)\right).$$

The result is an immediate consequence of Propositions 5.10, 5.11, 5.12, and 5.13. □

5.5 Tightness

Next, we argue that the scaled processes are tight. This result is a key step in proving Proposition 5.17, which in turn is used to extract the restorative drift of the limiting diffusion process.

Proposition 5.15 (Tightness of the scaled processes) *The processes $\{\tilde{Q}^n_\alpha, n \geq 1\}$ and $\{\tilde{W}^n_\alpha, n \geq 1\}$ are tight.*

Proof By Theorem 13.2 of [5], tightness follows from Lemma 5.1 and the fact that for any $\epsilon > 0$, we have that for either $\alpha \in \{S, T\}$,

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{u,v\in[0,t], v-u<\delta} \left|\tilde{Q}^n_\alpha(v) - \tilde{Q}^n_\alpha(u)\right| > \epsilon \right) = 0 \tag{5.18}$$

and

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{u,v\in[0,t], v-u<\delta} \left|\tilde{W}^n_\alpha(v) - \tilde{W}^n_\alpha(u)\right| > \epsilon \right) = 0. \tag{5.19}$$

We will show that (5.19) holds. Then (5.18) follows from Theorem 5.7 and (5.19). This will conclude the proof.

Fix $\alpha \in \{S, T\}$ and arbitrary constants $\epsilon, \eta > 0$. We will show that there exists a $\delta_0 > 0$ such that for any $\delta \le \delta_0$,

$$\mathbb{P}\left(\sup_{u,v\in[0,t], v-u<\delta} \left| W_\alpha^n(v) - W_\alpha^n(u) \right| > \frac{\epsilon}{\sqrt{n}} \right) < \eta \tag{5.20}$$

for sufficiently large n. Changes in the workload are due to the arrival of work and the processing of work. That is, for any $v \ge u \ge 0$,

$$\begin{aligned} W_\alpha^n(v) - W_\alpha^n(u) = &\sum_{i=A^n(u)+1}^{A^n(v)} v_i^n \cdot 1(b_i > Q_\alpha^n(t_i^n-)/\mu^n) \cdot 1(d_i > W_\alpha^n(t_i^n-)) \\ &- \left(T_\alpha^n(v) - T_\alpha^n(u)\right). \end{aligned} \tag{5.21}$$

By (3.11), the change in the service allocation process can be written in terms of the increase in the idle process, which in turn can be bounded by the shortfall in the arrival of workload relative to the potential server effort. For any $v \ge u \ge 0$,

$$\begin{aligned} I_\alpha^n(v) - I_\alpha^n(u) &= (v-u) - \left(T_\alpha^n(v) - T_\alpha^n(u)\right) \\ &\le \max\left(0, -\inf_{s\in[u,v]} \left(\sum_{i=A^n(u)+1}^{A^n(s)} v_i^n \cdot 1(b_i > Q_\alpha^n(t_i^n-)/\mu^n) \cdot 1(d_i \right.\right. \\ &\quad \left.\left. > W_\alpha^n(t_i^n-)) - (s-u) \right) \right). \end{aligned} \tag{5.22}$$

It follows from (5.21) and (5.22) that for any $0 < \delta \le \delta_0$,

$$\begin{aligned} \sup_{u,v\in[0,t], v-u<\delta} \left| W_\alpha^n(v) - W_\alpha^n(u) \right| &\le 2 \sup_{u,v\in[0,t], v-u<\delta} \left| \sum_{i=A^n(u)+1}^{A^n(v)} v_i^n - (v-u) \right| \\ &\quad + 2 \sup_{u\in[0,t]} \sum_{i=A^n(u)+1}^{A^n(u+\delta)} v_i^n \left(1(b_i \le Q_\alpha^n(t_i^n-)/\mu^n) \right. \\ &\quad \left. + 1(d_i \le W_\alpha^n(t_i^n-))\right) \\ &\le 2 \sup_{u,v\in[0,t], v-u<\delta_0} \left| \sum_{i=A^n(u)+1}^{A^n(v)} v_i^n - (v-u) \right| \\ &\quad + 2 \sup_{u\in[0,t]} \sum_{i=A^n(u)+1}^{A^n(u+\delta_0)} v_i^n \left(1(b_i \le Q_\alpha^n(t_i^n-)/\mu^n) \right. \\ &\quad \left. + 1(d_i \le W_\alpha^n(t_i^n-))\right). \end{aligned} \tag{5.23}$$

Finally,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{u,v\in[0,t],v-u<\delta}\left|W_\alpha^n(v)-W_\alpha^n(u)\right|>\frac{\epsilon}{\sqrt{n}}\right)\\
&\quad\le\mathbb{P}\left(2\sup_{u,v\in[0,t],v-u<\delta_0}\left|\sum_{i=A^n(u)+1}^{A^n(v)}v_i^n-(v-u)\right|>\frac{\epsilon}{2\sqrt{n}}\right)\\
&\quad+\mathbb{P}\left(2\sup_{u\in[0,t]}\sum_{i=A^n(u)+1}^{A^n(u+\delta_0)}v_i^n\left(1(b_i\le Q_\alpha^n(t_i^n-)/\mu^n)+1(d_i\le W_\alpha^n(t_i^n-))\right)>\frac{\epsilon}{2\sqrt{n}}\right)\\
&\le\mathbb{P}\left(2\sup_{u,v\in[0,t],v-u<\delta_0}\left|\sum_{i=A^n(u)+1}^{A^n(v)}v_i^n-(v-u)\right|>\frac{\epsilon}{2\sqrt{n}}\right)\\
&\quad+\mathbb{P}\left(\sup_{s\le t+\delta_0}Q_\alpha^n(s)/\mu^n>K/\sqrt{n}\right)\\
&\quad+\mathbb{P}\left(2\sup_{u\in[0,t]}\sum_{i=A^n(u)+1}^{A^n(u+\delta_0)}v_i^n\left(1(b_i\le K/\sqrt{n})+1(d_i\le K/\sqrt{n})\right)>\frac{\epsilon}{2\sqrt{n}}\right)\\
&\quad+\mathbb{P}\left(\sup_{s\le t+\delta_0}W_\alpha^n(s)>K/\sqrt{n}\right).
\end{aligned}
$$

By Lemmas 3.8, 5.1, Proposition 5.5, and (5.23), we can choose a $K>0$ and a $\delta_0>0$ such that (5.20) follows easily. □

5.6 Convergence to diffusion processes

In this subsection we pull together the last ingredients for our diffusion process. We start with the driving Brownian motions from our centered and scaled interarrival and service time processes. We derive the restorative drift term of our limiting diffusion process from the derivatives of the balking and abandonment distributions evaluated at zero. Finally, we complete the proof of our main result.

Let B_a and B_s be independent standard Brownian motions, i.e., $B_a(0)=B_s(0)=0$ and the processes have zero drift and unitary infinitesimal variance. The diffusion-scaled arrival processes and service completion processes converge to scaled versions of these Brownian motions. These are standard results—see, for example, [12].

Proposition 5.16 *Under the assumptions of Theorem 3.1,*

$$\tilde{A}^n\Rightarrow\mu\sigma_a B_a\quad\text{and}\quad\tilde{S}_\alpha^n\Rightarrow\mu\sigma_s B_s$$

as $n\to\infty$.

The process $\tilde{\delta}^n_\alpha$ swaps the sum of the abandonment and balking distributions, evaluated at the scaled queue length at the times of arrivals, with a smooth function involving the derivatives of the distribution functions evaluated at zero and then multiplied by the scaled queue lengths. The following proposition justifies this step and is proven in the Appendix.

Proposition 5.17 *For each $\alpha \in \{S, T\}$, and under the assumptions of Theorem 3.1,*

$$\tilde{\delta}^n_\alpha \to 0$$

in probability as $n \to \infty$.

We conclude with a proof of the main result.

Proof of Theorem 3.1 Fix $\alpha \in \{S, T\}$. Recall the expression for the scaled queue length and idleness process given in (3.27). The elements of the process $\tilde{X}^n_\alpha$ (see (3.28)) converge to either zero, a drift, or a Brownian motion. By Lemma 3.6 and Proposition 5.6, we have that

$$\tilde{M}^n_{b,\alpha} \circ \bar{A}^n \to 0 \quad \text{and} \quad \tilde{M}^n_{d,\alpha} \circ \bar{A}^n \to 0$$

as $n \to \infty$. Furthermore, Propositions 5.14 and 5.17 inform us that

$$\tilde{\epsilon}^n_\alpha \to 0 \quad \text{and} \quad \tilde{\delta}^n_\alpha \to 0,$$

respectively, as $n \to \infty$. Proposition 5.16 provides the convergence of the scaled and centered arrival and potential departure processes. Coupled with the service allocation process, we have

$$\tilde{A}^n + \tilde{S}^n_\alpha \circ T^n_\alpha \to \mu\sigma_a B_a + \mu\sigma_b B_b \overset{D}{=} \sigma B.$$

Finally, (3.24) provides the drift term

$$\frac{(\lambda^n - \mu^n)}{\sqrt{n}} \to \beta$$

as $n \to \infty$. Hence,

$$\tilde{X}^n_\alpha \Rightarrow \tilde{X}.$$

From (3.30), we have that

$$(\tilde{Q}^n(0), \tilde{X}^n_\alpha) \Rightarrow (\tilde{Q}_0, \tilde{X}), \quad \text{as} \quad n \to \infty,$$

and hence, by the continuous mapping theorem,

$$(\tilde{Q}^n_\alpha, \tilde{Y}^n_\alpha) = (\Phi_\theta, \Psi_\theta)(\tilde{Q}^n(0), \tilde{X}^n_\alpha) \Rightarrow (\Phi_\theta, \Psi_\theta)(\tilde{Q}(0), \tilde{X}) = (\tilde{Q}, \tilde{Y}), \quad \text{as} \quad n \to \infty,$$

where we recall that $\tilde{Q}(0)$ is equal in distribution to $\tilde{Q}_0$.

As α was chosen arbitrarily, this limit holds for both the ticket queue- and standard queue-related processes; both scaled queue length processes converge to the same limit in distribution. By (5.7) of Theorem 5.8, these processes are coupled as follows:

$$((\tilde{Q}_S^n, \tilde{Y}_S^n), (\tilde{Q}_T^n, \tilde{Y}_T^n)) \Rightarrow ((\tilde{Q}, \tilde{Y}), (\tilde{Q}, \tilde{Y})), \quad \text{as} \quad n \to \infty.$$

Finally, Proposition 5.7 demonstrates that asymptotically the scaled workload and scaled queue lengths are scalar multiples of each other and thus converge together to scalar multiples of the same process. Hence (3.31) holds. This concludes the proof. □

6 Numerical results

We have tested the heuristics forwarded in this paper extensively. Tables capturing these tests are given in the Appendix. We describe some of the highlights and trends in the results here. In Tables 1, 2, 3, and 4, we see that our approximations generated from the heavy traffic limit theorems are quite good, especially when the arrival and service rates are large, i.e., $\mu \approx 100$. In Table 1, we simulate a scenario where all of the parameters are generated from exponential distributions. We see that the approximations are good and accurate, except when the rates are small and the drift parameter β is very negative.

In Table 2 we simulate a non-Markovian model where the service rate follows a lognormal distribution and the balking and reneging distributions follow a uniform distribution. Unlike the exponential distribution, the uniform distribution is bounded and we see that the simulated values in Table 1 are very similar to the values of Table 2 even though they differ in the types of distributions that generate the queueing dynamics.

In Table 3 we simulate a scenario where all of the parameters are generated from exponential distributions. This table is different than Table 1 since we use different balking and reneging parameter values. Once again we see that the approximations are good and accurate, except when the rates are small and the drift parameter β is very negative. In Table 4 we simulate a non-Markovian model where the service rate follows a lognormal distribution and the balking and reneging distributions follow a uniform distribution. Once again the simulated results of Table 3 are very similar to Table 4 even though the distributions are not Markovian in Table 4.

In addition to our heavy traffic approximations accurately estimating the performance measures, we also notice that in every simulation the ticket queue is larger than the standard queue. This fact is irrespective of the distributions that are used to generate the queueing dynamics. Moreover, the fraction of balking in the ticket queue is also larger than the standard queue. Even though these two processes converge as the rates tend to infinity, when the rates are finite the ticket queue is perceived as being larger, which causes more people to balk from the system. This difference in queue length is also a function of the parameter β. When the β parameter is larger and positive, the difference is larger than when β is negative. Thus, when the rates are not infinite, the ticket queue exhibits interesting behavior that is to be expected.

7 Conclusions and extensions

In this paper we studied the dynamics of a critically loaded queueing system subject to customers who might either balk because the line is too long or abandon the queue from waiting too long in-queue. We consider two types of abandonment protocols. In the conventional approach to capturing abandonment, customers immediately leave the system when their patience time has been exceeded by their time waiting for service; we refer to the model in this setting as the standard queue. In the standard queue, everyone is aware of an abandoning customer's departure at the time of abandonment. We compare the standard queue to the ticket queue. In the ticket queue, customers whose patience has run out leave the queue in an unnoticed fashion. Their departure is only detected when their hypothetical service time would have begun. The paper is complementary to the study of [22] who study a heavily loaded system with impatient customers; in comparison, our customers are relatively patient. Our method of analysis is also fundamentally different from that of [22].

We prove a heavy traffic limit theorem for the diffusion-scaled queue length and workload processes. A key result in the theorem is that the standard and ticket queues are asymptotically coupled under diffusion scaling. The managerial interpretation is that regardless of how you implement your queue—whether with a physical line, which is best modeled with a standard queue, or as a ticket queue—the dynamics of the systems will not differ by much. In addition to this insight based on the sensitivity of the diffusion scaling, we provide some heuristics for calculating certain performance metrics of operational importance. These heuristics are beyond the sensitivity of diffusion scaling. Nonetheless, we assess the accuracy of these heuristics through simulation. We find that in a broad range of parameter and distributional settings, the heuristics perform well.

Appendix

The Appendix is split into two parts. The first half provides proofs of results stated earlier in the paper. The second half provides an extensive collection of numerical examples.

Proofs

This appendix contains many of the proofs from Sect. 5. We start with a proof of Proposition 5.4.

Proof of Proposition 5.4 Fix $\epsilon, \eta, t, K > 0$ and set $\delta = \frac{\epsilon}{4\mu K\theta}$. By Lemma 5.2 and our choice of δ, we have that

$$\mu n\delta \left(F_b(K\sqrt{n}) + F_d(K/\sqrt{n})\right) < 2\mu\sqrt{n}K\theta\delta \leq \frac{\epsilon\sqrt{n}}{2}$$

for sufficiently large n. It follows that, by Lemmas 3.3 and 3.6,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\le t}\sum_{i=A^n(s)+1}^{A^n(s+\delta)}\left(1(b_i\le K/\sqrt{n})+1(d_i\le K/\sqrt{n})\right)>\epsilon\sqrt{n}\right)\\
&\le \mathbb{P}\left(\sup_{0\le j\le\lfloor 2\mu nt\rfloor}\sum_{i=j+1}^{j+\lfloor 2\mu n\delta\rfloor}\left(1(b_i\le K/\sqrt{n})+1(d_i\le K/\sqrt{n})\right)>\epsilon\sqrt{n}\right)\\
&\quad+\mathbb{P}\left(A^n(t)>2\mu nt\right)+\mathbb{P}\left(\sup_{s\le t}\left(A^n(s+\delta)-A^n(s)\right)>2\mu n\delta\right)\\
&\le \frac{\eta}{2}+\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor 2\mu n\delta\rfloor+1}^{(j+1)\lfloor 2\mu n\delta\rfloor}\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)>\frac{\epsilon\sqrt{n}}{6}\right)\\
&\quad+\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor 2\mu n\delta\rfloor+1}^{(j+1)\lfloor 2\mu n\delta\rfloor}\left(1(d_i\le K/\sqrt{n})-F_d(K/\sqrt{n})\right)>\frac{\epsilon\sqrt{n}}{6}\right)\\
&\quad+\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor 2\mu n\delta\rfloor+1}^{(j+1)\lfloor 2\mu n\delta\rfloor}F_b(K/\sqrt{n})+F_d(K/\sqrt{n})>\frac{\epsilon\sqrt{n}}{6}\right) \qquad (8.1)
\end{aligned}$$

for sufficiently large n. □

We will show that, for sufficiently large n,

$$\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor\mu n\delta\rfloor+1}^{(j+1)\lfloor\mu n\delta\rfloor}\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)>\frac{\epsilon\sqrt{n}}{6}\right)<\frac{\eta}{6}, \qquad (8.2)$$

$$\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor\mu n\delta\rfloor+1}^{(j+1)\lfloor\mu n\delta\rfloor}F_b(K/\sqrt{n})+F_d(K/\sqrt{n})>\frac{\epsilon\sqrt{n}}{6}\right)<\frac{\eta}{6}, \qquad (8.3)$$

and by symmetry it will also follow that

$$\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor\mu n\delta\rfloor+1}^{(j+1)\lfloor\mu n\delta\rfloor}\left(1(d_i\le K/\sqrt{n})-F_d(K/\sqrt{n})\right)>\frac{\epsilon\sqrt{n}}{6}\right)<\frac{\eta}{6}. \qquad (8.4)$$

Consider (8.2). By Kolmogorov's inequality (see, for example, [4]),

$$\begin{aligned}
&\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor\mu n\delta\rfloor+1}^{(j+1)\lfloor\mu n\delta\rfloor}\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)>\frac{\epsilon\sqrt{n}}{6}\right)\\
&\le\left(\frac{2t}{\delta}+2\right)\left(\frac{144}{\epsilon^2 n}\right)\mathbb{E}\left[\left(\sum_{i=1}^{\lfloor\mu n\delta\rfloor}\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)\right)^2\right].
\end{aligned}$$

The indicators above are independent, so the cross-terms all have expectations of zero. Hence

$$\begin{aligned}
&\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2\mu nt/\delta\rfloor+1}\sum_{i=j\lfloor\mu n\delta\rfloor+1}^{(j+1)\lfloor\mu n\delta\rfloor}\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)>\frac{\epsilon\sqrt{n}}{6}\right)\\
&\le\left(\frac{2t}{\delta}+2\right)\left(\frac{144}{\epsilon^2 n}\right)\mathbb{E}\left[\sum_{i=1}^{\lfloor\mu n\delta\rfloor}\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)^2\right]\\
&<\left(\frac{2\mu t}{\delta}+2\right)\left(\frac{144}{\epsilon^2 n}\right)(\mu n\delta)\left(\frac{2K\theta}{\sqrt{n}}\right)\\
&\le\frac{\eta}{6}
\end{aligned}$$

for sufficiently large n. So (8.2) holds, as does (8.4) by symmetry. The result (5.3) follows from (8.1)–(8.4). □

Next we prove Proposition 5.5.

Proof of Proposition 5.5 We will alter (5.4) to an expression that, without loss of generality, excludes the abandonment random variables. We will show that for any $\eta, t>0$ and $K>0$, there exists a $\delta>0$ such that

$$\mathbb{P}\left(\sup_{s\le t}\sum_{i=A^n(s)+1}^{A^n(s+\delta)}v_i^n\cdot 1(b_i\le K/\sqrt{n})>\frac{\epsilon}{\sqrt{n}}\right)<\eta, \tag{8.5}$$

for sufficiently large n. Fix η, t, and K and set $\delta=\frac{\epsilon}{12K\theta}$ and notice that by Lemma 5.2,

$$(\mu\delta n)\left(\frac{1}{\mu^n}\right)F_b(K/\sqrt{n})<\frac{4\delta K\theta}{\sqrt{n}}\le\frac{\epsilon}{3\sqrt{n}}$$

for sufficiently large n. Proceeding in a manner similar to that of the proof of Proposition 5.4, it follows by Lemmas 3.3 and 3.6 that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\le t}\sum_{i=A^n(s)+1}^{A^n(s+\delta)}v_i^n\cdot 1(b_i\le K/\sqrt{n})>\frac{\epsilon}{\sqrt{n}}\right)\\
&\le\mathbb{P}\left(\sup_{0\le j\le\lfloor 2\mu nt\rfloor}\sum_{i=j+1}^{j+\lfloor\mu n\delta\rfloor}v_i^n\cdot\left(1(b_i\le K/\sqrt{n})+1(d_i\le K/\sqrt{n})\right)>\frac{\epsilon}{\sqrt{n}}\right)\\
&\quad+\mathbb{P}\left(A^n(t)>2\mu nt\right)+\mathbb{P}\left(\sup_{s\le t}\left(A^n(s+\delta)-A^n(s)\right)>2\mu n\delta\right)\\
&\le\frac{\eta}{4}+\mathbb{P}\left(\sup_{0\le j\le\lfloor 2\mu nt\rfloor}\sum_{i=j+1}^{j+\lfloor\mu n\delta\rfloor}\left(v_i^n-\frac{1}{\mu^n}\right)\cdot 1(b_i\le K/\sqrt{n})>\frac{\epsilon}{3\sqrt{n}}\right)\\
&\quad+\mathbb{P}\left(\sup_{0\le j\le\lfloor 2\mu nt\rfloor}\sum_{i=j+1}^{j+\lfloor\mu n\delta\rfloor}\frac{1}{\mu^n}\cdot\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)>\frac{\epsilon}{3\sqrt{n}}\right)
\end{aligned}\tag{8.6}$$

for sufficiently large n. Consider the third term on the far right-hand side. By Kolmogorov's inequality [4],

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\le j\le\lfloor 2\mu nt\rfloor}\sum_{i=j+1}^{j+\lfloor\mu n\delta\rfloor}\frac{1}{\mu^n}\cdot\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)>\frac{\epsilon}{3\sqrt{n}}\right)\\
&\le\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor\mu n\delta\rfloor+1}^{(j+1)\lfloor\mu n\delta\rfloor}\frac{1}{\mu^n}\cdot\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)>\frac{\epsilon}{3\sqrt{n}}\right)\\
&\le\left(\frac{2t}{\delta}+2\right)\left(\frac{36n}{\epsilon^2}\right)\left(\frac{1}{\mu^n}\right)^2\mathbb{E}\left[\left(\sum_{i=1}^{\lfloor\mu n\delta\rfloor}\left(1(b_i\le K/\sqrt{n})-F_b(K/\sqrt{n})\right)\right)^2\right]\\
&\le\left(\frac{2t}{\delta}+2\right)\left(\frac{36n}{\epsilon^2}\right)\left(\frac{2}{n\mu}\right)^2(n\mu\delta)\,F_b(K/\sqrt{n})\\
&<(2t+2\delta)\left(\frac{36}{\epsilon^2}\right)\left(\frac{8K\theta}{\sqrt{n}}\right)\\
&<\frac{\eta}{3}
\end{aligned}\tag{8.7}$$

for sufficiently large n. As for the second term on the far right-hand side of (8.6), applying Kolmogorov's inequality a second time yields

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\le j\le\lfloor 2\mu nt\rfloor}\sum_{i=j+1}^{j+\lfloor\mu n\delta\rfloor}\left(v_i^n-\frac{1}{\mu^n}\right)\cdot 1(b_i\le K/\sqrt{n})>\frac{\epsilon}{3\sqrt{n}}\right)\\
&\le\mathbb{P}\left(2\sup_{0\le j\le\lfloor 2t/\delta\rfloor+1}\sum_{i=j\lfloor\mu n\delta\rfloor+1}^{(j+1)\lfloor\mu n\delta\rfloor}\left(v_i^n-\frac{1}{\mu^n}\right)\cdot 1(b_i\le K/\sqrt{n})>\frac{\epsilon}{3\sqrt{n}}\right)\\
&\le\left(\frac{2t}{\delta}+2\right)\left(\frac{36n}{\epsilon^2}\right)\mathbb{E}\left[\left(\left(v_i^n-\frac{1}{\mu^n}\right)\cdot 1(b_i\le K/\sqrt{n})\right)^2\right]\\
&\le\left(\frac{2t}{\delta}+2\right)\left(\frac{36n}{\epsilon^2}\right)\mathbb{E}\left[\left(v_i^n-\frac{1}{\mu^n}\right)^2\cdot 1(b_i\le K/\sqrt{n})\right],
\end{aligned}\tag{8.8}$$

where the last inequality follows because the independence of the cross-terms makes their expectations zero. Furthermore, the service times are independent of the balking random variables. Hence,

$$\begin{aligned}
&\left(\frac{2t}{\delta}+2\right)\left(\frac{36n}{\epsilon^2}\right)\mathbb{E}\left[\sum_{i=1}^{\lfloor\mu n\delta\rfloor}\left(v_i^n-\frac{1}{\mu^n}\right)^2\cdot 1(b_i\le K/\sqrt{n})\right]\\
&\quad\le\left(\frac{2t}{\delta}+2\right)\left(\frac{36n}{\epsilon^2}\right)(n\mu\delta)\left(\frac{\sigma_b}{\mu^n}\right)^2F_b(K/\sqrt{n})
\end{aligned}$$

$$< (2t + 2\delta)\left(\frac{36}{\epsilon^2}\right)\left(\frac{2\sigma_b^2}{\mu}\right)\left(\frac{2K\theta}{\sqrt{n}}\right)$$
$$\leq \frac{\eta}{3} \tag{8.9}$$

for sufficiently large n. The result (8.5) follows from (8.6)–(8.9). □

Proof of Proposition 5.6 It suffices to prove that the centered and scaled abandonment process is asymptotically negligible. The analogous property for the balking process can be proved in an identical fashion. Fix $\varepsilon, \eta, t > 0$. By Kolmogorov's inequality we have that

$$\begin{aligned}
\mathbb{P}\left(\sup_{s\in[0,t]} \tilde{M}^n_{d,\alpha}(s) > \epsilon\right) &= \mathbb{P}\left(\sup_{s\in[0,t]} M^n_{d,\alpha}(s) > \sqrt{n}\cdot\epsilon\right) \\
&\leq \frac{1}{n\epsilon^2}\mathbb{E}\left[\left(M^n_{d,\alpha}(t)\right)^2\right] \\
&= \frac{1}{n\epsilon^2}\mathbb{E}\left[\left(\sum_{i=1}^{\lfloor nt\rfloor}\left[1(d_i < Q^n_\alpha(t_i^n-)/\mu^n)\right.\right.\right. \\
&\qquad \left.\left.\left. - F_d(Q^n_\alpha(t_i^n-)/\mu^n)\right]\right)^2\right]
\end{aligned}$$

Now, by Burkholder's inequality and bounds for indicator functions, there exists a $c > 0$ such that

$$\begin{aligned}
\mathbb{P}\left(\sup_{s\in[0,t]} \tilde{M}^n_{d,\alpha}(s) > \epsilon\right) &\leq \frac{c}{n\epsilon^2}\mathbb{E}\left[\sum_{i=1}^{\lfloor nt\rfloor}\left[1(d_i < Q^n_\alpha(t_i^n-)/\mu^n) - F_d(Q^n_\alpha(t_i^n-)/\mu^n)\right]^2\right] \\
&\leq \frac{c}{n\epsilon^2}\mathbb{E}\left[\sum_{i=1}^{\lfloor nt\rfloor}\left[1(d_i < Q^n_\alpha(t_i^n-)/\mu^n) + F_d(Q^n_\alpha(t_i^n-)/\mu^n)\right]\right] \\
&\leq \frac{2ct}{\epsilon^2}\mathbb{P}\left(d_1 < \max_{s\leq t} Q^n_\alpha(s)/\mu^n\right) \\
&< \eta
\end{aligned}$$

for sufficiently large n. The last inequality follows from Lemma 5.1. This completes the proof. □

Proof of Proposition 5.7 Fix $\alpha \in \{S, T\}$ and $t > 0$. For each $s \geq 0$, let $\breve{Q}^n_\alpha(s)$ denote the difference between the index of the last arriving job and the index of the job currently in service, plus any of the initial jobs present at time zero that remain in the system. The jobs present at time zero do not have indices. Note that, for each $s \geq 0$, $\breve{Q}^n_\alpha(s) \geq Q^n_\alpha(s)$, for both ticket and standard queues. The process $\breve{Q}^n_\alpha$ ignores the

balking and abandonment that has taken place since the arrival of the job currently in service. One can think of the process as progressing in a manner similar to a ticket queue for which, in addition to the abandoned tickets, balking is not accounted for until service would have begun for the departing job. We can bound the difference between the processes. Fix an arbitrary $\eta, \epsilon > 0$ and choose an $L > 0$ such that, by Lemma 5.1,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t]}\left(\breve{Q}^n_\alpha(s) - Q^n_\alpha(s)\right) > \frac{\epsilon\sqrt{n}}{3}\right)\\
&\quad\le \mathbb{P}\left(\sup_{s\in[0,t]} Q^n_\alpha(s) > L\sqrt{n}\right) + \mathbb{P}\left(\sup_{s\in[0,t]} W^n_\alpha(s) > L/\sqrt{n}\right)\\
&\qquad + \mathbb{P}\left(\sup_{j\le A^n(t)} \sum_{i=j+1}^{j+\lfloor(L+\epsilon/3)\sqrt{n}\rfloor} \left(1\left(b_i \le L\sqrt{n}/\mu^n\right) + 1\left(d_i \le L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{3}\right)\\
&\quad< \frac{\eta}{5} + \mathbb{P}\left(\sup_{j\le A^n(t)} \sum_{i=j+1}^{j+\lfloor(L+\epsilon/3)\sqrt{n}\rfloor} \left(1\left(b_i \le L/\sqrt{n}\right) + 1\left(d_i \le L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{3}\right)
\end{aligned} \tag{8.10}$$

for sufficiently large n. To appreciate the above inequality, note that $\breve{Q}^n_\alpha$ is an inflated version of Q^n_α. The jobs in the former not accounted for in the latter must have abandoned or balked, or will have eventually abandoned. So if there exists an $s \in [0, t]$ such that $\breve{Q}^n_\alpha(s) - Q^n_\alpha(s)$ exceeds $(L + \epsilon/3)\sqrt{n}$ and $Q^n_\alpha(u) \le L\sqrt{n}$ for all $u \in [0, t]$, then there must be at least $\epsilon\sqrt{n}/3$ abandoned or balked jobs within some $(L+\epsilon/3)\sqrt{n}$ consecutively arriving jobs. We place upper bounds on the queue length and workload and this makes our abandonment and balking indicators i.i.d. random variables.

For any $\delta > 0$, it is true that $(L + \epsilon/3)\sqrt{n} < \mu\delta n$ for sufficiently large n. Hence, by Proposition 5.4,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{j\le A^n(t)} \sum_{i=j+1}^{j+\lfloor(L+\epsilon/3)\sqrt{n}\rfloor} \left(1\left(b_i \le L/\sqrt{n}\right) + 1\left(d_i \le L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{3}\right)\\
&\quad\le \mathbb{P}\left(\sup_{j\le A^n(t)} \sum_{i=j+1}^{j+\lfloor\mu\delta n\rfloor} \left(1\left(b_i \le L/\sqrt{n}\right) + 1\left(d_i \le L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{3}\right)\\
&\quad\le \mathbb{P}\left(\sup_{s\le t} \sum_{i=A^n(s)+1}^{A^n(s+2\delta)} \left(1\left(b_i \le L/\sqrt{n}\right) + 1\left(d_i \le L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{3}\right)\\
&\qquad + \mathbb{P}\left(\inf_{s\le t}\left(A^n(s+2\delta) - A^n(s)\right) < \mu\delta n\right)\\
&\quad< \frac{\eta}{5}.
\end{aligned} \tag{8.11}$$

Additionally, define for each $s \geq 0$ the quantity $\breve{W}_\alpha^n(s)$ that tracks the service times of the jobs associated with $\breve{Q}_\alpha^n(s)$. This process has the *entire* service time of the job currently in service and the service times of all jobs that arrive after the arrival time of the job in service; jobs that abandon or balk contribute the process $\breve{W}_\alpha^n$ nonetheless. Just as with Q_α^n and its augmented version $\breve{Q}_\alpha^n$, it is also the case that $\breve{W}_\alpha^n(s) \geq W_\alpha^n(s)$ for each $s \geq 0$. Unlike the process W_α^n, the augmented process $\breve{W}_\alpha^n$ experiences both upward and downward jumps. Upward jumps are the size of the would-be service time of each arriving job, even those that balk or abandon, and occur at the time of arrival of the corresponding job. The downward jumps occur at service completion times. The downward jump size is equal to the service time of the job that was in service plus the would-be service times of jobs that arrived between the arrival time of the job just served and the arrival time of the next job to be served. Just as we did in (8.10) for the queue lengths, we can bound the difference between these processes. By Lemmas 3.3, 3.4, and 5.1, and Eq. (8.10),

$$\begin{aligned}
&\mathbb{P}\left(\mu^n \sup_{s\in[0,t]}\left(\breve{W}_\alpha^n(s) - W_\alpha^n(s)\right) > \frac{\epsilon\sqrt{n}}{3}\right) \\
&\leq \mathbb{P}\left(\mu^n \cdot \sup_{i\leq A^n(t)} v_i^n > \frac{\epsilon\sqrt{n}}{6}\right) + \mathbb{P}\left(\sup_{s\in[0,t]} \hat{Q}_\alpha^n(s) > (L+\epsilon/3)\sqrt{n}\right) \\
&\quad + \mathbb{P}\left(\sup_{s\in[0,t]} Q_\alpha^n(s)/\mu^n > L/\sqrt{n}\right) + \mathbb{P}\left(\sup_{s\in[0,t]} W_\alpha^n(s) > L/\sqrt{n}\right) \\
&\quad + \mathbb{P}\left(\mu^n \cdot \sup_{j\leq A^n(t)} \sum_{i=j+1}^{j+\lfloor(L+\epsilon/3)\sqrt{n}\rfloor} v_i^n \cdot \left(1\left(b_i \leq L/\sqrt{n}\right)\right.\right. \\
&\quad \left.\left. + 1\left(d_i \leq L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{6}\right) \\
&< \frac{\eta}{5} + \mathbb{P}\left(\mu^n \cdot \sup_{j\leq A^n(t)} \sum_{i=j+1}^{j+\lfloor(L+\epsilon/3)\sqrt{n}\rfloor} v_i^n \cdot \left(1\left(b_i \leq L/\sqrt{n}\right)\right.\right. \\
&\quad \left.\left. + 1\left(d_i \leq L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{6}\right)
\end{aligned} \tag{8.12}$$

for sufficiently large n. As with (8.10), we replace $(L+\epsilon)\sqrt{n}$ with a bigger quantity $\mu\delta n$ (provided n is sufficiently large), where, by Proposition 5.5, $\delta > 0$ is chosen such that

$$\mathbb{P}\left(\mu^n \sup_{j\leq A^n(t)} \sum_{i=j+1}^{j+\lfloor(L+\epsilon/3)\sqrt{n}\rfloor} v_i^n \cdot \left(1\left(b_i \leq L/\sqrt{n}\right) + 1\left(d_i \leq L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{6}\right)$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\mu^n \sup_{j\leq A^n(t)} \sum_{i=j+1}^{j+\lfloor \mu\delta n\rfloor} v_i^n \cdot \left(1\left(b_i \leq L/\sqrt{n}\right)+1\left(d_i \leq L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{6}\right)\\
&\leq \mathbb{P}\left(\mu^n \sup_{s\leq t} \sum_{i=A^n(s)+1}^{A^n(s+2\delta)} v_i^n \cdot \left(1\left(b_i \leq L/\sqrt{n}\right)+1\left(d_i \leq L/\sqrt{n}\right)\right) > \frac{\epsilon\sqrt{n}}{6}\right)\\
&\quad + \mathbb{P}\left(\inf_{s\leq t}\left(A^n(s+2\delta)-A^n(s)\right) < \mu\delta n\right)\\
&< \frac{\eta}{5}
\end{aligned}
\tag{8.13}
$$

for sufficiently large n.

Lastly, note that by the functional weak law of large numbers,

$$
\mathbb{P}\left(\sup_{s\in[0,t]}\left|\breve{Q}_\alpha^n(s)-\mu^n\breve{W}_\alpha^n(s)\right| > \frac{\varepsilon\sqrt{n}}{3}\right) < \frac{\eta}{5} \tag{8.14}
$$

for sufficiently large n.

It now follows from the triangle inequality and (8.10)–(8.14) that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t]}\left|\tilde{Q}_\alpha^n(s)-\mu\tilde{W}_\alpha^n(s)\right| > \varepsilon\right)\\
&\leq \mathbb{P}\left(\sup_{s\in[0,t]}\left|Q_\alpha^n(s)-\breve{Q}_\alpha^n(s)\right| > \frac{\varepsilon\sqrt{n}}{3}\right)+\mathbb{P}\left(\mu^n\sup_{s\in[0,t]}\left|W_\alpha^n(s)-\breve{W}_\alpha^n(s)\right| > \frac{\varepsilon\sqrt{n}}{3}\right)\\
&\quad +\mathbb{P}\left(\sup_{s\in[0,t]}\left|\breve{Q}_\alpha^n(s)-\mu^n\breve{W}_\alpha^n(s)\right| > \frac{\varepsilon\sqrt{n}}{3}\right)\\
&< \eta
\end{aligned}
$$

for sufficiently large n. This completes the proof. □

Proof of Proposition 5.9 Fix $\varepsilon, \eta, t > 0$ and α. The server cannot work faster than rate one. It follows that for each $s \leq t$,

$$
T_\alpha^n(s) \leq s. \tag{8.15}
$$

Using (3.10) we can provide a lower bound on the service allocation process for any $s \geq 0$,

$$
T_\alpha^n(s) \geq -W_\alpha^n(s)+\sum_{i=1}^{A^n(s)} v_i^n - \sum_{i=1}^{A^n(s)} v_i^n\left(1(b_i < Q_\alpha^n(t_i-)/\mu^n)+1(d_i < W_\alpha^n(t_i-))\right). \tag{8.16}
$$

It follows from (8.15), and (8.16) that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t]}\left|T_\alpha^n(s)-s\right|>\varepsilon\right)\\
&=\mathbb{P}\left(\sup_{s\in[0,t]}T_\alpha^n(s)<s-\varepsilon\right)\\
&\le\mathbb{P}\left(\sup_{s\in[0,t]}W_\alpha^n(s)>\frac{\varepsilon}{3}\right)+\mathbb{P}\left(\sup_{s\in[0,t]}\left|\sum_{i=1}^{A^n(s)}v_i^n-s\right|>\frac{\varepsilon}{3}\right)\\
&\quad+\mathbb{P}\left(\sum_{i=1}^{A^n(t)}v_i^n\cdot\left(1(b_i<Q_\alpha^n(t_i-)/\mu^n)+1(d_i<W_\alpha^n(t_i-))\right)>\frac{\varepsilon}{3}\right). \qquad (8.17)
\end{aligned}$$

From Lemmas 3.7 and 5.1, we can bound the first two terms on the right-hand side,

$$\mathbb{P}\left(\sup_{s\in[0,t]}W_\alpha^n(s)>\frac{\varepsilon}{3}\right)+\mathbb{P}\left(\sup_{s\in[0,t]}\left|\sum_{i=1}^{A^n(s)}v_i^n-s\right|>\frac{\varepsilon}{3}\right)<\frac{\eta}{2}. \qquad (8.18)$$

For the third term, by Lemma 3.3 and Proposition 5.5,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{A^n(t)}v_i^n\cdot\left(1(b_i<Q_\alpha^n(t_i-)/\mu^n)+1(d_i<W_\alpha^n(t_i-))\right)>\frac{\varepsilon}{3}\right) \qquad (8.19)\\
&\le\mathbb{P}\left(A^n(t)>2\mu nt\right)\\
&\quad+\mathbb{P}\left(\sum_{i=1}^{2\mu nt}v_i^n\cdot\left(1(b_i<Q_\alpha^n(t_i-)/\mu^n)+1(d_i<W_\alpha^n(t_i-))\right)>\frac{\epsilon}{3}\right)\\
&\le\mathbb{P}\left(\sum_{i=1}^{2\mu nt}\left(v_i^n-\frac{1}{\mu n}\right)\cdot\left(1(b_i<Q_\alpha^n(t_i-)/\mu^n)+1(d_i<W_\alpha^n(t_i-))\right)>\frac{\epsilon}{9}\right)\\
&\quad+\mathbb{P}\left(\sum_{i=1}^{2\mu nt}\frac{1}{\mu n}\cdot\left(1(b_i<Q_\alpha^n(t_i-)/\mu^n)-F_b(Q_\alpha^n(t_i-)/\mu^n)\right.\right.\\
&\qquad\left.\left.+1(d_i<W_\alpha^n(t_i-))-F_d(W_\alpha^n(t_i-))\right)>\frac{\epsilon}{9}\right)\\
&\quad+\mathbb{P}\left(\sum_{i=1}^{2\mu nt}\frac{1}{\mu n}\cdot\left(F_b(Q_\alpha^n(t_i-)/\mu^n)+F_d(W_\alpha^n(t_i-))\right)>\frac{\epsilon}{9}\right)+\mathbb{P}\left(A^n(t)>2\mu nt\right)\\
&\le\frac{\eta}{8}+\frac{\eta}{8}+\frac{\eta}{8}+\frac{\eta}{8}\\
&<\frac{\eta}{2}. \qquad (8.20)
\end{aligned}$$

The result follows from (8.16)–(8.19) and a modification of the proofs of Propositions 5.4–5.6. □

Proof of Proposition 5.10 Fix ϵ, η, and $t > 0$. First notice that for any $n \geq 0$, $s \geq 0$, and $\alpha \in \{S.T\}$, we have that $R_\alpha^{2,n}(s) - R_\alpha^n(s) \geq 0$. It is instructive to expand this difference for each of the α values. For the standard queue,

$$\begin{aligned} R_S^{2,n}(s) - R_S^n(s) &= \sum_{i=1}^{A^n(s)} 1(b_i > Q_S^n(t_i^n-)/\mu^n) \cdot 1(d_i \leq W_S^n(t_i^n)) \cdot 1(d_i > s - t_i^n) \\ &\leq \sum_{i=1}^{A^n(s)} 1(d_i \leq W_S^n(t_i^n)) \cdot 1(W_S^n(t_i^n) > s - t_i^n). \end{aligned}$$

Note that we have eliminated the indicator associated with balking. Moreover, for an abandoning customer the workload upon arrival must exceed the patience quantity. Hence, we can replace the patience quantity in the last of the indicators with the workload upon arrival. Now we consider the ticket queue:

$$\begin{aligned} R_T^{2,n}(s) - R_T^n(s) &= \sum_{i=1}^{A^n(s)} 1(b_i > Q_T^n(t_i^n-)/\mu^n) \cdot 1(d_i \leq W_T^n(t_i^n)) \cdot 1(W_T^n(t_i^n) > s - t_i^n) \\ &\leq \sum_{i=1}^{A^n(s)} 1(d_i \leq W_T^n(t_i^n)) \cdot 1(W_T^n(t_i^n) > s - t_i^n). \end{aligned}$$

Both the standard and the ticket queue have the same bounds:

$$\begin{aligned} &\mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{R}_\alpha^{2,n}(s) - \tilde{R}_\alpha^n(s) \right| > \epsilon \right) \\ &\quad \leq \mathbb{P}\left(\sup_{s\in[0,t]} \sum_{i=1}^{A^n(s)} 1(d_i \leq W_\alpha^n(t_i^n)) \cdot 1(W_\alpha^n(t_i^n) > s - t_i^n) > \sqrt{n}\epsilon \right) \end{aligned}$$

for each $\alpha \in \{S, T\}$. For the remainder of the proof, fix α. Next we replace the workload quantities with an upper bound:

$$\begin{aligned} &\mathbb{P}\left(\sup_{s\in[0,t]} \left| \tilde{R}_\alpha^{2,n}(s) - \tilde{R}_\alpha^n(s) \right| > \epsilon \right) \\ &\quad \leq \mathbb{P}\left(\sup_{s\in[0,t]} \sum_{i=1}^{A^n(s)} 1(d_i \leq K/\sqrt{n}) \cdot 1(K/\sqrt{n} > s - t_i^n) > \sqrt{n}\epsilon \right) \\ &\qquad + \mathbb{P}\left(\sup_{s\in[0,t]} W_\alpha^n(s) > K/\sqrt{n} \right). \end{aligned}$$

Notice that in the first term on the right-hand side above, the only jobs that contribute positively to the summation are those jobs i whose arrival time is after $s - K/\sqrt{n}$; that is $t_i^n > s - K/\sqrt{n}$. By Lemma 5.1,

$$\begin{aligned}&\mathbb{P}\left(\sup_{s\in[0,t]}\left|\tilde{R}_\alpha^{2,n}(s) - \tilde{R}_\alpha^n(s)\right| > \epsilon\right)\\ &\quad\le \frac{\eta}{2} + \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-K/\sqrt{n})^+)}^{A^n(s)} 1(d_i \le K/\sqrt{n}) > \sqrt{n}\epsilon\right)\end{aligned}$$

for sufficiently large n. For any arbitrarily chosen $\delta > 0$, it is true that $\delta > K/\sqrt{n}$ for sufficiently large n. It follows then by Proposition 5.4 that we can choose a $\delta > 0$ so that

$$\begin{aligned}&\mathbb{P}\left(\sup_{s\in[0,t]}\left|\tilde{R}_\alpha^{2,n}(s) - \tilde{R}_\alpha^n(s)\right| > \epsilon\right)\\ &\quad\le \frac{\eta}{2} + \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-K/\sqrt{n})^+)}^{A^n(s)} 1(d_i \le K/\sqrt{n}) > \sqrt{n}\epsilon\right)\\ &\quad\le \frac{\eta}{2} + \mathbb{P}\left(\sup_{s\in[0,t]}\sum_{i=A^n((s-\delta)^+)}^{A^n(s)} 1(d_i \le K/\sqrt{n}) > \sqrt{n}\epsilon\right)\\ &\quad< \eta\end{aligned}$$

for sufficiently large n. This completes the proof. □

Proof of Proposition 5.11 Fix $\epsilon, \eta, t > 0$, and $\alpha \in \{S, T\}$. Notice that $R_\alpha^{1,n} - R_\alpha^{2,n}$ is non-decreasing. Hence, replacing the workload and queue length with an upper bound yields

$$\begin{aligned}&\mathbb{P}\left(\sup_{s\in[0,t]}\left|\tilde{R}_\alpha^{1,n}(s) - \tilde{R}_\alpha^{2,n}(s)\right| > \epsilon\right)\\ &\quad\le \mathbb{P}\left(\sum_{i=1}^{A^n(t)} 1(b_i \le Q_\alpha^n(t_i^n-)/\mu^n)\cdot 1(d_i \le W_T^n(t_i^n-)) > \sqrt{n}\epsilon\right)\\ &\quad\le \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu n t\rfloor} 1(b_i \le K/\sqrt{n})\cdot 1(d_i \le K/\sqrt{n}) > \sqrt{n}\epsilon\right) + \mathbb{P}\left(A^n(t) > 2\mu n t\right)\\ &\qquad+\mathbb{P}\left(\sup_{s\in[0,t]} Q_\alpha^n(s)/\mu^n > K/\sqrt{n}\right) + \mathbb{P}\left(\sup_{s\in[0,t]} W_\alpha^n(s) > K/\sqrt{n}\right).\end{aligned}$$

Choose a $K > 0$ such that by Lemma 3.3 and (5.1) and (5.2) of Lemma 5.1,

$$\mathbb{P}\left(\sup_{s\in[0,t]}\left|\tilde{R}_\alpha^{1,n}(s)-\tilde{R}_\alpha^{2,n}(s)\right|>\epsilon\right) < \frac{\eta}{2}+\mathbb{P}\left(\sum_{i=1}^{\lceil 2\mu nt\rceil} 1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})>\sqrt{n}\epsilon\right). \tag{8.21}$$

As for the second term on the right-hand side above, we resort to adding and subtracting the mean of each summand:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor} 1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})>\sqrt{n}\epsilon\right)\\
&\le \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\mathbb{E}\left[1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})\right]>\frac{\sqrt{n}\epsilon}{2}\right)\\
&\quad+\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})\right.\right.\\
&\quad\left.\left.-\mathbb{E}\left[1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})\right]\right)>\frac{\sqrt{n}\epsilon}{2}\right)\\
&\le \mathbb{P}\left((2\mu nt)F_b(K/\sqrt{n})F_d(K/\sqrt{n})>\frac{\sqrt{n}\epsilon}{2}\right)\\
&\quad+\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})-F_b(K/\sqrt{n})F_d(K/\sqrt{n})\right)>\frac{\sqrt{n}\epsilon}{2}\right).
\end{aligned} \tag{8.22}$$

By Lemma 5.2 we can bound the first term on the right-hand side:

$$(2\mu nt)F_b(K/\sqrt{n})F_d(K/\sqrt{n})<\frac{\sqrt{n}\epsilon}{2} \tag{8.23}$$

for sufficiently large n. As for the second term, by Chebyshev's inequality and (8.23)

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})-F_b(K/\sqrt{n})F_d(K/\sqrt{n})\right)>\frac{\sqrt{n}\epsilon}{2}\right)\\
&\le \frac{4}{n\epsilon^2}\mathbb{E}\left[\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})-F_b(K/\sqrt{n})F_d(K/\sqrt{n})\right)\right)^2\right]\\
&\le \frac{4}{n\epsilon^2}\mathbb{E}\left[\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(b_i\le K/\sqrt{n})\cdot 1(d_i\le K/\sqrt{n})-F_b(K/\sqrt{n})F_d(K/\sqrt{n})\right)^2\right]
\end{aligned}$$

$$< \frac{8}{n\epsilon^2}(2\mu nt)F_b(K/\sqrt{n})F_d(K/\sqrt{n})$$
$$< \frac{\eta}{2}. \tag{8.24}$$

The result follows from (8.21)–(8.24). □

Proof of Proposition 5.12 Fix $\epsilon, \eta, t > 0$, $\alpha \in \{S, T\}$, and set $\delta = \epsilon/(16\mu t)$. By Lemma 5.1 and Proposition 5.7, respectively, we can choose a $K > 0$ such that for sufficiently large n,

$$\mathbb{P}\left(\sup_{s\in[0,t]} W_\alpha^n(s) > K/\sqrt{n}\right) < \min\left(\frac{\eta}{4}, \frac{\epsilon^2\eta}{128\mu t}\right) \tag{8.25}$$

and

$$\mathbb{P}\left(\sup_{s\in[0,t]} \left|Q_\alpha^n(s)/\mu^n - W_\alpha^n(s)\right| > \delta/\sqrt{n}\right) < \frac{\eta}{8}. \tag{8.26}$$

Fix such a K. We start by replacing the weighted queue length with the workload process:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in[0,t]} \left|\tilde{R}_\alpha^{0,n}(s) - \tilde{R}_\alpha^{1,n}(s)\right| > \epsilon\right) \\
&= \mathbb{P}\left(\sup_{s\in[0,t]} \left|\sum_{i=1}^{A^n(s)} 1(d_i \le Q_\alpha^n(t_i^n-)/\mu^n) - 1(d_i \le W_\alpha^n(t_i^n-))\right| > \sqrt{n}\epsilon\right) \\
&\le \mathbb{P}\left(\sum_{i=1}^{A^n(t)} 1(d_i \in [W_\alpha^n(t_i^n-) - \delta/\sqrt{n}, W_\alpha^n(t_i^n-) + \delta/\sqrt{n}]) > \sqrt{n}\epsilon\right) \\
&\quad + \mathbb{P}\left(\sup_{s\in[0,t]} \left|Q_\alpha^n(s)/\mu^n - W_\alpha^n(s)\right| > \delta/\sqrt{n}\right).
\end{aligned} \tag{8.27}$$

The second term of the far right-hand side is handled by (8.26). As for the first term, we construct a martingale. By Lemma 3.3,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{A^n(t)} 1(d_i \in [W_\alpha^n(t_i^n-) - \delta/\sqrt{n}, W_\alpha^n(t_i^n-) + \delta/\sqrt{n}]) > \sqrt{n}\epsilon\right) \\
&\le \frac{\eta}{4} \\
&\quad + \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor} \mathbb{E}\left[1(d_i \in [W_\alpha^n(t_i^n-) - \delta/\sqrt{n}, W_\alpha^n(t_i^n-) + \delta/\sqrt{n}])\middle|\mathcal{F}_{i-1}^n\right] > \frac{\sqrt{n}\epsilon}{2}\right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu nt\rfloor} \left(1(d_i \in [W_\alpha^n(t_i^n-) - \delta/\sqrt{n}, W_\alpha^n(t_i^n-) + \delta/\sqrt{n}]\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}\left[1(d_i \in [W_\alpha^n(t_i^n-) - \delta/\sqrt{n},\, W_\alpha^n(t_i^n-) + \delta/\sqrt{n}]) \,\middle|\, \mathcal{F}_{i-1}^n\right]) > \frac{\sqrt{n}\epsilon}{2}\Bigg) \\
&\le \frac{\eta}{4} + \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu n t\rfloor} \left(F_d(W_\alpha^n(t_i^n-) + \delta/\sqrt{n}) - F_d(W_\alpha^n(t_i^n-) - \delta/\sqrt{n})\right) > \frac{\sqrt{n}\epsilon}{2}\right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu n t\rfloor} \left(1(d_i \in [W_\alpha^n(t_i^n-) - \delta/\sqrt{n},\, W_\alpha^n(t_i^n-) + \delta/\sqrt{n}]) \right.\right. \\
&\quad \left.\left. - \left(F_d(W_\alpha^n(t_i^n-) + \delta/\sqrt{n}) - F_d(W_\alpha^n(t_i^n-) - \delta/\sqrt{n})\right)\right) > \frac{\sqrt{n}\epsilon}{2}\right). \qquad (8.28)
\end{aligned}
$$

As for the second term on the far right-hand side of (8.28), notice that each summand contributes an amount equal to the increase of the abandonment distribution function over an interval of length 2δ. By (8.25) and Lemma 5.3,

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu n t\rfloor} \left(F_d(W_\alpha^n(t_i^n-) + \delta/\sqrt{n}) - F_d(W_\alpha^n(t_i^n-) - \delta/\sqrt{n}])\right) > \frac{\sqrt{n}\epsilon}{2}\right) \\
&\le \mathbb{P}\left(\sup_{s\in[0,t]} \tilde{W}^n(s) > K\right) \\
&\quad + 1\left(2\mu n t \sup_{s\in[0,K]} \left(F_d((s+2\delta)/\sqrt{n}) - F_d(s/\sqrt{n})\right) > \frac{\sqrt{n}\epsilon}{2}\right) \\
&< \frac{\eta}{4} \qquad (8.29)
\end{aligned}
$$

for sufficiently large n. Now consider the third term on the right-hand side of (8.28). The following steps are similar to those in the proof of Proposition 5.6. By (8.25) and Lemma 5.2,

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\lfloor 2\mu n t\rfloor} \left(1(d_i \in [W_\alpha^n(t_i^n-) - \delta/\sqrt{n},\, W_\alpha^n(t_i^n-) + \delta/\sqrt{n}]) \right.\right. \\
&\quad \left.\left. - \left(F_d(W_\alpha^n(t_i^n-) + \delta/\sqrt{n}) - F_d(W_\alpha^n(t_i^n-) - \delta/\sqrt{n})\right)\right) > \frac{\sqrt{n}\epsilon}{2}\right) \\
&\le \frac{4}{n\epsilon^2}\mathbb{E}\left[\left(\sum_{i=1}^{\lfloor 2\mu n t\rfloor} \left(1(d_i \in [W_\alpha^n(t_i^n-) - \delta/\sqrt{n},\, W_\alpha^n(t_i^n-) + \delta/\sqrt{n}]) \right.\right.\right. \\
&\quad \left.\left.\left. - \left(F_d(W_\alpha^n(t_i^n-) + \delta/\sqrt{n}) - F_d(W_\alpha^n(t_i^n-) - \delta/\sqrt{n})\right)\right)\right)^2\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\le \frac{4}{n\epsilon^2}\mathbb{E}\left[\sum_{i=1}^{\lfloor 2\mu nt\rfloor}\left(1(d_i \in [W_\alpha^n(t_i^n-)-\delta/\sqrt{n}, W_\alpha^n(t_i^n-)+\delta/\sqrt{n}])\right.\right.\\
&\qquad \left.\left.-\left(F_d(W_\alpha^n(t_i^n-)+\delta/\sqrt{n})-F_d(W_\alpha^n(t_i^n-)-\delta/\sqrt{n})\right)\right)^2\right]\\
&\le \frac{8}{n\epsilon^2}\mathbb{E}\left[\sum_{i=1}^{\lfloor 2\mu nt\rfloor} F_d(W_\alpha^n(t_i^n-)+\delta/\sqrt{n})\right]\\
&\le \frac{16\mu t}{\epsilon^2}\left(\mathbb{P}\left(\sup_{s\in[0,t]} W_\alpha^n(s) > K/\sqrt{n}\right)+F_d((K+\delta)/\sqrt{n})\right)\\
&< \frac{\eta}{4}, \qquad (8.30)
\end{aligned}
$$

for sufficiently large n greater than $\left(\frac{256\mu t\theta(K+\delta)}{\epsilon^2\eta}\right)^2$.

The result follows from (8.26)–(8.30). □

Proof of Proposition 5.13 Fix $\eta, t > 0$, and $\alpha \in \{S, T\}$. By Lemma 5.1 there exists a $K > 1$ such that, for sufficiently large n,

$$
\mathbb{P}\left(\sup_{s\in[0,t]} Q_\alpha^n(s) > K\sqrt{n}\right)+\mathbb{P}\left(\sup_{s\in[0,t]} W_\alpha^n(s) > \frac{K}{\sqrt{n}}\right) < \frac{\eta}{2}.
$$

We use these constants to replace the workload and queue length:

$$
\begin{aligned}
\mathbb{P}\left(\hat{R}_\alpha^n(t) > L\right) &\le \mathbb{P}\left(\sum_{i=1}^{Q^n(0)} 1(\hat{d}_i \le \hat{w}_{i-1}^n) > L\right)\\
&\le \mathbb{P}\left(\sum_{i=1}^{\lfloor K\sqrt{n}\rfloor} 1(\hat{d}_i \le K/\sqrt{n}) > L\right)+\mathbb{P}\left(Q^n(0) \ge K\sqrt{n}\right)\\
&\quad +\mathbb{P}\left(W^n(0) > \frac{K}{\sqrt{n}}\right)\\
&< \mathbb{P}\left(\sum_{i=1}^{\lfloor K\sqrt{n}\rfloor} 1(\hat{d}_i \le K/\sqrt{n}) > L\right)+\frac{\eta}{2} \qquad (8.31)
\end{aligned}
$$

for sufficiently large n.

Recall that the residual deadlines of the initial jobs may have different distributions, $\hat{F}_{d,i}$, but those distributions have a common bound near the origin. Namely, there exists an $\hat{f} \ge 1$ and an $h_0 \in (0, 1/\hat{f})$ such that $\hat{F}_{d,i}(h) \le h\hat{f}$ for each $h \le h_0$. Given such an h_0 and $\hat{f}$, set L so that $L \ge \max(2\hat{f}K^2, 4\hat{f}K/\sqrt{\eta})$. We replace the initial jobs' residual deadlines with random variables which are i.i.d. In particular, on the same

probability space, each $\hat{d}_i$ is replaced by $\check{d}_i$. Whenever $\hat{d}_i < h_0$ then $\check{d}_i \geq \hat{d}_i$. Moreover, the common distribution of each $\check{d}_i$ is

$$\check{F}(x) = \begin{cases} x\hat{f}, & x < h_0 \\ 1, & x \geq h_0. \end{cases}$$

When $K/\sqrt{n} \leq h_0$ we have that

$$\left(K\sqrt{n}\right)\check{F}(K/\sqrt{n}) = \hat{f}K^2 < L/2$$

so that by Kolmogorov's inequality [4],

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\lfloor K\sqrt{n}\rfloor} 1(\hat{d}_i \leq K/\sqrt{n}) > L\right) \leq \mathbb{P}\left(\sum_{i=1}^{\lfloor K\sqrt{n}\rfloor} 1(\check{d}_i \leq K/\sqrt{n}) > L\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{\lfloor K\sqrt{n}\rfloor} \left(1(\check{d}_i \leq K/\sqrt{n}) - \check{F}(K/\sqrt{n})\right) > \frac{L}{2}\right) \\
&\leq \left(\frac{4}{L^2}\right)(K\sqrt{n})\mathbb{E}\left[\left(1(\check{d}_1 \leq K/\sqrt{n}) - \check{F}(K/\sqrt{n})\right)^2\right] \\
&\leq \left(\frac{4}{L^2}\right)(K\sqrt{n})\check{F}(K/\sqrt{n}) < \frac{\eta}{2}
\end{aligned} \tag{8.32}$$

for sufficiently large n. The result follows from (8.31) and (8.32). □

Proof of Proposition 5.17 We can break δ_α^n into four parts. For every $t \geq 0$ and $\alpha \in \{S, T\}$,

$$\begin{aligned}
\delta_\alpha^n(t) = &\frac{1}{\sqrt{n}}\sum_{i=1}^{A^n(t)} F_b(Q_\alpha^n(t_i^n-)/\mu^n) - F_b(Q_\alpha^n(t_i^n-)/(\mu n)) + F_d(Q_\alpha^n(t_i^n-)/\mu^n) \\
&- F_d(Q_\alpha^n(t_i^n-)/(\mu n)) \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{A^n(t)} \left(F_b(Q_\alpha^n(t_i^n-)/(\mu n)) + F_d(Q_\alpha^n(t_i^n-)/(\mu n)) - \theta\frac{Q_\alpha^n(t_i^n-)}{\mu n}\right) \\
&+ \theta\left(\int_0^t \tilde{Q}_\alpha^n(s-)d\left(\frac{\bar{A}^n(s)}{\mu}\right) - \int_0^t \tilde{Q}_\alpha^n(s)\mathrm{d}s\right) \\
&+ \frac{1}{\sqrt{n}}\left(\hat{S}^n(t) - \mu^n \min(t, W^n(0))\right).
\end{aligned} \tag{8.33}$$

We will show that each of these converges to zero in probability as $n \to \infty$.

Fix $t > 0$ and $\alpha \in \{S, T\}$ and select arbitrary constants $\epsilon, \eta > 0$. We first show that both

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{A^n(t)} \left|F_b(Q_\alpha^n(t_i^n-)/(\mu n)) + F_b(Q_\alpha^n(t_i^n-)/(\mu n))\right| > \frac{\epsilon}{4}\right) < \frac{\eta}{4} \tag{8.34}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \left| F_d(Q_\alpha^n(t_i^n-)/(\mu n)) + F_d(Q_\alpha^n(t_i^n-)/(\mu n)) \right| > \frac{\epsilon}{4} \right) < \frac{\eta}{4}. \tag{8.35}$$

We will prove (8.34), and the proof of (8.35) follows trivially. The right-hand side of (8.34) can be expanded as follows:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \left| F_b(Q_\alpha^n(t_i^n-)/(\mu n)) + F_b(Q_\alpha^n(t_i^n-)/(\mu n)) \right| > \frac{\epsilon}{4} \right) \\
&\quad \le \mathbb{P}\left(A^n(t) \left(\sup_{s\in[0,t]} \left| F_b(Q_\alpha^n(s)/\mu^n) + F_b(Q_\alpha^n(s)/(\mu n)) \right| \right) > \frac{\epsilon\sqrt{n}}{4} \right) \\
&\quad \le \mathbb{P}\left(\sup_{s\in[0,t]} Q_\alpha^n(s) > K\sqrt{n} \right) + \mathbb{P}\left(A^n(t) > 2\mu n t \right) \\
&\qquad + \mathbb{P}\left(\sup_{s\le K} \left| F_b(s\sqrt{n}/(\mu n)) + F_b(s\sqrt{n}/(\mu n)) \right| > \frac{\epsilon}{8\mu t\sqrt{n}} \right).
\end{aligned} \tag{8.36}$$

Fix $K > 0$ so that by (5.1) of Lemma 5.1,

$$\mathbb{P}\left(\sup_{s\in[0,t]} Q_\alpha^n(s) > K\sqrt{n} \right) < \frac{\eta}{8}. \tag{8.37}$$

Notice that for any $\delta > 0$, we have that

$$\sup_{s\le K} \left| \frac{s\sqrt{n}}{\mu n} - \frac{s\sqrt{n}}{\mu n} \right| \le \left| \frac{K\sqrt{n}}{\mu^n} - \frac{K\sqrt{n}}{\mu n} \right| < \frac{\delta}{\sqrt{n}}$$

for sufficiently large n. Set $\delta = \epsilon/(16\mu t\theta)$. The first result, (8.34), follows from (8.36), (8.37), and Lemmas 3.3 and 5.3. Likewise, (8.35) follows from an identical argument.

Next, we show that

$$\limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \left| F_b(Q_\alpha^n(t_i^n-)/(\mu n)) + F_d(Q_\alpha^n(t_i^n-)/(\mu n)) - \theta \frac{Q_\alpha^n(t_i^n-)}{\mu n} \right| > \frac{\epsilon}{4} \right) < \frac{\eta}{4}. \tag{8.38}$$

We explore the derivative of the abandonment and balking distributions:

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \left| F_b(Q_\alpha^n(t_i^n-)/(\mu n)) + F_d(Q_\alpha^n(t_i^n-)/(\mu n)) - \theta \frac{Q_\alpha^n(t_i^n-)}{\mu n} \right| > \frac{\epsilon}{4} \right)$$

$$\leq \mathbb{P}\left(A^n(t) \sup_{s\in[0,t]} \left| F_b(Q_\alpha^n(s)/(\mu n)) + F_d(Q_\alpha^n(s)/(\mu n)) - \theta \frac{Q_\alpha^n(s)}{\mu n} \right| > \frac{\epsilon\sqrt{n}}{4} \right)$$

$$\leq \mathbb{P}\left(\sup_{s\in[0,t]} Q_\alpha^n(s) > K\sqrt{n} \right) + \mathbb{P}\left(A^n(t) > 2\mu n t \right)$$

$$+ 1\left(\sup_{s\in[0,K/\mu]} \left| F_b(s/\sqrt{n}) + F_d(s/\sqrt{n}) - \frac{\theta s}{\sqrt{n}} \right| > \frac{\epsilon}{8\mu t\sqrt{n}} \right). \tag{8.39}$$

Recall that the derivatives at zero of both F_b and F_d exist and sum to θ. Hence, for any given $\delta > 0$, there exists an h_0 such that

$$\sup_{s\leq h} \left| \frac{F_b(s)}{s} + \frac{F_d(s)}{s} - \theta \right| < \delta$$

for all $h \leq h_0$. Let $\delta = \frac{\epsilon}{8Kt}$ and let h_0 be a constant so that the above inequality holds. Now choose an n_0 such that $K/(\mu\sqrt{n}) \leq h_0$ for all $n \geq n_0$. It follows that for each $n \geq n_0$,

$$\sup_{s\in[0,K/(\mu\sqrt{n})]} |F_b(s) + F_d(s) - \theta s| = \sup_{s\in[0,K/\mu]} \left| F(s/\sqrt{n}) - \frac{\theta s}{\sqrt{n}} \right| < \delta \frac{K}{\mu\sqrt{n}}$$

$$= \frac{\epsilon}{8\mu t\sqrt{n}}. \tag{8.40}$$

The result (8.38) follows from (8.37), (8.39), (8.40), and Lemma 3.3.

Third, we show that

$$\limsup_{n\to\infty} \mathbb{P}\left(\left| \int_0^t \tilde{Q}_\alpha^n(s-)\mathrm{d}\left(\frac{\bar{A}^n(s)}{\mu} \right) - \int_0^t \tilde{Q}_\alpha^n(s)\mathrm{d}s \right| > \epsilon \right) < \eta. \tag{8.41}$$

By Proposition 5.15, the process $\{\tilde{Q}_\alpha^n(s), s \leq t\}$ is tight. Consider a subsequence $\{n'\}$ over which the process $\tilde{Q}_\alpha^{n'}$ has a limit, say $\tilde{Q}_\alpha$. By the Skorohod Representation Theorem, there exists an alternative probability space on which are defined a sequence $\{(\hat{\tilde{Q}}_\alpha^{n'}, \hat{\bar{A}}_\alpha^{n'}), n \geq 1\}$ and, by Lemma 3.6, a limit process $(\hat{\tilde{Q}}_\alpha, \hat{\bar{A}}_\alpha)$ such that

$$(\hat{\tilde{Q}}_\alpha^{n'}, \hat{\bar{A}}_\alpha^{n'}) \stackrel{D}{=} (\tilde{Q}_\alpha^{n'}, \bar{A}_\alpha^{n'})$$

for each n' and such that $(\hat{\tilde{Q}}_\alpha^{n'}, \hat{\bar{A}}_\alpha^{n'}) \to (\hat{\tilde{Q}}_\alpha, \hat{\bar{A}}_\alpha)$ almost surely as $n' \to \infty$. It is also true that

$$\int_0^u \hat{\tilde{Q}}_\alpha^{n'}(s-)\mathrm{d}\left(\frac{\hat{\bar{A}}^{n'}(s)}{\mu} \right) \stackrel{D}{=} \int_0^u \tilde{Q}_\alpha^{n'}(s-)\mathrm{d}\left(\frac{\bar{A}^{n'}(s)}{\mu} \right),$$

and

$$\int_0^t \hat{\tilde{Q}}_\alpha^{n'}(s)\mathrm{d}s \overset{D}{=} \int_0^t \tilde{Q}_\alpha^{n'}(s)\mathrm{d}s$$

for each n'. Applying Lemma 8.3 from [6] twice we have

$$\sup_{u\in[0,t]} \left| \int_0^u \hat{\tilde{Q}}_\alpha^{n'}(s-)d\left(\frac{\hat{\tilde{A}}^{n'}(s)}{\mu}\right) - \int_0^u \hat{\tilde{Q}}(s)\mathrm{d}s \right| \to 0$$

and

$$\sup_{u\in[0,t]} \left| \int_0^t \hat{\tilde{Q}}_\alpha^{n'}(s)\mathrm{d}s - \int_0^t \hat{\tilde{Q}}(s)\mathrm{d}s \right| \to 0$$

almost surely as $n' \to \infty$, so that

$$\sup_{u\in[0,t]} \left| \int_0^u \hat{\tilde{Q}}_\alpha^{n'}(s-)d\left(\frac{\hat{\tilde{A}}^{n'}(s)}{\mu}\right) - \int_0^t \hat{\tilde{Q}}_\alpha^{n'}(s)\mathrm{d}s \right| \to 0$$

almost surely as $n' \to \infty$. It follows that in our original probability space

$$\sup_{u\in[0,t]} \left| \int_0^u \tilde{Q}_\alpha^{n'}(s-)d\left(\frac{\hat{\tilde{A}}^{n'}(s)}{\mu}\right) - \int_0^t \tilde{Q}_\alpha^{n'}(s)\mathrm{d}s \right| \to 0$$

as $n' \to \infty$. This limit holds on the arbitrarily chosen subsequence $\{n'\}$. Hence (8.41) holds.

Consider the fourth term on the right-hand side of (8.33). The service times of initial jobs that are actually served constitute an i.i.d sequence. This sequence obeys a weak law of large numbers, as does the renewal process $\hat{S}^n$, constructed by these service times over intervals of time that are of the order $1/\sqrt{n}$. Hence, by Proposition 5.15,

$$\mathbb{P}\left(\sup_{t\le W^n(0)} \left| \hat{S}^n(t) - \mu_n t \right| > \epsilon\sqrt{n} \right) < \mathbb{P}\left(\sup_{t\le K/\sqrt{n}} \left| \hat{S}^n(t) - \mu_n t \right| > \epsilon\sqrt{n} \right) + \mathbb{P}\left(W^n(0) > K/\sqrt{n} \right) < \eta, \tag{8.42}$$

for sufficiently large n.

Finally, our result follows from (8.33), (8.34), (8.35), (8.38), (8.41), and Lemma 3.3. □

Numerical tables

See Tables 1, 2, 3, and 4.

Table 1 Simulated results vs. heavy traffic approximations: arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), service = Exp($\mu$), balking = Exp($\theta_b$), reneging = Exp($\theta_r$)

Q_α	ρ	μ	β	θ_b	θ_r	Q	Q_{ROU}	W	W_{ROU}	R	R_{ROU}	B	B_{ROU}
Q_S	1.1	100	1	0.1	0.1	49.48 ± 0.199	50.74	0.481 ± 0.002	0.5074	0.044 ± 0.0002	0.0461	0.048 ± 0.0002	0.0461
Q_T	1.1	100	1	0.1	0.1	50.18 ± 0.203	50.74	0.475 ± 0.002	0.5074	0.043 ± 0.0002	0.0461	0.049 ± 0.0002	0.0461
Q_S	1.1	25	0.5	0.1	0.1	15.25 ± 0.071	15.24	0.587 ± 0.003	0.6099	0.0519 ± 0.0002	0.0554	0.0585 ± 0.0003	0.0554
Q_T	1.1	25	0.5	0.1	0.1	15.59 ± 0.073	15.24	0.578 ± 0.003	0.6099	0.0510 ± 0.0002	0.0554	0.0597 ± 0.0003	0.0554
Q_S	1.1	4	0.2	0.1	0.1	4.47 ± 0.021	4.40	1.04 ± 0.0006	1.100	0.0813 ± 0.0004	0.1000	0.102 ± 0.0006	0.1000
Q_T	1.1	4	0.2	0.1	0.1	4.67 ± 0.023	4.40	1.02 ± 0.0006	1.100	0.0791 ± 0.0004	0.1000	0.106 ± 0.0006	0.1000
Q_S	1.01	100	0.1	0.1	0.1	19.79 ± 0.094	19.78	0.195 ± 0.001	0.1978	0.0186 ± 0.0001	0.0195	0.0195 ± 0.0001	0.0195
Q_T	1.01	100	0.1	0.1	0.1	19.98 ± 0.095	19.78	0.194 ± 0.001	0.1978	0.0186 ± 0.0001	0.0195	0.0196 ± 0.0001	0.0195
Q_S	1.01	25	0.05	0.1	0.1	9.34 ± 0.045	9.39	0.363 ± 0.0018	0.3956	0.0333 ± 0.0002	0.03718	0.0362 ± 0.0002	0.03718
Q_T	1.01	25	0.05	0.1	0.1	9.51 ± 0.046	9.39	0.360 ± 0.0018	0.3956	0.0330 ± 0.0002	0.03718	0.0368 ± 0.0002	0.03718
Q_S	1.01	4	0.02	0.1	0.1	3.61 ± 0.019	3.64	0.855 ± 0.005	0.9104	0.0681 ± 0.0005	0.0901	0.0843 ± 0.0006	0.0901
Q_T	1.01	4	0.02	0.1	0.1	3.77 ± 0.020	3.64	0.842 ± 0.005	0.9104	0.0667 ± 0.0004	0.0901	0.0873 ± 0.0006	0.0901
Q_S	1	100	0	0.1	0.1	17.82 ± 0.083	17.84	0.176 ± 0.0009	0.1784	0.0168 ± 0.0001	0.0178	0.0176 ± 0.0001	0.0178
Q_T	1	100	0	0.1	0.1	17.98 ± 0.085	17.84	0.175 ± 0.0009	0.1784	0.0168 ± 0.0001	0.0178	0.0177 ± 0.0001	0.0178
Q_S	1	25	0	0.1	0.1	8.86 ± 0.046	8.92	0.345 ± 0.002	0.3568	0.0317 ± 0.0002	0.0356	0.0344 ± 0.0002	0.0356
Q_T	1	25	0	0.1	0.1	9.03 ± 0.047	8.92	0.343 ± 0.002	0.3568	0.0314 ± 0.0002	0.0356	0.0350 ± 0.0002	0.0356
Q_S	1	4	0	0.1	0.1	3.52 ± 0.019	3.56	0.832 ± 0.005	0.8920	0.0670 ± 0.0005	0.0892	0.0817 ± 0.0005	0.0892
Q_T	1	4	0	0.1	0.1	3.67 ± 0.021	3.56	0.820 ± 0.005	0.8920	0.0657 ± 0.0005	0.0892	0.0848 ± 0.0005	0.0892
Q_S	0.99	100	−0.1	0.1	0.1	15.95 ± 0.067	16.14	0.157 ± 0.0007	0.1614	0.0152 ± 0.0001	0.0163	0.0157 ± 0.0001	0.0163
Q_T	0.99	100	−0.1	0.1	0.1	16.10 ± 0.068	16.14	0.157 ± 0.0007	0.1614	0.0151 ± 0.0001	0.0163	0.0159 ± 0.0001	0.0163

Table 1 continued

Q_α	ρ	μ	β	θ_b	θ_r	Q	Q_{ROU}	W	W_{ROU}	R	R_{ROU}	B	B_{ROU}
Q_S	0.99	25	−0.05	0.1	0.1	8.40 ± 0.046	8.48	0.328 ± 0.002	0.3392	0.0302 ± 0.0002	0.0342	0.0326 ± 0.0002	0.0342
Q_T	0.99	25	−0.05	0.1	0.1	8.55 ± 0.046	8.48	0.325 ± 0.002	0.3392	0.0299 ± 0.0002	0.0342	0.0332 ± 0.0002	0.0342
Q_S	0.99	4	−0.02	0.1	0.1	3.44 ± 0.020	3.49	0.814 ± 0.0058	0.8741	0.0657 ± 0.0005	0.0882	0.0801 ± 0.0005	0.0882
Q_T	0.99	4	−0.02	0.1	0.1	3.58 ± 0.022	3.49	0.803 ± 0.0055	0.8741	0.0644 ± 0.0005	0.0882	0.0830 ± 0.0005	0.0882
Q_S	0.9	100	−1	0.1	0.1	7.12 ± 0.0264	7.77	0.0708 ± 0.0003	0.0777	0.00692 ± 0.0001	0.0086	0.0071 ± 0.0001	0.0086
Q_T	0.9	100	−1	0.1	0.1	7.16 ± 0.0267	7.77	0.0707 ± 0.0003	0.0777	0.00691 ± 0.0001	0.0086	0.0071 ± 0.0001	0.0086
Q_S	0.9	25	−0.5	0.1	0.1	5.19 ± 0.024	5.61	0.204 ± 0.001	0.2246	0.0192 ± 0.0001	0.0249	0.0204 ± 0.0001	0.0249
Q_T	0.9	25	−0.5	0.1	0.1	5.27 ± 0.024	5.61	0.203 ± 0.001	0.2246	0.0192 ± 0.0001	0.0249	0.0207 ± 0.0001	0.0249
Q_S	0.9	4	−0.2	0.1	0.1	2.74 ± 0.013	2.93	0.653 ± 0.0036	0.7328	0.0542 ± 0.0004	0.0814	0.0645 ± 0.0005	0.0814
Q_T	0.9	4	−0.2	0.1	0.1	2.84 ± 0.014	2.93	0.647 ± 0.0035	0.7328	0.0534 ± 0.0004	0.0814	0.0667 ± 0.0005	0.0814
Q_S	0.8	100	−2	0.1	0.1	3.73 ± 0.0098	4.59	0.0372 ± 0.0001	0.0459	0.0037 ± 0.0001	0.0057	0.0037 ± 0.0001	0.0057
Q_T	0.8	100	−2	0.1	0.1	3.75 ± 0.0099	4.59	0.0372 ± 0.0001	0.0459	0.0037 ± 0.0001	0.0057	0.0037 ± 0.0001	0.0057
Q_S	0.8	25	−1	0.1	0.1	3.22 ± 0.011	3.88	0.127 ± 0.0005	0.1555	0.0122 ± 0.0001	0.0194	0.0127 ± 0.0001	0.0194
Q_T	0.8	25	−1	0.1	0.1	3.25 ± 0.012	3.88	0.127 ± 0.0005	0.1555	0.0122 ± 0.0001	0.0194	0.0129 ± 0.0001	0.0194
Q_S	0.8	4	−0.4	0.1	0.1	2.11 ± 0.012	2.44	0.508 ± 0.0036	0.6113	0.0432 ± 0.0003	0.0764	0.0496 ± 0.0003	0.0764
Q_T	0.8	4	−0.4	0.1	0.1	2.18 ± 0.013	2.44	0.505 ± 0.0035	0.6113	0.0428 ± 0.0003	0.0764	0.0511 ± 0.0003	0.0764

Table 2 Simulated results vs. heavy traffic approximations: arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), service = LogNormal($1/\mu, 1/\mu^2$), balking = Uniform(0,$1/\theta_b$), reneging = Uniform(0,$1/\theta_r$)

Q_α	ρ	μ	β	θ_b	θ_r	Q	Q_{ROU}	W	W_{ROU}	R	R_{ROU}	B	B_{ROU}
Q_S	1.1	100	1	0.1	0.1	48.17 ± 0.174	50.74	0.468 ± 0.002	0.5074	0.044 ± 0.0002	0.0461	0.048 ± 0.0002	0.0461
Q_T	1.1	100	1	0.1	0.1	48.85 ± 0.177	50.74	0.462 ± 0.002	0.5074	0.043 ± 0.0002	0.0461	0.049 ± 0.0002	0.0461
Q_S	1.1	25	0.5	0.1	0.1	14.65 ± 0.073	15.24	0.563 ± 0.003	0.6099	0.0517 ± 0.0002	0.0554	0.0587 ± 0.0003	0.0554
Q_T	1.1	25	0.5	0.1	0.1	14.97 ± 0.076	15.24	0.555 ± 0.003	0.6099	0.0508 ± 0.0002	0.0554	0.0599 ± 0.0003	0.0554
Q_S	1.1	4	0.2	0.1	0.1	4.17 ± 0.022	4.40	0.968 ± 0.0006	1.100	0.0813 ± 0.0004	0.1000	0.104 ± 0.0006	0.1000
Q_T	1.1	4	0.2	0.1	0.1	4.35 ± 0.024	4.40	0.946 ± 0.0006	1.100	0.0786 ± 0.0004	0.1000	0.108 ± 0.0006	0.1000
Q_S	1.01	100	0.1	0.1	0.1	19.35 ± 0.086	19.78	0.191 ± 0.0009	0.1978	0.0185 ± 0.0001	0.0195	0.0194 ± 0.0001	0.0195
Q_T	1.01	100	0.1	0.1	0.1	19.55 ± 0.087	19.78	0.189 ± 0.0009	0.1978	0.0184 ± 0.0001	0.0195	0.0196 ± 0.0001	0.0195
Q_S	1.01	25	0.05	0.1	0.1	9.04 ± 0.047	9.39	0.352 ± 0.0020	0.3956	0.0331 ± 0.0002	0.03718	0.0362 ± 0.0002	0.03718
Q_T	1.01	25	0.05	0.1	0.1	9.22 ± 0.048	9.39	0.349 ± 0.00189	0.3956	0.0328 ± 0.0002	0.03718	0.0369 ± 0.0002	0.03718
Q_S	1.01	4	0.02	0.1	0.1	3.40 ± 0.019	3.64	0.795 ± 0.006	0.9104	0.0684 ± 0.0005	0.0901	0.0849 ± 0.0005	0.0901
Q_T	1.01	4	0.02	0.1	0.1	3.54 ± 0.020	3.64	0.792 ± 0.005	0.9104	0.0667 ± 0.0005	0.0901	0.0884 ± 0.0006	0.0901
Q_S	1	100	0	0.1	0.1	17.45 ± 0.082	17.84	0.172 ± 0.0009	0.1784	0.0167 ± 0.0001	0.0178	0.0174 ± 0.0001	0.0178
Q_T	1	100	0	0.1	0.1	17.62 ± 0.084	17.84	0.171 ± 0.0009	0.1784	0.0167 ± 0.0001	0.0178	0.0176 ± 0.0001	0.0178
Q_S	1	25	0	0.1	0.1	8.62 ± 0.037	8.92	0.335 ± 0.002	0.3568	0.0317 ± 0.0002	0.0356	0.0344 ± 0.0002	0.0356
Q_T	1	25	0	0.1	0.1	8.78 ± 0.038	8.92	0.333 ± 0.002	0.3568	0.0314 ± 0.0002	0.0356	0.0351 ± 0.0002	0.0356

Table 2 continued

Q_α	ρ	μ	β	θ_b	θ_r	Q	Q_{ROU}	W	W_{ROU}	R	R_{ROU}	B	B_{ROU}
Q_S	1	4	0	0.1	0.1	3.30 ± 0.016	3.56	0.773 ± 0.006	0.8920	0.0663 ± 0.0005	0.0892	0.0825 ± 0.0005	0.0892
Q_T	1	4	0	0.1	0.1	3.44 ± 0.017	3.56	0.761 ± 0.005	0.8920	0.0647 ± 0.0005	0.0892	0.0860 ± 0.0005	0.0892
Q_S	0.99	100	−0.1	0.1	0.1	15.78 ± 0.077	16.14	0.156 ± 0.0008	0.1614	0.0152 ± 0.0001	0.0163	0.0157 ± 0.0001	0.0163
Q_T	0.99	100	−0.1	0.1	0.1	15.92 ± 0.078	16.14	0.155 ± 0.0008	0.1614	0.0151 ± 0.0001	0.0163	0.0159 ± 0.0001	0.0163
Q_S	0.99	25	−0.05	0.1	0.1	8.18 ± 0.040	8.48	0.319 ± 0.002	0.3392	0.0301 ± 0.0002	0.0342	0.0327 ± 0.0002	0.0342
Q_T	0.99	25	−0.05	0.1	0.1	8.32 ± 0.042	8.48	0.317 ± 0.002	0.3392	0.0298 ± 0.0002	0.0342	0.0333 ± 0.0002	0.0342
Q_S	0.99	4	−0.02	0.1	0.1	3.24 ± 0.018	3.49	0.761 ± 0.006	0.8741	0.0655 ± 0.0005	0.0882	0.0812 ± 0.0005	0.0882
Q_T	0.99	4	−0.02	0.1	0.1	3.38 ± 0.019	3.49	0.750 ± 0.006	0.8741	0.0640 ± 0.0004	0.0882	0.0846 ± 0.0005	0.0882
Q_S	0.9	100	−1	0.1	0.1	7.09 ± 0.0252	7.77	0.0705 ± 0.0003	0.0777	0.00694 ± 0.0001	0.0086	0.0071 ± 0.0001	0.0086
Q_T	0.9	100	−1	0.1	0.1	7.13 ± 0.0256	7.77	0.0704 ± 0.0003	0.0777	0.00693 ± 0.0001	0.0086	0.0071 ± 0.0001	0.0086
Q_S	0.9	25	−0.5	0.1	0.1	5.10 ± 0.027	5.61	0.200 ± 0.001	0.2246	0.0193 ± 0.0001	0.0249	0.0205 ± 0.0001	0.0249
Q_T	0.9	25	−0.5	0.1	0.1	5.18 ± 0.028	5.61	0.200 ± 0.001	0.2246	0.0192 ± 0.0001	0.0249	0.0207 ± 0.0001	0.0249
Q_S	0.9	4	−0.2	0.1	0.1	2.61 ± 0.015	2.93	0.619 ± 0.004	0.7328	0.0541 ± 0.0004	0.0814	0.0655 ± 0.0005	0.0814
Q_T	0.9	4	−0.2	0.1	0.1	2.71 ± 0.016	2.93	0.612 ± 0.004	0.7328	0.0532 ± 0.0004	0.0814	0.0680 ± 0.0005	0.0814
Q_S	0.8	100	−2	0.1	0.1	3.72 ± 0.0098	4.59	0.0371 ± 0.0001	0.0459	0.0037 ± 0.0001	0.0057	0.0037 ± 0.0001	0.0057
Q_T	0.8	100	−2	0.1	0.1	3.74 ± 0.0099	4.59	0.0371 ± 0.0001	0.0459	0.0037 ± 0.0001	0.0057	0.0037 ± 0.0001	0.0057
Q_S	0.8	25	−1	0.1	0.1	3.19 ± 0.013	3.88	0.126 ± 0.0006	0.1555	0.0123 ± 0.0001	0.0194	0.0127 ± 0.0001	0.0194
Q_T	0.8	25	−1	0.1	0.1	3.22 ± 0.014	3.88	0.126 ± 0.0006	0.1555	0.0122 ± 0.0001	0.0194	0.0129 ± 0.0001	0.0194
Q_S	0.8	4	−0.4	0.1	0.1	2.02 ± 0.011	2.44	0.481 ± 0.0034	0.6113	0.0429 ± 0.0004	0.0764	0.0505 ± 0.0004	0.0764
Q_T	0.8	4	−0.4	0.1	0.1	2.08 ± 0.011	2.44	0.478 ± 0.0033	0.6113	0.0424 ± 0.0004	0.0764	0.0522 ± 0.0004	0.0764

Table 3 Simulated results vs. heavy traffic approximations: arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), service = Exp($\mu$), balking = Exp($\theta_b$), reneging = Exp($\theta_r$)

Q_α	ρ	μ	β	θ_b	θ_r	Q	Q_{ROU}	W	W_{ROU}	R	R_{ROU}	B	B_{ROU}
Q_S	1.1	100	1	1	1	7.96 ± 0.012	7.88	0.075 ± 0001	0.0788	0.0637 ± 0.0001	0.0717	0.075 ± 0.0001	0.0717
Q_T	1.1	100	1	1	1	8.21 ± 0.013	7.88	0.074 ± 0.0001	0.0788	0.0624 ± 0.0001	0.0717	0.077 ± 0.0001	0.0717
Q_S	1.1	25	0.5	1	1	3.38 ± 0.006	3.32	0.125 ± 0.0002	0.133	0.122 ± 0.0001	0.1209	0.091 ± 0.0002	0.1209
Q_T	1.1	25	0.5	1	1	3.56 ± 0.006	3.32	0.122 ± 0.0003	0.133	0.0088 ± 0.0002	0.1209	0.127 ± 0.0002	0.1209
Q_S	1.1	4	0.2	1	1	1.24 ± 0.003	1.20	0.277 ± 0.0009	0.301	0.124 ± 0.0004	0.2736	0.240 ± 0.0005	0.2736
Q_T	1.1	4	0.2	1	1	1.37 ± 0.003	1.20	0.268 ± 0.0009	0.301	0.117 ± 0.0004	0.2736	0.255 ± 0.0005	0.2736
Q_S	1.01	100	0.1	1	1	5.78 ± 0.009	5.82	0.055 ± 0.0001	0.0582	0.0483 ± 0.0001	0.0576	0.0552 ± 0.0001	0.0576
Q_T	1.01	100	0.1	1	1	5.94 ± 0.009	5.82	0.055 ± 0.0001	0.0582	0.0476 ± 0.0001	0.0576	0.0567 ± 0.0001	0.0576
Q_S	1.01	25	0.05	1	1	2.83 ± 0.005	2.86	0.105 ± 0.0002	0.1146	0.079 ± 0.0002	0.1135	0.103 ± 0.0002	0.1135
Q_T	1.01	25	0.05	1	1	2.97 ± 0.005	2.86	0.103 ± 0.0002	0.1146	0.077 ± 0.0002	0.1135	0.107 ± 0.0002	0.1135
Q_S	1.01	4	0.02	1	1	1.14 ± 0.003	1.13	0.255 ± 0.0009	0.2839	0.118 ± 0.0005	0.2811	0.222 ± 0.0006	0.2811
Q_T	1.01	4	0.02	1	1	1.25 ± 0.003	1.13	0.249 ± 0.0009	0.2839	0.112 ± 0.0004	0.2811	0.235 ± 0.0006	0.2811
Q_S	1	100	0	1	1	5.58 ± 0.009	5.64	0.0537 ± 0.0001	0.0564	0.047 ± 0.0001	0.0564	0.053 ± 0.0001	0.0564
Q_T	1	100	0	1	1	5.74 ± 0.009	5.64	0.0531 ± 0.0001	0.0564	0.046 ± 0.0001	0.0564	0.055 ± 0.0001	0.0564
Q_S	1	25	0	1	1	2.78 ± 0.005	2.82	0.104 ± 0.0002	0.1128	0.078 ± 0.0002	0.1128	0.102 ± 0.0002	0.1128
Q_T	1	25	0	1	1	2.92 ± 0.005	2.82	0.102 ± 0.0002	0.1128	0.076 ± 0.0002	0.1128	0.106 ± 0.0002	0.1128
Q_S	1	4	0	1	1	1.12 ± 0.003	1.12	0.253 ± 0.0010	0.2820	0.117 ± 0.0005	0.2820	0.220 ± 0.0006	0.2820
Q_T	1	4	0	1	1	1.23 ± 0.003	1.12	0.247 ± 0.0009	0.2820	0.110 ± 0.0004	0.2820	0.233 ± 0.0006	0.2820
Q_S	0.99	100	−0.1	1	1	5.39 ± 0.009	5.46	0.052 ± 0.0001	0.0551	0.045 ± 0.0001	0.0546	0.052 ± 0.0001	0.0546
Q_T	0.99	100	−0.1	1	1	5.54 ± 0.009	5.46	0.051 ± 0.0001	0.0551	0.045 ± 0.0001	0.0546	0.053 ± 0.0001	0.0546

Table 3 continued

Q_α	ρ	μ	β	θ_b	θ_r	Q	Q_{ROU}	W	W_{ROU}	R	R_{ROU}	B	B_{ROU}
Q_S	0.99	25	−0.05	1	1	2.73 ± 0.005	2.77	0.102 ± 0.0002	0.1110	0.077 ± 0.0002	0.1121	0.100 ± 0.0002	0.1121
Q_T	0.99	25	−0.05	1	1	2.86 ± 0.006	2.77	0.100 ± 0.0002	0.1110	0.075 ± 0.0002	0.1121	0.104 ± 0.0002	0.1121
Q_S	0.99	4	−0.02	1	1	1.11 ± 0.003	1.12	0.251 ± 0.0009	0.2802	0.117 ± 0.0005	0.2831	0.219 ± 0.0006	0.2831
Q_T	0.99	4	−0.02	1	1	1.22 ± 0.003	1.12	0.245 ± 0.0009	0.2802	0.110 ± 0.0005	0.2831	0.232 ± 0.0006	0.2831
Q_S	0.9	100	−1	1	1	3.88 ± 0.005	4.16	0.0377 ± 0.0001	0.0416	0.034 ± 0.0001	0.0462	0.037 ± 0.0001	0.0462
Q_T	0.9	100	−1	1	1	3.97 ± 0.006	4.16	0.0375 ± 0.0001	0.0416	0.034 ± 0.0001	0.0462	0.038 ± 0.0001	0.0462
Q_S	0.9	25	−0.5	1	1	2.25 ± 0.005	2.41	0.085 ± 0.0002	0.0964	0.066 ± 0.0002	0.1071	0.083 ± 0.0002	0.1071
Q_T	0.9	25	−0.5	1	1	2.36 ± 0.005	2.41	0.084 ± 0.0002	0.0964	0.065 ± 0.0002	0.1071	0.087 ± 0.0002	0.1071
Q_S	0.9	4	−0.2	1	1	1.01 ± 0.003	1.05	0.229 ± 0.0009	0.2646	0.109 ± 0.0005	0.2940	0.200 ± 0.0006	0.2940
Q_T	0.9	4	−0.2	1	1	1.10 ± 0.003	1.05	0.224 ± 0.0009	0.2646	0.103 ± 0.0004	0.2940	0.211 ± 0.0007	0.2940
Q_S	0.8	100	−2	1	1	2.70 ± 0.005	3.19	0.026 ± 0.0001	0.0319	0.024 ± 0.0001	0.0399	0.026 ± 0.0001	0.0399
Q_T	0.8	100	−2	1	1	2.75 ± 0.005	3.19	0.026 ± 0.0001	0.0319	0.024 ± 0.0001	0.0399	0.027 ± 0.0001	0.0399
Q_S	0.8	25	−1	1	1	1.80 ± 0.003	2.08	0.069 ± 0.0002	0.0832	0.055 ± 0.0002	0.1040	0.067 ± 0.0002	0.1040
Q_T	0.8	25	−1	1	1	1.87 ± 0.003	2.08	0.068 ± 0.0002	0.0832	0.054 ± 0.0002	0.1040	0.069 ± 0.0002	0.1040
Q_S	0.8	4	−0.4	1	1	0.89 ± 0.002	0.9947	0.204 ± 0.0008	0.2486	0.100 ± 0.0005	0.3108	0.178 ± 0.0007	0.3108
Q_T	0.8	4	−0.4	1	1	0.97 ± 0.002	0.9947	0.200 ± 0.0008	0.2486	0.095 ± 0.0004	0.3108	0.188 ± 0.0007	0.3108

Table 4 Simulated results vs. heavy traffic approximations: arrivals = Exp($\mu + \sqrt{\mu} \cdot \beta$), service = LogNormal($1/\mu, 1/\mu^2$), balking = Uniform($0, 1/\theta_b$), reneging = Uniform($0, 1/\theta_r$)

Q_α	ρ	μ	β	θ_b	θ_r	Q	Q_{ROU}	W	W_{ROU}	R	R_{ROU}	B	B_{ROU}
Q_S	1.1	100	1	1	1	7.56 ± 0.011	7.88	0.072 ± 0001	0.0788	0.0636 ± 0.0001	0.0717	0.076 ± 0.0001	0.0717
Q_T	1.1	100	1	1	1	7.79 ± 0.012	7.88	0.070 ± 0.0001	0.0788	0.0622 ± 0.0001	0.0717	0.078 ± 0.0001	0.0717
Q_S	1.1	25	0.5	1	1	3.11 ± 0.005	3.32	0.114 ± 0.0003	0.133	0.091 ± 0.0001	0.1209	0.124 ± 0.0002	0.1209
Q_T	1.1	25	0.5	1	1	3.27 ± 0.006	3.32	0.111 ± 0.0002	0.133	0.088 ± 0.0002	0.1209	0.131 ± 0.0002	0.1209
Q_S	1.1	4	0.2	1	1	1.06 ± 0.002	1.20	0.232 ± 0.001	0.301	0.118 ± 0.0005	0.2736	0.266 ± 0.0006	0.2736
Q_T	1.1	4	0.2	1	1	1.15 ± 0.002	1.20	0.225 ± 0.001	0.301	0.105 ± 0.0004	0.2736	0.289 ± 0.0007	0.2736
Q_S	1.01	100	0.1	1	1	5.57 ± 0.009	5.82	0.052 ± 0.0001	0.0582	0.0470 ± 0.0001	0.0576	0.0557 ± 0.0001	0.0576
Q_T	1.01	100	0.1	1	1	5.73 ± 0.009	5.82	0.051 ± 0.0001	0.0582	0.0463 ± 0.0001	0.0576	0.0573 ± 0.0001	0.0576
Q_S	1.01	25	0.05	1	1	2.65 ± 0.004	2.86	0.098 ± 0.0002	0.1146	0.080 ± 0.0002	0.1135	0.106 ± 0.0002	0.1135
Q_T	1.01	25	0.05	1	1	2.78 ± 0.005	2.86	0.096 ± 0.0002	0.1146	0.078 ± 0.0002	0.1135	0.111 ± 0.0002	0.1135
Q_S	1.01	4	0.02	1	1	0.99 ± 0.002	1.13	0.218 ± 0.001	0.2839	0.113 ± 0.0005	0.2811	0.247 ± 0.0006	0.2811
Q_T	1.01	4	0.02	1	1	1.07 ± 0.002	1.13	0.211 ± 0.001	0.2839	0.101 ± 0.0004	0.2811	0.268 ± 0.0007	0.2811
Q_S	1	100	0	1	1	5.38 ± 0.008	5.64	0.0516 ± 0.0001	0.0564	0.047 ± 0.0001	0.0564	0.054 ± 0.0001	0.0564
Q_T	1	100	0	1	1	5.53 ± 0.009	5.64	0.0510 ± 0.0001	0.0564	0.046 ± 0.0001	0.0564	0.055 ± 0.0001	0.0564
Q_S	1	25	0	1	1	2.59 ± 0.005	2.82	0.095 ± 0.0003	0.1128	0.079 ± 0.0002	0.1128	0.104 ± 0.0002	0.1128
Q_T	1	25	0	1	1	2.72 ± 0.005	2.82	0.094 ± 0.0002	0.1128	0.077 ± 0.0002	0.1128	0.109 ± 0.0002	0.1128
Q_S	1	4	0	1	1	0.98 ± 0.002	1.12	0.217 ± 0.001	0.2820	0.113 ± 0.0005	0.2820	0.245 ± 0.0006	0.2820
Q_T	1	4	0	1	1	1.06 ± 0.002	1.12	0.210 ± 0.001	0.2820	0.100 ± 0.0004	0.2820	0.265 ± 0.0006	0.2820

Table 4 continued

Q_α	ρ	μ	β	θ_b	θ_r	Q	Q_{ROU}	W	W_{ROU}	R	R_{ROU}	B	B_{ROU}
Q_S	0.99	100	−0.1	1	1	5.18 ± 0.008	5.46	0.049 ± 0.0001	0.0551	0.045 ± 0.0001	0.0546	0.052 ± 0.0001	0.0546
Q_T	0.99	100	−0.1	1	1	5.33 ± 0.008	5.46	0.049 ± 0.0001	0.0551	0.045 ± 0.0001	0.0546	0.053 ± 0.0001	0.0546
Q_S	0.99	25	−0.05	1	1	2.54 ± 0.005	2.77	0.094 ± 0.0002	0.1110	0.078 ± 0.0002	0.1121	0.102 ± 0.0002	0.1121
Q_T	0.99	25	−0.05	1	1	2.67 ± 0.005	2.77	0.093 ± 0.0002	0.1110	0.075 ± 0.0002	0.1121	0.108 ± 0.0002	0.1121
Q_S	0.99	4	−0.02	1	1	0.97 ± 0.002	1.12	0.214 ± 0.0009	0.2802	0.112 ± 0.0005	0.2831	0.242 ± 0.0006	0.2831
Q_T	0.99	4	−0.02	1	1	1.05 ± 0.002	1.12	0.208 ± 0.0009	0.2802	0.101 ± 0.0005	0.2831	0.263 ± 0.0006	0.2831
Q_S	0.9	100	−1	1	1	3.77 ± 0.007	4.16	0.0364 ± 0.0001	0.0416	0.034 ± 0.0001	0.0462	0.038 ± 0.0001	0.0462
Q_T	0.9	100	−1	1	1	3.86 ± 0.007	4.16	0.0346 ± 0.0001	0.0416	0.034 ± 0.0001	0.0462	0.039 ± 0.0001	0.0462
Q_S	0.9	25	−0.5	1	1	2.13 ± 0.004	2.41	0.080 ± 0.0002	0.0964	0.068 ± 0.0002	0.1071	0.085 ± 0.0002	0.1071
Q_T	0.9	25	−0.5	1	1	2.23 ± 0.004	2.41	0.078 ± 0.0002	0.0964	0.065 ± 0.0002	0.1071	0.089 ± 0.0002	0.1071
Q_S	0.9	4	−0.2	1	1	0.89 ± 0.002	1.05	0.198 ± 0.0009	0.2646	0.106 ± 0.0005	0.2940	0.223 ± 0.0006	0.2940
Q_T	0.9	4	−0.2	1	1	0.96 ± 0.002	1.05	0.192 ± 0.0009	0.2646	0.096 ± 0.0004	0.2940	0.241 ± 0.0007	0.2940
Q_S	0.8	100	−2	1	1	2.65 ± 0.004	3.19	0.026 ± 0.0001	0.0319	0.024 ± 0.0001	0.0399	0.026 ± 0.0001	0.0399
Q_T	0.8	100	−2	1	1	2.70 ± 0.004	3.19	0.026 ± 0.0001	0.0319	0.024 ± 0.0001	0.0399	0.027 ± 0.0001	0.0399
Q_S	0.8	25	−1	1	1	1.72 ± 0.003	2.08	0.065 ± 0.0002	0.0832	0.056 ± 0.0002	0.1040	0.069 ± 0.0002	0.1040
Q_T	0.8	25	−1	1	1	1.79 ± 0.003	2.08	0.064 ± 0.0002	0.0832	0.055 ± 0.0002	0.1040	0.072 ± 0.0002	0.1040
Q_S	0.8	4	−0.4	1	1	0.79 ± 0.002	0.9947	0.180 ± 0.0008	0.2486	0.090 ± 0.0005	0.3108	0.199 ± 0.0007	0.3108
Q_T	0.8	4	−0.4	1	1	0.86 ± 0.002	0.9947	0.176 ± 0.0008	0.2486	0.099 ± 0.0004	0.3108	0.215 ± 0.0007	0.3108

References

1. Allon, G., Bassamboo, A.: The impact of delaying the delay announcements. Oper. Res. **59**(5), 1198–1210 (2011)
2. Armony, M., Shimkin, N., Whitt, W.: The impact of delay announcements in many-server queues with abandonment. Oper. Res. **57**(1), 66–81 (2009)
3. Baccelli, F., Hebuterne, G.: On queues with impatient customers. In: Kylstra, F.J. (ed.) Performance 81. North-Holland, Amsterdam (1981)
4. Bass, R.F.: Probabilistic Techniques in Analysis. Springer, New York (1995)
5. Billingsley, P.: Convergence of Probability Measures, vol. 493. Wiley, New York (2009)
6. Dai, J., Dai, W.: A heavy traffic limit theorem for a class of open queueing networks with finite buffers. Queueing Syst. **32**(1–3), 5–40 (1999)
7. Gromoll, H.C., Robert, P., Zwart, B.: Fluid limits for processor-sharing queues with impatience. Math. Oper. Res. **33**(2), 375–402 (2008)
8. Guo, P., Zipkin, P.: Analysis and comparison of queues with different levels of delay information. Manag. Sci. **53**(6), 962–970 (2007)
9. Ibrahim, R., Whitt, W.: Real-time delay estimation based on delay history. Manuf. Serv. Oper. Manag. **11**(3), 397–415 (2009)
10. Jennings, O.B.: Averaging principles for a diffusion-scaled, heavy-traffic polling station with k job classes. Math. Oper. Res. **35**(3), 669–703 (2010)
11. Jennings, O.B., Puha, A.L.: Fluid limits for overloaded multiclass FIFO single-server queues with general abandonment. Stoch. Syst. **3**(1), 262–321 (2013)
12. Jennings, O.B., Reed, J.E.: An overloaded multiclass FIFO queue with abandonments. Oper. Res. **60**(5), 1282–1295 (2012)
13. Liu, Y., Whitt, W.: The Gt/GI/St+GI many-server fluid queue. Queueing Syst. **71**(4), 405–444 (2012)
14. Massey, W.A., Pender, J.: Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. Queueing Syst. **75**(2–4), 243–277 (2013)
15. Massey, W., Pender, J.: Approximating and stabilizing Jackson networks with abandonment. Technical Report. Working Paper (2014)
16. Pender, J.: Gram Charlier expansion for time varying multiserver queues with abandonment. SIAM J. Appl. Math. **74**(4), 1238–1265 (2014)
17. Pender, J.: The truncated normal distribution: applications to queues with impatient customers. Oper. Res. Lett. **43**(1), 40–45 (2015)
18. Reed, J., Ward, A.R.: Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic. Math. Oper. Res. **33**(3), 606–644 (2008)
19. Ward, A.R., Glynn, P.W.: A diffusion approximation for a GI/GI/1 queue with balking or reneging. Queueing Syst. **50**(4), 371–400 (2005)
20. Whitt, W.: Improving service by informing customers about anticipated delays. Manag. Sci. **45**(2), 192–207 (1999)
21. Whitt, W.: Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Manag. Sci. **50**(10), 1449–1461 (2004)
22. Xu, S.H., Gao, L., Ou, J.: Service performance analysis and improvement for a ticket queue with balking customers. Manag. Sci. **53**(6), 971–990 (2007)
23. Zeltyn, S., Mandelbaum, A.: Call centers with impatient customers: many-server asymptotics of the M/M/n + G queue. Queueing Syst. **51**(3–4), 361–402 (2005)