Queueing Systems 50, 371–400, 2005

A Diffusion Approximation for a GI/GI/1 Queue with Balking or Reneging

AMY R. WARD amy@isye.gatech.edu
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

PETER W. GLYNN glynn@stanford.edu
Department of Management Science & Engineering, Stanford University, Stanford, CA 94305, USA

Received 31 October 2003; Revised 7 January 2005

Abstract. Consider a single-server queue with a renewal arrival process and generally distributed processing times in which each customer independently reneges if service has not begun within a generally distributed amount of time. We establish that both the workload and queue-length processes in this system can be approximated by a regulated Ornstein-Uhlenbeck (ROU) process when the arrival rate is close to the processing rate and reneging times are large. We further show that a ROU process also approximates the queue-length process, under the same parameter assumptions, in a balking model. Our balking model assumes the queue-length is observable to arriving customers, and that each customer balks if his or her conditional expected waiting time is too large.

Keywords: deadlines, reneging, balking, impatience, GI/GI/1-GI queue, Ornstein-Uhlenbeck process, regulated diffusion, reflected diffusion

1. Introduction

Queueing models in which customers may renege, or abandon the system before receiving service, arise in many diverse applications domains. For example, in the multi-billion dollar call center industry, impatient customers often hang-up before reaching a service agent. Furthermore, in the manufacturing context, customer order cancellation was one factor in the $2.2 billion component inventory write-off of Cisco Systems in the third quarter of 2001 [2]. Finally, in the networking context, certain packets of data (for example, location data in wireless networks) lose their value unless transmitted or received within a given time interval.

In the case of a call center, it is tempting to assume a purely exponential model. However, the work of Brown et al. [5] that analyzes a data set consisting of over 1,200,000 calls to a bank call center concludes that service times follow a lognormal distribution and reneging times are not exponential. Hence it is desirable to obtain performance

measure approximations for queueing models with reneging in the most general framework possible.

The main contribution of this paper is to show that a regulated Ornstein-Uhlenbeck (ROU) process plays a similar role in the context of queues with reneging or balking as does regulated Brownian motion (RBM) in the setting of conventional queues. Specifically, we show that a ROU arises as a heavy traffic diffusion limit for the queue-length and workload processes when the customer arrival rate is close to the service rate and reneging times are large; see Theorems 1, 2, 3, and 4. This extends the work of Ward and Glynn [23], where heavy-traffic limits for a single-server queue with reneging were developed in a purely exponential setting.

We develop our limit theorems via a result for the *offered waiting time process*. The offered waiting time process consists of the workload of all customers in the queue that will eventually reach the server. This process differs from the *observed workload process*, which consists of the workload of **all** customers in the queue—regardless of whether or not they eventually renege. We show that the behavior of both of these processes is identical in the heavy-traffic diffusion limit; see Theorems 1 and 2.

Our limit theorems show the behavior of the derivative of the deadline distribution function at the point zero is important. In particular, it is this value that controls the strength of the linear state-dependence in the limiting ROU process.

The remainder of this paper is organized as follows. We conclude this section with a review of relevant literature. In Section 2, we formulate our balking and reneging models, and provide stability conditions. In Section 3, we establish the weak convergence of the diffusion scaled offered waiting time and workload processes. We then show the weak convergence of the diffusion scaled queue-length processes in both our reneging and balking models in Section 4. Finally, in Section 5, we provide a simulation study showing the mean of an appropriate ROU process provides a good approximation to the steady-state mean queue-length in a reneging or balking model.

1.1. Literature review

Palm [18] introduced reneging as a means of modeling the behavior of telephone switchboard customers more than 60 years ago. Later, for a purely exponential model, Ancker and Gafarian [3] explicitly computed stationary probabilities for the number of customers in the system. Such explicit expressions are not possible for GI/GI/1 models with generally distributed reneging times. It *is* true that expressions for many steady-state performance measures of interest can be written in terms of the limiting distribution of the work seen by an arbitrary arrival to the system (assuming the limit distribution exists); see Stanford [22] and Baccelli, Boyer, and Hebuterne [4]. However, this limit distribution is given in the form of an integral equation, which may be difficult to solve.

The situation for obtaining explicit expressions for transient performance measures is hopeless. Even for purely exponential models, computations for transient performance measures generally involve numerical procedures to invert transforms, as shown in Whitt [25]. Hence it is desirable to develop diffusion approximations (with

analytically tractable steady-state **and** transient behavior) for the queue-length and workload processes in systems with reneging. Ward and Glynn [23] show a ROU process well approximates an M/M/1 queue with exponential reneging when the arrival rate is close to the processing rate and the reneging rate is small. For a queue with many servers, Poisson arrivals, and exponential processing and reneging, Garnett, Mandelbaum, and Reiman [12] develop a different diffusion approximation as the number of servers grows large in the Halfin-Whitt regime. For a many server queue with a general reneging distribution having Poisson arrivals and exponential processing times, Mandelbaum and Zeltyn [17] analyze the robustness of a linear relationship between customer abandonment probability and average waiting time observed in telephone call center data.

For models with general distributions, Coffman, Puhalskii, Reiman, and Wright [9] conjecture the appropriate diffusion limit for a system with reneging operating under a processor-sharing discipline when the arrival rate is large. Doytchinov, Lehoczky, and Shreve [10] establish the heavy traffic limit for a GI/GI/1 system with generally distributed customer reneging times operating under the earliest-deadline-first queue discipline. In the case of constant deadlines (but a renewal arrival process and general service time distribution), Plambeck, Kumar, and Harrison [19], devise input and scheduling controls that are asymptotically optimal in the heavy-traffic limit regime.

2. Balking and reneging models

Our goal in this paper is to develop performance measure approximations for a class of GI/GI/1 queueing models with either reneging or balking. For all our models, we assume the server works at rate 1 under the FIFO discipline. The model primitives are three independent sequences of non-negative, mean 1, i.i.d. random variables $\{u_i : i \geq 1\}$, $\{v_i : i \geq 1\}$ and $\{w_i : i \geq 1\}$, which are all defined on a common probability space $(\Omega, \mathcal{F}, P)$. For a given inter-arrival rate ρ and mean deadline time m, the inter-arrival time between the $(i-1)$st and ith customer is $\rho^{-1}u_i$, the service time of the ith customer is v_i, and the maximum time the customer will wait for processing is $d_i \equiv mw_i$. For $t > 0$, the renewal process

$$A(t) = \max\{i \geq 0 : u_1 + \cdots + u_i \leq \rho t\}$$

represents the cumulative number of arrivals in the first t time units, while the renewal process

$$S(t) = \max\{i \geq 0 : v_1 + \cdots + v_i \leq t\}$$

represents the number of customers processed in the first t units of server busy time.

2.1. *Evolution equations*

In our reneging model, all customers initially join the queue. However, if the ith customer, who arrives at time

$$t_i \equiv \sum_{j=1}^{i} \rho^{-1} u_j,$$

must wait longer than d_i time units to reach the server, he departs at time $t_i + d_i$ without receiving service. As observed by Baccelli, Boyer, and Hebuterne [4], for such a system, the time a served customer waits for service depends only on the processing times of the "non-reneging" customers in the queue at his arrival time. In other words, customers that eventually renege from the queue do not contribute to the delay served customers experience. We track the waiting time an arriving customer would experience at time $t > 0$ using the *offered waiting time process*

$$V(t) = \sum_{i=1}^{A(t)} v_i \mathbf{1}\{V(t_i^-) < d_i\} - B(t), \tag{2.1}$$

where $B(t) \equiv \int_0^t \mathbf{1}\{V(s) > 0\} ds$ is the cumulative busy time.

The offered waiting time process differs from the process that measures the total processing time of all the customers in the system at each point in time. We call this process the *observed workload process* and define it to be

$$W(t) = V(t) + \sum_{i=1}^{A(t)} v_i \mathbf{1}\{V(t_i^-) \geq d_i \text{ and } t_i \leq t < t_i + d_i\}. \tag{2.2}$$

We also desire to track the number of customers in the system over time, including both reneging and non-reneging customers. By looking at the offered waiting time process at each customer arrival time, we can count the cumulative number of arrivals that eventually receive service, $\sum_{i=1}^{A(t)} \mathbf{1}\{V(t_i^-) < d_i\}$, and the number of customers currently in queue that eventually renege, $\sum_{i=1}^{A(t)} \mathbf{1}\{V(t_i^-) < d_i \text{ and } t_i \leq t < t_i + d_i\}$. Counting the cumulative number of departing customers requires more work because processing times are associated with customers, and so we must first determine the indices of served jobs. Let $\pi(i)$ denote the index of the ith served job for $i \geq 1$. The system starts empty so that $\pi(1) = 1$. For $i \geq 2$ we define $\pi(i)$ recursively as

$$\pi(i) = \min\{n > \pi(i-1) : V(t_n^-) < d_n\}.$$

Then, the cumulative number of departures that have received service before time t is

$$S(t) = \max\left\{ i \geq 0 : \sum_{j=1}^{i} v_{\pi(j)} \leq t \right\}.$$

Finally, the queue-length at time t is

$$Q(t) = \sum_{i=1}^{A(t)} \mathbf{1}\{V(t_i^-) < d_i\} + \sum_{i=1}^{A(t)} \mathbf{1}\{V(t_i^-) \geq d_i \text{ and } t_i \leq t < t_i + d_i\} - S(B(t)). \tag{2.3}$$

The offered waiting time process also arises as the workload process for a system in which each customer balks if at his arrival time the offered waiting time exceeds his deadline. Such a balking assumption is realistic for models in which the processing time of each customer in the queue is known. However, customer processing times often are not observable, meaning the above balking model does not apply. In the case that queue lengths are observable (but customer processing times are not), one reasonable balking model formulation assumes each customer joins the queue if the customer's conditional expected waiting time at his arrival time exceeds his deadline. The queue-length process in this system is

$$Q_B(t) = \sum_{i=1}^{A(t)} \mathbf{1}\{Q_B(t_i^-) - 1 < d_i\} - S(B_B(t)), \tag{2.4}$$

where $B_B(t) = \int_0^t \mathbf{1}\{Q_B(s) > 0\}ds$ is the cumulative busy time in the balking system.

It is useful for our analysis to represent the offered waiting time process in terms of a stochastic integral and the following two martingales:

$$\left\{ \left(M_v(i) \equiv \sum_{j=1}^{i} (v_j - 1)\mathbf{1}\{V(t_j^-) < d_j\}, \mathcal{F}_i \right), i \geq 0 \right\}$$

and

$$\left\{ \left(M_d(i) \equiv \sum_{j=1}^{i} \mathbf{1}\{V(t_j^-) \geq d_j\} - E[\mathbf{1}\{V(t_j^-) \geq d_j\} \mid \mathcal{F}_{j-1}], \mathcal{F}_i \right), \ i \geq 0 \right\},$$

where

$$\mathcal{F}_i = \sigma((u_1, v_1, w_1), \ldots (u_i, v_i, w_i), u_{i+1}) \subset \mathcal{F}.$$

Our definition of $\mathcal{F}_i$ implies

$$P(V(t_i^-) \geq d_i \mid \mathcal{F}_{i-1}) = F\left(\frac{V(t_i^-)}{m}\right), \quad i = 1, 2, \ldots,$$

almost surely, where F is the cdf of w_1, because $V(t_i^-)$ is $\mathcal{F}_{i-1}$-measurable and w_i is independent of $\mathcal{F}_{i-1}$. A little algebra shows

$$V(t) + \sum_{i=1}^{A(t)} F\left(\frac{V(t_i^-)}{m}\right) = A(t) + M_v(A(t)) - M_d(A(t)) - t + I(t),$$

where $I(t) = t - B(t)$ is the cumulative idle time in the reneging system. Finally, we write the evolution equation for V in terms of a stochastic integral

$$V(t) + \int_0^t F\left(\frac{V(s^-)}{m}\right) dA(s) = A(t) + M_v(A(t)) - M_d(A(t)) - t + I(t). \quad (2.5)$$

2.2. *Stability*

A sufficient condition for the existence of a non-degenerate limiting distribution for the offered waiting time process, V, observed workload process, W, and queue-length process in our balking system, Q_B is:

$$\rho P(d_1 = \infty) < 1. \quad (2.6)$$

The result for V follows from Lemma 2 in Baccelli, Boyer, and Hebuterne [4]; the result for Q_B follows from Theorem 4.1 in Lillo and Martin [16]. Our next proposition establishes (2.6) is the appropriate sufficient stability condition for W.

Proposition 1. Let τ be the length of the initial busy cycle (the initial idle period plus the first busy period). If condition (2.6) holds, then $E[\tau] < \infty$ and

$$\lim_{t\to\infty} P(W(t) \le x) = \frac{E\int_0^\tau \mathbf{1}\{W(t) \le x\}\, dt}{E\tau}. \quad (2.7)$$

Proof. Let $\{T_n^V : n \ge 1\}$ and $\{T_n^W : n \ge 1\}$ be the times at which busy cycles begin for the processes V and W respectively, so that the sequences are of regeneration time points. In order to establish (2.7), by Smith's renewal theorem (see, for example, Theorem 3.12.1 in Resnick) it is enough to show $ET_1^W < \infty$. If no jobs renege in $[0, T_1^V]$, then clearly $T_1^V = T_1^W$. So, assume otherwise and let t_{n_R} be the arrival time of the last reneging job, indexed by n_R, during $[0, T_1^V]$. Since this job does not reach the server, $d_{n_R} \le V(t_{n_R}^-)$. Since the system empties linearly at rate 1, $V(t_{n_R}^-) \le (T_1^V - t_{n_R})$. Since job n_R departs at time $t_{n_R} + d_{n_R} \le T_1^V$,

$$W(t) = V(t) \text{ a.s.} \quad \text{for } t_{n_R} + d_{n_R} \le t \le T_1^V,$$

which guarantees $T_1^V = T_1^W$ a.s. The consequences of Lemma 2 in Baccelli, Boyer, and Heburterne [4] detailed in Section 3.2 guarantee $ET_1^V < \infty$, and so $ET_1^W < \infty$ also. □

2.3. *Heavy traffic conditions*

Because obtaining closed-form expressions for performance measures in our balking and reneging models appears impossible, we perform an asymptotic analysis in heavy traffic. We consider a sequence of systems, indexed by n, in which the arrival rate is ρ^n and mean deadline lengths are m^n. Any process or quantity associated with the nth system we superscript by n. Then, the random variable d_i^n that represents the maximum time a customer will wait for processing in the nth system has cdf

$$F^n(x) \equiv P\left(d_i^n \le x\right) = P\left(w_i \le \frac{x}{m^n}\right) = F\left(\frac{x}{m^n}\right).$$

Our analysis requires the following assumptions.

Assumption 1 (Heavy Traffic Requirements)**.**

(a) As $n \to \infty$, $\sqrt{n}(\rho^n - 1) \to c$, where c is a finite constant.

(b) The variances $\mathrm{var}(u_1)$ and $\mathrm{var}(v_1)$ are finite.

(c) The cdf F used to determine a particular customer's deadline is differentiable about 0, and $F_d'(0) < \infty$.

We also state the following technicalities. All random variables are defined on a common probability space $(\Omega, \mathcal{F}, P)$. Let $D([0, \infty), \Re)$ be the space of right continuous functions with left limits (RCLL) in $\Re$ having time domain $[0, \infty)$. We endow $D([0, \infty), \Re)$ with the usual Skorokhod J_1 topology, and let M denote the Borel σ-algebra associated with the J_1 topology. All stochastic processes are measurable functions from $(\Omega, \mathcal{F}, P)$ into $(D([0, \infty), \Re), M)$. Suppose $\{\xi^n\}_{n=1}^{\infty}$ is a sequence of stochastic processes. The notation $\xi^n \Rightarrow \xi$ means that the probability measures induced by the ξ^n's on $(D([0, \infty), \Re), M)$ converge weakly to the probability measure on $(D([0, \infty), \Re), M)$ induced by the stochastic process ξ. For $x \in (D([0, \infty), \Re), M)$ and $t > 0$, let

$$\|x\|_t \equiv \sup_{0 \le s \le t} |x(s)|$$

be the uniform norm and note that ξ^n converges almost surely (a.s.) to a continuous limit process ξ in the J_1 topology if and only if the convergence is uniform on compact sets (u.o.c.)

$$\|\xi^n - \xi\|_t \to 0 \text{ a.s.}$$

for every $t > 0$.

Finally, it is convenient to define the following diffusion and fluid-scaled quantities

$$\tilde{A}^n(t) \equiv \sqrt{n}\left(\frac{A^n(nt)}{n} - \rho^n t\right) \quad \bar{A}^n(t) \equiv \frac{A^n(nt)}{n}$$
$$\tilde{S}^n(t) \equiv \sqrt{n}\left(\frac{S^n(nt)}{n} - t\right) \quad \bar{S}^n(t) \equiv \frac{S^n(nt)}{n}.$$

Let B_1 and B_2 be independent, standard Brownian motions, and let $e(t) \equiv t$ for $t > 0$ be the identity function. Our analysis relies on the weak convergences guaranteed by the functional central limit theorem (FCLT; see, for example Theorem 5.11 in Chen and Yao [8])

$$\tilde{A}^n \Rightarrow \sqrt{\text{var}(u_1)}B_1 \quad \text{and} \quad \tilde{S}^n \Rightarrow \sqrt{\text{var}(v_1)}B_2,$$

as $n \to \infty$, and the almost sure convergences guaranteed by the functional strong law of large numbers (FSLLN; see, for example Theorem 5.10 in Chen and Yao [8])

$$\bar{A}^n \to e \text{ and } \bar{S}^n \to e, \text{ a.s., u.o.c.},$$

as $n \to \infty$. We additionally use the following diffusion-scaled quantities

$$\tilde{V}^n(t) \equiv \frac{1}{\sqrt{n}} V^n(nt)$$
$$\tilde{Q}^n(t) \equiv \frac{1}{\sqrt{n}} Q^n(nt)$$
$$\tilde{Q}^n_B(t) \equiv \frac{1}{\sqrt{n}} Q^n_B(nt)$$
$$\tilde{M}^n_v(t) \equiv \frac{1}{\sqrt{n}} M^n_v(\lfloor nt \rfloor)$$
$$\tilde{M}^n_d(t) \equiv \frac{1}{\sqrt{n}} M^n_d \lfloor nt \rfloor)$$
$$\tilde{I}^n(t) \equiv \frac{1}{\sqrt{n}} I^n(nt)$$
$$\tilde{I}^n_B(t) \equiv \frac{1}{\sqrt{n}} I^n_B(nt).$$

3. Asymptotic behavior of the offered waiting time and observed workload processes

We show that mean deadlines should be of order n in order that the limiting diffusion approximations capture both the effects of limited service capacity and customer reneging. In particular, if deadlines are larger than order n, our heavy traffic limits are identical

to those for the waiting time process in a conventional GI/GI/1 queue (without reneging). On the other hand, when deadlines are shorter than order n, customers renege often enough that offered waiting times are lower than order $\sqrt{n}$, meaning the diffusion-scaled offered waiting time process weakly converges to the zero process. Finally, we show that the asymptotic behavior of the offered waiting time and observed workload processes is identical under diffusion scaling.

In preparation for our results, let $Z = (Z(t) : t \geq 0)$ be the solution to the stochastic differential equation (SDE)

$$dZ(t) = (\alpha - \gamma Z(t))dt + \sigma dB(t) + dL(t), \tag{3.1}$$

subject to $Z(0) \geq 0$, where $L = (L(t) : t \geq 0)$ is the minimal non-decreasing process which makes $Z(t) \geq 0$ for $t \geq 0$. The process L increases only when Z is 0, so that

$$\int_{[0,\infty)} \mathbf{1}(Z(t) > 0)dL(t) = 0. \tag{3.2}$$

The existence of a unique strong solution to the SDE (3.1) is guaranteed by Proposition 2 in Section A.1 of the appendix, because

$$(Z, L) = (\Phi_\gamma, \Psi_\gamma)(\alpha e + \sigma B(t)),$$

where $(\Phi_\gamma, \Psi_\gamma)$ is the linearly generalized regulator mapping (having $\Gamma = \gamma$ and $M = 1$) given in the appendix. We refer to Z as a regulated Ornstein-Uhlenbeck (ROU) process with infinitesimal drift $\alpha - \gamma z$ and infinitesimal variance σ^2. Such a process has analytically tractable steady-state and transient behavior; see Ward and Glynn [24].

When $\gamma = 0$, the SDE in (3.1) yields the familiar regulated Brownian motion (RBM) process with infinitesimal drift α and infinitesimal variance σ^2. The process (Z, L) is now represented using the conventional regulator mapping so that

$$(Z, L) = (\Phi, \Psi)(\alpha e + \sigma B(t)),$$

where $(\Phi, \Psi) = (\Phi_0, \Psi_0)$. For the steady-state and transient behavior of RBM, see Harrison [14].

Theorem 1 (Weak convergence of the offered waiting time process)**.**

(a) Suppose that $m_n = n$. If $\tilde{V}^n(0) \Rightarrow Z, (0)$ as $n \to \infty$, then

$$(\tilde{V}^n, \tilde{I}^n) \Rightarrow (Z, L),$$

as $n \to \infty$, where Z is a ROU with initial position $Z(0)$, infinitesimal drift $(c - F_d'(0)z)$, infinitesimal variance $\text{var}(u_1) + \text{var}(v_1)$, and L satisfies (3.2).

(b) Suppose that $m_n = n^{1+\epsilon}$ for epsilon > 0. If $\tilde{V}^n(0) \Rightarrow Z^R(0)$ as $n \to \infty$, then

$$(\tilde{V}^n, \tilde{I}^n) \Rightarrow (Z^R, L^R),$$

as $n \to \infty$, where Z^R is a RBM with infinitesimal drift c, infinitesimal variance $\text{var}(u_1) + \text{var}(v_1)$, and L^R satisfies (3.2) with Z^R substituted for Z.

(c) Suppose that $m_n = n^{1-\epsilon}$ for $1 > \epsilon > 1/2$. If $\tilde{V}^n(0) \Rightarrow 0$ as $n \to \infty$, then

$$\tilde{V}^n \Rightarrow 0,$$

as $n \to \infty$.

The proof of Theorem 1 uses the following two Lemmas, whose proofs can be found in Section A.2 of the appendix.

Lemma 1. For any $t > 0$, given $\epsilon > 0$, there exists K such that for all n

$$P\left(\max_{j=1,\ldots,\lfloor nt \rfloor} \frac{1}{\sqrt{n}} V^n\left(t_j^{n,-}\right) > K \right) < \epsilon.$$

Lemma 2. Suppose $m_n = n^p$ for $p > 1/2$. Then

$$\tilde{M}_d^n \Rightarrow 0$$

as $n \to \infty$.

Proof of Theorem 1.

Part (a): From the pathwise equation for V in (2.5) and our assumption that $m_n = n$,

$$\tilde{V}^n(t) + F'(0) \int_0^t \tilde{V}^n(s) ds = \tilde{X}^n(t) + \tilde{I}^n(t),$$

where

$$\begin{aligned} \tilde{X}^n(t) \equiv\ & \tilde{A}^n(t) + \tilde{M}_v^n(\bar{A}^n(t)) + \sqrt{n}(\rho^n - 1)t \\ & - \tilde{M}_d^n(\bar{A}^n(t)) + F'(0) \int_0^t \tilde{V}^n(s) ds - \int_0^t \sqrt{n} F(n^{-1/2} \tilde{V}^n(s)) d\bar{A}^n(s). \end{aligned}$$

Since $\tilde{I}^n(0) = 0$, $\tilde{I}^n$ is non-decreasing, $\tilde{I}^n$ increases only when $\tilde{V}^n = 0$, and Proposition 2 guarantees the uniqueness of the linearly generalized regulator mapping, for $\Gamma \equiv F_d'(0)$,

$$(\tilde{V}^n, \tilde{I}^n) = (\Phi_\Gamma, \Psi_\Gamma)(\tilde{X}^n). \tag{3.3}$$

Let B_1, B_2, and B be standard, independent Brownian motions. By the FCLT,

$$\tilde{A}^n \Rightarrow \sqrt{\text{var}(u_1)} B_1 \tag{3.4}$$

as $n \to \infty$. Next, observe that

$$\tilde{M}_v^n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) - \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) \mathbf{1}\{V(t_j^{n,-}) \geq d_j^n\}. \tag{3.5}$$

By Donsker's theorem, the first term in (3.5) weakly converges to $\sqrt{\text{var}(v_1)} B_2$. Therefore, to show

$$\tilde{M}_v^n \Rightarrow \sqrt{\text{var}(v_1)} B_2,$$

it is sufficient to show the second term in (3.5) weakly converges to 0. Let $\delta > 0$ and $\epsilon > 0$. By Lemma 1, we can choose K so that

$$P\left(\max_{j=1,\ldots,\lfloor nt \rfloor} n^{-1/2} V^n(t_j^{n,-}) > K \right) < \delta/2. \tag{3.6}$$

Also, for n large enough,

$$\frac{(18 \times 2\sqrt{2})^2 E|v_1 - 1|^2 \lfloor nt \rfloor}{n\epsilon^2} F\left(\frac{K}{\sqrt{n}} \right) < \frac{\delta}{2}, \tag{3.7}$$

since $F(n^{-1/2}K) \to 0$ as $n \to \infty$. Finally, the following chain of inequalities

$$\begin{aligned}
&P\left(\sup_{0 \leq s \leq t} n^{-1/2} \sum_{j=1}^{\lfloor ns \rfloor} (v_j - 1) \mathbf{1}\{V(t_j^{n,-}) \geq d_j^n\} > \epsilon \right) \\
&\leq P\left(\begin{array}{c} \sup_{0 \leq s \leq t} n^{-1/2} \sum_{j=1}^{\lfloor ns \rfloor} (v_j - 1) \mathbf{1}\{V(t_j^{n,-}) \geq d_j^n\} > \epsilon \\ \cap \max_{j=1,\ldots,\lfloor nt \rfloor} n^{-1/2} V^n(t_j^{n,-}) \leq K \end{array} \right) + \frac{\delta}{2} \\
&\leq P\left(\sup_{0 \leq s \leq t} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor ns \rfloor} (v_j - 1) \mathbf{1}\{d_j^n \leq \sqrt{n} K\} \right| > \epsilon \right) + \frac{\delta}{2} \\
&\leq \frac{E\left| \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1) \mathbf{1}\{d_j^n \leq \sqrt{n} K\} \right|^2}{n\epsilon^2} + \frac{\delta}{2} \text{ (Kolmogorov's submartingale inequality)} \\
&\leq (18 \times 2\sqrt{2})^2 \frac{E \sum_{j=1}^{\lfloor nt \rfloor} (v_j - 1)^2 \mathbf{1}\{d_j^n \leq \sqrt{n} K\}}{n\epsilon^2} + \frac{\delta}{2} \text{ (Burkholder's inequality)} \\
&= \frac{(18 \times 2\sqrt{2})^2 E(v_1 - 1)^2 \lfloor nt \rfloor}{n\epsilon^2} F\left(\frac{K}{\sqrt{n}} \right) + \frac{\delta}{2} \\
&\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta
\end{aligned}$$

shows the second term in (3.5) weakly converges to 0. Therefore, $\tilde{M}_v^n \Rightarrow \sqrt{\text{var}(v_1)} B_2$ and, by the random time change theorem, since $\bar{A}^n \to e$ as $n \to \infty$ a.s., u.o.c. by the

FSLLN,

$$\tilde{M}_v^n \circ \bar{A}^n \Rightarrow \sqrt{\text{var}(v_1)}B, \tag{3.8}$$

as $n \to \infty$. Next, using Lemma 1, it is straightforward to check the conditions of Theorem 15.2 in Billingsley showing $\{\tilde{V}^n\}$ is tight. Let $\{n_k\}$ be a subsequence along which $\{\tilde{V}^{n_k}\}$ converges in distribution to V. The FSLLN implies $\bar{A}^{n_k} \to e$ a.s., u.o.c., and so $\bar{A}^{n_k} \Rightarrow e$ as $n_k \to \infty$. By the Skorokhod representation theorem, there exists

$$(\underline{\tilde{V}}^n, \underline{\bar{A}}^n) \stackrel{D}{=} (\tilde{V}^n, \bar{A}^n)$$

such that

$$(\underline{\tilde{V}}^n, \underline{\bar{A}}^n) \to (V, e) \text{ a.s, u.o.c.},$$

as $n \to \infty$. Since

$$\sqrt{n}F(n^{-1/2}x) \to F'(0)x \text{ u.o.c.},$$

it follows that

$$\sqrt{n}F(n^{-1/2}\underline{\tilde{V}}^n(s)) \to F'(0)V(s) \text{ a.s. u.o.c.},$$

Therefore, by Lemma 8.3 in Dai and Dai, for any $t > 0$,

$$\int_0^t \sqrt{n}F(n^{-1/2}\underline{\tilde{V}}^n(s))d\bar{A}^n(s) \to \int_0^t F'(0)V(s)ds,$$

as $n \to \infty$, a.s., u.o.c., which implies

$$F'(0)\int_0^{\cdot} \tilde{V}^n(s)ds - \int_0^{\cdot} \sqrt{n}F(n^{-1/2}\tilde{V}^n(s))d\bar{A}^n(s) \Rightarrow 0\,, \tag{3.9}$$

as $n \to \infty$. Finally, Lemma 2 and the random time change theorem show

$$\tilde{M}_d^n \circ \bar{A}^n \Rightarrow 0, \tag{3.10}$$

as $n \to \infty$.

The convergences (3.4), (3.8), (3.9), and (3.10) establish

$$\tilde{X}^n \Rightarrow ce + (\sqrt{\text{var}(u_1)} + \sqrt{\text{var}(v_1)})B,$$

as $n \to \infty$. Use of the representation for $(\tilde{V}^n, \tilde{I}^n)$ in (3.3), the continuity of the linearly generalized regulator mapping stated in Proposition 2, and the continuous mapping

theorem shows (recalling that $\Gamma = F_d'(0)$)

$$(\tilde{V}^n, \tilde{I}^n) \Rightarrow (\Phi_\Gamma, \Psi_\Gamma)\,(ce + (\sqrt{\text{var}(u_1)} + \sqrt{\text{var}(v_1)})B),$$

a ROU with infinitesimal drift $c - \Gamma z$ and infinitesimal variance $\text{var}(u_1) + \text{var}(v_1)$.

Part (b): When $m_n = n^{1+\epsilon}$, the pathwise equation for V in (3.3) can be written as

$$\tilde{V}^n(t) = \tilde{X}^n(t) + \tilde{I}^n(t),$$

where

$$\tilde{X}^n(t) \equiv \bar{A}^n(t) + \tilde{M}_v^n(\bar{A}^n(t)) + \sqrt{n}(\rho^n - 1)t - \tilde{M}_d^n(\bar{A}^n(t)) - \int_0^t \sqrt{n} F(n^{-1/2-\epsilon}\tilde{V}^n(s))\, d\bar{A}^n(s).$$

Similar to the proof of part (a), by the properties of $\tilde{I}^n$ and the uniqueness of the conventional regulator mapping,

$$(\tilde{V}^n, \tilde{I}^n) \,(\Phi, \Psi)\, (\tilde{X}^n). \tag{3.11}$$

For any $x \geq 0$,

$$\sqrt{n} F(n^{-1/2-\epsilon} x) \to 0,$$

as $n \to \infty$, and so it is straightforward to use the tightness of $\tilde{V}^n$ established in Lemma 1 to show

$$\int_0^\cdot \sqrt{n} F(n^{-1/2-\epsilon}\tilde{V}^n(s))\, d\bar{A}^n(s) \Rightarrow 0.$$

Identical arguments to those in part (a) show the weak convergences of $\tilde{A}^n$, $\tilde{M}_v^n \circ \bar{A}^n$, and $\tilde{M}_d^n \circ \bar{A}^n$, and so

$$\tilde{X}^n \Rightarrow ce + (\sqrt{\text{var}(u_1)} + \sqrt{\text{var}(v_1)})B,$$

as $n \to \infty$. Therefore, from (3.11), the continuity of the conventional regulator mapping, and the continuous mapping theorem,

$$(\tilde{V}^n, \tilde{I}^n) \Rightarrow (\Phi, \Psi)\,(ce + (\sqrt{\text{var}(u_1)} + \sqrt{\text{var}(v_1)})B),$$

a RBM with infinitesimal drift c and infinitesimal variance $\text{var}(u_1) + \text{var}(v_1)$.

Part (c): When $m_n = n^{1-\epsilon}$, the pathwise equation for V in (2.5) can be written as

$$\begin{aligned}\tilde{V}^n(t) &= \tilde{A}^n(t) + \tilde{M}_v^n(\bar{A}^n(t)) + \sqrt{n}(\rho^n - 1)t - \tilde{M}_d^n(\bar{A}^n(t)) \\ &\quad - \int_0^t \sqrt{n} F(n^{-1/2+\epsilon}\tilde{V}^n(s)) d\bar{A}^n(s) + \tilde{I}^n(t) \\ &\geq 0.\end{aligned}$$

Since, as in the proofs of parts (a) and (b), for B a standard Brownian motion,

$$\tilde{A}^n + \tilde{M}_v^n \circ \bar{A}^n + \sqrt{n}(\rho^n - 1)e - \tilde{M}_d^n \circ \bar{A}^n \Rightarrow ce + (\sqrt{\text{var}(u_1)} + \sqrt{\text{var}(v_1)})B,$$

a proper random variable, $\tilde{I}^n$ only increases when $\tilde{V}^n = 0$, and for any $x \geq 0$,

$$\sqrt{n} F(n^{-1/2+\epsilon}x) \to +\infty$$

as $n \to \infty$, arguments similar to those in Theorem 2 in Section 3.2 in Reiman [21] show

$$\tilde{V}^n \Rightarrow 0$$

as $n \to \infty$. □

Our next theorem shows that customers who abandon the system without receiving service do not influence queueing fluctuations too much. In particular, the limiting behavior of the workload process is the same regardless of whether or not the workload of customers who eventually renege is included.

Theorem 2 (Weak convergence of the observed workload process)**.** Theorem 1 holds with $\tilde{W}^n$ replacing $\tilde{V}^n$, and the requirement that $\epsilon > 1/6$ in part (c).

Proof. From Theorem 1 and the definition of W in (2.2), we must show that for any given $t, \epsilon, \delta > 0$, and for $d_i^n = n^p w_i$ with $p > 5/6$, for large enough n

$$P\left(\sup_{0\leq s\leq t} \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(ns)} v_i \mathbf{1}\{V^n(t_i^{n,-}) \geq d_i^n \text{ and } t_i^n \leq ns \leq t_i^n + d_i^n\} > \epsilon\right) < \delta.$$

Since the supremum in the above expression occurs at t and $n^{-1}A^n \to e$ a.s., u.o.c., by the random time change theorem, it is enough to show, for large n,

$$P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} v_i \mathbf{1}\{V^n(t_i^{n,-}) \geq d_i^n \text{ and } t_i^n \leq nt \leq t_i^n + d_i^n\} > \epsilon\right) < \delta.$$

By Lemma 1, we can choose K large enough so that

$$P\left(\max_{j=1,\ldots,\lfloor nt \rfloor} \frac{1}{\sqrt{n}} V^n(t_j^{n,-}) > K\right) < \frac{\delta}{2}$$

for all n, and so

$$
\begin{aligned}
&P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor} v_i \mathbf{1}\left\{V^n\left(t_i^{n,-}\right)\geq d_i^n \text{ and } t_i^n \leq nt \leq t_i^n + d_i^n\right\} > \epsilon\right)\\
&\leq P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor} v_i \mathbf{1}\left\{V^n\left(t_i^{n,-}\right)\geq d_i^n \text{ and } t_i^n \leq nt \leq t_i^n + d_i^n\right\}\right.\\
&\left.> \epsilon \cap \max_{i=1,\ldots,\lfloor nt\rfloor}\frac{1}{\sqrt{n}}V^n\left(t_j^{n,-}\right)\leq K\right) + \frac{\delta}{2}\\
&\leq P\left(\frac{1}{\sqrt{n}}\sum_{i=\rho^n nt - n^{5/6}}^{nt} v_i \mathbf{1}\{n^{p-1/2}w_i \leq K\} > \frac{\epsilon}{2}\right) \qquad (3.12)\\
&+ P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{\rho^n nt - n^{5/6}} v_i \mathbf{1}\left\{nt - \sqrt{n}K \leq t_i^n\right\} > \frac{\epsilon}{2}\right) + \frac{\delta}{2}. \qquad (3.13)
\end{aligned}
$$

To finish the proof, we show terms (3.12) and (3.13) are both less than or equal to $\delta/4$. Since $p > 5/6$, $\sqrt{n}(\rho^n - 1) \to c < \infty$ as $n \to \infty$, $F_d'(0) < \infty$, and $\mathrm{var}(u_1) < \infty$, we can choose n large enough so that

$$
\frac{2K}{\epsilon}n^{5/6-p}(n^{1/6}t(1-\rho^n)+1)\frac{F(n^{1/2-p}K)}{n^{1/2-p}K} < \frac{\delta}{4} \qquad (3.14)
$$

$$
n^{-1/6}\frac{2(t-n^{-1/6})}{\epsilon}\frac{(\rho^n t - n^{-1/6})E(u_1-1)^2}{(1-n^{-1/3}K\rho^n)^2} < \frac{\delta}{4}. \qquad (3.15)
$$

For term (3.12), by Markov's inequality and (3.14)

$$
\begin{aligned}
(3.12) &\leq \frac{2}{\epsilon\sqrt{n}}E\left[\sum_{i=\rho^n nt - n^{5/6}}^{nt} v_i \mathbf{1}\{n^{p-1/2}w_i \leq K\}\right]\\
&= \frac{2}{\epsilon\sqrt{n}}(nt - \rho^n nt + n^{5/6})F(n^{1/2-p}K)\\
&= \frac{2K}{\epsilon}n^{5/6-p}(n^{1/6}t(1-\rho^n)+1)\frac{F(n^{1/2-p}K)}{n^{1/2-p}K}\\
&< \frac{\delta}{4}.
\end{aligned}
$$

For term (3.13), by two applications of Markov's inequality and (3.15),

$$
\begin{aligned}
(3.13) &\le \frac{2}{\epsilon\sqrt{n}} E\left[\sum_{i=1}^{\rho^n nt - n^{5/6}} v_i \mathbf{1}\left\{nt - \sqrt{n}K \le t_i^n\right\} > \frac{\epsilon}{2}\right] \\
&= \frac{2}{\epsilon\sqrt{n}} \sum_{i=1}^{\rho^n nt - n^{5/6}} P\left(\rho^n(nt - \sqrt{n}K) \le \sum_{j=1}^{i} u_j\right) \\
&\le \frac{2}{\epsilon\sqrt{n}}(\rho^n nt - n^{5/6}) P\left(\rho^n(nt - \sqrt{n}K) \le \sum_{j=1}^{\rho^n nt - n^{5/6}} u_j\right) \\
&= \frac{2(\rho^n nt - n^{5/6})}{\epsilon\sqrt{n}} P\left(n^{5/6} - \sqrt{n}K\rho^n \le \sum_{j=1}^{\rho^n nt - n^{5/6}} (u_j - 1)\right) \\
&\le \frac{2(\rho^n nt - n^{5/6})}{\epsilon\sqrt{n}} \frac{E\left(\sum_{j=1}^{\rho^n nt - n^{5/6}} (u_j - 1)\right)^2}{(n^{5/6} - \sqrt{n}K\rho^n)^2} \\
&= n^{-1/6} \frac{2(\rho^n t - n^{-1/6})}{\epsilon} \frac{(\rho^n t - n^{-1/6}) E(u_1 - 1)^2}{(1 - n^{-1/3}K\rho^n)^2} \\
&< \frac{\delta}{4}.
\end{aligned}
$$

□

4. Asymptotic behavior of the reneging and balking queue-length processes

We first develop a heavy traffic limit theorem for the queue-length process in our reneging model that shows a state-space collapse. In particular, the limiting behavior of the queue-length process is close to the offered waiting time (and the observed workload process by use of Theorem 2). We next show that the behavior of the queue-length process in our balking model, in which customer processing times are not observable is identical in heavy-traffic.

Theorem 3 (Weak convergence of the queue-length process for the reneging model)**.** Theorem 1 holds with $\tilde{Q}^n$ replacing $\tilde{V}^n$, and the requirement that $\epsilon > 1/6$ in part (c).

Proof. We leverage off the weak convergence result for the offered waiting time process in Theorem 1. Our proof is similar to Theorem 4 in Reiman [21], which shows the weak convergence of the diffusion-scaled queue-length process in a GI/GI/1 queue by using a weak convergence result for the waiting time process. However, we must account for reneging customers.

Let $a^n(t)$ be the arrival time of the customer in service at time t in the nth system. If the server is idle, let $a^n(t) = t$. Since the service discipline is FIFO, the number of customers currently in queue is less than the number that have arrived after the customers

currently in service plus one, $A^n(t) - A^n(a^n(t)) + 1$. Additionally, the current queue-length exceeds $A^n(t) - A^n(a^n(t))$ minus the number of customers that have arrived after the one currently in service that will eventually renege, and so

$$A^n(t) - A^n(a^n(t)) - \sum_{i=A^n(a^n(t))}^{A^n(t)} \mathbf{1}\{V^n(t_i^{n,-}) \geq d_i^n\} \leq Q^n(t) \leq A^n(t) - A^n(a^n(t)) + 1,$$

or

$$|Q^n(t) - V^n(t)| \leq |A^n(t) - A^n(a^n(t)) - V^n(t)| + 1 + \sum_{i=A^n(a^n(t))}^{A^n(t)} 1\{V^n(t_i^{n,-}) \geq d_i^n\}.$$

Time and space-scaling the above equation along with some algebra yields

$$\begin{aligned} |\tilde{Q}^n(t) - \tilde{V}^n(t)| \leq\ & |\tilde{A}^n(t) - \bar{\tilde{A}}^n(\bar{a}^n(t))| + n^{-1/2} + |\rho^n(\tilde{V}^n(\bar{a}^n(t)^-) - \tilde{V}^n(t))| \\ & + |\rho^n(\sqrt{n}(t - \bar{a}^n(t)) - \tilde{V}^n(\bar{a}^n(t)^-))| + |\tilde{V}^n(t)(\rho^n - 1)| \\ & + n^{-1/2} \sum_{i=A^n(a^n(nt))}^{A^n(nt)} \mathbf{1}\{V^n(t_i^{n,-}) \geq d_i^n\}, \end{aligned} \tag{4.1}$$

where $\bar{a}^n(t) = n^{-1}a(nt)$.

We now argue the right-hand side of (4.1) weakly converges to 0, which is sufficient to complete the proof. For $t \geq 0$ and $n \geq 1$, since the server works at rate 1 and the system is FIFO,

$$V^n(a^n(t)^-) \leq t - a^n(t) \leq V^n(a^n(t)^-) + v_{A^n(a^n(t))}. \tag{4.2}$$

From Lemma 3 in Iglehart and Whitt [15], for any $t \geq 0$

$$\sup_{k=1,\ldots,nt} n^{-1/2} v_k \to 0 \quad \text{in probability} \tag{4.3}$$

as $n \to \infty$. Combining (4.2) and (4.3) shows

$$\sup_{0 \leq s \leq t} |\tilde{V}^n(\bar{a}^n(s)^-) - \sqrt{n}(s - \bar{a}^n(s))| \to 0 \quad \text{in probability}, \tag{4.4}$$

as $n \to \infty$. Since $\tilde{V}^n \Rightarrow Z$ as $n \to \infty$ by Theorem 1 and $\bar{a}^n(s) \leq s$,

$$\frac{1}{\sqrt{n}}(\tilde{V}^n \circ \bar{a}^n) \Rightarrow 0, \tag{4.5}$$

as $n \to \infty$. Together (4.4) and (4.5) imply

$$\bar{a}^n \Rightarrow e, \tag{4.6}$$

as $n \to \infty$. Since $\tilde{A}^n$ and $\tilde{V}^n$ both weakly converge to a continuous limit process,

$$\tilde{A}^n - \tilde{A}^n \circ \bar{a}^n \Rightarrow 0 \quad \text{and} \quad \tilde{V}^n - \tilde{V}^n \circ \bar{a}^n \Rightarrow 0, \tag{4.7}$$

as $n \to \infty$. Since $\rho^n \to 1$ as $n \to \infty$ and for any $t > 0$, $\sup_{0 \le s \le t} \tilde{V}^n(s)$ is tight by Lemma 1,

$$\tilde{V}^n(\rho^n - 1) \Rightarrow 0, \tag{4.8}$$

as $n \to \infty$. Therefore, by (4.1) and the convergences in (4.4), (4.7), and (4.8), to complete the proof, it is sufficient to show

$$n^{-1/2} \sum_{i=A^n(a^n(n\cdot))}^{A^n(n\cdot)} \mathbf{1}\{V^n(t_i^{n,-}) < d_i^n\} \Rightarrow 0,$$

as $n \to \infty$.

Since

$$\begin{aligned} & n^{-1/2} \sum_{i=A^n(a^n(nt))}^{A^n(nt)} \mathbf{1}\{V^n(t_i^{n,-}) < d_i^n\} \\ &= n^{-1/2} \sum_{i=A^n(a^n(nt))}^{A^n(nt)} P(V^n(t_i^{n,-}) \ge d_i^n) \\ &\quad + n^{-1/2} \sum_{i=A^n(a^n(n\cdot))}^{A^n(n\cdot)} \left(1\{V^n(t_i^{n,-}) < d_i^n\} - E[\mathbf{1}\{V^n(t_i^{n,-}) \ge d_i^n\} | \mathcal{F}_{i-1}]\right) \end{aligned}$$

and arguments similar to Lemma 2 show the second term in the right-hand side of the above expression weakly converges to 0, to complete the proof, we must show

$$n^{-1/2} \sum_{i=A^n(a^n(n\cdot))}^{A^n(n\cdot)} P(V^n(t_i^{n,-}) \ge d_i^n) \Rightarrow 0,$$

as $n \to \infty$. Given ϵ and δ, choose K large enough so that

$$P\left(\max_{j=0,1,\ldots,\lfloor nt \rfloor} n^{-1/2} V^n(t_j^{n,-}) > K \right) < \delta/2. \tag{4.9}$$

for all n. (Lemma 1 guarantees such a K exists.) Also, the FSLLN, the fact that $p = 1$ implies $n^{1/2} F(n^{1/2-p} K) \to f(0)K$ and $p > 1$ implies $n^{1/2} F(n^{1/2-p} K) \to 0$ as $n \to \infty$, and the convergence in (4.6) imply we can choose n large enough so that

$$P\left(\sup_{0 \le s \le t} (\bar{A}^n(s) - \bar{A}^n(\bar{a}^n(s))) n^{1/2} F(K n^{1/2-p}) > \epsilon \right) < \delta/2. \tag{4.10}$$

From (4.9) and (4.10)

$$\begin{aligned} P\left(\left(\sup_{0\le s\le t}\frac{1}{\sqrt{n}}\sum_{i=A^n(a^n(ns))}^{A^n(ns)} P\left(V^n\left(t_i^{n,-}\right)\ge d_i^n\right)\right)>\epsilon\right) \\ \le\ P\left(\sup_{0\le s\le t}\left(\bar{A}^n(s)-\bar{A}^n(\bar{a}^n(s))\right)n^{1/2}F(Kn^{1/2-p})>\epsilon\right)+\delta/2 \\ <\ \delta. \end{aligned}$$

Part 3 follows as in the proof of Theorem 1 (c) because a little algebra shows

$$\begin{aligned} \tilde{Q}^n(t) = \tilde{A}^n(t) - \tilde{S}^n(\bar{B}^n(t)) + \sqrt{n}t(\rho^n - 1) - \tilde{M}_d^n(\bar{A}^n(t)) + \tilde{\Delta}^n(t) \\ - \int_0^t \sqrt{n}F\left(n^{-1/2+\epsilon}\tilde{V}^n(s)\right)d\bar{A}^n(s) + \tilde{I}^n(t), \end{aligned}$$

and $\sqrt{n}F(n^{-1/2+\epsilon}x) \to +\infty$ as $n \to \infty$. (Note that we needed $p > 5/6$ to conclude, similar to the proof of Theorem 2, that $\tilde{\Delta}^n \Rightarrow 0$ as $n \to \infty$.) □

Our final theorem shows the diffusion-scaled queue-length process in our balking model has the same limiting behavior as the observed queue-length process in our reneging model.

Theorem 4 (Weak convergence of the queue-length process for the balking model)**.** Theorem 1 holds with $\tilde{Q}_B^n$ replacing $\tilde{V}^n$.

Proof. Define

$$M_B^n(i) \equiv \sum_{j=1}^{i} \mathbf{1}\left\{Q_B^n\left(t_i^{n,-}\right) - 1 \ge d_i^n\right\} - E\left[\mathbf{1}\left\{Q_B^n\left(t_i^{n,-}\right) - 1 \ge d_i^n\right\} \,\middle|\, \mathcal{F}_{i-1}\right],$$

and observe that (1) M_B^n is a martingale and (2) $E[\mathbf{1}\{Q_B^n(t_i^-) - 1 \ge d_i^n\} \,|\, \mathcal{F}_{i-1}] = F^n(Q_B^n(t_i^{n,-}) - 1)$. (Recall the definition of $\mathcal{F}_i$ at the end of Section 1.1.) Algebraic manipulations of the pathwise equation for Q_B in (2.4) show

$$\tilde{Q}_B^n(t) + \int_0^t \sqrt{n}F\left(\frac{Q_B(ns) - 1}{m_n}\right)d\bar{A}^n(s) = \tilde{X}^n(t) + \tilde{I}_B^n(t), \tag{4.11}$$

where

$$\tilde{X}_B^n(t) \equiv \tilde{A}^n(t) - \tilde{S}^n\left(\bar{B}_B^n(t)\right) + \sqrt{n}t(\rho^n - 1) - \tilde{M}_B^n(\bar{A}^n(t)),$$

for $\bar{B}_B^n(t) \equiv n^{-1}B_B^n(nt)$, $\tilde{M}_B^n(t) = n^{-1/2}M_B^n(\lfloor nt \rfloor)$, and all other scaled quantities are as defined in Section 2.3. We can also write the pathwise equation for Q_B in terms of

the fluid-scaled quantities $\bar{Q}_B^n(t) \equiv n^{-1} Q_B^n(nt)$, $\bar{A}^n$, $\bar{S}^n$, $\bar{B}_B^n$, and $\bar{I}_B^n(t) \equiv n^{-1} I_B^n(nt)$ as follows

$$\bar{Q}_B^n(t) = \bar{X}_B^n(t) + \bar{I}_B^n(t), \tag{4.12}$$

where

$$\bar{X}_B^n(t) \equiv \bar{A}^n(t) - \rho^n t - \bar{S}^n\left(\bar{B}_B^n(t)\right) + \bar{B}_B^n(t) + t(\rho^n - 1) - n^{-1} \sum_{i=1}^{A^n(nt)} \mathbf{1}\left\{Q_B\left(t_i^{n,-}\right) - 1 \geq d_i^n\right\}.$$

Since $\bar{I}_B^n$ and $\bar{Q}_B^n$ satisfy conditions (C1) and (C2) in Section A.1 for $\Gamma = 0$, we can write $(\bar{Q}_B^n, \bar{I}_B^n)$ using the conventional regulator mapping

$$\left(\bar{Q}_B^n, \bar{I}_B^n\right) = (\Phi, \Psi)\left(\bar{X}_B^n\right). \tag{4.13}$$

By the FSLLN, the fact that $\bar{B}_B^n(t) \leq t$ for all n, and the fact that $\rho^n \to 1$ as $n \to \infty$, we can conclude

$$\bar{X}_B^n \Rightarrow 0,$$

as $n \to \infty$, provided we can show

$$\frac{1}{n} \sum_{i=1}^{A^n(n\cdot)} \mathbf{1}\left\{Q_B\left(t_i^{n,-}\right) - 1 \geq d_i^n\right\} \Rightarrow 0, \tag{4.14}$$

as $n \to \infty$. Let $\epsilon > 0$, $\delta > 0$, and $t > 0$. Similar arguments to those in Lemma 1 show there exists a K such that for all n

$$P\left(\max_{j=1,\ldots,\lfloor nt \rfloor} n^{-1/2} Q_B^n\left(t_j^{n,-}\right) > K + 1\right) < \frac{\delta}{2}. \tag{4.15}$$

Also, for large enough n, since $p \geq 1/2$,

$$F(Kn^{1/2-p}) < \frac{\delta\epsilon}{2t}. \tag{4.16}$$

Recalling that $d_i^n = n^p w_i$, Markov's inequality, (4.15), and (4.16) show

$$P\left(\sup_{0 \leq s \leq t} \left| n^{-1} \sum_{i=1}^{ns} \mathbf{1}\left\{Q_B\left(t_t^{n,-}\right) - 1 \geq d_i^n\right\} \right| > \epsilon\right) = P\left(n^{-1} \sum_{i=1}^{nt} \mathbf{1}\left\{Q_B^n\left(t_i^{n,-}\right) - 1 \geq d_i^n\right\} > \epsilon\right)$$

$$
\begin{aligned}
&\le P\left(\frac{1}{n}\sum_{i=1}^{nt}\mathbf{1}\left\{Q_B^n\left(t_i^{n,-}\right)-1\ge d_i^n\right\}>\epsilon\cap\max_{j=1,\ldots,\lfloor nt\rfloor}\frac{1}{\sqrt{n}}Q_B^n\left(t_t^{n,-}\right)>K\right)+\frac{\delta}{2}\\
&\le \frac{\sum_{i=1}^{nt}E[\mathbf{1}\{d_i^n\le\sqrt{n}K\}]}{n\epsilon}+\frac{\delta}{2}\\
&= \frac{t}{\epsilon}F(n^{1/2-p}K)+\frac{\delta}{2}\\
&= \delta,
\end{aligned}
$$

which shows (4.14) using the random time change theorem. Since $\bar{X}_B^n\Rightarrow 0$ as $n\to\infty$, from (4.13),

$$\left(\bar{Q}_B^n,\bar{I}_B^n\right)\Rightarrow(0,0),$$

as $n\to\infty$, which implies

$$\bar{B}_B^n\to e, \tag{4.17}$$

as $n\to\infty$.

By (4.17), the FCLT, the random time change theorem, and an argument similar to Lemma 2 establishing $\tilde{M}_B^n\Rightarrow 0$,

$$\tilde{X}_B^n\Rightarrow ce+\sqrt{\operatorname{var}(u_1)+\operatorname{var}(v_1)}B$$

as $n\to\infty$, where B is a standard Brownian motion. Similar arguments to those in the proof of Theorem 1 establish the asymptotic behavior of the integral term

$$\int_0^t\sqrt{n}F\left(\frac{Q_B(ns)-1}{m_n}\right)d\bar{A}^n(s),$$

and complete the proof. □

5. Proposed approximations and simulation results

Our weak convergence results in Theorems 3 and 4 suggest approximating the queue-length process in either our balking or reneging model with a ROU process. Specifically, consider a system having arrival rate ρ, mean service time 1, and deadline distribution function F. We propose to approximate the queue-length process with the ROU Z having infinitesimal drift $\rho-1-F'(0)z$, infinitesimal variance $\rho\operatorname{var}(u_1)+(\rho\wedge 1)\operatorname{var}(v_1)$, and steady-state mean (as given in Proposition 18.3 in Browne and Whitt [6])

$$\bar{Z}=E\left[N\left(m_\rho,b_\rho^2\right)\,\middle|\,0\le N\left(m_\rho,b_\rho^2\right)\right],$$

where $N(m, b^2)$ is a normal random variable with mean m and variance b^2, and

$$m_\rho = \frac{\rho - 1}{F'(0)} \quad \text{and} \quad b_\rho^2 = \frac{\rho \text{var}(u_1) + (\rho \wedge 1)\text{var}(v_1)}{2F'(0)}.$$

To understand the intuition behind the proposed approximation, first rewrite the queue-length process evolution equation (2.3) as

$$Q(t) + \int_0^t F\left(\frac{V(s)}{m}\right) dA(s) = X(t) + I(t),$$

where

$$\begin{aligned} X(t) &\equiv [A(t) - \rho t] - [S(B(t)) - B(t)] + t(\rho - 1) + \epsilon(t) \\ \epsilon(t) &\equiv -M_d(A(t)) + \sum_{i=1}^{A(t)} \mathbf{1}\{V(t_i^-) \geq d_i \text{ and } t_i \leq t \leq t_i + d_i\}. \end{aligned}$$

For any $x \in \Re$, as $n \to \infty$, $nF(n^{-1}x) \to F'(0)x$ and $\bar{A}^n \to e$ a.s., u.o.c. When $m \equiv n$ is large (meaning the system has traffic intensity ρ close to 1 and mean deadlines of order $(1 - \rho)^{-2}$), for each $t > 0$, Theorem 3 and the preceding observation suggest

$$\begin{aligned} \int_0^t F\left(\frac{V^n(s)}{n}\right) dA^n(s) &= \int_0^t nF(n^{-1}V^n(s)) d\bar{A}^n(s) \\ &\overset{D}{\approx} \int_0^t nF(n^{-1}Q^n(s)) d\bar{A}^n(s) \\ &\overset{D}{\approx} \int_0^t F'(0)Q^n(s) ds, \end{aligned}$$

where the symbol $\overset{D}{\approx}$ denotes approximately equal in distribution. Hence when ρ is close to 1 and mean deadlines are of order $(1 - \rho)^{-2}$, for $\Gamma = F'(0)$,

$$(Q, I) \overset{D}{\approx} (\Phi_\Gamma, \Psi_\Gamma)(X). \tag{5.1}$$

The FCLT shows for $t > 0$

$$\begin{aligned} A(t) - \rho t &\overset{D}{\approx} \sqrt{n} B_1\left(\frac{t}{n}\right) \overset{D}{=} B_1(t) \\ S(t) - t &\overset{D}{\approx} B_2(t), \end{aligned}$$

where B_1 and B_2 are two independent zero-mean Brownian motions with infinitesimal variances $\rho\text{var}(u_1)$ and $\text{var}(v_1)$ respectively. The assumption that the cumulative busy

time in $[0, t]$, $B(t)$ is proportional to $(\rho \wedge 1)t$ implies for $t > 0$

$$S(B(t)) - B(t) \overset{D}{\approx} B_2((\rho \wedge 1)t).$$

Therefore, because Lemma 2 and a slight modification of the proof of Theorem 2 show $n^{-1/2}\epsilon^n(\cdot n) \Rightarrow 0$ as $n \to \infty$, the process X can be approximated by a Brownian motion W with drift $\rho - 1$ and variance $\rho \text{var}(u_1) + (\rho \wedge 1)\text{var}(v_1)$. The representation (5.1) then establishes

$$(Q, I) \overset{D}{\approx} (\Phi_\Gamma, \Psi_\Gamma)(W),$$

a ROU satisfying the stochastic equation (3.1) with $\alpha = \rho - 1$ and $\sigma^2 = \rho \text{var}(u_1) + (\rho \wedge 1)\text{var}(v_1)$. Observe that when $\Gamma = 0$ (meaning the effects of reneging/balking are not accounted for in the limiting diffusion), our proposed approximation is a RBM R with infinitesimal drift $\rho - 1$, infinitesimal variance $\rho \text{var}(u_1) + (\rho \wedge 1)\text{var}(v_1)$, and steady state mean (for $\rho < 1$)

$$\bar{R} = \frac{\rho(\text{var}(u_1) + \text{var}(v_1))}{2(1 - \rho)},$$

the standard heavy traffic approximation for a conventional GI/GI/1 queue when the primitive inter-arrival and service time sequences $\{u_i : i \geq 0\}$ and $\{v_i : i \geq 0\}$ have mean 1; see Section 6.5 in Chen and Yao [8].

Of course, our proposal to approximate the steady-state mean queue-length with the ROU steady state mean $\bar{Z}$ assumes the limit interchange

$$\lim_{t\to\infty} \lim_{n\to\infty} \tilde{Q}^n(t) \overset{(?)}{=} \lim_{n\to\infty} \lim_{t\to\infty} \tilde{Q}^n(t)$$

is valid. Proposition 1 in Ward and Glynn verifies the interchange in a purely exponential setting. In general, we conjecture an argument similar to that in Gamarnik and Zeevi [11] verifies the interchange.

We compare the steady-state of the regulated O-U process Z with the steady-state mean queue-length that results from simulating our reneging and balking models. For the simulation results displayed in Tables 1 and 2, we present 95% confidence intervals for the mean queue-length found after 5 simulation runs (performed using the Extend simulation language [1]), each of 500,000 time units. We also present the percent of reneging and balking customers for both models, averaged over the 5 runs. As one would expect, the percent of reneging customers is consistently slightly higher than the percent of balking customers, and so the mean queue-length for the balking queue is generally slightly higher than that for the reneging queue with the same parameters. Although this is too fine a difference to evidence itself in our diffusion limits, it makes sense to keep this in mind when applying our approximations.

Table 1 shows that under various assumptions on the variability of inter-arrival and service times, our approximation predicts mean queue-lengths reasonably accurately in

Table 1
Simulated mean queue lengths, $\bar{Q}$ and $\bar{Q}^B$, for our reneging and balking models when $\rho = 1$. We assume inter-arrival and service times have distribution ρ^{-1} Gamma (m, m) and Gamma (m, m) respectively and that deadline lengths are distributed Uniform $(0, d)$.

	Reneging model		Balking model		
	% renege	$\bar{Q}$	% balk	$\bar{Q}^B$	$\bar{Z}$
$m = 2\,d$					
100,000	0.04	(33.630, 46.350)	0.04	(40.527, 50.667)	45.50
10,000	0.32	(29.495, 37.615)	0.30	(26.220, 34.240)	31.71
1000	1.42	(14.695, 15.199)	1.39	(14.208, 15.862)	14.56
100	4.98	(5.569, 5.730)	4.55	(5.696, 5.814)	5.27
10	14.58	(2.071, 2.080)	12.26	(2.243, 2.259)	1.74
1	32.86	(0.860, 0.861)	23.90	(1.075, 1.076)	0.56
$R = 49.5$					
$m = 1d$					
100,000	0.10	(52.823, 118.423)	0.08	(75.012, 98.812)	85.27
10,000	0.50	(48.091, 54.851)	0.51	(48.683, 55.323)	52.15
1000	2.09	(20.422, 22.502)	2.09	(20.930, 22.656)	21.80
100	6.99	(7.369, 7.551)	6.66	(7.542, 7.626)	7.59
10	19.90	(2.369, 2.382)	17.38	(2.542, 2.556)	2.47
1	40.28	(0.817, 0.822)	32.98	(0.991, 0.997)	0.79
$R = 99$					
$m = \frac{1}{2}d$					
100,000	0.17	(113.050, 199.450)	0.16	(109.520, 256.520)	154.26
10,000	0.76	(67.873, 83.753)	0.85	(70.971, 97.571)	82.73
1000	3.16	(30.567, 32.443)	3.13	(29.795, 32.955)	32.11
100	9.82	(9.625, 9.969)	9.40	(9.862, 9.988)	10.87
10	26.60	(2.665, 2.698)	23.81	(2.865, 2.884)	3.51
1	47.91	(0.773, 0.780)	42.67	(0.896, 0.902)	1.12
$R = 198$					

systems with balking and reneging customers. The approximation loses accuracy as the percentage of balking or reneging customers increases.

Still, our approximation turns out to be robust over a wide range of traffic intensities in a fairly practical setting. Motivated by the statistical analysis of a bank call center data set in Brown et al. [5], we assume service times follow a lognormal distribution and inter-arrival times are exponentially distributed. (Actually, they show arrivals are well modelled as an inhomogeneous Poisson process. However, our theory cannot handle this non-stationarity.) Although their analysis does not show abandonment times to follow a particular "named" distribution, it is clear that abandonment times do not follow an

Table 2
Simulated mean queue lengths, $\bar{Q}$ and $\bar{Q}^B$, for our reneging and balking models. We assume inter-arrival and service times have distribution ρ^{-1} Exponential (1) and Lognormal (1, 1) respectively and that deadline lengths are distributed Uniform (0, 1000).

	Reneging model		Balking model		Approximations	
ρ	% renege	$\bar{Q}$	% balk	$\bar{Q}^B$	$\bar{Z}$	$\bar{R}$
0.99	2.09	(20.36, 23.00)	2.06	(20.85, 23.75)	21.79	99.00
0.95	1.23	(12.22, 12.88)	1.13	(12.81, 13.38)	12.98	19.00
0.90	0.78	(7.54, 8.41)	0.64	(7.54, 8.41)	7.84	9.00
0.80	0.38	(3.78, 3.91)	0.25	(3.84, 3.93)	3.88	4.00
0.70	0.22	(2.26, 2.34)	0.12	(2.28, 2.34)	2.31	2.33
0.60	0.14	(1.47, 1.49)	0.06	(1.48, 1.52)	1.49	1.50
0.50	0.09	(0.99, 1.01)	0.03	(0.98, 1.00)	1.00	1.00

exponential distribution. We assume deadline times follow a uniform distribution. Table 2 shows that our approximation is accurate across traffic intensities ranging between 0.5 and 1—much lower than one might guess from our supporting limit theorems.

In many real-world service industry applications of queueing theory (such as call centers, fast food restaurants, etc.), reneging and/or balking is present. If reneging and/or balking percentages are small, one might be tempted to model the system as a conventional GI/GI/1 queue. In such settings, the appropriate diffusion approximation to the queue is the RBM R defined a couple paragraphs earlier. Therefore, we take this opportunity to present the mean steady-state queue-length predicted by the RBM R.

It is striking to observe the degree to which small percentages of reneging and balking customers affect queue-lengths. For most all parameter combinations presented in Tables 1 and 2, the RBM approximation does not provide reasonable predictions. It is only when $d = 100,000$ in Table 1, and when $\rho \leq 0.8$ in Table 2 that the RBM approximation is reasonably accurate. This seems to confirm a heuristic that can be inferred from Theorems 1–4; that when deadline lengths are generally larger than $(1-\rho)^{-2}$, the presence of balking or reneging may be effectively ignored. Of course, the safe approach is to always model balking and reneging behavior since this heuristic may be hard to confirm in practical situations.

A. Appendix

A.1. A regulator mapping with state dependence

For the reader's convenience, we reproduce results on the existence, uniqueness, and continuity of a linearly generalized regulator mapping. For d a positive integer, $x \in D([0,\infty), \Re^d)$ having $x(0) \geq 0$, and Γ, M square matrices of dimension $d \times d$, the

linearly generalized regulator mapping

$$(\Phi_\Gamma, \Psi_\Gamma)(x) : D([0, \infty), \Re^d) \to D([0, \infty), [0, \infty)^{2d})$$

is defined by

$$(\Phi_\Gamma, \Psi_\Gamma)(x) \equiv (z, l),$$

where

(C1) $z(t) + \int_0^t \Gamma_z(s)ds = x(t) + Ml(t) \geq 0$ for all $t \geq 0$
(C2) $l(0) = 0$, l is non-decreasing, and $\int_0^\infty z_j(t)dl_j(t) = 0$, $j = 1, \ldots, J$.

Observe that if Γ is the zero matrix, we have the conventional regulator mapping discussed in Section 7.2 of Chen and Yao [8]. We write Φ, Ψ, where $\Phi = \Phi_0$ and $\Psi = \Psi_0$ to emphasize when we are referring to the conventional regulator mapping.

The key to establishing the existence, uniqueness, and Lipschitz continuity of Φ_Γ and Ψ_Γ is understanding the properties of the following integral equation

$$u(t) = x(t) - \int_0^t \Gamma\Phi(u)(s)ds. \tag{A.1}$$

In particular, as in Chen [7], define the mapping $\mathcal{M} : D([0, \infty), \Re^d) \to D([0, \infty), \Re^d)$ (which exists uniquely by Lemma 3) by $\mathcal{M}(x) \equiv u$, and observe that conditions (C1)–(C2) are satisfied when

$$(\Phi_\Gamma, \Psi_\Gamma)(x) = (\Phi, \Psi)(\mathcal{M}(x)). \tag{A.2}$$

The following Lemma, whose proof can be found in Reed and Ward [20], establishes the basic properties of integral equations having the form (A.1).

Lemma 3. Suppose $\eta : D([0, \infty), \mathcal{R}^d) \to D([0, \infty), \mathcal{R}^d)$ is Lipschitz continuous. Then for any given $x \in D([0, \infty), \mathcal{R}^d)$, there exists a unique $u \in D([0, \infty), \mathcal{R}^d)$ that satisfies the integral equation

$$u(t) = x(t) - \int_0^t n(u)(s)ds, \tag{A.3}$$

and has initial condition $u(0) = x(0)$. Furthermore, the mapping $\mathcal{M}_\eta : D([0, \infty), \mathcal{R}^d) \to D([0, \infty), \mathcal{R}^d)$ defined by $\mathcal{M}_\eta(x) \equiv u$ is Lipschitz continuous.

Using the representation (A.2), Lemma 3, and Theorem 7.2 of Chen and Yao [8], which establishes the existence, uniqueness, and Lipschitz continuity of the mapping (Φ, Ψ), it is immediate to prove the following proposition.

Proposition 2. Suppose M has positive diagonal elements, non-positive off-diagonal elements, and a non-negative inverse. Then, for each $x \in D([0,\infty), \Re^d)$ having $x(0) \geq 0$, there exists a unique (z, l) satisfying (C1)–(C2). Furthermore, the mappings Φ_Γ and Ψ_Γ are Lipschitz continuous.

A.2. Lemma proofs

Proof of Lemma 1. First observe that

$$P\left(\max_{j=1,\ldots,\lfloor nt\rfloor} \frac{1}{\sqrt{n}} V^n\left(t_j^{n,-}\right) > K\right) \leq P\left(\sup_{0\leq s\leq t} \frac{1}{\sqrt{n}} V^n\left((ns)\frac{\sum_{i=1}^{\lfloor ns\rfloor} u_i}{\lfloor ns\rfloor}\right) > K\right).$$

Since the strong law of large numbers establishes for any $0 < s \leq t$

$$\frac{\sum_{i=1}^{\lfloor ns\rfloor} u_i}{\lfloor ns\rfloor} \to 1 \text{ a.s.},$$

by the random time change theorem, for a given $\epsilon > 0$, it is enough to show there exists K such that

$$P\left(\sup_{0\leq s\leq t} \tilde{V}^n(s) > K\right) < \epsilon. \tag{A.4}$$

Define

$$R(t) \equiv \sum_{i=1}^{A(t)} v_i - B(t).$$

The process R is the workload process in a conventional GI/GI/1 queue (without reneging), for which the following convergence is known (see, for example, Theorem 1 in Section 3.2 in Reiman [21])

$$\tilde{R}^n \Rightarrow Z^R, \tag{A.5}$$

as $n \to \infty$, where $\tilde{R}^n(t) = n^{-1/2} R^n(nt)$, and Z^R is a RBM with infinitesimal drift c and infinitesimal variance $\text{var}(u_1) + \text{var}(v_1)$ (as in part (b) of Theorem 1). The weak convergence in (A.5) implies there exists K such that for all n

$$P\left(\sup_{0\leq s\leq t} \tilde{R}^n(s) > K\right) < \epsilon,$$

which implies (A.4) since $R(t) \geq V(t)$ for all $t \geq 0$. □

Proof of Lemma 2. For any given $t, \epsilon, \delta > 0$, we must show

$$P\left(\sup_{0\leq s\leq t} \left|\bar{M}_d^n(s)\right| > \epsilon\right) = P\left(\max_{i=1,\ldots,\lfloor nt\rfloor} \left|M_d^n(i)\right| > \epsilon\sqrt{n}\right) < \delta, \tag{A.6}$$

for large enough n. By a generalization of Kolmogorov's inequality, and by Burkholder's inequality (see, for example, Corollary 2.1 and Theorem 2.10 in Hall and Heyde [13]),

$$\begin{aligned} &P\left(\max_{i=1,\ldots,\lfloor nt\rfloor}\left|M_d^n(i)\right| > \epsilon\sqrt{n}\right) \\ &\quad\le \frac{E\left|M_d^n(\lfloor nt\rfloor)\right|^2}{n\epsilon^2} \\ &\quad\le \frac{\bar{c}}{n\epsilon^2}E\left[\sum_{j=1}^{\lfloor nt\rfloor}\left(\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge d_j^n\right\} - E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge d_j^n\right\}\,\middle|\,\mathcal{F}_{j-1}\right]\right)^2\right], \end{aligned}$$

where $\bar{c}$ is a finite constant-that does not depend on n.

Since

$$\begin{aligned} &\left(\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge d_j^n\right\} - E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge d_j^n\right\}\,\middle|\,\mathcal{F}_{j-1}\right]\right)^2 \\ &\quad\le \mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge d_j^n\right\} + E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge d_j^n\right\}\,\middle|\,\mathcal{F}_{j-1}\right] \end{aligned}$$

and

$$E\left[\mathbf{1}\left\{V^n\left(t_j^{n,-}\right)\ge d_j^n\right\} + E\left[1\left\{V^n\left(t_j^{n,-}\right)\ge d_j^n\right\}\,\middle|\,\mathcal{F}_{j-1}\right]\right] = 2P\left(V^n\left(t_j^{n,-}\right)\ge d_j^n\right),$$

the inequality

$$P\left(\max_{i=1,\ldots,\lfloor nt\rfloor}\left|M_d^n(i)\right| > \epsilon\sqrt{n}\right) \le \frac{2\bar{c}}{n\epsilon^2}\sum_{j=1}^{\lfloor nt\rfloor}P\left(V^n\left(t_j^{n,-}\right)\ge d_j^n\right) \tag{A.7}$$

holds.

By Lemma 1, we can choose K large enough so that for all n

$$P\left(\max_{j=1,\ldots,\lfloor nt\rfloor}\frac{1}{\sqrt{n}}V^n\left(t_j^{n,-}\right) > K\right) < \frac{\delta\epsilon^2}{4\bar{c}t}.$$

Since F is the distribution function of a positive random variable and $p > 1/2$, we can choose n large enough so that

$$F(Kn^{1/2-p}) < \frac{\delta\epsilon^2}{4\bar{c}t}.$$

Therefore, for each $j = 1, \ldots, \lfloor nt\rfloor$,

$$\begin{aligned} P\left(V^n\left(t_j^{n,-}\right)\ge d_j^n\right) &= P\left(\frac{1}{\sqrt{n}}V^n(t_j^-)\ge n^{p-1/2}w_j\right) \\ &\le P(n^{p-1/2}w_j\le K) + \frac{\delta\epsilon^2}{4\bar{c}t} \\ &\le \frac{\delta\epsilon^2}{2\bar{c}t}, \end{aligned}$$

and so, from (A.7)

$$P\left(\max_{i=1,\ldots,\lfloor nt\rfloor}\left|M_d^n(i)\right| > \epsilon\sqrt{n}\right) \leq \frac{\bar{c2}}{n\epsilon^2}\lfloor nt\rfloor\frac{\delta\epsilon^2}{2\bar{c}t} = \delta,$$

as required to complete the proof. □

Acknowledgments

We would like to thank Hong Chen, Tom Kurtz and Shane Henderson for helpful discussions relating to the issues discussed in this paper. We would also like to thank Dave Krahl for help with building our simulation models.

References

[1] : 2000, Extend: Professional simulation tools. Imagine That, 6830 Via Del Oro, Suite 230, San Jose, CA 95119, version 5 edition.
[2] : 2002, Cisco: Behind the hype. *Business Week*.
[3] C.J. Ancker and A.V. Gafarian, Queueing with impatient customers who leave at random, Journal of Industrial Engineering 13 (1962) 84–90.
[4] F. Baccelli, P. Boyer and G. Hebuterne, Single-server queues with impatient customers, Adv. Appl. Prob. 16 (1984) 887–905.
[5] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao, Statistical analysis of a telephone call center: A queueing-science perspective, Working Paper, (2002).
[6] S. Browne and W. Whitt, Piecewise-linear diffusion processes, in: *Advances in Queueing: Theory, Methods, and Open Problems*, ed. J. Dshalalow (CRC Press, 1995) pp. 463–480.
[7] H. Chen, Generalized regulated mapping: Fluid and diffusion limits, Notes prepared for Avi Mandelbaum, 1990.
[8] H. Chen, and D.D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization* (Springer-Verlag, New York, 2001).
[9] E. Coffman, A. Puhalskii, M. Reiman and P. Wright, Processor-shared buffers with reneging, Performance Evaluation 19 (1994) 25–46.
[10] B. Doytchinov, J. Lehoczky and S. Shreve, Real-time queues in heavy traffic with earliest-deadline-first queue discipline, Annals of Applied Probability 11 (2001) 332–378.
[11] D. Gamarnik and A. Zeevi, Validity of heavy traffic steady-state approximations in open queueing networks, Working Paper, 2004.
[12] O. Garnett, A. Mandelbaum and M. Reiman, Designing a call center with impatient customers, Manufacturing and Service Operations Management 4 (2002) 208–227.
[13] P. Hall and C.C. Heyde, Martingale Limit Theory and its Application (Academic Press, Inc., Boston, 1980).
[14] J.M. Harrison, *Brownian Motion and Stochastic Flow Systems* (John Wiley & Sons, New York, 1985).
[15] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, I and II, Adv. Appl. Prob. 2 (1970) 150–177 and 355–364.
[16] R. Lillo and M. Martin, Stability in queues with impatient customers, Stochastic Models 17 (2001).
[17] A. Mandelbaum and S. Zeltyn, The impact of customers' patience on delay and abandonment: Some empirically-driven experiments with the M/M/N + G queue, OR Spectrum 26(3) (2004) 377–411. Special Issue on Call Centers.

[18] C. Palm, Etude des delais d'attente, Ericson Technics 5 (1937) 37–56.
[19] E.L. Plambeck, S. Kumar and J.M. Harrison, A multiclass queue in heavy traffic with throughput time constraints: Asymptotically optimal dynamic controls, Queueing Systems 39 (2001) 23–54.
[20] J. Reed and A.R. Ward, A diffusion approximation for a generalized Jackson network with reneging, in: *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*, 2004.
[21] M.I. Reiman, Some diffusion approximations with state space collapse, in: *Lecture Notes in Control and Information Sciences*, eds. F. Baccelli and G. Fayolle (Springer, 1984) vol. 60 pp. 209–240.
[22] R.E. Stanford, Reneging phenomena in single channel queues, Mathematics of Operations Research 4 (1979) 162–178.
[23] A.R. Ward and P.W. Glynn, A diffusion approximation for a Markovian queue with reneging, Queueing Systems 43 (2003a) 103–128.
[24] A.R. Ward and P.W. Glynn, Properties of the reflected ornstein-uhlenbeck process, Queueing Systems 44 (2003b) 109–123.
[25] W. Whitt, Improving service by informing customers about anticipated delays, Management Science 45(2) (1999) 192–207.