HEAVY-TRAFFIC LIMITS FOR THE $G/H_2^*/n/m$ QUEUE

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027

July 5, 2002

Revision: February 13, 2004

(Submitted to *Mathematics of Operations Research*)

Abstract

We establish heavy-traffic stochastic-process limits for queue-length, waiting-time and overflow stochastic processes in a class of $G/GI/n/m$ queueing models with n servers and m extra waiting spaces. We let the arrival process be general, only requiring that it satisfy a functional central limit theorem. In order to capture the impact of the service-time distribution beyond its mean within a Markovian framework, we consider a special class of service-time distributions, denoted by H_2^*, which are mixtures of an exponential distribution with probability p and a unit point mass at 0 with probability $1-p$. These service-time distributions exhibit relatively high variability, having squared coefficients of variation greater than or equal to one. As in Halfin and Whitt (1981), Puhalskii and Reiman (2000) and Garnett, Mandelbaum and Reiman (2000), we consider a sequence of queueing models indexed by the number of servers, n, and let n tend to infinity along with the traffic intensities ρ_n so that $\sqrt{n}(1-\rho_n) \to \beta$ for $-\infty < \beta < \infty$. To treat finite waiting rooms, we let $m_n/\sqrt{n} \to \kappa$ for $0 < \kappa \leq \infty$. With the special H_2^* service-time distribution, the limit processes are one-dimensional Markov processes, behaving like diffusion processes with different drift and diffusion functions in two different regions, above and below zero. We also establish a limit for the $G/M/n/m+M$ model, having exponential customer abandonments.

American Mathematical Society 1991 subject classifications. Primary 60K25, 60F17; Secondary 60K30.

Keywords: queues, multiserver queues, stochastic-process limits, heavy-traffic, Halfin-Whitt regime, diffusion approximations, abandonments, reneging, customer impatience.

1. Introduction

Our goal in this paper is to establish new heavy-traffic stochastic-process limits for multi-server queues in which the number of servers is allowed to increase along with the traffic intensity. Such limits were established for the $GI/M/n/\infty$ queueing model (with renewal arrival process, exponential service times, n servers, unlimited waiting room and first-come first-served service discipline) by Halfin and Whitt (1981), for the more general $GI/PH/n/\infty$ model (with phase-type service times) by Puhalskii and Reiman (2000) and for the $M/M/n/\infty + M$ model with exponential customer abandonment by Garnett, Mandelbaum and Reiman (2000). They considered a sequence of models indexed by the number of servers, n, and let $n \to \infty$ with the traffic intensities ρ_n converging to 1, the critical value for stability. Interesting nondegenerate limits occur when

$$\sqrt{n}(1 - \rho_n) \to \beta \quad \text{for} \quad -\infty < \beta < \infty \ . \tag{1.1}$$

(The systems without customer abandonment are stable with proper steady-state distributions only when $\beta > 0$.)

We obtain more general results by allowing a non-renewal arrival process and a finite waiting room, but we only consider a special class of GI service-time distributions: The non-exponential service-time distribution we consider is the mixture of an exponential distribution with probability p and a unit point mass at 0 with probability $1 - p$. This special service-time distribution is mathematically appealing because, just like the exponential service-time distribution, it makes appropriate queue-length processes Markov processes in the renewal-arrival case, and because it leads to a one-dimensional limiting Markov process in the stochastic-process limit. Interestingly, the limit process is not directly a diffusion process, because of anomalous behavior at an interior boundary point, but it is a convex piecewise-linear function of a diffusion process, which is quite tractable.

We want to analyze the $G/GI/n/m$ model with the special H_2^* service-time distribution because, even though the service-time distribution is special, it may provide insight into the way performance depends on the service-time distribution beyond its mean. Indeed, we exploit the heavy-traffic stochastic-process limits here in a companion paper, Whitt (2004), to support a heuristic approximation for the queue-length process and its steady-state distribution in the more general $G/GI/n/m$ model with general service-time distribution. That approximation is asymptotically correct in the regime (1.1) for the $G/H_2^*/n/m$ special case. Whitt (2004) examines the quality of approximations for basic steady-state performance measures, using

results from simulations and numerical algorithms.

Since the special service-time distribution is an extremal distribution among the class of hyperexponential (H_2, mixtures of two exponentials) distributions, see Whitt (1984b), we denote this class by H_2^*. Whitt (1983) observed that H_2^* service-time distributions are convenient for developing explicit closed-form expressions for performance measures in the $M/GI/n/\infty$ model. For example, he showed that the steady-state delay probability with the H_2^* service-time distribution is independent of the parameter p, provided that the mean service time is held fixed.

Puhalskii and Reiman (2000) already established many-server heavy-traffic limits for the $GI/PH/n/\infty$ model with phase-type service-time distributions, but the limit process there is a complicated multidimensional diffusion process, whose steady-state distribution remains to be determined. The standard H_2 distributions are a subclass of the PH distributions, and so are covered by the results in Puhalskii and Reiman (2000), but the case H_2^* is not covered, because their analysis makes use of the fact that the component exponential distributions have positive mean (and thus finite rate). Indeed, going from H_2 to H_2^* lowers the dimension of the limiting Markov process from two-dimensional to one-dimensional.

To treat a finite waiting room in the heavy-traffic regime (1.1), it is necessary to let $m_n \to \infty$ as $n \to \infty$ so that

$$m_n/\sqrt{n} \to \kappa \quad \text{for} \quad 0 < \kappa \leq \infty \ . \tag{1.2}$$

The case of a finite waiting room is not discussed in Halfin and Whitt (1981). Even for $GI/M/n/m$, a different proof is required for the heavy-traffic limit, because the finite waiting room introduces a reflecting upper barrier in the diffusion process, which cannot be represented simply as a reflection map applied to an unreflected free process. For the $M/M/n/m$ model, related heavy-traffic limits have been established by Massey and Wallace (2002).

Motivated by Garnett, Mandelbaum and Reiman (2003) and Ward and Glynn (2001), in this paper we also establish a stochastic-process limit for the $G/M/n/m$ model with exponential customer abandonment (the $G/M/n/m+M$ model): each customer that must wait in queue before beginning service abandons after an exponential time with mean θ^{-1} if service has not begun by that time. (The extension to H_2^* service times remains an open problem.) The stochastic-process limit is similar to the previous $G/M/n/m$ limit: The exponential customer abandonment only changes the drift for $x > 0$ from constant to linear.

Here is how the rest of the present paper is organized: We state the stochastic-process limits for the $G/H_2^*/n/m$ model in Section 2 and the extension to allow exponential customer

abandonment in Section 3. We provide proofs in Sections 4–6.

2. The Stochastic-Process Limit with H_2^* Service Times

In this section we formulate the heavy-traffic stochastic-process limits for the $G/H_2^*/n/m$ model. We construct a sequence of these $G/H_2^*/n/m$ models indexed by the number of servers, n, and let $n \to \infty$. We let the associated sequence of traffic intensities $\{\rho_n : n \geq 1\}$ approach 1 and the associated sequence of waiting-room sizes $\{m_n : n \geq 1\}$ approach infinity so that (1.1) and (1.2) hold.

We start with a rate-1 arrival counting process $C \equiv \{C(t) : t \geq 0\}$ with associated interarrival times $\{U_k : k \geq 1\}$. Our key assumption about C is that it satisfies a functional central limit theorem (FCLT). To state the assumed limit, let $\Rightarrow$ denote convergence in distribution and let $D \equiv D([0,\infty),\mathbb{R})$ be the function space of right-continuous real-valued functions on the positive halfline with left limits, endowed with the customary Skorohod (J_1) topology; see Billingsley (1999) and Whitt (2002). Since we frequently refer to Whitt (2002), we refer to it by its title initials "SPL".

Let $\mathbf{C}_n$ be the random element of D defined by

$$\mathbf{C}_n(t) \equiv [C(nt) - nt]/\sqrt{n}, \quad t \geq 0. \tag{2.1}$$

We assume that

$$\mathbf{C}_n \Rightarrow \mathbf{C} \equiv \sqrt{c_a^2}\mathbf{B} \quad \text{in} \quad (D, J_1) \tag{2.2}$$

for some nonnegative scaling constant c_a^2, where $\mathbf{B}$ is standard (zero drift, unit diffusion coefficient) Brownian motion. When the arrival process is a renewal process, the limit (2.2) holds with c_a^2 being the squared coefficient of variation (SCV, variance divided by the square of the mean) of an interarrival time (then assumed to be finite), but the limit (2.2) holds much more generally; see Corollary 7.3.1 of SPL.

When the number of servers is n, we scale time in the arrival process, letting the arrival process be

$$C_n(t) \equiv C(\lambda_n t), \quad t \geq 0 \ , \tag{2.3}$$

where λ_n is the arrival rate in model n (with n servers). Equivalently, the interarrival times in model n are

$$U_{n,k} \equiv U_k/\lambda_n \ . \tag{2.4}$$

As a consequence, assuming that $\lambda_n/n\mu \to 1$, we have the associated limit

$$\mathbf{C}'_n \Rightarrow \mathbf{C}' \equiv \sqrt{\mu c_a^2}\mathbf{B} \quad \text{in} \quad (D, J_1) \tag{2.5}$$

where

$$\mathbf{C}'_n(t) \equiv \frac{C_n(t) - \lambda_n t}{\sqrt{n}}, \quad t \geq 0 \ . \tag{2.6}$$

Let the H_2^* service-time distribution be independent of n. With probability p, it is an exponential with mean ν^{-1}; with probability $1-p$ it is 0. It has mean $\mu^{-1} = p\nu^{-1}$, so that the traffic intensity as a function of n is $\rho_n = \lambda_n/\mu n$. The second moment of a service time is thus $2p\nu^{-2}$, so that the SCV is $c_s^2 = (2/p) - 1$. Equivalently, $p^{-1} = (c_s^2 + 1)/2$. The SCV c_s^2 ranges from 1 to ∞ as p decreases from 1 to 0. Hence, the variability of the H_2 distribution is greater than or equal to that of an exponential distribution.

Let $Q_n(t)$ be the queue length at time t, by which we mean the number in system, including both waiting and in service. We assume that the stochastic process Q_n almost surely has sample paths in the function space D; in particular, the process Q_n provides no record of an arrival with zero service time that can enter service upon arrival and depart immediately. Let $Q_n^a(k)$ be the queue length just before the k^{th} (potential) arrival, including all arrivals up to number $k-1$ if there are batch arrivals. The arrival is a potential arrival, because it may leave immediately upon arrival if it has a zero service time and there is a free server or if the system has finite capacity and is full at that arrival epoch, in which case the customer is blocked and lost (without affecting future arrivals). Customers with zero service times are all counted by the discrete-time process Q_n^a.

For the stochastic-process limit, we construct scaled random elements of D by letting

$$\begin{aligned} \mathbf{Q}_n(t) &\equiv [Q_n(t) - n]/\sqrt{n}, \\ \mathbf{Q}_n^a(t) &\equiv [Q_n^a(\lfloor nt \rfloor) - n]/\sqrt{n}, \quad t \geq 0 \ . \end{aligned} \tag{2.7}$$

There is no time scaling for $\mathbf{Q}_n$ in (2.7) because the arrival rate λ_n is allowed to grow directly.

We also must specify the initial conditions, which could be complicated because of the general arrival process. In standard heavy-traffic limits for the $G/GI/n/\infty$ model with a fixed number of servers, it is common to start the system empty. However, with the scaling in (2.7), where $n \to \infty$, it is convenient to let $Q_n(0) = n$. Alternatively, we could let $Q_n(0) = \lfloor n + x\sqrt{n} \rfloor \vee 0$ for some real number x, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x and $x \vee 0 = \max\{0, x\}$. More generally, we let $Q_n(0)$ be an integer-valued random variable

with

$$0 \le Q_n(0) \le n + m_n \tag{2.8}$$

that is independent of the arrival process $\{C_n(t) : t \ge 0\}$ and we assume that

$$\mathbf{Q}_n(0) \Rightarrow \mathbf{Q}(0) \quad \text{as} \quad n \to \infty \ , \tag{2.9}$$

where $\mathbf{Q}(0)$ is a proper random variable and

$$\mathbf{Q}_n(0) \equiv [Q_n(0) - n]/\sqrt{n} \ . \tag{2.10}$$

We also let $Q_n^a(0) = Q_n(0)$ and $\mathbf{Q}_n^a(0) = \mathbf{Q}_n(0)$. Moreover, we assume that the $\min\{n, Q_n(0)\}$ customers initially in service have exponential service times with mean ν^{-1}, while the $[Q_n(0) - n]^+$ customers initially waiting in queue have the H_2^* cdf. (That is, we assume that customers with zero service times would already have left if they could be in service.) Finally, given that specification, we assume that all service times are independent of the initial state $Q_n(0)$ and the arrival process.

Let $D^2 \equiv D \times D$ be the product space with the associated product topology. As indicated above, we use the standard J_1 topology on each coordinate, but the specific Skorohod topology (e.g., J_1 or M_1) does not matter because the limit process has continuous sample paths. Indeed, the topology could be the J_1 or M_1 topology on $D([0,\infty), \mathbb{R}^2)$; see Sections 3.3 and 11.5 and Chapter 12 of SPL. Let $\mathbf{e}$ be the identity function in D, i.e., $\mathbf{e}(t) = t$, $t \ge 0$. Let $\circ$ be the composition map, defined by $(x \circ y)(t) \equiv x(y(t))$; see Section 13.2 of SPL.

Theorem 2.1. *For the family of $G/H_2^*/n/m$ models specified above, where the rate-1 arrival process obeys the FCLT in* (2.2)*, suppose that the arrival rate λ_n and the number of waiting spaces, m_n, change with n so that* (1.1) *and* (1.2) *hold with $-\infty < \beta < \infty$ and $0 < \kappa \le \infty$. In addition, suppose that the initial conditions are as specified above with* (2.8)-(2.10)*. Then*

$$(\mathbf{Q}_n, \mathbf{Q}_n^a) \Rightarrow (\mathbf{Q}, \mathbf{Q}^a) \quad \textit{in} \quad (D, J_1)^2 \quad \textit{as} \quad n \to \infty \ , \tag{2.11}$$

where

$$\mathbf{Q}(t) \equiv g(\mathbf{Q}^p(t)), \quad t \ge 0, \tag{2.12}$$

$$g(x) \equiv \begin{cases} x, & x < 0, \\ x/p, & 0 \le x \le p\kappa \ , \end{cases} \tag{2.13}$$

$$\mathbf{Q}^a \equiv \mathbf{Q} \circ \mu^{-1}\mathbf{e} \quad \textit{and} \ , \tag{2.14}$$

and $\mathbf{Q}^p$ *is a diffusion process starting at* $\mathbf{Q}^p(0) = g^{-1}(\mathbf{Q}(0))$ *with a reflecting upper barrier at* $p\kappa$ *if* $\kappa < \infty$ *and an inaccessible upper boundary at infinity if* $\kappa = \infty$. *The diffusion process* $\mathbf{Q}^p$ *has infinitesimal mean* (*drift function*)

$$m_{\mathbf{Q}^p}(x) = \begin{cases} -p\mu\beta, & 0 \le x < p\kappa, \\ -p\mu(x+\beta), & x < 0 \ , \end{cases} \tag{2.15}$$

and infinitesimal variance (*diffusion function*)

$$\sigma^2_{\mathbf{Q}^p}(x) = p^2\mu(c_a^2 + (2/p) - 1) = p^2\mu(c_a^2 + c_s^2), -\infty \le x < p\kappa \ . \tag{2.16}$$

Remark 2.1. *The superscript p.* The limit process $\mathbf{Q}^p$ in Theorem 2.1 has a natural physical interpretation: It is the limit process for the scaled version of the queue-length process $\{Q^p(t) : t \ge 0\}$ containing only the customers with positive (non-zero) service times, ignoring the customers with zero service times. When all servers are not busy, we can ignore the customers with zero service times because they leave immediately upon arrival, and $Q(t)$ does not record their appearance. Except for the upper barrier at m_n, the customers with zero service times have no impact on other customers. To obtain the limit process $\mathbf{Q}^p$ directly in the case $m_n = \kappa = \infty$, we ignore the customers with zero service times, giving us the stochastic process $\{Q^p(t) : t \ge 0\}$, which corresponds to the queue-length process $\{Q(t) : t \ge 0\}$ in the $G/M/n/\infty$ model where $p = 1$, but with different parameters. Thus, in the special case of GI arrivals and unlimited waiting space, the limit for the scaled version of $\{Q^p(t) : t \ge 0\}$ is a consequence of Halfin and Whitt (1981) and Puhalskii and Reiman (2000). For the limiting diffusion process $\mathbf{Q}^p$, the extension to a finite upper barrier κ is obtained by inserting a reflecting upper barrier at κ; see Remark 2.5 for a discussion of the construction. We do not actually prove that $\mathbf{Q}^p_n \to \mathbf{Q}^p$ in the finite-waiting-room case here; that remains an open problem.

Remark 2.2. $\mathbf{Q}$ *and* $\mathbf{Q}^a$ *are not diffusion processes.* Since the function g in (2.13) is not differentiable at 0 (and has a discontinuous derivative using one-sided derivatives), the limit processes $\mathbf{Q}$ and $\mathbf{Q}^a$ are *not* diffusion processes with the common definitions; e.g., see p. 110 of Rogers and Williams (1987) and p. 159 of Karlin and Taylor (1981). The limit processes $\mathbf{Q}$ and $\mathbf{Q}^a$ are strong Markov processes with continuous sample paths, but the infinitesimal mean and variance are not well defined in state 0. However, the function g is a convex function, so that the limit processes $\mathbf{Q}$ and $\mathbf{Q}^a$ can be characterized as stochastic integrals, using a generalized Itô rule for convex functions based on Tanaka's formula; e.g., see Sections 43, 45 and 47 of Rogers and Williams (1987). Indeed, by Theorem 45.1 of Rogers and Williams (1987), $\mathbf{Q}$ can

be represented as the stochastic integral

$$\begin{aligned}\mathbf{Q}(t) &= g(\mathbf{Q}^p(t)) \\ &= g(\mathbf{Q}^p(0)) + \int_0^t [1_{(-\infty,0]}(\mathbf{Q}^p(s)) + (1/p)1_{(0,p\kappa]}(\mathbf{Q}^p(s))]d\mathbf{Q}^p(s) \\ &\quad + \frac{1-p}{2p} L_{\mathbf{Q}^p}(t,0) \ , \end{aligned} \tag{2.17}$$

where $1_A(x)$ is the indicator function, with $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ otherwise, and $L_{\mathbf{Q}^p}(t,0)$ is the local time in state 0 of the diffusion process $\mathbf{Q}^p$ (with infinitesimal parameters in (2.15) and (2.16)). In turn, by Theorem 49.1 of Rogers and Williams (1987), the local time of the diffusion process $\mathbf{Q}^p$ is a time change of the local time of standard Brownian motion $\mathbf{B}$, i.e.,

$$L_{\mathbf{Q}^p}(t,0) = L_{\mathbf{B}}(\gamma(t),0) \tag{2.18}$$

for appropriate time-change function $\gamma(t)$ fully specified there.

Remark 2.3. *Tractability.* It is evident that the limit processes $\mathbf{Q}$ and $\mathbf{Q}^a$ are quite tractable due to the representation in (2.12) - (2.14). First, it is easy to obtain the steady-state distributions from the steady-state distribution of $\mathbf{Q}^p$. We do not give details here, because the steady-state distribution is discussed extensively in Whitt (2004). It also follows that the limit processes $\mathbf{Q}$ and $\mathbf{Q}^a$ act like diffusion processes away from the origin. Away from the origin, the process $\mathbf{Q}$ has infinitesimal mean (drift function)

$$m_{\mathbf{Q}}(x) = \begin{cases} -\mu\beta, & 0 < x < \kappa, \\ -p\mu(x+\beta), & x < 0 \ , \end{cases} \tag{2.19}$$

and infinitesimal variance (diffusion function)

$$\sigma^2_{\mathbf{Q}}(x) = \begin{cases} \mu(c_a^2 + (2/p) - 1) = \mu(c_a^2 + c_s^2), & 0 < x < \kappa, \\ p^2\mu(c_a^2 + 2 - p) = p^2\mu(c_a^2 + c_s^2), & x < 0 \ . \end{cases} \tag{2.20}$$

However, the infinitesimal parameters are not well defined at 0. For example, in the $M/H_2^*/n/m$ special case, from state n the process $Q_n(t)$ has a drift up of λ_n, but a drift down of $p\mu n$. If we could regard the process $\mathbf{Q}$ as a diffusion process with the infinitesimal parameters in (2.19) and (2.20), extended to 0, then the diffusion process would be a piecewise-linear diffusion (like $\mathbf{Q}^p$) as in Browne and Whitt (1995), and we could directly write down the steady-state distribution. However, since $\mathbf{Q}$ is not actually a diffusion, that alleged steady-state distribution for $\mathbf{Q}$ is not correct.

Remark 2.4. *Different speeds in different regions.* The infinitesimal variance $\sigma^2_{\mathbf{Q}}(x)$ in (2.20) is discontinuous at $x = 0$ when $p < 1$: $\sigma^2_{\mathbf{Q}}(0-) = p^2\sigma^2_{\mathbf{Q}}(0+)$, so that the limit process $\mathbf{Q}$ "moves

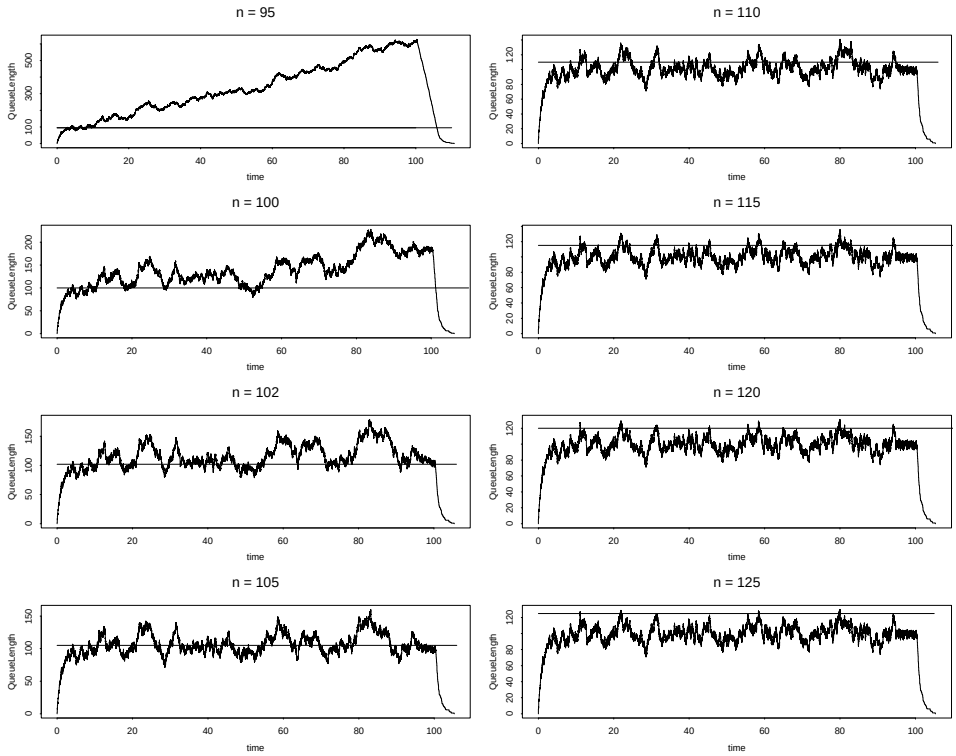

Figure 1: Sample paths of the queue-length process for 10^4 arrivals in the $M/M/n$ queue with arrival rate $\lambda = 100$, service rate $\mu = 1$ and several values of n. A common realization of the arrival process and service times is used for all n.

faster" when $x > 0$. That difference in the infinitesimal variances is evident from plots of queue-length sample paths obtained from simulations. To demonstrate that property, we plot sample paths of the queue-length process for 10^4 arrivals in the $M/M/n/\infty$ and $M/H_2^*/n/\infty$ models with $\lambda = 100$, $\mu = 1$, $p = 0.1$ and several values of n in Figures 1 and 2. For the $M/H_2^*/n/\infty$ model with $p = 0.1$, the infinitesimal variance of $\mathbf{Q}$ is $\sigma^2_{\mathbf{Q}}(x) = 2p\mu$ for $x < 0$ and $\sigma^2_{\mathbf{Q}}(x) = 2\mu/p$ for $x > 0$. Hence the ratio of the infinitesimal variances in the two regions is $p^2 = 0.01$. The difference is striking in the plots.

For the simulation, the same arrival process sample path is used for all plots, and the same service-time realizations are used for different n in each separate queueing system. Consistent with the steady-state distribution described in Whitt (2004), the steady-state probability that all servers are busy tends to be no greater for the more highly-variable H_2^* service times than for the exponential service times. Indeed, for $n = 120$ in these plots, no customers are delayed for H_2^* service times, whereas some are for exponential (M) service times.

Remark 2.5. *Constructing* $\mathbf{Q}^p$. The key limit process $\mathbf{Q}^p$ in Theorem 2.1 is a diffusion process on the interval $(-\infty, p\kappa)$ with reflection at the upper barrier when $\kappa < \infty$. It is of course

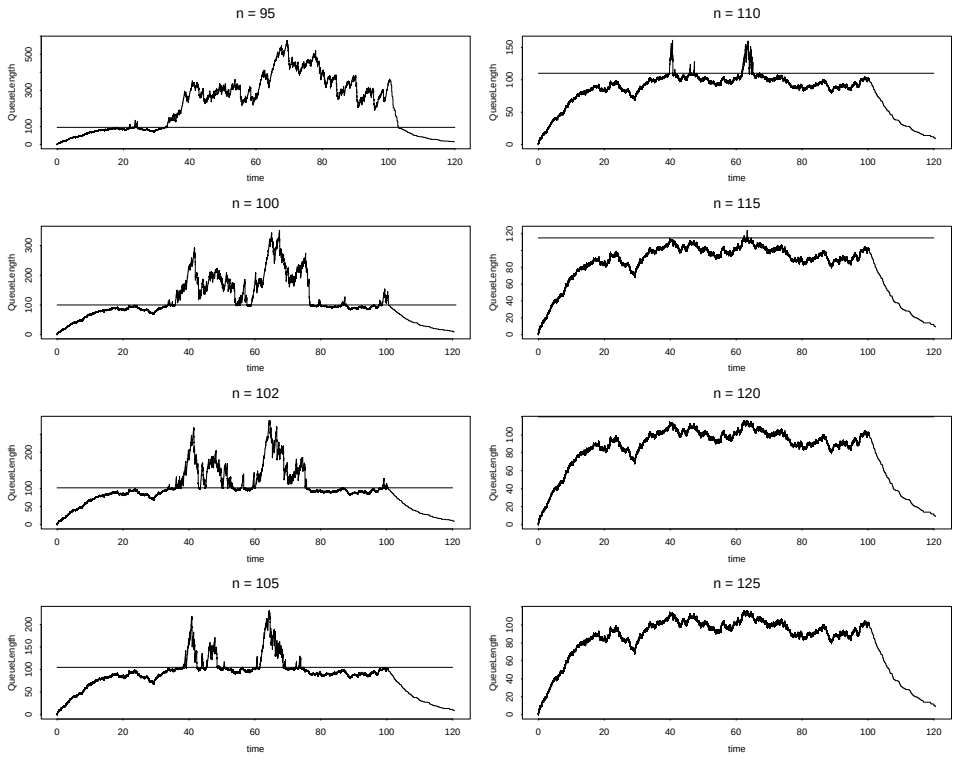

Figure 2: Sample paths of the queue-length process for 10^4 arrivals in the $M/H_2^*/n$ queue with arrival rate $\lambda = 100$, $\mu = 1$ and $p = 0.1$ (SCV $c_s^2 = 19$) and several values of n. A common realization of the arrival process and service times is used for all n.

important that this limiting diffusion process be well defined. Constructing this diffusion process is somewhat complicated when $\kappa < \infty$, because it cannot be regarded simply as the image of an "unreflected free process" under a reflection map, as in Sections 5.2, 9.3, 13.5 and 14.8 of SPL. There are several ways to do the construction. One is to rely on an asymptotic construction of the reflected process from an associated unreflected process on $(-\infty, \infty)$ as in the proof of Theorem 4.1 of Srikant and Whitt (1996). That construction characterizes the probability law of the reflected process as the common limit (in distribution) of two converging sequences of bounding processes. These bounding processes have small jumps into the interior of the state space the instant the boundary is hit.

A second approach is to directly apply the standard reflection map in the neighborhood of the upper barrier. That second approach is useful to construct an approximation for the overflow process in the queueing model (recording arrivals turned away because the waiting room is full), which we do in the next corollary. To do that construction, we can use the following "alternating-renewal-process" construction: We let the reflected diffusion process be distributed as the unreflected diffusion process until the first time the upper barrier is hit. Since the diffusion process has constant drift for states in the interval $(0, p\kappa)$, we can then let the reflected diffusion process be reflected Brownian motion (with one-sided reflection down from the upper barrier) until a state b is next hit, with $0 < b < p\kappa$, using the usual construction involving the reflection map; see Chapters 5 and 9 in SPL. The approximation for the losses in the queueing model is determined by the upper-barrier regulator map associated with the reflection map in the random intervals during which the process acts as reflected Brownian motion. After hitting the state b again, we repeat the construction above. For further discussion about constructing diffusion processes, see Lions and Sznitman (1984), Stroock and Varadhan (1979) and Rogers and Williams (1987). ∎

We now state some corollaries. Our first is for the loss processes when $\kappa < \infty$. Let $L_n(t)$ be the number of customers lost (blocked) in the interval $[0, t]$ and let $L_n^a(k)$ be the number of customers lost among the first k arrivals. Paralleling (2.7), let the associated scaled processes be

$$\begin{aligned} \mathbf{L}_n(t) &\equiv L_n(t)/\sqrt{n}, \\ \mathbf{L}_n^a(t) &\equiv L_n^a(\lfloor nt \rfloor)/\sqrt{n}, \quad t \geq 0 \ . \end{aligned} \tag{2.21}$$

We construct the loss process associated with the limiting diffusion process in Theorem 2.1 by using the reflection map in the "alternating-renewal-process" framework specified in Remark

2.5.

Corollary 2.1. *If, in addition to the assumptions of Theorem* 2.1, $\mathbf{L}_n(0) = \mathbf{L}_n^a(0) = 0$ *w.p.1 and* $0 < \kappa < \infty$, *then*

$$(\mathbf{L}_n, \mathbf{L}_n^a) \Rightarrow (\mathbf{L}, \mathbf{L}^a) \quad in \quad (D, J_1)^2 \tag{2.22}$$

jointly with (2.11), *where* $\mathbf{L}_n$ *and* $\mathbf{L}_n^a$ *are as in* (2.21), $\mathbf{L}^a = \mathbf{L} \circ \mu^{-1}\mathbf{e}$ *and* $\mathbf{L}$ *is constructed as indicated above in Remark* 2.5.

We now state a corollary for the waiting time and virtual waiting time. Let $W_n(k)$ be the waiting time of the k^{th} *admitted* customer (before beginning service) and let W_n^v be the virtual waiting time (the time required for all the customers in the queue at time t to begin service) in model n. Since there are n servers, the waiting time $W_n(k)$ tends to be about $[Q_n(l) - n]^+/n\mu$, where l is the index of the k^{th} admitted customer. (In the limit the proportion of admitted customers approaches 1, so the shift in index is asymptotically negligible.) Thus, for the stochastic-process limit, we need to scale the waiting times by *multiplying* by $\sqrt{n}$ instead of dividing by $\sqrt{n}$ as in (2.7).

Let

$$\begin{aligned} \mathbf{W}_n(t) &\equiv \sqrt{n} W_n(\lfloor nt \rfloor), \\ \mathbf{W}_n^v(t) &\equiv \sqrt{n} W_n^v(t), \quad t \geq 0 \ . \end{aligned} \tag{2.23}$$

For $x \in D$, let $x \vee 0$ be the element of D defined by

$$(x \vee 0)(t) \equiv x(t) \vee 0 \equiv \max\{x(t), 0\}, \quad t \geq 0 \ . \tag{2.24}$$

The following result is established very similarly to Corollary 2.3 of Puhalskii and Reiman (2000); we give details in Section 4.

Corollary 2.2. *Under the conditions of Theorem* 2.1,

$$(\mathbf{W}_n, \mathbf{W}_n^v) \Rightarrow (\mu^{-1}\mathbf{Q}^a \vee 0, \mu^{-1}\mathbf{Q} \vee 0) \quad in \quad (D, J_1)^2 \quad as \quad n \to \infty \ , \tag{2.25}$$

where $(\mathbf{Q}^a, \mathbf{Q})$ *is as in Theorem* 2.1.

3. Extension for Customer Abandonments

As in Garnett et al. (2003), suppose that each customer that joins the queue before receiving service abandons, independently of all other events, after an exponential time with

mean θ^{-1} if service has not begun before that time. We now extend Theorem 2.1 to this setting for the special case of exponential service times. (The extension to H_2^* service times remains an open problem.)

Theorem 3.1. *For the $G/M/n/m_n + M$ model with exponential customer abandonment as specified above, under the conditions of Theorem* 2.1*, the conclusions of Theorem* 2.1 *hold with two modifications: First, here $p = 1$; second, the infinitesimal mean for $\mathbf{Q} \equiv \mathbf{Q}^p$ should be changed to*

$$m_{\mathbf{Q}}(x) = \begin{cases} -\beta\mu - \theta x, & 0 \le x < \kappa, \\ -\beta\mu - \mu x, & x < 0 \ . \end{cases} \tag{3.1}$$

4. Proof of Theorem 2.1

4.1. Outline of the Proof

In this subsection we give a high-level view. Our proof of Theorem 2.1 has three steps:

Step 1. $G/M/n/\infty$**.** We first establish a stochastic-process limit under two extra restrictions: (i) We consider only the customers with positive service times, and (ii) we assume an unlimited waiting room. We show that this first step is equivalent to establishing the heavy-traffic stochastic-process limit for the $G/M/n/\infty$ model, which then requires only a slight generalization of the results by Halfin and Whitt (1981) and Puhalskii and Reiman (2000) (restricted to the special case of exponential service times). The $G/M/n/\infty$ result is only more general because the arrival process need not be a renewal process, but that is a useful generalization for applications.

We actually give two different proofs of the $G/M/n/\infty$ result. Since the service times are exponential in this step, the second proof extends to the $G/M/n/m_n + M$ model with exponential abandonments and finite waiting room, thus yielding a proof of Theorem 3.1.

Proof 1. Piecewise Construction for $G/M/n/m_n$ Exploiting Previous Results. Our first proof is a piecewise construction for $G/M/n/m_n$, with finite waiting room, exploiting established results for the more elementary $G/M/\infty$ infinite-server and $G/M/1/m_n$ single-server models. The $G/M/\infty$ model applies below state n when not all servers are busy, while the $G/M/1/m_n$ model applies above n when all servers are busy. In each separate region we can apply previous results for these more elementary systems. We recursively establish limits in the different regions, letting the end of the previous excursion in the other region serve as

the initial distribution for the next excursion in the new region. Then we show that the pieces can be put together to imply convergence for the entire process. The piecewise construction is interesting in part because it can be applied in other contexts. Indeed, our proof in Step 3 uses a variant of the same argument.

Proof 2. Martingale Proof for $M/M/n/m_n + M$ Extended to G Arrivals. In Section 5 we also give a second proof for the $G/M/n/\infty$ model. This second proof is a martingale proof for the $M/M/n/\infty$ model in the spirit of Puhalskii and Reiman (2000), but extended to a general G arrival process. Since the service times are exponential, the model can also have finite waiting room and exponential customer abandonments, so we obtain a proof of Theorem 3.1 at the same time. The logic for the extension to G arrival processes also applies to many other contexts. In particular, the same reasoning lets us extend the results in Puhalskii and Reiman (2000) from $GI/PH/n/\infty$ to $G/PH/n/\infty$. For the general arrival process, we assume only a FCLT, as in (2.2).

Open Problems. We conjecture that the proof for the $G/M/n/m_n + M$ model (Proof 2 above) can be extended to yield corresponding direct proofs for the $G/H_2^*/n/m_n$ model (Theorem 2.1) and generalizations to, first, the $G/H_2^*/n/m_n + M$ model and, more generally the $G/H_2^*/n/m_n + H_2^*$ model. Those remain interesting open problems. Even if those direct proofs can be done, we believe that the piecewise constructions are interesting.

Step 2. $G/H_2^*/n/\infty$. We apply the $G/M/n/\infty$ result in Step 1 (not considering the extensions to finite waiting rooms and customer abandonment) to obtain the stochastic-process limit for the more general $G/H_2^*/n/\infty$ model, having H_2^* service times instead of M service times, but still an unlimited waiting room. For the case of an unlimited waiting room, we show that the distance between $\mathbf{Q}_n$ and $\mathbf{Q}_n^p$ is asymptotically negligible. Thus for the $G/H_2^*/n/\infty$ model we establish joint convergence of $(\mathbf{Q}_n, \mathbf{Q}_n^p)$.

Step 3. $G/H_2^*/n/m_n$. Finally, we apply the result for the $G/H_2^*/n/\infty$ model in Step 2 to obtain the desired stochastic-process limit for the associated $G/H_2^*/n/m_n$ model, having finite waiting room. The finite-waiting-room proof is by no means a simple extension, such as directly applying the continuous-mapping theorem with a reflection map. We use a piecewise construction as in the first proof of Step 1. It requires the same rather complicated recursive or inductive proof. To treat the finite waiting rooms, we consider two regions with boundary

above n. In the upper region, the system behaves like a $G/M^X/1/m_n$ single-server system with batch service, while in the lower region the system behaves like the $G/H_2^*/n/\infty$ system treated in Step 2. We can apply the standard one-sided reflection map associated with the upper barrier to treat each piece in the upper region, and we can apply Step 2 to treat each piece in the lower region. Thus, in both steps 2 and 3, we make strong use of the result established in the previous step. (By our argument, it is not possible to skip any steps.)

4.2. Positive Customers and an Unlimited Waiting Room.

We start with the case of an unlimited waiting room, which produces a limit process without an upper barrier. Establishing the desired limit with H_2^* service times is complicated even with an infinite waiting room because of different system behavior in two regions of the state space. If the servers are not all busy, then customers with zero service times depart immediately upon arrival. however, if all severs are busy, then customers with zero service times must join the queue. Subsequently, upon any service completion, there is a random batch of departures, because customers with zero service times that enter service at that time will also depart immediately. Hence there may be several simultaneous departures at each of these departure epochs.

With an infinite waiting room, the situation simplifies if we focus on the customers with positive (non-zero) service times. With an infinite waiting room (but not with a finite waiting room), the customers with zero service times have absolutely no impact on the customers with positive service times. Thus, with an infinite waiting room, we can focus on the customers with positive service times, by simply ignoring the customers with zero service times. We initially establish a limit for the queue-length process consisting only of the customers with positive service times. Afterwards, in Step 2, we use the limit for customers with positive service times to establish the limit for all customers (in the setting with unlimited waiting room).

When we look only at the customers with positive service times, the system behaves like a $G/M/n/\infty$ model with a new G arrival process and a new initial condition. To reduce our problem to the $G/M/n/\infty$ model, we need to show that the assumed FCLT for the arrival process and the assumed initial conditions imply corresponding behavior for the positive customers alone. We first show that the assumed FCLT for the full arrival process in (2.2) implies a corresponding FCLT for the arrival process of customers with positive service times.

Let $C^p(t)$ count the number of arrivals in the interval $[0,t]$ that have positive service times. We first observe that an analog of the FCLT assumed for the full arrival process C in (2.2)

holds for C^p under the assumption (2.2). Let $\mathbf{C}_n^p$ be the random element of D defined by

$$\mathbf{C}_n^p(t) \equiv [C^p(nt) - pnt]/\sqrt{nc_{p,a}^2} \ , \tag{4.1}$$

where the new scaling parameter is

$$c_{p,a}^2 \equiv pc_a^2 + p(1-p) \ . \tag{4.2}$$

Lemma 4.1. *If the FCLT in* (2.2) *holds, then*

$$\mathbf{C}_n^p \Rightarrow \mathbf{B} \quad in \quad (D, J_1) \ . \tag{4.3}$$

Proof. Recall that $C^p(t)$ can be written as the random sum

$$C^p(t) = \sum_{i=1}^{C(t)} Y_i \ , \tag{4.4}$$

where $\{Y_i : i \geq 1\}$ is a sequence of IID Bernoulli random variables, independent of the stochastic process C, with $P(Y_i = 1) = 1 - P(Y_i = 0) = p$, so that Y_i has mean p and variance $p(1-p)$. Hence,

$$(\sqrt{c_a^2}\mathbf{C}_n, \mathbf{S}_{Y,n}) \Rightarrow (\sqrt{c_a^2}\mathbf{B}_1, \sqrt{p(1-p)}\mathbf{B}_2) \quad \text{in} \quad (D, J_1)^2 \ , \tag{4.5}$$

where $\mathbf{C}_n$ is given in (2.1), $\mathbf{B}_1$ and $\mathbf{B}_2$ are independent standard Brownian motions and

$$\mathbf{S}_{Y,n}(t) \equiv n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} (Y_i - p), \quad t \geq 0 \ . \tag{4.6}$$

Hence we can apply the continuous mapping theorem with composition and addition to obtain the desired conclusion; specifically, we can Theorem 9.5.1 of SPL with (4.12) to obtain (4.3).

■

In the same spirit, we need to show that the initial conditions specified for $Q_n(t)$ imply corresponding initial conditions for $Q_n^p(t)$. For that purpose, let

$$\mathbf{Q}_n^p(0) \equiv [Q_n^p(0) - n]/\sqrt{n} \ . \tag{4.7}$$

Lemma 4.2. *If*

$$\mathbf{Q}_n(0) \equiv [Q_n(0) - n]/\sqrt{n} \Rightarrow \mathbf{Q}(0) \quad in \quad \mathbb{R} \quad as \quad n \to \infty \ , \tag{4.8}$$

then

$$|\mathbf{Q}_n^p(0) - g^{-1}(\mathbf{Q}_n(0))| \Rightarrow 0 \tag{4.9}$$

so that

$$\mathbf{Q}_n^p(0) \Rightarrow g^{-1}(\mathbf{Q}(0)) \quad in \quad \mathbb{R} \quad as \quad n \to \infty \ , \tag{4.10}$$

Proof. Note that $[Q_n^p(0) - n]^+$ can be written as the random sum

$$[Q_n^p(0) - n]^+ = \sum_{i=1}^{[Q_n(0)-n]^+} Y_i \ , \tag{4.11}$$

where $\{Y_i : i \geq 1\}$ is the sequence of IID Bernoulli random variables we introduced to prove Lemma 4.1. Hence,

$$\mathbf{Q}_n^p(0) - g^{-1}(\mathbf{Q}_n(0)) = n^{-1/2} \sum_{i=1}^{[Q_n(0)-n]^+} (Y_i - EY_i) \ . \tag{4.12}$$

We have $\mathbf{Q}_n^p(0) - g^{-1}(\mathbf{Q}_n(0)) = 0$ where $\mathbf{Q}_n(0) \leq 0$. Otherwise, $\mathbf{Q}_n^p(0) - g^{-1}(\mathbf{Q}_n(0))$ is asymptotically negligible. To see that, use the Skorohod representation theorem to replace the convergence $\mathbf{Q}_n(0) \Rightarrow \mathbf{Q}(0)$ by convergence w.p.1. For the case $\mathbf{Q}(0) \leq 0$, we have $\mathbf{Q}_n^p(0) = \mathbf{Q}_n(0) \to \mathbf{Q}(0) \leq 0$. Henceforth focus on the case $\mathbf{Q}(0) > 0$. For that case, we can write

$$n^{-1/2} \sum_{i=1}^{[Q_n(0)-n]^+} (Y_i - EY_i) = \frac{[Q_n(0) - n]^+}{\sqrt{n}} \frac{\sum_{i=1}^{[Q_n(0)-n]^+} (Y_i - EY_i)}{[Q_n(0) - n]^+} \ , \tag{4.13}$$

and then apply the LLN together with the assumed limit for $Q_n(0)$. That w.p.1 convergence implies convergence in law for the original versions, which is equivalent to convergence in probability because the limit is deterministic. ∎

Hence, establishing the limit for the customers with positive service times is actually equivalent to proving Theorem 2.1 for the special case of the $G/M/n/\infty$ model, i.e., with an unlimited waiting room and exponential service times.

4.3. Proof 1 in Step 1: The Piecewise-Construction Proof

Thus, to complete Step 1, it suffices to prove Theorem 2.1 for the special case with exponential (M) service and unlimited waiting room ($m_n = \infty$). However, given that we now consider exponential service times, we are also able to treat finite waiting rooms. As mentioned earlier, we give two different proofs for the $G/M/n/m_n$ result, because each is of some independent interest. The second proof appears in Section 5. There we treat customer abandonments as well; there is no customer abandonment here.

Our main result in this subsection is

Theorem 4.1. *If, in addition to the conditions of Theorem* 2.1*, the service-time distribution is exponential* ($p = 1$)*, then*

$$\mathbf{Q}_n \Rightarrow \mathbf{Q} \quad \textit{in} \quad (D, J_1) \quad \textit{as} \quad n \to \infty \ , \tag{4.14}$$

where $\mathbf{Q}$ *is the diffusion process specified in Theorem* 2.1 *with* $p = 1$*, so that* $\mathbf{Q} = \mathbf{Q}^p$*.*

As a consequence, we obtain the desired result for the customers with positive service times. To state it, let $Q_n^p(t)$ be the queue length of customers with positive service times at time t in the n^{th} model. Let $\mathbf{Q}_n^p$ be the associated random element of D defined by

$$\mathbf{Q}_n^p(t) \equiv [Q_n^p(t) - n]/\sqrt{n}, \quad t \geq 0 \ . \tag{4.15}$$

Corollary 4.1. *If, in addition to the conditions of Theorem* 2.1, $m_n = \infty$ *for all* n, *then*

$$\mathbf{Q}_n^p \Rightarrow \mathbf{Q}^p \quad in \quad (D, J_1) \quad as \quad n \to \infty \ , \tag{4.16}$$

where $\mathbf{Q}^p$ *is the diffusion process specified in Theorem* 2.1 (*with* $\kappa = \infty$).

The Main Ideas in the Proof of Theorem 4.1. Our proof is based on the recognition that the $G/M/n/m_n$ model and the limiting diffusion process have different character in two regions, with the state-dependent rates being piecewise-linear, as discussed in Halfin and Whitt (1981) and Browne and Whitt (1995). When $Q_n(t) < n$, the system behaves like the $G/M/\infty$ model; when $Q_n(t) \geq n$, the system behaves like the $G/M/1/m_n$ model. For both those component models, limits have already been established. (For the $G/M/\infty$ model, we could employ a simplification (special case) of the second (martingale) proof in Section 5 below.) Similarly, the limiting diffusion process acts like simple reflected Brownian motion (RBM) above 0 and like the Ornstein-Uhlenbeck (OU) diffusion process below 0.

The idea, then is to apply the previous limits in successive excursions above and below n. Suppose that we start above n. Then we use the known convergence for the $G/M/1/m_n$ model during the excursion above n, until $Q_n(t)$ falls below n. Then we switch over to the other region, using the terminal distribution of the process in the upper region to serve as the initial distribution for the excursion in the lower region. We apply induction to deduce that the limits hold for all such excursions, and we use the continuous mapping theorem to show that we can put all the pieces together to obtain the originally desired convergence for the full process.

However, there is a complication in the piecewise argument as just described: As stated, there are too many excursions, because the process changes back and forth quickly in the neighborhood of the boundary n (which will become 0 for the limiting diffusion process). (There will be no such difficulty in Step 3 later, because the switchover points are widely separated.) In order to circumvent this difficulty here, we modify the original processes at the boundary n. When we hit level n from above, we insert a jump down of size $\lfloor \epsilon\sqrt{n} \rfloor$; and

when we hit level n from below, we insert a jump up of size $\lfloor \epsilon\sqrt{n} \rfloor$. By inserting those jumps in the original process, we ensure that the excursions above and below n asymptotically have length of order $O(1)$. (Without inserting the jumps, the expected lengths of the excursions are of order $O(1/\sqrt{n})$.) We then carry out the piecewise constructions for the processes with these extra jumps. Afterwards, we obtain the desired result by letting $\epsilon \downarrow 0$. Conceptually, the argument is relatively simple (should already be crystal clear), but there are several technical details, which we will try to treat carefully.

Overview of the Detailed Proof of Theorem 4.1. Recall that we are now considering the $G/M/n/m_n$ model. We will establish the desired convergence $\mathbf{Q}_n \Rightarrow \mathbf{Q}$ in (4.14) by approximating the processes $\mathbf{Q}_n$ by related processes that are easier to analyze. For each $\epsilon > 0$, we will define processes $\mathbf{Q}_n^\epsilon$ such that, for all $t > 0$ and $n \geq 1$,

$$\|\mathbf{Q}_n - \mathbf{Q}_n^\epsilon\|_t \leq \epsilon, \tag{4.17}$$

where

$$\|x\|_t = \sup_{0 \leq s \leq t} \{|x(s)|\} \ , \tag{4.18}$$

and, for each $\epsilon > 0$,

$$\mathbf{Q}_n^\epsilon \Rightarrow \mathbf{Q}^\epsilon \quad \text{in} \quad (D, J_1) \quad \text{as} \quad n \to \infty \ . \tag{4.19}$$

We form $\mathbf{Q}_n^\epsilon$ by deliberately introducing jumps, so the limit processes $\mathbf{Q}^\epsilon$ do not have continuous sample paths, but they only have jumps of size ϵ.

Since the limit processes $\mathbf{Q}^\epsilon$ in (4.19) have jumps, we will need to use the nonuniform J_1 topology on D in the convergence. Given the nonuniform J_1 convergence in (4.19), it is useful to measure distance on D using a J_1 metric over the interval $[0, t]$, say d_{J_1} as in (3.2) of SPL. Let π be the Prohorov metric on the space of all probability measures on (D, J_1), using the time interval $[0, t]$ and the metric d_{J_1} on D; see (2.2) of SPL. The main property for our purposes is that it induces weak convergence. For random elements X_1 and X_2, let $\pi(X_1, X_2)$ denote the Prohorov metric applied to the probability laws of the random elements. We can apply the triangle inequality to deduce that

$$\pi(\mathbf{Q}_n, \mathbf{Q}) \leq \pi(\mathbf{Q}_n, \mathbf{Q}_n^\epsilon) + \pi(\mathbf{Q}_n^\epsilon, \mathbf{Q}^\epsilon) + \pi(\mathbf{Q}^\epsilon, \mathbf{Q}) \ . \tag{4.20}$$

Now we use the fact that

$$\begin{aligned} \pi(X_1, X_2) &\leq \inf\{c > 0 : P(d_{J_1}(X_1, X_2) > c) < c\} \\ &\leq \inf\{c > 0 : P(\|X_1 - X_2\|_t > c) < c\} \end{aligned} \tag{4.21}$$

for any random elements $X_1, X_2 \in D$, by virtue of the Strassen representation theorem, Theorem 11.3.5 of SPL. As a consequence, (4.17) implies that

$$\pi(\mathbf{Q}_n, \mathbf{Q}_n^\epsilon) \leq \epsilon \quad \text{for all} \quad n \ . \tag{4.22}$$

Hence we can apply (4.17) and (4.19) to treat the first two terms on the right in (4.20). We complete the proof by showing that

$$\mathbf{Q}^\epsilon \Rightarrow \mathbf{Q} \quad \text{as} \quad \epsilon \downarrow 0 \ , \tag{4.23}$$

which is equivalent to $\pi(\mathbf{Q}^\epsilon, \mathbf{Q}) \to 0$.

Thus, we can apply (4.22) and (4.23) to first pick ϵ to make the first and third terms on the right in (4.20) small, uniformly in n. Then, by (4.19), given that ϵ, we can choose n to make the second term arbitrarily small. In that way, we succeed in establishing the desired convergence.

Verifying (4.17): Constructing the Approximation with Jumps. To establish (4.17), we modify the unscaled process Q_n by inserting a jump up of $\lfloor \epsilon\sqrt{n} \rfloor$ whenever the sample path reaches level n from below, and a jump down of $\lfloor -\epsilon\sqrt{n} \rfloor$ whenever the sample path reaches level $n-1$ from above.

Let the associated scaled processes be

$$\mathbf{Q}_n^\epsilon \equiv [Q_n^\epsilon(t) - n]/\sqrt{n}, \quad t \geq 0 \ . \tag{4.24}$$

Clearly the scaled processes have jumps of size ϵ, at least asymptotically as $n \to \infty$.

We construct the unscaled processes Q_n^ϵ on the same sample space as Q_n so that (4.17) holds. First, we give all these processes the same sample path of arrivals. We cannot give the processes the same sample paths of departures, because they are in different states with different rates. However, we can exploit the special form of our exponential service-time distribution to perform a stochastic coupling construction with the desired property, drawing on Whitt (1981); see especially Theorem 10.

For simplicity, suppose we start at time 0 with an arrival from state n. Then Q_n^ϵ immediately has a jump up of $\lfloor \epsilon\sqrt{n} \rfloor$. It thus starts out $\lfloor \epsilon\sqrt{n} \rfloor$ above Q_n; i.e., initially we have the relation

$$Q_n^\epsilon(t) \geq Q_n(t) \geq Q_n^\epsilon(t) - \lfloor \epsilon\sqrt{n} \rfloor \ . \tag{4.25}$$

Above the level n, the servers are all busy, so that the processes can be given identical service completions, which occur in a Markovian manner. Specifically, the departure events can be

generated by a Poisson process with rate $n\mu$. At each departure event, there is a single departure, which occurs in both processes as long as $Q_n(t) \geq n$.

We have yet to account for the upper barrier at state $n + m_n$. If both processes are equal, then the common new arrivals will be lost in both processes. Otherwise, the higher process may hit the upper barrier, while the lower process does not. That may cause losses to occur in the higher process that are not matched in the lower process. But that causes no problem; that just brings the two ordered sample paths closer together. Even with the upper barrier, we maintain the relation (4.25) throughout the excursion in the upper region.

Now consider what happens when Q_n first hits level $n-1$ from above. Because all servers are no longer busy, its departure rate decreases. However, below level $n-1$, the departure process is a pure death process with rate $k\mu$ in level k. We can thus generate all departures from the common Poisson process with rate $n\mu$ by thinning: If the queue-length process Q_n is at level k $(< n)$ at a departure epoch, then the candidate departure event generated from the Poisson process with rate $n\mu$ is an actual departure with probability k/n; otherwise the candidate departure event has no effect. Since, we construct the departures for the two queue-length processes from a common Poisson process, whenever a departure occurs in Q_n a corresponding departure necessarily occurs in Q_n^ϵ, but there may be departures in Q_n^ϵ that are not matched in Q_n. Those departures may bring the two sample paths closer together, but the relation (4.25) is necessarily maintained. Moreover, the construction gives each of the two processes their correct probability law.

Now we come to the time that the process Q_n^ϵ first hits the level $n-1$. As indicated before, that process immediately is given a jump down of $\lfloor \epsilon\sqrt{n} \rfloor$. Since prior to that jump, the relation (4.25) held, after the jump the order of the processes is switched, i.e., we have the relation

$$Q_n^\epsilon(t) \leq Q_n(t) \leq Q_n^\epsilon(t) + \lfloor \epsilon\sqrt{n} \rfloor \ . \tag{4.26}$$

Going forward in time, the processes get no further apart, because we do the sample path construction so that the higher process Q_n has a departure whenever the lower process Q_n^ϵ does. The lower process may fail to match departures with the upper process, either because of hitting the lower barrier at 0 or, probabilistically, because of the difference in the service rates. That could cause the processes to couple, after which the sample paths would be identical until level n is first hit from below. In any case, relation (4.26) is maintained until Q_n^ϵ again hits level n from below.

When Q_n^ϵ again hits level n from below, it experiences a jump up of $\lfloor \epsilon\sqrt{n} \rfloor$, causing relation

(4.26) to be replaced by relation (4.25), with the subsequent reasoning repeated (leading to a formal proof by induction on the successive hitting of level n from below and level $n-1$ from above). From the scaling in (2.7), we have thus established the inequality in (4.17), which implies the inequality in (4.22) uniformly in n.

Verifying (4.19): Establishing Convergence of the Approximations. To establish the convergence in (4.19), we focus on the successive intervals during which the unscaled processes Q_n^ϵ spend above n and below $n-1$. Equivalently, we focus on the successive intervals during which the scaled processes $\mathbf{Q}_n^\epsilon$ spend above 0 and below 0. Because of the jumps of size $\epsilon > 0$ at each crossing of 0 by the scaled processes, those intervals are asymptotically of positive finite (but random) length. (Without the size-ϵ jumps, the average excursion interval length would be of order $1/\sqrt{n}$ and would be harder to analyze.)

The convergence in each of the two regions follows easily from previous heavy-traffic limits, because the unscaled processes Q_n^ϵ behave like queue-length processes in previously studied queueing systems in their excursions above and below level n. Above level n, the queue-length process Q_n^ϵ behaves like the queue-length process in a $G/M/1/m_n$ queue; below level n, the queue-length process Q_n^ϵ behaves like the queue-length process in a $G/M/\infty$ queue. The assumed FCLT for the arrival process in (2.2) implies associated convergence over random subintervals.

One step of the proof is to treat the successive excursions between each successive crossing. Another step is to show that we can put together all the pieces and establish convergence of the overall process. We address that second step first.

Putting the Pieces Together. It should be apparent that convergence of all the pieces implies convergence of the overall process. To demonstrate it, we apply the continuous mapping theorem; see Section 3.4 of SPL. We now define the function that puts together all the converging pieces. Let $\mathbf{t} \equiv \{t_k : k \geq 0\}$ be a sequence of numbers with $0 = t_0 < t_1 < \ldots < t_k < t_{k+1} < \ldots$ such that $t_k \to \infty$ as $k \to \infty$. Let Δ be the subset of such sequences in $\mathbb{R}^\infty$. (The subset Δ is well defined, being an intersection of open subsets in $\mathbb{R}^\infty$: The subset $A \equiv \{\mathbf{t} : t_0 = 0, \quad t_k \leq t_{k+1} \quad \text{for all} \quad k\}$ is a closed subset of $\mathbb{R}^\infty$. The subset $A_m \equiv \{\mathbf{t} : t_k \leq m \quad \text{for all} \quad k\}$ is a closed subset, the subset $B_m \equiv \{\mathbf{t} : t_m = t_{m+1}\}$ is a closed subset, and Δ is the (countable) intersection of the complements A_m^c and B_m^c within A. Thus Δ is metrizable as a complete separable metric space with the relative topology from $\mathbb{R}^\infty$; see

p. 371 of SPL.)

We also consider a sequence of elements of D: let $\mathbf{x} \equiv \{x_k : k \geq 0\}$ be an element of the product space D^∞ (with the product topology; see Section 11.4 of SPL). Let $h : (D^\infty \times \Delta) \to (D, J_1)$ be the function defined by

$$h((\mathbf{x}, \mathbf{t}))(s) \equiv x_k(s) : t_k \leq s < t_{k+1}, \quad k \geq 0 \ . \tag{4.27}$$

Note that we need to restrict h to $(D^\infty \times \Delta)$ (instead of just $(D^\infty \times \mathbb{R}^\infty)$ in order for $h((\mathbf{x}, \mathbf{t}))$ to be a legitimate element of D. We use the following lemma.

Lemma 4.3. *The function $h : (D^\infty \times \Delta) \to (D, J_1)$ defined in* (4.27) *is continuous at all $(\mathbf{x}, \mathbf{t})$ such that, for all k, x_k is continuous everywhere except possibly at the points $t_1, \ldots, t_k$.*

Proof. Suppose that $(\mathbf{x}, \mathbf{t})_n \equiv (\mathbf{x}_n, \mathbf{t}_n) \to (\mathbf{x}, \mathbf{t}))$ as $n \to \infty$ in $(D^\infty \times \Delta)$ with the continuity condition holding. We want to show that $y_n \equiv h((\mathbf{x}_n, \mathbf{t}_n)) \to h((\mathbf{x}, \mathbf{t}) \equiv y$ as $n \to \infty$ in (D, J_1). Because of the discontinuities at the transition points t_k, we need the J_1 topology on the range. It suffices to focus on bounded intervals $[0, t]$, where t is not one of the limit points t_k. Suppose such a t is given. We fix k by also supposing that $t_k < t < t_{k+1}$. Hence it suffices to work with the first $k + 1$ limits $(x_{j,n}, t_{j,n}) \to (x_j, t_j) : 0 \leq j \leq k$.

To treat the convergence in D with time domain $[0, t]$, we make use of the J_1 metric on $D([0, t], \mathbb{R})$; see (3.2) on p. 79 of SPL. We let $\lambda_n : [0, t] \to [0, t]$ be the strictly increasing time transformations used there (not to be confused with the arrival rate in the queue). We want to construct time transformations λ_n such that

$$\|\lambda_n - \mathbf{e}\|_t \to 0 \quad \text{as} \quad n \to \infty \ , \tag{4.28}$$

where $\mathbf{e}$ is the identity map and

$$\|y_n \circ \lambda_n - y\|_t \to 0 \quad \text{as} \quad n \to \infty \ . \tag{4.29}$$

Note that the only difficulties (discontinuities of the functions in D) occur at the points t_k. (There is local uniform convergence elsewhere.) We thus construct λ_n by requiring that

$$\lambda_n(s) = s \tag{4.30}$$

for $s = 0$, $s = (t_{j-1} + t_j)/2$, $1 \leq j \leq k$, and for $s = t$. We let λ_n be defined on the subinterval $[0, (t_0 + t_1)/2]$ as required to get the convergence $x_{1,n} \to x_1$ for the restrictions to $[0, (t_0 + t_1)/2]$. We let λ_n be defined on the subinterval $[(t_{j-1} + t_j)/2, (t_j + t_{j+1})/2]$ as required to get the

convergence $x_{j+1,n} \to x_{j+1}$ for the restrictions to $[(t_{j-1}+t_j)/2, (t_j+t_{j+1})/2]$. This continues up to the last component process $x_{k,n}$ and the last interval, which is $[(t_{k-1}+t_k)/2, (t_k+t)/2]$. It is easy to see that this construction produces the desired asymptotic behavior in (4.28) and (4.29). ∎

Analyzing the Pieces. Now we construct the processes that let us analyze the different pieces. We define a sequence of queue-length processes $\{\mathbf{Q}^\epsilon_{n,k} : k \geq 0\}$ and an associated sequence of first passage times $\{T^\epsilon_{n,k} : k \geq 0\}$. The process $\mathbf{Q}^\epsilon_{n,k}$ is constructed to agree with the process $\mathbf{Q}^\epsilon_n$ up to the random time $T^\epsilon_{n,k}$.

As before, for simplicity, we assume that there is an initial jump up, so that the scaled queue length starts off at $+\epsilon$. In particular, let

$$\mathbf{Q}^\epsilon_{n,0}(0) = \epsilon \quad \text{and} \quad T^\epsilon_{n,0} = 0 \ . \tag{4.31}$$

(It is easy to modify this with some other initial condition as specified before Theorem 2.1.)

For $t > 0$, we let $\mathbf{Q}^\epsilon_{n,0}(t)$ be the scaled queue-length process in the $G/M/1/m_n$ model with arrival process $C_n(\lambda_n t)$ and service rate μn. Hence we can apply established heavy-traffic limits for single-server queues in Chapter 9 of SPL, modified to treat the upper barrier to deduce that

$$\mathbf{Q}^\epsilon_{n,0} \Rightarrow \mathbf{Q}^\epsilon_0 \quad \text{in} \quad (D, J_1) \quad \text{as} \quad n \to \infty \ . \tag{4.32}$$

The assumption of exponential service times allows us to directly apply the continuous mapping theorem with the one-sided reflection map to treat the upper barrier. Alternatively, we could use the two-sided reflection map, as in Chapter 5 and Section 14.8 of SPL.

We now define the remaining random times and processes recursively. For $k \geq 1$, let

$$\begin{aligned}
T^\epsilon_{n,2k-1} &\equiv \inf\{t > T^\epsilon_{n,2k-2} : \mathbf{Q}^\epsilon_{n,2k-2}(t) \leq 0\} \ , \\
\mathbf{Q}^\epsilon_{n,2k-1}(T^\epsilon_{n,2k-1}) &\equiv -\epsilon \ , \\
\mathbf{Q}^\epsilon_{n,2k-1}(t) &\equiv \mathbf{Q}^\epsilon_{n,2k-2}(t), \quad 0 \leq t < T^\epsilon_{n,2k-1} \ , \\
T^\epsilon_{n,2k} &\equiv \inf\{t > T^\epsilon_{n,2k-1} : \mathbf{Q}^\epsilon_{n,2k-1}(t) \geq 0\} \ , \\
\mathbf{Q}^\epsilon_{n,2k}(T^\epsilon_{n,2k}) &\equiv +\epsilon \ , \\
\mathbf{Q}^\epsilon_{n,2k}(t) &\equiv \mathbf{Q}^\epsilon_{n,2k-1}(t), \quad 0 \leq t < T^\epsilon_{n,2k} \ .
\end{aligned} \tag{4.33}$$

As part of the recursive definition, we also must define the scaled queue-length processes $\mathbf{Q}^\epsilon_{n,k}$ after the random time $T^\epsilon_{n,k}$. For $t > T^\epsilon_{n,2k-1}$, we let $\mathbf{Q}^\epsilon_{n,2k-1}(t)$ be the scaled queue-length

process associated with the $G/M/\infty$ model with individual service rate μ and scaled arrival process $C_n(\lambda_n t)$ starting at level $-\epsilon$ at the random time $T^\epsilon_{n,2k-1}$. Just as for $\mathbf{Q}^\epsilon_{n,0}$, for $t > T^\epsilon_{n,2k}$, we let $\mathbf{Q}^\epsilon_{n,2k}(t)$ be the scaled queue-length process in the $G/M/1/m_n$ model with service rate μn and arrival process $C_n(\lambda_n t)$, starting at lever $+\epsilon$ at the random time $T^\epsilon_{n,2k}$.

Having established the convergence in (4.32), we next show that

$$T^\epsilon_{n,1} \Rightarrow T^\epsilon_1 \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty \ , \tag{4.34}$$

jointly with the limit in (4.32), where T^ϵ_1 is the first passage time to the origin for the limiting diffusion process $\mathbf{Q}^\epsilon_0$, i.e.,

$$T^\epsilon_1 \equiv \inf\{t > T^\epsilon_0 = 0 : \mathbf{Q}^\epsilon_0(t) \le 0\} \ . \tag{4.35}$$

We obtain the convergence in (4.34) by applying the continuous mapping theorem with the first-passage-time function, see Section 13.6.3 of SPL. We use the fact that the first-passage-time function is measurable and continuous almost surely with respect to the limit process. The almost sure continuity follows because the limiting diffusion process is almost surely not flat on any interval. That property for general diffusions can be reduced to the familiar property of Brownian motion because the diffusion process can be expressed as a time and space transformation of Brownian motion involving strictly increasing functions; see Chapter 7 of Rogers and Williams (1987).

We next turn to $\mathbf{Q}^\epsilon_{n,1}$. As indicated above, the process is defined after random time $T^\epsilon_{n,1}$ by treating it as the queue-length process in a $G/M/\infty$ model with the scaled arrival process starting after $T^\epsilon_{n,1}$. The previous results imply that the initial conditions satisfy the conditions needed for a stochastic-process limit after the random times $T^\epsilon_{n,1}$. (It is perhaps helpful to think of the processes as having domain $[0,\infty)$, shifting time in the n^{th} process by $T^\epsilon_{n,1}$ and shifting time in the limit process by T^ϵ_1. Afterwards, we can shift time back to obtain the desired construction.

Just as in Srikant and Whitt (1996), we can thus apply a previous FCLT for the $G/M/\infty$ system, specifically Theorem 1 on p. 103 of Borovkov (1984). An especially transparent argument to show that the limit should apply to G arrival processes only under the FCLT condition (2.2) is given in Glynn and Whitt (1991) for $G/GI/\infty$ queues for the special case of discrete service-time distributions having only finitely many point masses; see Section 10.3 of SPL. An alternative direct proof is provided in Proof 2 in Section 5 below; the $G/M/\infty$ result is an easier special case. Also see Krichagina and Puhalskii (1997), which treats the

more difficult case of general service times, but again with infinite waiting room. Here, the established $G/M/\infty$ FCLT applies to each "below 0" interval, yielding convergence to the Ornstein-Uhlenbeck diffusion process starting each such random interval in state $-\epsilon$.

In order to obtain the desired convergence, we use the established convergence of $\mathbf{Q}^\epsilon_{n,1}$ before time $T^\epsilon_{n,1}$. To obtain the joint convergence of all random quantities considered, we exploit the map $h_1 : D \times C \times \mathbb{R} \to D \times D \times \mathbb{R}$ defined by

$$h_1(x,y,t)(s) \equiv \begin{cases} (x(s), x(s), t), & 0 \le s < t, \\ (x(s), y(s), t), & s \ge t \ . \end{cases} \tag{4.36}$$

This map h_1 is continuous at all $(x,y,t) \in D \times C \times \mathbb{R}$ such that x is continuous at t (our case).

We thus obtain the joint convergence

$$(\mathbf{Q}^\epsilon_{n,0}, \mathbf{Q}^\epsilon_{n,1}, T^\epsilon_{n,1}) \Rightarrow (\mathbf{Q}^\epsilon_0, \mathbf{Q}^\epsilon_1, T^\epsilon_1) \tag{4.37}$$

in $(D, J_1)^2 \times \mathbb{R}$ as $n \to \infty$, where $\mathbf{Q}^\epsilon_1$ is an OU process after the random time T^ϵ_1. Since T^ϵ_1 is obtained as a first passage time relative to $\mathbf{Q}^\epsilon_0$, it is a stopping time relative to $\mathbf{Q}^\epsilon_0$. Hence the limit process $\mathbf{Q}^\epsilon_1$ is a (Markov) diffusion process.

Paralleling (4.33) and (4.35), we recursively define the other limit processes by

$$\begin{aligned} T^\epsilon_{2k-1} &\equiv \inf\{t > T^\epsilon_{2k-2} : \mathbf{Q}^\epsilon_{2k-2}(t) \le 0\} \ , \\ \mathbf{Q}^\epsilon_{2k-1}(T^\epsilon_{2k-1}) &\equiv -\epsilon \ , \\ \mathbf{Q}^\epsilon_{2k-1}(t) &\equiv \mathbf{Q}^\epsilon_{2k-2}(t) \ , \quad 0 \le t < T^\epsilon_{2k-1} \\ T^\epsilon_{2k} &\equiv \inf\{t > T^\epsilon_{2k-1} : \mathbf{Q}^\epsilon_{2k-1}(t) \ge 0\} \ , \\ \mathbf{Q}^\epsilon_{2k}(T^\epsilon_{2k}) &\equiv +\epsilon \ , \\ \mathbf{Q}^\epsilon_{2k}(t) &\equiv \mathbf{Q}^\epsilon_{2k-1}(t) \ , \quad 0 \le t < T^\epsilon_{2k} \ . \end{aligned} \tag{4.38}$$

As before, we also need to define the processes $\mathbf{Q}^\epsilon_k$ after the random times T^ϵ_k; we let the process evolve after T^ϵ_k according to the appropriate diffusion process, depending on whether we are above zero or below zero. As before, since T^ϵ_k is a first passage time, T^ϵ_k is a stopping time relative to $\mathbf{Q}^\epsilon_{k-1}$ for each k, so that $\mathbf{Q}^\epsilon_k$ is a diffusion process for all k. Moreover, $\{T^\epsilon_{2k-1} - T^\epsilon_{2k-2} : k \ge 1\}$ and $\{T^\epsilon_{2k} - \{T^\epsilon_{2k-1} : k \ge 1\}$ are independent sequences of IID positive random variables.

We then apply the arguments above to recursively establish the limits

$$\begin{aligned} T^\epsilon_{n,2k-1} &\Rightarrow T^\epsilon_{2k-1} \quad \text{in} \quad \mathbb{R} \ , \\ \mathbf{Q}^\epsilon_{n,2k-1} &\Rightarrow \mathbf{Q}^\epsilon_{2k-1} \quad \text{in} \quad (D, J_1) \ , \\ T^\epsilon_{n,2k} &\Rightarrow T^\epsilon_{2k} \quad \text{in} \quad \mathbb{R} \ , \\ \mathbf{Q}^\epsilon_{n,2k} &\Rightarrow \mathbf{Q}^\epsilon_{2k} \quad \text{in} \quad (D, J_1) \end{aligned} \tag{4.39}$$

for all $k \geq 1$, where the convergence is joint. In order to get the joint convergence, we need to modify the map h_1 in (4.36) as k increases. In particular, for each k, we construct an analogous map $h_k : (D \times \mathbb{R})^{k-1} \times C \times \mathbb{R} \to (D \times \mathbb{R})^k$ and apply induction to obtain joint convergence for all k. That joint convergence for all k then implies convergence of the entire sequence in the product space $(D \times \mathbb{R})^\infty$.

Finally, we apply the continuous map h in (4.27) to establish the overall desired convergence stated in Theorem 4.1. We can apply Lemma 4.3 because

$$\mathbf{Q}_n^\epsilon = h((\{\mathbf{Q}_{n,k}^\epsilon : k \geq 0\}, \{T_{n,k}^\epsilon : k \geq 0\})) \tag{4.40}$$

and

$$\mathbf{Q}^\epsilon = h((\{\mathbf{Q}_k^\epsilon : k \geq 0\}, \{T_k^\epsilon : k \geq 0\})) \ . \tag{4.41}$$

We use the fact that $\{T_{2k-1}^\epsilon - T_{2k-2}^\epsilon : k \geq 1\}$ and $\{T_{2k}^\epsilon - \{T_{2k-1}^\epsilon : k \geq 1\}$ are independent sequences of IID positive random variables to deduce that the single sequence $\{T_k^\epsilon : k \geq 0\}$ almost surely belongs to Δ: By the strong law of large numbers, the averages converge to the positive means, which implies that $T_k^\epsilon \to \infty$ w.p.1.

Verifying (4.23): $\mathbf{Q}^\epsilon \Rightarrow \mathbf{Q}$. We now complete the proof of Theorem 4.1 by establishing (4.23), i.e., by showing that $\mathbf{Q}^\epsilon \Rightarrow \mathbf{Q}$ as $\epsilon \downarrow 0$. We give two different proofs.

The first proof exploits previously established limits for the special $GI/M/n/\infty$ model in Halfin and Whitt (1981) or Puhalskii and Reiman (2000), where the arrival process as assumed to be renewal. The previous results imply, first, that $\mathbf{Q}_n \Rightarrow \mathbf{Q}$ and, second, as a consequence of that, $\mathbf{Q}_n^\epsilon \Rightarrow \mathbf{Q}^\epsilon$ as $n \to \infty$ for each $\epsilon > 0$. Then, by doing the special construction to establish (4.17), we obtain $\pi(\mathbf{Q}_n^\epsilon, \mathbf{Q}_n) \leq \epsilon$ for all n and $\epsilon > 0$. As a consequence, we obtain $\pi(\mathbf{Q}^\epsilon, \mathbf{Q}) \leq \epsilon$ for each $\epsilon > 0$, which implies the desired conclusion.

The second proof works directly with the limiting diffusion processes $\mathbf{Q}^\epsilon$ and $\mathbf{Q}$. As in comparison theorems for diffusion processes, such as in Theorem 43.1 of Rogers and Williams (1987), we construct the two diffusions on the same sample space by using a common Brownian motion in the definition of the stochastic differential equations. In this way, we show that analogs of the two relations (4.25) and (4.26) hold for the processes $\mathbf{Q}^\epsilon$ and $\mathbf{Q}$ over excursions above and below 0.

First, suppose that we start with a jump up to ϵ in $\mathbf{Q}^\epsilon$. We then construct the two processes on the same space using the stochastic integrals

$$\mathbf{Q}(t) \quad = \quad \mathbf{Q}(0) + \int_0^t \sigma_{\mathbf{Q}}^2 \, d\mathbf{B}(s) + \int_0^t m(\mathbf{Q}(s)) \, ds$$

$$\mathbf{Q}^\epsilon(t) \quad = \quad \epsilon + \int_0^t \sigma_{\mathbf{Q}}^2 \, d\mathbf{B}(s) + \int_0^t m(\mathbf{Q}^\epsilon(s)) \, ds, \tag{4.42}$$

where $\sigma_{\mathbf{Q}}^2$ is the constant diffusion coefficient of $\mathbf{Q}$ in (2.16), $0 \leq \mathbf{Q}(0) \leq \epsilon$ w.p.1 and we use a common standard Brownian motion $\mathbf{B}$ in both cases. Since the diffusion coefficient is constant, we can simplify the component stochastic integrals with respect to Brownian motion, to obtain $\sigma_{\mathbf{Q}}^2 \mathbf{B}(t)$. This construction remains valid until $\mathbf{Q}^\epsilon$ next hits zero, after which there is a jump down to $-\epsilon$.

Referring to the two stochastic integrals in (4.42), we see that the drifts are identical when $\mathbf{Q}(t) > 0$, but the drift of $\mathbf{Q}$ is greater than the drift of $\mathbf{Q}^\epsilon$ whenever $\mathbf{Q}^\epsilon(t) > 0 > \mathbf{Q}(t)$. As a consequence, with the special construction, the distance between the two stochastic processes is a nonincreasing function until the two sample paths coincide, i.e., until they couple. In particular,

$$\mathbf{Q}^\epsilon(t) - \mathbf{Q}(t) = \epsilon + \int_0^t (m(\mathbf{Q}^\epsilon) - m(\mathbf{Q}(s))) \, ds \ . \tag{4.43}$$

Hence we have the relation

$$\mathbf{Q}^\epsilon(t) \geq \mathbf{Q}(t) \geq \mathbf{Q}^\epsilon(t) - \epsilon \tag{4.44}$$

during each excursion of $\mathbf{Q}^\epsilon$ above 0. Essentially the same argument works for excursions of $\mathbf{Q}^\epsilon$ below 0, yielding the relation

$$\mathbf{Q}^\epsilon(t) \leq \mathbf{Q}(t) \leq \mathbf{Q}^\epsilon(t) + \epsilon \tag{4.45}$$

during each excursion of $\mathbf{Q}^\epsilon$ below 0. From these constructions, we obtain

$$\|\mathbf{Q}^\epsilon - \mathbf{Q}\|_t \leq \epsilon \tag{4.46}$$

for the special processes on the same sample space. That in turn implies that

$$\pi(\mathbf{Q}^\epsilon, \mathbf{Q}) \leq \epsilon \ , \tag{4.47}$$

which implies the claimed convergence. ∎

4.4. Step 2: $G/H_2^*/n/\infty$; Relating $\mathbf{Q}_n$ to $\mathbf{Q}_n^p$.

We now return attention to the $G/H_2^*/n/\infty$ model. We now show that the limit for $\mathbf{Q}_n^p$ established in Corollary 4.1, the Corollary to to Theorem 4.1, implies a corresponding limit for the primary processes of interest $\mathbf{Q}_n$ when there are H_2^* service times, in the case of an unlimited waiting room. We do this by establishing the following result, which goes beyond

Theorem 2.1 to establish joint convergence of $\mathbf{Q}_n$ and $\mathbf{Q}_n^p$ (only in the case of unlimited waiting room). Let $\hat{g} : D \to D$ be the function defined by

$$\hat{g}(x)(t) = g(x(t)) \quad \text{for all} \quad t \geq 0 \ , \tag{4.48}$$

where $g : \mathbb{R} \to \mathbb{R}$ is the function defined in (2.13). Clearly $\hat{g}$ is a continuous function.

Theorem 4.2. *Consider the $G/H_2^*/n/\infty$ model under the assumptions of Theorem* 2.1*. For each $t > 0$,*

$$\|\mathbf{Q}_n - \hat{g}(\mathbf{Q}_n^p)\|_t \Rightarrow 0 \quad as \quad n \to \infty \ , \tag{4.49}$$

so that

$$(\mathbf{Q}_n, \mathbf{Q}_n^p) \Rightarrow (\hat{g}(\mathbf{Q}^p), \mathbf{Q}^p) \quad in \quad (D, J_1)^2 \tag{4.50}$$

where $\mathbf{Q}^p$ is the limit process in Theorem 2.1 *and $\hat{g}$ is the mapping in* (4.48) *and* (2.13)*.*

Proof. We exploit the infinite-waiting room assumption. Let $D_n^p(t)$ be the number of departures of customers with positive service times in the interval $[0, t]$ in the n^{th} system. Since there is unlimited waiting space, we have the basic relation

$$Q_n^p(t) = Q_n^p(0) + C_n^p(t) - D_n^p(t) \quad \text{for} \quad t \geq 0 \ . \tag{4.51}$$

We now relate Q_n to Q_n^p, C_n^p, D_n^p and a single sequence of IID geometric random variables $\{X_i : i \geq 1\}$. The random variable X_i represents 1 (for the i^{th} arrival with positive service times) plus the number of customers with zero service times that arrive after the i^{th} arrival with positive service times but before the $(i+1)^{st}$ arrival with positive service times. For example, $X_1 + \cdots + X_k$ represents the total number of arrivals before the $(k+1)^{st}$ arrival with a positive service time.

First, if $Q_n^p(t) < n$, then $Q_n(t) = Q_n^p(t)$. Next, if $Q_n^p(t) \geq n$, then we have the bound

$$\sum_{i=D_n^p(t)+n}^{Q^p(0)+C_n^p(t)-1} X_i \leq Q_n(t) \leq \sum_{i=D_n^p(t)+n}^{Q^p(0)+C_n^p(t)} X_i \ . \tag{4.52}$$

The bound applies for all $Q_n^p(t)$ if we understand the sum to be zero whenever the lower index exceeds the upper index. The upper bound includes all the possible arrivals with zero service times that could occur following the $(C_n^p(t))^{th}$ arrival with positive service times, while the lower bound omits the last batch, allowing for the possibility that some of those customers have not arrived yet.

Hence we can write

$$(Q_n(t)-n)-g(Q_n^p(t)-n)) = 1_{\{Q_n^p(t)\geq n\}}[\sum_{i=D_n^p(t)+n}^{Q_n^p(0)+C_n^p(t)}(X_i-p^{-1})+R_n(t)] \quad \text{for} \quad t\geq 0 \ , \ (4.53)$$

where $EX_i = p^{-1}$ and $R_n(t)$ is a remainder term involving the last batch, as can be seen from the bounds in (4.53). We obtain the desired convergence in (4.49) because both the partial sums of the summands $X_i - p^{-1}$ satisfy a FCLT and the random number of terms, within 1 of $[Q_n^p(t)-n]^+$, also satisfies a FCLT.

First, we apply Donsker's theorem for the IID geometric random variables, i.e.,

$$\mathbf{S}_n \Rightarrow \sqrt{(1-p)/p^2}\mathbf{B} \ , \tag{4.54}$$

where

$$\mathbf{S}_n(t) = n^{-1/2}\sum_{i=1}^{\lfloor nt\rfloor}(X_i-p^{-1}) \ . \tag{4.55}$$

As a first consequence of this FCLT, we can deduce that the remainder term $R_n(t)$ in (4.53) is asymptotically negligible after dividing by $\sqrt{n}$, uniformly over the interval $[0,t]$. (We first obtain the related FCLT for the random sum of $(X_i - p^{-1})$ up to $C_n^p(t)$, Corollary 13.3.2 of SPL, and then we apply the continuous mapping theorem with the maximum jump functional; p. 119 of SPL.) To state that result, let $\mathbf{R}_n(t) = n^{-1/2}R_n(t)$, $t \geq 0$; we are concluding that $\|\mathbf{R}_n\|_t \Rightarrow 0$.

As a second consequence of the limit in (4.54), we can apply Prohorov's theorem to obtain tightness, so that we have an associated bound on the oscillations of $\mathbf{S}_n$: For each $u > 0$, $\epsilon > 0$ and $\eta > 0$, there exists a δ with $0 < \delta < 1$ and an n_0 such that

$$P(w(\mathbf{S}_n,\delta,u) > \epsilon) \leq \eta \quad \text{for all} \quad n \geq n_0 \ , \tag{4.56}$$

where $w(x,\delta,u)$ is the modulus of continuity of x over the interval $[0,u]$, i.e.,

$$w(x,\delta,u) \equiv \sup\{|x(s)-x(t)| : 0\leq s\leq u, \quad 0\leq t\leq u, \quad |s-t|<\delta\} \ ; \tag{4.57}$$

see Section 11.6 of SPL.

As a consequence of the continuous mapping theorem and Corollary 4.1, for each $t > 0$,

$$\|\mathbf{Q}_n^p\|_t \equiv \sup_{0\leq s\leq t}\{|\mathbf{Q}_n^p(s)|\} \Rightarrow \sup_{0\leq s\leq t}\{|\mathbf{Q}^p(s)|\} \equiv \|\mathbf{Q}^p\|_t \ . \tag{4.58}$$

Finally, combining (4.53), (4.56) and (4.58), we obtain (4.49), which together with Corollary 4.1 implies the desired limit (4.50), using the convergence-together theorem, Theorem 11.4.7

of SPL. In particular, defining events

$$\begin{aligned} A_{n,\epsilon} &\equiv \{\|\mathbf{Q}_n - \hat{g}(\mathbf{Q}_n^p)\|_t > \epsilon\}, \\ B_{n,\epsilon} &\equiv \{w(\mathbf{S}_n, \delta, t) > \epsilon/2\}, \\ C_{n,\epsilon} &\equiv \{\|\mathbf{R}_n\|_t > \epsilon/2\}, \\ D_n &\equiv \{\|\mathbf{Q}_n^p\|_t > \eta\} \ , \end{aligned} \tag{4.59}$$

we have, for any $\eta > 0$ and then all sufficiently large n, that

$$A_{n,\epsilon} \subseteq B_{n,\epsilon} \cup C_{n,\epsilon} \cup D_n \ , \tag{4.60}$$

so that

$$P(A_{n,\epsilon}) \leq P(B_{n,\epsilon}) + P(C_{n,\epsilon}) + P(D_n) \ , \tag{4.61}$$

Since, for each $\epsilon > 0$ and $\eta > 0$, each of the terms on the right converges to 0 as $n \to \infty$, the result is established. ∎

We have now completed the proof of Theorem 2.1 for the stochastic process $\mathbf{Q}_n$ in the case of an unlimited waiting room. We treat the stochastic process $\mathbf{Q}_n^a$ later.

4.5. Step 3. $G/H_2^*/n/m_n$; Piecewise Construction for a Finite Waiting Room.

We now apply the previous results, in particular, Corollary 4.1 and Theorem 4.2, to establish the desired limit for $\mathbf{Q}_n$ in the $G/H_2^*/1/m_n$ model; i.e., we now treat $\mathbf{Q}_n$ in the case of a finite waiting room. Our proof here is recursive or inductive, exploiting a piecewise construction, just as in the first proof of Step 1 above. In particular, our proof here is like the verification of (4.19) previously. By that, we mean that we use a similar piecewise construction. The overall argument now is much easier than Proof 1 of Step 1, however, because (i) we now make no special distinction between the customers with positive service times and the customers with zero service times, and (ii) we do not need to introduce any approximating processes, such as we did before by adding the jumps away from the boundary. Now we are able to construct the necessary pieces directly. However, the proof now closely follows part of Proof 1 of Step 1. In particular, the specific construction here is an obvious modification of (4.31) - (4.41), so we will be brief here.

We break up the construction of the processes, and the justification of convergence, into pieces, just like we did for $\mathbf{Q}_n^\epsilon$ in the verification of (4.19). Here we consider two levels a and b with $0 < a < b < \kappa$. Assuming that the scaled queue-length (number in system) process starts below level b, we first consider the scaled queue-length process until it first hits or passes level

b. Up to that time, we use the result for an infinite waiting room established in Step 2 above. That obviously is reasonable, because whenever the scaled number in system is below level b, the actual finite waiting room plays no role.

After the process hits level b from below, we switch over to another process, in particular, to a reflected version of the scaled queue-length process, using the standard reflection map with a reflecting upper barrier at κ. It is natural for arrivals to occur one at a time, so that the scaled process will indeed pass the level b at a well defined time. However, our assumptions permit batch arrivals. In that event, the batch sizes necessarily are of order $O(1)$ before scaling, and become asymptotically negligible after scaling. So, without loss of generality, it suffices to assume that the scaled process is asymptotically at level b when the switch occurs.

New treatment is required for the pieces starting when level b is first hit or passed from below. In each of these "upper" pieces, starting at a hitting time of b from below, the queue-length process behaves like the queue-length process in a single-server $G/H_2^*/1/m_n$ model and a finite waiting room, which in turn is equivalent to a $G/M^X/1/m_n$ model with batch service (a geometric batch at exponential intervals) as long as all servers remain busy. In order to make this equivalence clear, we elaborate on two points:

First note that, without loss of generality, we can let customer service times be determined the moment that a customer starts service. Thus the identification of customers – specifying whether they have positive service time or zero service time – can be determined by independent Bernoulli random variables, with each customer having a positive service time with probability p, independent of the system history prior to the instant the customer enters service. (That construction is not actually required, but it can help understanding. It is then easy to see that there is no dependence between the number in system and the type of the customers waiting in queue.)

Second, we should explain what we mean by the batch service. As usual, in the batch-service queue we have in mind, there is a batch of potential service times, with the number of potential service times being geometrically distributed. At a service epoch, service is simultaneously performed on that many (the batch size) customers if that many customers are in the system; otherwise, all available customers are served. Thus the number of *potential* customers served at a service epoch has a geometric distribution, but the actual number of customers served at a service epoch does *not* have a geometric distribution. The M in M^X means that the intervals between successive service epochs are exponentially distributed, provided there are customers to be served. Since all servers are busy in the $G/H_2^*/n/m_n$ model when the scaled process is

above level a, that will always (asymptotically) be the case.

The upper pieces start when they hit or pass level b from below, and they end when they hit or pass level a from above. Since $0 < a < b < \kappa$, the length of these pieces is asymptotically of order $O(1)$. Moreover, all servers are always busy when we are considering the upper pieces. We can analyze each piece starting after hitting or passing b from below and ending when the lower level a is hit or passed from above by applying the continuous-mapping theorem together with the one-sided reflection map associated with the upper barrier at κ together with established limits for the $G/M^X/1/\infty$ queue; see Sections 5.2 and 13.5 of SPL. Since the arrival process is exogenous and the service times are Markovian, the construction for each of these pieces starting at level b is routine. That is, the reflection map applies directly; no extra approximation step is needed. Expressed differently, we can treat each upper piece starting at b and ending at a by applying known results for the $G/M^X/1/m_n$ model.

We use the reflection construction just specified until the scaled queue-length process next hits or passes below level a from above. In general, the scaled queue-length process will jump below level a because of batch services, but with the scaling, the batch sizes are asymptotically negligible. The fact that the heavy-traffic limit for the $G/M^X/1/m_n$ model is RBM, which almost surely has continuous paths, allows us to prove the point by applying the maximum jump functional; see p. 119 of SPL. Thus, asymptotically, no servers in the full $G/H_2^*/n/m_n$ model will become idle at these transition epochs. Moreover, it suffices to assume that the level a is actually hit at the transition points.

For an upper piece, after the scaled queue-length process hits or passes level a from above, we transition to a lower piece; i.e., we revert back to the previous construction involving the $G/H_2^*/n/\infty$ model without an upper barrier, discussed above. We use the lower piece starting at a until the process next hits or passes level b again. We switch back and forth between successive visits to b from below and a from above.

Just as for (4.19), each successive piece requires a new construction. The overall construction and proof is inductive. The limit for each successive piece provides the convergence for the initial conditions in the next piece. The initial weak convergence of the arrival process implies weak convergence for the new arrival processes after the random times. Since the switching times are specified as first passage times, they are again stopping times. As before, that stopping time property implies that the overall limit process is a diffusion process.

The argument just sketched justifies the convergence, but how do we know that the limit process has the properties stated in Theorem 2.1? We know that because the full diffusion

limit has the local character of the diffusion limits for the pieces. First, below level a, the new limit must agree with the previously established limit for the $G/H_2^*/n/\infty$ model. Necessarily, we have thus captured the difficult behavior at 0 without directly addressing the issue again when there is a finite waiting room. (Indeed, it should be evident that the addition of a finite waiting room cannot alter the behavior at 0.) At the same time, above level b, the new limit must coincide with the RBM limit for the upper piece; and it is easy to see that that is the case. In particular, the RBM limit for the upper piece determines the reflection map applying at the upper barrier and the infinitesimal mean and variance of the diffusion process, which we have displayed in (2.19) and (2.20). Finally, since the switching levels a and b are arbitrary, they obviously play no role in the final result. ∎

4.6. Establishing convergence of $(\mathbf{Q}_n^a, \mathbf{Q}_n)$

We now show that the stochastic-process limit established for $\mathbf{Q}_n$ in (2.7) implies a corresponding stochastic-process limit for $\mathbf{Q}_n^a$ and the joint convergence in (2.11). Just as in Halfin and Whitt (1981), we apply a random-time-change argument to connect the two limits; i.e., we apply the continuous mapping theorem with the composition map.

Recall that $C_n(t)$ counts the number of arrivals in the interval $[0,t]$ and form the scaled process

$$\hat{\mathbf{C}}_n(t) \equiv C_n(t)/n, \quad t \geq 0 \ . \tag{4.62}$$

Since $\lambda_n/n \to \mu$ as $n \to \infty$, it is an elementary consequence of the assumed FCLT in (2.2) that

$$\hat{\mathbf{C}}_n \Rightarrow \mu \mathbf{e} \quad \text{in} \quad (D, J_1) \quad \text{as} \quad n \to \infty. \tag{4.63}$$

Now let $T_n(k)$ be the arrival time of the k^{th} customer in model n and let $\hat{\mathbf{T}}_n$ be the scaled random element of D defined by

$$\hat{\mathbf{T}}_n \equiv T_n(\lfloor nt \rfloor), \quad t \geq 0 \ . \tag{4.64}$$

By the continuous mapping theorem with the inverse map, see Section 13.6 of SPL,

$$\hat{\mathbf{T}}_n \Rightarrow \mu^{-1} \mathbf{e} \quad \text{in} \quad (D, J_1) \quad \text{as} \quad n \to \infty. \tag{4.65}$$

Given that we have established $\mathbf{Q}_n \Rightarrow \mathbf{Q}$, we can invoke Theorem 11.4.5 of SPL to obtain the joint convergence

$$(\mathbf{Q}_n, \hat{\mathbf{T}}_n) \Rightarrow (\mathbf{Q}, \mu^{-1}\mathbf{e}) \quad \text{in} \quad (D, J_1)^2 \ , \tag{4.66}$$

from which we deduce (by applying the continuous mapping theorem with the composition map) that

$$\mathbf{Q}_n \circ \hat{\mathbf{T}}_n \Rightarrow \mathbf{Q} \circ \mu^{-1}\mathbf{e} = \mathbf{Q}^a \ . \tag{4.67}$$

We now show that

$$\|\mathbf{Q}_n^a - \mathbf{Q}_n \circ \hat{\mathbf{T}}_n\|_t \Rightarrow 0 \tag{4.68}$$

for each $t > 0$, which implies the desired conclusion. The limit in (4.68) follows because the difference there is bounded by the maximum batch size among all arrivals up to time t divided by $\sqrt{n}$. However, we can apply the assumed convergence in (2.2) to deduce that this scaled maximum batch size is asymptotically negligible. In particular, we can apply the continuous mapping theorem with the maximum jump function, as on p. 119 of SPL, with the limit in (2.2) to obtain (4.68). ∎

That finally completes the proof of Theorem 2.1. We now turn to the alternative proof of Step 1 in our proof of Theorem 2.1 and Theorem 3.1.

5. Proof of Theorem 3.1: Martingale Proof for $G/M/n/m_n + M$

We now present the proof of the limit with customer abandonments in Theorem 3.1. The special case with an infinite waiting room and without customer abandonments yields the second proof of Step 1 in the proof of Theorem 2.1.

In particular, here we prove the following result, which extends Theorem 3.1 by giving an alternative characterization of the limit process (which is equivalent).

Theorem 5.1. *Under the conditions of Theorem* 3.1, $\mathbf{Q}_n \Rightarrow \mathbf{Q}$, *where* $\mathbf{Q}$ *is a reflected diffusion process, defined by*

$$\mathbf{Q}(t) = \mathbf{Q}(0) + \mathbf{C}'(t) - \mu\beta t - \theta \int_0^t (\mathbf{Q}(s) \vee 0)\, ds - \mu \int_0^t (\mathbf{Q}(s) \wedge 0)\, ds + \sqrt{\mu}\mathbf{W}(t) - \Phi(t) \ , \tag{5.1}$$

where $\mathbf{C}'$ *is the limit in* (2.5), $\mathbf{W}$ *is a Brownian motion independent of* $\mathbf{Q}(0)$ *and* $\mathbf{C}'$, *and* Φ *is the regulator process recording the time spent at the upper barrier* κ, *i.e.,*

$$\Phi(t) = \int_0^t 1(\mathbf{Q}(s) = \kappa)\, d\Phi(s) \ . \tag{5.2}$$

Proof. Much of the proof can follow Puhalskii and Reiman (2000), so we will be brief. We will start by indicating how we do the extension to G arrival processes. Given that the martingale proof is the "standard modern" argument, the extension to G arrival processes seems to be

the most interesting part. The fact that the arrival process is exogenous allows us to condition on it and then afterwards uncondition, and establish convergence.

As in Theorem 2.1, the arrival processes is created from a single rate-one arrival process by scaling according to (2.3). The scaled rate-one process in (2.1) satisfies the FCLT in (2.2). Thus we have the FCLT in (2.5). We now condition on possible realizations of these processes. For that purpose, for each n, let $\mathbf{c}_n$ be a possible realization of the scaled stochastic process $\mathbf{C}'_n$ in (2.6), and let $\mathbf{c}$ be a possible realization of the limit process $\mathbf{C}'$. Let $\mathbf{Q}_n^{\mathbf{c}_n}$ be the conditional scaled queue-length stochastic process $\mathbf{Q}_n$ in the $G/M/n/m_n + M$ model given that $\mathbf{C}'_n = \mathbf{c}_n$, and let $\mathbf{Q}^{\mathbf{c}}$ be the conditional limit process $\mathbf{Q}$ given that $\mathbf{C}' = \mathbf{c}$. Technically, it is significant that these conditional probabilities can be regular conditional probabilities; see Chapter 5 of Parthasarathy (1967). The martingale proof below establishes that

$$\mathbf{Q}_n^{\mathbf{c}_n} \Rightarrow \mathbf{Q}^{\mathbf{c}} \quad \text{in} \quad D \quad \text{whenever} \quad \mathbf{c}_n \to \mathbf{c} \quad \text{in} \quad D \ . \tag{5.3}$$

We establish the desired convergence $\mathbf{Q}_n \Rightarrow \mathbf{Q}$ by showing that

$$E[f(\mathbf{Q}_n)] \to E[f(\mathbf{Q})] \quad \text{as} \quad n \to \infty \tag{5.4}$$

for each continuous bounded real-valued function on D. For that purpose, observe that the limit (5.3) can be restated as

$$h_n(\mathbf{c}_n) \equiv E[f(\mathbf{Q}_n^{\mathbf{c}_n})] \to E[f(\mathbf{Q}^{\mathbf{c}})] \equiv h(\mathbf{c}) \quad \text{as} \quad n \to \infty \tag{5.5}$$

for all such f. We obtain the desired limit in (5.4) by combining (2.5) and (5.5); i.e.,

$$E[f(\mathbf{Q}_n)] = E[h_n(\mathbf{C}'_n)] \to E[h(\mathbf{C}')] = E[f(\mathbf{Q})] \quad \text{as} \quad n \to \infty \ , \tag{5.6}$$

by virtue of the generalized continuous-mapping theorem; Theorem 3.4.4 of SPL.

It remains to establish the limit in (5.3). For that purpose, it suffices to establish the limit under the condition that $\mathbf{C}'_n$ converges to $\mathbf{C}'$ with probability one, and, for that, we use the martingale proof, following the line of reasoning in Puhalskii and Reiman (2000). What makes the simple global conditioning argument work is the fact that the arrival process is exogenous in the queueing model.

Let $L_k(t)$ and $N_k(t)$ be mutually independent Poisson processes with rates θ and μ, respectively, for each k, $k \geq 1$. The unscaled number in system can be written as

$$Q_n(t) = Q_n(0) + A_n(t) - D_{n,1}(t) - D_{n,2}(t) \ , \tag{5.7}$$

where

$$A_n(t) \equiv \sum_{s:s\le t} (n + m_n - Q_n(s-)) \wedge \Delta C_n(s) \ ,$$

with the sum being over the jumps of the arrival process C_n, and

$$\begin{aligned} D_{n,1}(t) &\equiv \sum_{k=1}^{\infty} \int_0^t 1(Q_n(s-) \ge k+n)\, dL_k(s) \ , \\ D_{n,2}(t) &\equiv \sum_{k=1}^{n} \int_0^t 1(Q_n(s-) \ge k)\, dN_k(s) \ . \end{aligned}$$

Let

$$\begin{aligned} \Phi_n(t) &\equiv \sum_{s:s\le t} (\Delta C_n(s) + Q_n(s-) - n - m_n)^+, \\ \mathbf{M}_{n,1}(t) &\equiv n^{-1/2} \sum_{k=1}^{\infty} \int_0^t 1(Q_n(s-) \ge k+n)\, d(L_k(s) - \theta s) \ , \\ \mathbf{M}_{n,2}(t) &\equiv n^{-1/2} \sum_{k=1}^{n} \int_0^t 1(Q_n(s-) \ge k)\, d(N_k(s) - \mu s) \ . \end{aligned}$$

Then, by (5.7), we have the following equation for the scaled process

$$\begin{aligned} \mathbf{Q}_n(t) &= \mathbf{Q}_n(0) + \mathbf{C}_n'(t) - \theta \int_0^t (\mathbf{Q}_n(s) \vee 0)\, ds - \mu \int_0^t (\mathbf{Q}_n(s) \wedge 0)\, ds \\ &\quad -\mathbf{M}_{n,1}(t) - \mathbf{M}_{n,2}(t) - \Phi_n(t), \end{aligned} \tag{5.8}$$

where

$$\Phi_n(t) = \int_0^t 1(Q_n(s) = m_n + n)\, d\Phi_n(s) = \int_0^t 1(\mathbf{Q}_n(s) = m_n/\sqrt{n})\, d\Phi_n(s) \ . \tag{5.9}$$

To apply the martingale argument, we need to specify the filtration. Let the filtration $\mathbf{F}_n \equiv \{\mathcal{F}_n(t) : t \ge 0\}$ be defined by

$$\mathcal{F}_n(t) = \sigma(Q_n(0); L_k(s), 0 \le s \le t; N_k(s), 0 \le s \le t; k \ge 1) \ .$$

Then the processes $M_{n,1}$ and $M_{n,2}$ are orthogonal $\mathbf{F}_n$-locally-square-integrable martingales with predictable quadratic variation processes

$$\langle \mathbf{M}_{n,1} \rangle(t) = \theta \int_0^t (\frac{Q_n(s)}{n} - 1)^+ \, ds \tag{5.10}$$

and

$$\langle \mathbf{M}_{n,2} \rangle(t) = \mu \int_0^t (\frac{Q_n(s)}{n} \wedge 1)\, ds \ . \tag{5.11}$$

We next deduce limits in the fluid scale (when dividing by n instead of $\sqrt{n}$). By (5.8), the fact that $\mathbf{Q}_n(0)/\sqrt{n} \Rightarrow 0$, (5.10) and (5.11), we deduce that $\mathbf{Q}_n(t)/\sqrt{n} \Rightarrow 0$ uniformly in t over bounded intervals. Hence, by (5.10) and (5.11),

$$\langle \mathbf{M}_{n,1} \rangle(t) \Rightarrow 0 \quad \text{and} \quad \langle \mathbf{M}_{n,2} \rangle(t) \Rightarrow \mu t$$

uniformly in t over bounded intervals. Also the jumps of $\mathbf{M}_{n,1}$ and $\mathbf{M}_{n,2}$ uniformly go to 0. Thus, by the martingale central limit theorem, see Ethier and Kurtz (1986) or Liptser and Shiryaev (1989),

$$\mathbf{M}_{n,1} \Rightarrow \underline{0} \quad \text{and} \quad \mathbf{M}_{n,2} \Rightarrow \sqrt{\mu}\mathbf{W} \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ ,$$

where $\underline{0}(t) \equiv 0$, $t \geq 0$. Then we can apply the continuous mapping theorem with the reflection map in (5.8) and (5.9) to deduce the claimed limit. ∎

6. Proofs of the Corollaries

We conclude by proving the two corollaries in Section 2.

Proof of Corollary 2.1. We exploit the alternating-renewal-process construction used in the definition of the limit process $\mathbf{L}$ before the statement of Corollary 2.1 (without requiring the independence in the converging processes) and used in the proof of Theorem 2.1 in the case of a finite waiting room. With that explicit use of the reflecting upper barrier, we obtain converence of the upper-boundary regulator processes along with convergence of the content processes $\mathbf{Q}_n$ by an application of the continuous mapping theorem; see Sections 3.4, 5.2 and 13.5 of SPL. The same argument can be used for $\mathbf{Q}_n^a$.

Proof of Corollary 2.2. We apply Lemma A.2 of Puhalskii and Reiman (2000), just as they do to establish their Corollary 2.3. (In the statement of Lemma A.2, the condition $\lambda_N/N \to \lambda$ should be replaced by the stronger condition $(\lambda_N - N\lambda)/\sqrt{N} \to \beta$, which holds in our application.) Their Lemma A.2 draws upon Puhalskii (1994); see Theorem 13.7.4 of SPL and Section 5.4 of the Internet Supplement to SPL.

By (2.3), $C_n(t)$ counts the number of arrivals in the interval $[0,t]$ in model n. Let $C_n^{ad}(t)$ count the number of admitted customers in the interval $[0,t]$ in model n. Let

$$\begin{aligned} \mathbf{C}_n^1 &\equiv [C_n(t) - \lambda_n t]/\sqrt{\lambda_n c_a^2}, \\ \mathbf{C}_n^{ad} &\equiv [C_n^{ad}(t) - \lambda_n t]/\sqrt{\lambda_n c_a^2} \ . \end{aligned} \tag{6.1}$$

(We use the superscript in $\mathbf{C}_n^1$ to avoid confusion with $\mathbf{C}_n$ in (2.1).) By (2.2), $\mathbf{C}_n^1 \Rightarrow \mathbf{B}$, where $\mathbf{B}$ is standard Brownian motion. It is evidently possible, with some work, to extend Theorem 2.1 and Corollary 2.1 to obtain the joint convergence

$$(\mathbf{C}_n^1, \mathbf{Q}_n, \mathbf{L}_n) \Rightarrow (\mathbf{B}, \mathbf{Q}, \mathbf{L}) \quad \text{in} \quad (D, J_1)^3 \ , \tag{6.2}$$

but it is not necessary to do that. Tightness for the sequence $\{(\mathbf{C}_n^1, \mathbf{Q}_n, \mathbf{L}_n) : n \geq 1\}$ follows from the convergence of the component processes; see Theorems 11.6.1 and 11.6.7 of SPL. By Prohorov's theorem, that tightness implies relative compactness: Any subsequence has a convergent subsubsequence. Consider any convergent subsequence:

$$(\mathbf{C}_{n_k}^1, \mathbf{Q}_{n_k}, \mathbf{L}_{n_k}) \Rightarrow (\mathbf{C}^1, \mathbf{Q}, \mathbf{L}) \quad \text{in} \quad (D, J_1)^3 \ . \tag{6.3}$$

Since $\mathbf{C}_n^{ad} = \mathbf{C}_n^1 - \mathbf{L}_n$, we deduce that

$$(\mathbf{C}_{n_k}^{ad}, \mathbf{Q}_{n_k}) \Rightarrow (\mathbf{C}^1 - \mathbf{L}, \mathbf{Q}) \quad \text{in} \quad (D, J_1)^2 \ . \tag{6.4}$$

We can apply Lemma A.2 of Puhalskii and Reiman (2000) to (6.4) in order to obtain the limit (2.25) for that subsequence. Since the limit $\mathbf{Q}$ is independent of the subsequence chosen, we obtain the full convergence in (2.25). ∎

7. Acknowledgment

This research was done at AT&T Labs Research, Columbia University and Avaya Labs Research. At Columbia University, the author was supported by National Science Foundation Grant DMS-02-2340. The author thanks a Referee and an Associate Editor for helping to improve the paper. In particular, the martingale proof in Section 5, including the nice method for treating general G arrival processes, was suggested by the Referee. Theorem 3.1 had previously been proved by extending strong-approximation results in Mandelbaum and Pats (1995) to G arrival processes, assuming a strong approximation for the G arrival process (which itself is of some interest).

References

Billingsley, P. 1999. *Convergence of Probability Measures*, second ed., Wiley, New York.

Borovkov, A. A. 1984. *Asymptotic Methods in Queueing Theory*, Springer, New York.

Browne, S. and Whitt, W. 1995. Piecewise-linear diffusion processes. *Advances in Queueing*, J. Dshalalow (ed.), CRC Press, Boca Raton, FL, 463-480.

Ethier, S. N. and Kurtz, T. G. 1986. *Markov Processes, Characterization and Convergence*, Wiley, New York.

Garnett, O., Mandelbaum, A. and Reiman, M. I. 2003. Designing a call center with impatient customers. *Operations Res.*, to appear.

Glynn, P. W. and Whitt, W. 1991. A new view of the heavy-traffic limit theorem for the infinite-server queue. *Adv. Appl. Prob.* 23, 188-209.

Halfin, S. and Whitt, W. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29, 567-588.

Jelenkovic P., Mandelbaum A. and Momcilovic P. 2002. Heavy Traffic Limits for Queues with Many Deterministic Servers. Columbia University.

Karlin, S. and Taylor, H. M. 1981. *A Second Course in Stochastic Processes*, Academic Press, New York.

Krichagina, E. V. and A. A. Puhalskii. 1997. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* 25, 235–280.

Lions, P. L. and Sznitman, A. S. 1984. Stochastic differential equations with reflecting boundary conditions. *Commun. Pure Appl. Math.* 37, 511-537.

Liptser, R. Sh. and Shiryaev, A. N. 1989. *Theory of Martingales*, Kluwer.

Mandelbaum, A. and Pats, G. 1995. State-dependent queues: approximations and applications. *Stochastic Networks*, IMA Volumes in Mathematics and its Applications, F. P. Kelly and R. J. Williams, eds., Springer, Berlin, 239—282.

Massey, W. A., and Wallace, R. 2002. An asymptotically optimal design of the $M/M/c/k$ queue for call centers. Department of Operations Research and Financial Engineering, Princeton University.

Parthasarathy, K. R. 1967. *Probability Measures on Metric Spaces*, Academic, New York.

Puhalskii, A. A. 1994. On the invariance principle for the first passage time. *Math. Oper. Res.* 19, 946-954.

Puhalskii, A. A. and Reiman, M. I. 2000. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Prob.* 32, 564-595.

Rogers, L. C. G. and Williams, D. 1987. *Diffusions, Markov Processes and Martingales, Volume 2: Itô Calculus*, Wiley, New York.

Srikant, R. and Whitt, W. 1996. Simulation run lengths to estimate blocking probabilities. *ACM Trans. Modeling and Computer Simulation* 6, 7-52.

Stroock, D. and Varadhan, S. R. S. 1979. *Multidimensional Diffusion Processes*, Springer, New York.

Ward, A. R. and Glynn, P. W. 2003. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43, 103-128.

Whitt, W. 1981. Comparing counting processes and queues. *Adv. Appl. Prob.* 13, 207-220.

Whitt, W. 1983. Comparison conjectures about the $M/G/s$ queue. *Operations Research Letters* 2, 203-210.

Whitt, W. 1984. On approximations for queues, III: mixtures of exponential distributions. *AT&T Bell Lab. Tech. J.* 63, 163-175.

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

Whitt, W. 2004. A diffusion approximation for the $G/GI/n/m$ Queue. Department of Industrial Engineering and Operations Research, Columbia University. *Operations Res.*, to appear.