Applied Probability Trust (16 April 2010)

HEAVY-TRAFFIC LIMITS FOR NEARLY DETERMINISTIC QUEUES

KARL SIGMAN AND WARD WHITT,* *Columbia University*

Abstract

We establish heavy-traffic limits for nearly deterministic queues, such as the $G/D/n$ many-server queue. Since waiting times before starting service in the $G/D/n$ queue are equivalent to waiting times in an associated $G_n/D/1$ model, where the G_n interarrival times are the sum of n consecutive interarrival times in the original model, we focus on the $G_n/D/1$ model and the generalization to $G_n/G_n/1$, where "cyclic thinning" is applied to both the arrival and service processes. We establish different limits in two cases: (i) when $(1-\rho_n)\sqrt{n} \to \beta$ as $n \to \infty$ and (ii) $(1-\rho_n)n \to \beta$ as $n \to \infty$, where ρ_n is the traffic intensity in model n. The nearly deterministic feature leads to interesting nonstandard scaling.

Keywords: heavy traffic; nearly deterministic queues; Erlang queues; many-server queues; deterministic service times; Gaussian random walk; cyclic thinning; functional central limit theorems; invariance principles.

2000 Mathematics Subject Classification: Primary 60K25

Secondary 60F17; 90B22

1. Introduction

A primary cause of congestion in a queueing system is stochastic fluctuations in the arrival times and the service times. We say that a queueing system is *nearly deterministic* if these stochastic fluctuations are low. At customary loads, the congestion in a nearly deterministic queueing system will be negligible. However, if the system is nearly deterministic, then it is natural to operate the system at higher loads. In this paper we explore the interplay between low variability and high loads. In particular, we establish heavy-traffic (HT) limits for some nearly deterministic queueing models.

* Postal address: Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA; {ks20, ww2040}@columbia.edu

A classic example of a nearly deterministic queueing model is the $GI/D/n$ multiserver queue when n is large. The $GI/D/n$ model has n homogeneous servers working in parallel, an unlimited waiting room, the first-come first-served (FCFS) service discipline, identical deterministic service times and a renewal arrival process with a general interarrival-time distribution. It is well known that waiting times (before starting service) in this model can be identified with waiting times in the corresponding $GI_n/D/1$ model, where the GI_n means that the arrival process is the renewal process whose interarrival times are distributed as the sum of n interarrival times in the original renewal arrival process; e.g., see Theorem 4.6.1 of [13]. That occurs because, without loss of generality, the customers can be assigned to the servers in a round robin or cyclic order. For large n, these $GI_n/D/1$ models become nearly deterministic, approaching the $D/D/1$ queue. Of course, the service times are completely deterministic from the outset, but also the GI_n interarrival times become nearly deterministic as n increases by virtue of the law of large numbers. For example, if the original GI interarrival times have squared coefficient of variation (scv, variance divided by the square of the mean) c_a^2, then the GI_n interarrival times have scv c_a^2/n, which converges to 0 as $n \to \infty$.

In applications, a Poisson arrival process is often a realistic assumption. The reduction of $M/D/n$ to $E_n/D/1$ for the waiting times is often mentioned in textbooks. Otherwise, the renewal process assumption is not so realistic. Thus, it is important that both the reduction of the $GI/D/n$ model to the $GI_n/D/1$ model and our HT limits hold for more general "G" arrival processes. There are no algorithms available to compute the steady-state waiting time distribution or even only its mean in the new $G_n/D/1$ model. Thus, the simple approximations stemming from the HT limits we establish here can be very useful. Consistent with the large body of HT literature, we only assume that the general G arrival counting process or, equivalently, the associated partial sums of consecutive interarrival times, satisfies a functional central limit theorem (FCLT); see §4.4 of [14] for examples with dependence that are covered.

Motivated by the example above, we will consider the waiting time process in single-server queues with the $G_n/G_n/1$ structure, where cyclic thinning is applied to *both* the interarrival times and the service times. Our results cover the two models $D/G_n/1$ and $G_n/D/1$ as special cases, because G_n coincides with the original G when the G is D; i.e., $D_n = D$. That is so, because the deterministic renewal process (D) is the

unique fixed point among renewal processes of the operation mapping GI into GI_n, when we rescale to fix the mean. (Uniqueness follows immediately from the scv's.) When working with partial sums of interarrival times or service times, we let the new sequence of G_n partial sums $\{S_{n,k} : k \geq 1\}$ be defined in terms of the sequence of original G partial sums $\{S_{1,k} : k \geq 1\}$ by letting $S_{n,k} \equiv S_{1,kn}/n$, $k \geq 1$; i.e., we scale the index k in the original partial sums $S_{1,k}$ by n because we add n consecutive times, but we also divide by n in order to keep the mean fixed in the identically distributed case. It is easy to see that $D_n = D$ with this construction.

Although we primarily focus on the $G_n/G/n/1$ model, which includes $D/G_n/1$, $G_n/D/1$ and $G/D/n$ as special cases, our analysis also applies to more general nearly deterministic queueing models, provided that there is appropriate statistical regularity associated with the low variability. As we explain in §4.3, we actually only require that the sequence of arrival and service processes associated with the sequence of queueing models satisfy a FCLT, stemming from their being asymptotically deterministic. The cyclic thinning that converts a G point process into a G_n point process is one natural mechanism producing asymptotically deterministic behavior characterized by a FCLT. However, for clarity we focus on the concrete $G_n/G_n/1$ framework.

If the traffic intensity ρ_n in the $G_n/G_n/1$ model (assumed well defined) is held fixed at a stable value or, more generally, satisfies $\rho_n \to \rho < 1$ as $n \to \infty$, then the $G_n/G_n/1$ model approaches the purely deterministic $D/D/1$ model, and the stationary waiting time becomes asymptotically negligible. However, we let $\rho_n \uparrow 1$ as $n \to \infty$. We thus obtain an interesting *double limit*, in which the models approach $D/D/1$, while the traffic intensity increases. On the one hand, congestion should decrease, because the models are becoming less variable, approaching $D/D/1$. On the other hand, the congestion should increase because we let $\rho_n \uparrow 1$. We let ρ_n approach 1 at an appropriate rate so that we get revealing nondegenerate limits.

For the multiserver $G/D/n$ model mentioned at the outset, the double limit coincides with the familiar many-server HT limit, in which we let the traffic intensities ρ_n approach 1 as the number of servers, n, increases, e.g., see [6, 10]. We consider the so-called Halfin-Whitt or quality-and-efficiency-driven (QED) regime, in which $(1-\rho_n)\sqrt{n} \to \beta$, $0 < \beta < \infty$. However, we also consider the case in which $(1-\rho_n)n \to \beta$, $0 < \beta < \infty$. In that case, we obtain a nondegenerate limit for the *un-normalized*

waiting times.

The asymptotically-deterministic feature is critical for these new limits. For example, the HT limits for the $G_n/GI/1$ model as $n \to \infty$ with fixed service-time distribution are significantly different in the two cases: (i) when the GI service-time distribution is D and (ii) when the service-time distribution is not D (and we do not perform the cyclic thinning on the service times, replacing GI by GI_n). When the service-time distribution is not deterministic, the $G_n/GI/1$ model is not asymptotically deterministic as $n \to \infty$. As a consequence, the HT limit agrees with the conventional one for the corresponding $D/GI/1$ model, with the usual scaling, obtained by simply replacing the interarrival-time distribution in the G_n process by a deterministic interarrival times with the same mean. In contrast, that is *not* the case with the nearly deterministic $G_n/D/1$ model.

In many ways, the HT behavior of the $G_n/G_n/1$ models as $n \to \infty$ is different from the conventional HT behavior of the $GI/GI/1$ model, discussed in Chapters 5 and 9 of [14]. Unlike the conventional HT theory for the $GI/GI/1$ model, for the $G_n/G_n/1$ models there need not be any spatial scaling. In the conventional HT limit, the queue-length and waiting time processes have the same asymptotic behavior; both processes behave like reflected Brownian motion, after the same scaling. In contrast, for the $G_n/G_n/1$ models, the waiting-time and queue-length processes look very different.

To illustrate these differences, and to provide motivation for the results to follow, we now plot sample paths of the waiting times (before starting service) of successive arrivals and the continuous-time queue-length process from one simulation run of the $E_{100}/D/1$ queue with traffic intensity $\rho = 0.99$ and unit service times. Figure 1 shows the waiting times at arrival epochs and the continuous-time queue length process, starting empty, in the final subinterval of length 200 ending at $t = 5 \times 10^4$ from a single run over the time interval $[0, 5 \times 10^4]$. (The entire sample paths are displayed in the appendix.)

First, all values of both processes over the full time interval of length $50,000$ fall in the interval $[0, 5]$ without any spatial scaling. The waiting times are comparable to the unit service times, e.g., the average waiting time is about 0.5. Second the waiting-time plots look continuous, like a plot of reflected Brownian motion, which we will show is indeed its HT limit. In contrast, the queue-length process is integer-valued, making

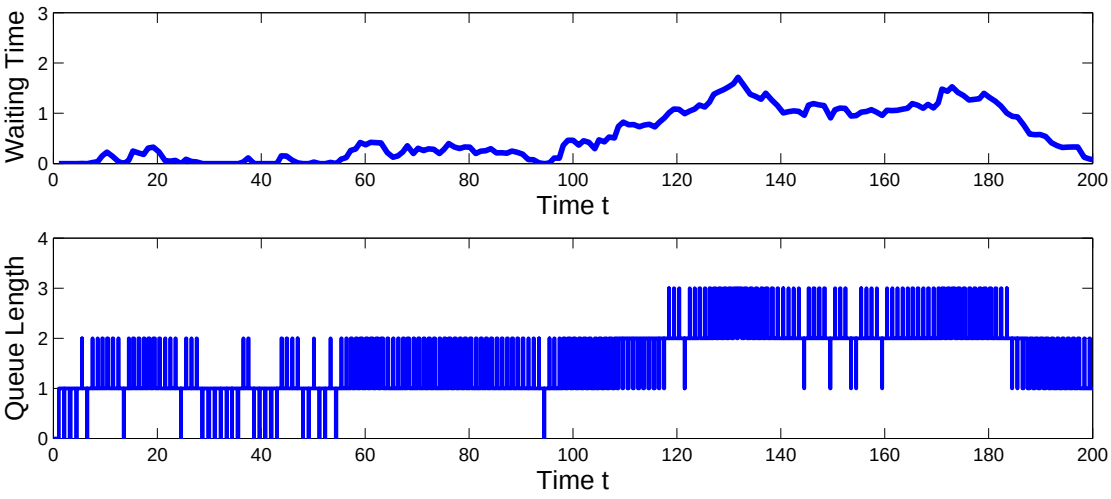

FIGURE 1: Simulation plots of the waiting times at arrival epochs and queue lengths at arbitrary times in the $E_{100}/D/1$ model with $\rho = 0.99$, starting empty, for a time interval of length 200 ending at $t = 50,000$.

frequent jumps of size 1. Evidently its limiting behavior is more complicated. We will explain these plots in the rest of this paper.

2. Related literature and organization

Three motivating precedents. In doing this work, we were motivated by three precedents in the literature: (i) the 1993 paper by Abate, Choudhury and Whitt [1], (ii) the 1996 paper by Song and Zipkin [12] and (iii) the 2004 paper by Jelenkovic, Mandelbaum and Momcilovic [7].

In [1], the authors developed an algorithm to compute the distribution and its cumulants of the steady-state waiting-time distribution in general $GI/GI/1$ queues. In order to demonstrate the power of this algorithm, examples of various models were considered that should be challenging by other methods. Example 3.1 of [1] considers the high-order Erlang model $E_n/E_n/1$. The algorithm was applied for $n = 10^k$ for $k = 1, \ldots, 4$. However, it was observed that the waiting time would become negligible unless the traffic intensity ρ_n in model n were allowed to increase with n. Numerical

results in Table 1 of [1] show that the waiting time distribution converges to a mean-1 exponential distribution as $n \to \infty$ when $(1 - \rho_n)n = 1$ as $n \to \infty$. Table 1 of [1] show that the HT approximation is remarkably accurate for the $E_n/E_n/1$ model with large n. At that time, this limiting result was confirmed mathematically using the transform method for establishing HT limits due to Kingman [8]. Here we show that result generalizes, first, to more general models, second, to transient as well as steady-state waiting times and, third, to other related processes, such as the queue length.

In [12], the authors studied a basic (r, q) inventory model, in which the demand forms a Poisson process at rate λ and the lead times are i.i.d. distributed as L. Every q^{th} demand from the Poisson process triggers an order requiring time L to arrive. Thus there is a $E_q/GI/\infty$ queue in the background. They were interested in the joint effect upon performance of $Var(L)$ and the lot size q. Because the model is intractable, they use a HT approximation. They first consider the case of deterministic lead times, yielding the $E_k/D/\infty$ queue. They then apply the standard HT limit for the $GI/D/\infty$ queue, as in [5]. They then make the observation (on p. 1356) that the "interesting" (e.g., optimal) value for q should be of $O(\sqrt{\lambda})$. That observation suggests considering the *joint* HT limit in which $\lambda \to \infty$ and $q \equiv q(\lambda) \to \infty$ with $q(\lambda) = \sqrt{\lambda}$. However, as in [5], in the $G/D/\infty$ model the queue length (number of busy servers) at time t, $Q(t)$, can be expressed directly in terms of the arrival counting process $N(t)$: with unit service times $Q_q(t) = N_q(t) - N_q(t-1)$. Thus, it is natural to ask about limits for the counting process in which $\lambda \to \infty$ and $q \equiv q(\lambda) \to \infty$ with $q(\lambda) = \sqrt{\lambda}$. The counting process $\{N_q(t) : t \geq 0\}$ itself is interesting, being a deterministic cyclic thinning of a base counting process. We investigate these counting processes here in §5.We show that a conventional HT limit does not exist, but unconventional ones do.

In [7], the authors establish a HT limit for the steady-state waiting time in the $GI/D/n$ model. They consider the conventional QED many-server regime in which $(1 - \rho_n)\sqrt{n} \to \beta$, $0 < \beta < \infty$. Under that scaling, they obtained a nondegnerate limit for the scaled steady-state waiting time $\sqrt{n}W_{n,\infty}$ (which implies that $W_{n,\infty} \Rightarrow 0$). This attracted our attention, because at first it seemed inconsistent with the previous results in [1]. At first, we thought that one must be incorrect, but later we discovered that was not so. We demonstrate that here, by obtaining HT limits in *both* regimes:

$(1-\rho_n)\sqrt{n} \to \beta$ and $(1-\rho_n)n \to \beta$.

Organization. Here is how the rest of this paper is organized. In §3 we review the framework for the conventional HT limit for the waiting times in the single server queue. We do this in a way that makes our limits for the $G_n/G_n/1$ model easy to establish. In §4 we establish the HT FCLT's for the waiting times in the $G_n/G_n/1$ model with the two different scalings. In §4.3 we show that the limits for $G_n/G_n/1$ model also extend to other nearly deterministic queueing models; it suffices to have appropriate FCLT's hold for the arrival and service processes in the sequence of queueing models. In §5 we establish limits for associated counting processes. These require further unconventional scaling plus an unconventional space and topology. In §6 we apply the results in previous sections to obtain heavy-traffic limits for associated continuous-time processes, such as the workload and queue length. In §7 we apply the results in §5.3 to obtain a HT FCLT for the $G_n/D/\infty$ model related to the motivating inventory problem from [12].

In a sequel to this paper, [11], we obtain additional heavy-traffic limits for the stationary distributions. In particular, we justify the limit interchange $\lim_{t\to\infty}\lim_{n\to\infty} = \lim_{n\to\infty}\lim_{t\to\infty}$. We first show how to place cyclically thinned point processes in a proper stationary framework. We also compare the approximations for steady-state distributions to simulation estimates for the $E_{100}/D/1$ ($M/D/100$), $D/E_{100}/1$ and $E_{100}/E_{100}/1$ models with traffic intensities $\rho = 0.99$ and $\rho = 0.90$. In another sequel [9] we study the heavily loaded $GI/D/n + GI$ many-server queue with customer abandonment (the $+GI$). We show that the nearly deterministic nature leads to periodic behavior, which can be understood through a careful study of the limiting deterministic fluid model.

3. Background for general single-server queues

We will obtain very short proofs of the HT limits for the $G_n/G_n/1$ models in §4 by relating the basic processes to those arising in the conventional HT limit for the G/G/1 model. Hence we review that framework in detail in this section; see Chapters 5 and 9 of [14] for background.

3.1. The standard double-sequence framework

The standard framework for HT limits involves a sequence of queueing models, which we take to be indexed by $n \geq 1$. We first consider general single-server queues with unlimited waiting room and the FCFS service discipline. For each $n \geq 1$, the model can be specified by a sequence of ordered pairs of random variables $\{(U_{n,k}, V_{n,k-1}) : k \geq 1\}$, with $U_{n,k}$ representing the interarrival time between customers $k-1$ and k and $V_{n,k}$ representing the service time of customer k. We assume that a 0^{th} customer arrives at time 0 and experiences an initial wait $W_{n,0}$. (That is due to customers initially in the system at time 0. To describe the waiting times of new customers, we do not need to identify these old customers and their service times.) Let $W_{n,k}$ be the waiting time (before beginning service) of customer k in model n. The waiting times can be defined recursively by

$$W_{n,k} \equiv [W_{n,k-1} + V_{n,k-1} - U_{n,k}]^+, \quad k \geq 1, \tag{3.1}$$

where $[x]^+ \equiv \max\{x, 0\}$ and $W_{n,0}$ is the initial wait. As a consequence, the waiting times can be expressed directly in terms of partial sums and the initial wait $W_{n,0}$ via

$$W_{n,k} = W_{n,0} + S_{n,k} - \min_{0 \leq j \leq k} \{(W_{n,0} + S_{n,j}) \wedge 0\}, \quad k \geq 0, \tag{3.2}$$

where $a \wedge b \equiv \min\{a, b\}$,

$$S_{n,k} \equiv X_{n,1} + \cdots + X_{n,k} \quad \text{for} \quad X_{n,k} \equiv V_{n,k-1} - U_{n,k}, \quad k \geq 1, \tag{3.3}$$

with $S_{n,0} \equiv 0$, so that $S_{n,k} = S^v_{n,k} - S^u_{n,k}$ with

$$S^u_{n,k} \equiv U_{n,1} + \cdots + U_{n,k}, \quad S^v_{n,k} \equiv V_{n,0} + \cdots + V_{n,k-1}, \quad k \geq 1, \tag{3.4}$$

$S^v_{n,0} \equiv 0$ and $S^u_{n,0} \equiv 0$; see §9.2 of [14].

Note that formula (3.2) can be regarded as a discrete reflection map, mapping the space $\mathbb{R}^\infty$ of sequences $x \equiv \{x_k : k \geq 0\}$ into itself; i.e., $W_n = \tilde{\phi}(W_{n,0} + S_n)$ for $W_n \equiv \{W_{n,k} : k \geq 0\}$, $S_n \equiv \{S_{n,k} : k \geq 0\}$ and $W_{n,0} + S_n \equiv \{W_{n,0} + S_{n,k} : k \geq 0\}$, where $\tilde{\phi} : \mathbb{R}^\infty \to \mathbb{R}^\infty$ is defined by

$$\tilde{\phi}(k) \equiv x_k - \min_{0 \leq j \leq k} \{x_j \wedge 0\}, \quad k \geq 0. \tag{3.5}$$

The standard HT limit is for a sequence of random elements in the function space $D \equiv D([0, \infty), \mathbb{R})$ of all right-continuous real-valued functions on the positive half line

with limits from the left everywhere (except at 0), endowed with the standard Skorohod (J_1) topology; see [3, 14]. The HT limit involves scaling space and time. For a real number t, let $\lfloor t \rfloor$ be the floor function, giving the greatest integer less than or equal to t. Let random elements associated with the sequences above be defined by

$$\begin{aligned} \mathbf{S}_n^u(t) &\equiv \frac{S^u_{n,\lfloor nt \rfloor} - \lfloor nt \rfloor}{\sqrt{n}}, \quad \mathbf{S}_n^v(t) \equiv \frac{S^v_{n,\lfloor nt \rfloor} - \lfloor nt \rfloor}{\sqrt{n}}, \\ \mathbf{S}_n(t) &\equiv \frac{S_{n,\lfloor nt \rfloor}}{\sqrt{n}} \quad \text{and} \quad \mathbf{W}_n(t) \equiv \frac{W_{n,\lfloor nt \rfloor}}{\sqrt{n}}. \end{aligned} \tag{3.6}$$

Let $D^k \equiv D \times \cdots \times D$ be the k-fold product space of D with itself; let C and C^k be the subsets of continuous functions in D and D^k, respectively, and let $\Rightarrow$ denote convergence in distribution. Let $\phi : D \to D$ be the one-dimensional reflection map, defined by

$$\phi(x)(t) \equiv x(t) - \inf_{0 \le s \le t} \{x(s) \wedge 0\}, \quad t \ge 0; \tag{3.7}$$

see §§3.5 and 13.5 of [14]. The following result is now well known; see Chapter 9 of [14], especially Theorems 9.3.1 and 9.3.3.

Theorem 3.1. (HT limit for the waiting times in $G/G/1$ models) *If*

$$(\mathbf{W}_n(0), \mathbf{S}_n^u, \mathbf{S}_n^v) \Rightarrow (\mathbf{W}(0), L^u, L^v) \quad \textit{in} \quad \mathbb{R} \times D^2, \tag{3.8}$$

where $P((L^u, L^v) \in C^2) = 1$, *then*

$$(\mathbf{S}_n^u, \mathbf{S}_n^v, \mathbf{S}_n, \mathbf{W}_n) \Rightarrow (L^u, L^v, L, \mathbf{W}) \quad \textit{in} \quad D^4 \quad \textit{as} \quad n \to \infty, \tag{3.9}$$

where $L \equiv L^v - L^u$, $\mathbf{W} \equiv \phi(\mathbf{W}(0) + L)$ *and the limit is in* C^4 *w.p.1.*

Proof. Apply the continuous mapping theorem with the addition and reflection functions, because $\mathbf{W}_n(0) + \mathbf{S}_n = \mathbf{W}_n(0) + \mathbf{S}_n^v - \mathbf{S}_n^u$ and $\mathbf{W}_n = \phi(\mathbf{W}_n(0) + \mathbf{S}_n)$.

3.2. Scaling unit-rate processes

For simplicity, and without practical loss of generality, we can construct the sequence of sequences $\{\{(U_{n,k}, V_{n,k-1}) : k \ge 1\} : n \ge 1\}$ specifying the sequence of queueing models starting from a single sequence of ordered pairs of random variables $\{(U_k, V_{k-1}) : k \ge 1\}$. This simplification is important for us, because we want to simultaneously consider different scaling in a common framework.

Paralleling (3.4), let

$$S_k^u \equiv U_1 + \cdots + U_k, \quad \text{and} \quad S_k^v \equiv V_0 + \cdots + V_{k-1}, \quad k \geq 1, \tag{3.10}$$

$S_0^v \equiv 0$ and $S_0^u \equiv 0$.

We have not yet specified any specific stochastic properties. As a canonical case, we have in mind the special case in which $\{U_k : k \geq 1\}$ and $\{V_{k-1} : k \geq 1\}$ are independent sequences of i.i.d. random variables with means $E[U_k] = E[V_k] = 1$. With that case in mind (but not assumed), we define the usual sequence of random elements of D associated with this sequence $(\hat{\mathbf{S}}^u, \hat{\mathbf{S}}^v) \equiv \{(\hat{\mathbf{S}}_k^u, \hat{\mathbf{S}}_k^v) : k \geq 0\}$ by

$$\hat{\mathbf{S}}_n^u(t) \equiv \frac{S_{\lfloor nt \rfloor}^u - \lfloor nt \rfloor}{\sqrt{n}}, \quad \text{and} \quad \hat{\mathbf{S}}_n^v(t) \equiv \frac{S_{\lfloor nt \rfloor}^v - \lfloor nt \rfloor}{\sqrt{n}}, \quad t \geq 0. \tag{3.11}$$

In this context, our basic assumption is that the sequence $\{(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v) : n \geq 1\}$ converges, i.e., the partial sums satisfy a joint FCLT.

To construct a sequence of $G/G/1$ models in which the arrival rate and, thus, the traffic intensity are ρ_n in model n, where $\rho_n \uparrow 1$ as $n \to \infty$, we use the given service-time sequence for all n and introduce extra scaling in the interarrival times; i.e., we let

$$V_{n,k} \equiv V_k \quad \text{and} \quad U_{n,k} \equiv \frac{U_k}{\rho_n} \quad \text{for all} \quad n, k \geq 1, \tag{3.12}$$

with the understanding that $0 < \rho_n < 1$ and that we intend to let $\rho_n \uparrow 1$ as $n \to \infty$. We have thus defined a sequence of queueing models as in §3.1.

We are now ready to establish the HT limit theorem for the waiting times in this context. For that purpose, let e be the identity function in D, i.e., $e(t) = t$, $t \geq 0$, and let $\stackrel{\mathrm{d}}{=}$ mean equality in distribution (as a process). We will also describe the standard special case, which involves standard Brownian motion (BM), which has zero drift and unit diffusion coefficient. Then the HT limit $\phi(L)$ for the waiting times becomes reflected Brownian motion (RBM) with negative drift.

Theorem 3.2. (HT limit in the single sequence framework) *Suppose that*

$$(\mathbf{W}_n(0), \hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v) \Rightarrow (\mathbf{W}(0), \hat{L}^u, \hat{L}^v) \quad \textit{in} \quad \mathbb{R} \times D^2 \tag{3.13}$$

for $(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v)$ *in* (3.11)*, where* $P((\hat{L}^u, \hat{L}^v) \in C^2) = 1$*. If*

$$(1 - \rho_n)\sqrt{n} \to \beta, \quad 0 < \beta < \infty, \quad \textit{as} \quad n \to \infty, \tag{3.14}$$

then the conditions and conclusions of Theorem 3.1 *hold with*

$$L^u = \hat{L}^u + \beta e \quad \textit{and} \quad L^v = \hat{L}^v. \tag{3.15}$$

If, in addition,

$$(\hat{L}^u, \hat{L}^v) = (\sigma_u B_u, \sigma_v B_v), \tag{3.16}$$

where $\mathbf{W}(0)$*,* B_u *and* B_v *are mutually independent, and* B_u *and* B_v *are standard Brownian motions, then* $L \equiv L^v - L^u \stackrel{\mathrm{d}}{=} \sigma B - \beta e$*,* B *is a standard BM and* $\sigma^2 + \sigma_u^2 + \sigma_v^2$*, so that the limit* (3.9) *holds with* $\mathbf{W}_n \Rightarrow \mathbf{W} \equiv \phi(\mathbf{W}(0) + L) = \phi(\mathbf{W}(0) + \sigma B - \beta e)$*, which is RBM with drift* $-\beta$ *starting at an independent random initial state* $\mathbf{W}(0)$*. Furthermore, if* (3.13) *also holds with* $\mathbf{W}(0)$ *exponentially distributed with mean* $\sigma^2/2\beta$*, then the limit* $\mathbf{W}$ *is a stationary RBM.*

Proof. Under condition (3.14),

$$\frac{1}{\rho_n} = \frac{1}{1 - (\beta/\sqrt{n}) + o(1/\sqrt{n})} = 1 + (\beta/\sqrt{n}) + o(1/\sqrt{n}) \tag{3.17}$$

as $n \to \infty$. Hence

$$\mathbf{S}_n^u(t) = \hat{\mathbf{S}}_n^u(t) + \left(\frac{S^u_{\lfloor nt \rfloor}}{n}\right)(\beta + o(1)) \tag{3.18}$$

as $n \to \infty$. The assumed FCLT implies a corresponding FWLLN, so that $\hat{\mathbf{S}}_n^u/\sqrt{n} \Rightarrow 0e$ as $n \to \infty$. Hence, the second term on the right in (3.18) converges to βe, so that the conditions of Theorem 3.1 are satisfied with (3.15). It is well known that the RBM converges to an exponential distribution with mean $\sigma^2/2\beta$ as $t \to \infty$.

The standard special case in which conditions (3.13) and (3.16) hold is the $GI/GI/1$ queue, where first $\{W_{n,0} : n \geq 1\}$ is independent of $\{(U_k, V_{k-1}) : k \geq 1\}$ and second $\{U_k : k \geq 1\}$ and $\{V_k : k \geq 1\}$ can be taken to be independent sequences of i.i.d. random variables with unit means, $E[U_1] = E[V_1] = 1$, and variances $Var(U_1) = \sigma_u^2$ and $Var(V_1) = \sigma_v^2$.

In a fixed $G/G/1$ system, for suitably large k (e.g., the stationary distribution), we would thus use the approximation based on setting $\beta = 1$ and $\sqrt{n} = 1/(1-\rho)$, i.e.,

$$P(W_k > x) \approx e^{-2(1-\rho)x/\sigma^2}, \quad x \geq 0, \quad \text{and} \quad E[W_k] \approx \frac{\sigma^2}{2(1-\rho)}. \tag{3.19}$$

However, Theorems 3.1 and 3.2 do not directly imply convergence of the stationary distributions; we discuss that issue in [11].

4. Two heavy-Traffic limits for $G_n/G_n/1$ models

In the setting of §3.1, involving a sequence of $G/G/1$ queueing models indexed by n, we can obtain a sequence of nearly deterministic queueing models if we assume that cyclic thinning is performed on both the interarrival times and the service times for the n^{th} queueing model, with the cycle length increasing as $n \to \infty$. With "cyclic thinning" of a point process of order n, we select every n^{th} point; i.e., the k^{th} point in the thinned process is point kn of the original process. In this context, we call n the cycle length. In this section we assume that the cycle length in model n is n, and refer to the model as the $G_n/G_n/1$ model. In this way, we map an original "base" sequence of $G/G/1$ models into a sequence of $G_n/G_n/1$ models.

In the framework of §3.1 above, we replace the partial sums $S^u_{n,k}$ and $S^v_{n,k}$ with new partial sums $S^{c,u}_{n,k}$ and $S^{c,v}_{n,k}$ defined by

$$S^{c,u}_{n,k} \equiv S^u_{n,kn}/n \quad \text{and} \quad S^{c,v}_{n,k} \equiv S^v_{n,kn}/n \quad \text{for all} \quad n \geq 1 \quad \text{and} \quad k \geq 1. \tag{4.20}$$

Then let the associated interarrival times and service times be defined in terms of the increments by

$$U^c_{n,k} \equiv S^{c,u}_{n,k} - S^{c,u}_{n,k-1} \quad \text{and} \quad V^c_{n,k-1} \equiv S^{c,v}_{n,k} - S^{c,v}_{n,k-1}, \tag{4.21}$$

From (4.20) and (4.21), we see that each new interarrival time is the sum of n of the original interarrival times in model n, but we also divide the sums by n to leave the means unchanged (in the case of identically distributed random variables).

We also must treat the initial conditions. We assume that does not get transformed by cyclic thinning. Hence, we have $S^c_{n,0} \equiv S_{n,0} \equiv W_{n,0}$ for each n. We will assume that the initial conditions scale differently. However, there would be no difference if the systems start empty.

The canonical example starts with base $M/M/1$ models; then the associated $G_n/G_n/1$ models are the Erlang $E_n/E_n/1$ models, where the mean interarrival times and service times are the same as in the corresponding base model. As discussed in §1, it is significant that the sequence of $G_n/G_n/1$ models constructed by applying definition (4.20) to a base sequence of $G/G/1$ models also includes as special cases the sequence of $G_n/D/1$ and $D/G_n/1$ models, where cyclic thinning is applied to only the interarrival

times alone or the service times alone, provided that the other component is D, where D is interpreted as deterministic and constant random variables.

In this section we will show that HT limits for a base sequence of $G/G/1$ "base" models translate into corresponding HT limits for the sequence of $G_n/G_n/1$ models. As n increases, the sequence of $G_n/G_n/1$ models becomes nearly deterministic. By the law of large numbers, which follows as a consequence of the assumed FCLT in Theorem 3.1, the interarrival times and service times in the thinned process tend to approach a deterministic limit. Thus, the sequence of $G_n/G_n/1$ queueing models approaches a purely deterministic $D/D/1$ model as n increases. However, we obtain nontrivial limiting behavior by letting the load increase as n increases. We will show that this can be done in two different ways, depending upon the scaling. We will get different limiting behavior in the two cases:

$$\begin{array}{ll} (i) & (1-\rho_n)\sqrt{n} \to \beta \quad \text{as} \quad n \to \infty, \quad \text{where} \quad 0 < \beta < \infty \quad \text{and} \\ (ii) & (1-\rho_n)n \to \beta \quad \text{as} \quad n \to \infty, \quad \text{where} \quad 0 < \beta < \infty. \end{array} \tag{4.22}$$

Case (i) in (4.22) is the traditional scaling used in §3. However, because of the nearly deterministic nature of the queueing models, we need to *scale up* the waiting times by $\sqrt{n}$ in order to get a nondegenerate limit in case (i). That is in stark contrast with (3.5), where we had to *scale down* the waiting times by $\sqrt{n}$. For the first case, the stationary waiting times were treated previously in [7] for the special case of the $GI_n/D/1$ model, which has the $GI/D/1$ base model. In case (ii), even with the more rapid increase of ρ_n, we obtain a nondegenerate limit for the waiting times without any spatial scaling.

4.1. Limits for scaled waiting times in case (i)

We will express the HT limit for case (i) in terms of random elements of $\mathbb{R}^\infty$, using the discrete reflection map $\tilde{\phi}$ defined in (3.5). For that purpose, we introduce the following random elements of $\mathbb{R}^\infty$: let

$$\begin{array}{rcl} \tilde{S}_n^{c,u}(k) & \equiv & \sqrt{n}(S_{n,k}^{c,u} - k), \quad \tilde{S}_n^{c,v}(k) \equiv \sqrt{n}(S_{n,k}^{c,v} - k), \\ \tilde{S}_n^{c}(k) & \equiv & \sqrt{n}S_{n,k}^{c}, \quad \text{and} \quad \tilde{W}_n^{c}(k) \equiv \sqrt{n}W_{n,k}^{c}, \quad k \geq 1, \quad n \geq 1, \end{array} \tag{4.23}$$

with $\tilde{S}_n^c(0) \equiv \sqrt{n}S_{n,0} \equiv \sqrt{n}W_{n,0}$, where $(S_{n,k}^{c,u}, S_{n,k}^{c,v})$ is defined in (4.20), $S_{n,k}^c \equiv S_{n,k}^{c,v} - S_{n,k}^{c,u}$ and $W_{n,k}^c$ is defined in terms of $\{S_{n,k}^c : k \geq 0\}$ as in (3.1). The scaling in which we multiply by $\sqrt{n}$ converts the HT problem into a model continuity problem. When we consider HT limits for the stationary waiting times in [11], we apply the model continuity results in §X.6 of [2] and Chapter 4 of [4].

Theorem 4.1. (HT limit for the scaled waiting times in the $G_n/G_n/1$ models) *Consider a sequence of $G_n/G_n/1$ models associated with a base sequence of $G/G/1$ models satisfying*

$$(\sqrt{n}W_{n,0}, \mathbf{S}_n^u, \mathbf{S}_n^v) \Rightarrow (\tilde{W}(0), L^u, L^v) \quad \textit{in} \quad \mathbb{R} \times D^2, \tag{4.24}$$

where $P((L^u, L^v) \in C^2) = 1$, as in Theorem 3.1*, but the scaling of the initial conditions is changed. Then*

$$(\tilde{W}_n^c(0), \tilde{S}_n^{c,u}, \tilde{S}_n^{c,v}, \tilde{S}_n^c, \tilde{W}_n^c) \Rightarrow (\tilde{W}^c(0), \tilde{L}^u, \tilde{L}^v, \tilde{L}, \tilde{W}) \quad \textit{in} \quad \mathbb{R} \times (\mathbb{R}^\infty)^3, \tag{4.25}$$

where $\tilde{W} \equiv \tilde{\phi}(\tilde{W}(0) + \tilde{L})$ for $\tilde{\phi}$ defined in (3.5)*, $\tilde{L} \equiv \tilde{L}^v - \tilde{L}^u$, $\tilde{L}^u(k) \equiv L^u(k)$ and $\tilde{L}^v(k) \equiv L^v(k)$, $k \geq 1$, with (L^u, L^v) from* (3.8)*.*

Proof. By (4.20) and (4.23), $(\tilde{W}_n^c(0), \tilde{S}_n^{c,u}(k), \tilde{S}_n^{c,v}(k)) = (\sqrt{n}W_{n,0}, \mathbf{S}_n^u(k), \mathbf{S}_n^v(k))$ for $k \geq 1$, where $(\mathbf{S}_n^u, \mathbf{S}_n^v)$ is the random element of D^2 defined in (3.6). By (4.24), $(\sqrt{n}W_{n,0}, \mathbf{S}_n^u, \mathbf{S}_n^v) \Rightarrow (\tilde{W}(0), L^u, L^v)$. Applying the continuous mapping theorem with the projection map for the last two components, $\pi_{1,2,\ldots,k} : \mathbb{R} \times D^2 \to \mathbb{R} \times (\mathbb{R}^2)^k \equiv \mathbb{R}^{2k+1}$, defined by $\pi_{1,2,\ldots,k}(a, x) \equiv (a, x(1), x(2), \ldots x(k))$, we deduce convergence on the initial segments, which implies convergence of the first three components of (4.25) in $\mathbb{R} \times (\mathbb{R}^\infty)^2$. We apply the continuous mapping theorem again with addition and discrete reflection to treat the final two components.

If we impose all the additional conditions in Theorem 3.2, then we obtain a reflected Gaussian random walk as a limit; we treat the stationary distributions in [11].

Corollary 4.1. (HT limit for the scaled waiting times in standard $G_n/G_n/1$ models) *Consider a sequence of $G_n/G_n/1$ models associated with a base sequence of $G/G/1$ models satisfying*

$$(\sqrt{n}W_{n,0}, \hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v) \Rightarrow (\tilde{W}(0), \hat{L}^u, \hat{L}^v) \quad \textit{in} \quad \mathbb{R} \times D^2 \tag{4.26}$$

for $(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v)$ *in* (3.11)*, where* $P((\hat{L}^u, \hat{L}^v) \in C^2) = 1$*. If* $(1-\rho_n)\sqrt{n} \to \beta$*,* $0 < \beta < \infty$ *as* $n \to \infty$*, as in* (3.14)*. then the limit in* (4.25) *holds with* $L^u = \hat{L}^u + \beta e \quad$ *and* $\quad L^v = \hat{L}^v$*. If, in addition, condition* (3.16) *of Theorem* 3.2 *holds, then* $L = \sigma B - \beta e$*, where* B *is a standard BM and* e *is the identity map in* D*, so that* $\tilde{W}$ *becomes a reflected Gaussian random walk with negative drift, starting at the independent initial state* $\tilde{W}(0)$*, in particular,*

$$\tilde{W} \equiv \{\tilde{W}(k) : k \geq 0\} = \{\tilde{\phi}(\tilde{W}(0) + \sigma B - \beta e)(k) : k \geq 1\} \quad \textit{in} \quad \mathbb{R}^\infty. \tag{4.27}$$

4.2. Limits for unscaled waiting times in case (ii)

We now obtain a different limit in case (ii) of (4.22). Here it will be convenient to exploit the single-sequence framework of §3. In this case, we express the HT limit in terms of random elements of D. For that purpose, let

$$\begin{aligned}
\mathbf{S}_n^{c,u}(t) &\equiv S_{n,\lfloor nt \rfloor}^{c,u} - \lfloor nt \rfloor = \frac{S_{n\lfloor nt \rfloor}^u}{\rho_n n} - \lfloor nt \rfloor, \\
\mathbf{S}_n^{c,v}(t) &\equiv S_{n,\lfloor nt \rfloor}^{c,v} - \lfloor nt \rfloor = \frac{S_{n,n\lfloor nt \rfloor}^v - n\lfloor nt \rfloor}{n} = \frac{S_{n\lfloor nt \rfloor}^v - n\lfloor nt \rfloor}{n}, \\
\mathbf{S}_n^{c}(t) &\equiv S_{n,\lfloor nt \rfloor}^{c} = \frac{S_{n,n\lfloor nt \rfloor}^v}{n} - \frac{S_{n,n\lfloor nt \rfloor}^u}{\rho_n n} \\
&\quad = (\mathbf{S}_n^{c,v} - \mathbf{S}_n^{c,u})(t), \\
\mathbf{W}_n^{c}(t) &\equiv W_{n,\lfloor nt \rfloor}^{c} = \frac{W_{n,n\lfloor nt \rfloor}}{n} = \phi(\mathbf{W}_n^c(0) + \mathbf{S}_n^c)(t).
\end{aligned} \tag{4.28}$$

Theorem 4.2. (HT limit for the unscaled waiting times in the $G_n/G_n/1$ models) *Consider a sequence of* $G_n/G_n/1$ *models associated with a single base* $G/G/1$ *model satisfying*

$$(W_{n,0}, \hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v) \Rightarrow (\mathbf{W}^c(0), \hat{L}^u, \hat{L}^v) \quad \textit{in} \quad \mathbb{R} \times D^2 \tag{4.29}$$

for $(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v)$ *in* (3.11)*, where* $P((\hat{L}^u, \hat{L}^v) \in C^2) = 1$ (*just as in* (3.13) *except for the initial conditions*)*. Instead of condition* (3.14) *of Theorem* 3.2*, assume that*

$$(1 - \rho_n)n \to \beta, \quad 0 < \beta < \infty, \quad \textit{as} \quad n \to \infty, \tag{4.30}$$

as in case (*ii*) *of* (4.22)*. Then, as* $n \to \infty$*,*

$$(\mathbf{S}_n^c(0), \mathbf{S}_n^{c,u}, \mathbf{S}_n^{c,v}, \mathbf{S}_n^c, \mathbf{W}_n^c) \Rightarrow (\mathbf{W}^c(0), \hat{L}^u + \beta e, \hat{L}^v, L, \mathbf{W}^c) \quad \textit{in} \quad \mathbb{R} \times D^4, \tag{4.31}$$

where $\mathbf{W}^c \equiv \phi(\mathbf{W}^c(0) + L)$*,* ϕ *is given in* (3.7)*,* $L \equiv \hat{L}^v - \hat{L}^u - \beta e$ *and* $(\hat{L}^u, \hat{L}^v)$ *comes from* (3.13)*. If, in addition condition* (3.16) *of Theorem* 3.2 *holds, then* $L \stackrel{\mathrm{d}}{=} \sigma B - \beta e$*, where* B *is a standard BM and* $\sigma^2 = \sigma_u^2 + \sigma_v^2$*.*

Proof. We are exploiting the single-sequence framework in §3.2, because the final expressions in the first two rows in (4.28) above involve only the single sequence. We will relate our processes $(\mathbf{S}_n^{c,u}, \mathbf{S}_n^{c,v})$ directly to these. From the formulas in (4.28), we see that $\mathbf{S}_n^{c,v}(t) = \hat{\mathbf{S}}_{n^2}^v(t)$ for $t = k/n$ for all nonnegative integers n and k, while $\mathbf{S}_n^{c,v}$ is constant in all intervals $[k/n, (k+1)/n)$. Now let $\|\cdot\|_t$ be the uniform norm for $\mathbb{R}^k$-valued functions on the interval $[0,t]$, using the maximum norm $|\cdot|$ in $\mathbb{R}^k$, and let the modulus of continuity be defined for any $x \in D$ by

$$w_x(\delta, t) \equiv \sup\{|x(t_1) - x(t_2)| : 0 \le t_1 < t_2 \le t, \quad |t_2 - t_1| < \delta\}. \tag{4.32}$$

Hence,

$$\|\mathbf{S}_n^{c,v} - \mathbf{S}_{n^2}^v\|_t \le w_{\mathbf{S}_{n^2}^v}(1/n, t) \quad \text{for each} \quad n \ge 1 \quad \text{and} \quad t > 0. \tag{4.33}$$

By (3.13) and the fact that $P(\hat{L}^v \in C^2) = 1$, we have $w_{\hat{\mathbf{S}}_{n^2}^v}(1/n, t) \Rightarrow 0$ as $n \to \infty$, which with (4.33) implies that

$$\|\mathbf{S}_n^{c,v} - \hat{\mathbf{S}}_{n^2}^v\|_t \to 0 \quad \text{as} \quad n \to \infty. \tag{4.34}$$

A minor modification of the same reasoning applies to $\mathbf{S}_n^{c,u}(t)$. We need to be more careful because of the extra scaling by ρ_n. For that purpose, we introduce the fluid scaled process

$$\bar{\mathbf{S}}_n^u(t) \equiv \frac{S^u_{\lfloor nt \rfloor}}{n}, \quad t \ge 0. \tag{4.35}$$

Given the assumed limit in (3.13), we have the extension $(\hat{\mathbf{S}}_n^u, \bar{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v) \Rightarrow (\hat{L}^u, e, \hat{L}^v)$ in D^3. Now, using the reasoning in (3.17) and (3.18), we observe that

$$\mathbf{S}_n^{c,u}(t) = Z_{n^2}(t) \equiv \hat{\mathbf{S}}_{n^2}^u(t) + (\beta + o(1))\bar{\mathbf{S}}_{n^2}^u(t) \tag{4.36}$$

for $t = k/n$ for all integers $k \ge 0$ and $n \ge 1$, while $\mathbf{S}_n^{c,u}$ remains constant in each interval $[k/n, (k+1)/n)$. Hence, reasoning as for $\mathbf{S}_n^{c,v}$ above, we have

$$\|\mathbf{S}_n^{c,u} - Z_{n^2}\|_t \le w_{Z_{n^2}}(1/n, t) \quad \text{for each} \quad n \ge 1 \quad \text{and} \quad t > 0. \tag{4.37}$$

Since $Z_n \Rightarrow \hat{L}^u + \beta e$, where $P(\hat{L}^u + \beta e \in C) = 1$, we have $w_{Z_{n^2}}(1/n, t) \Rightarrow 0$ as $n \to \infty$ for all $t > 0$. Combining the results above, we have

$$\|(W_{n,0}, \mathbf{S}_n^{c,u}, \mathbf{S}_n^{c,v}) - (W_{n,0}, Z_{n^2}, \mathbf{S}_{n^2}^v)\| \Rightarrow 0. \tag{4.38}$$

That in turn, with the "convergence-together theorem," Theorem 11.4.7 of [14], implies convergence for the first three terms in (4.31). Finally, we apply the continuous mapping theorem with subtraction and reflection to get the full limit in (4.31).

Remark 4.1. (*practical significance of the scaling*) We want to emphasize the practical significance of the scaling of space and time in Theorem 4.2. To do so, it is helpful to focus on a single system with traffic intensity ρ, which we can relate to n by replacing the assumed growth condition by an equality. For the original $G/G/1$ model, with $(1-\rho)\sqrt{n} \to \beta$, we have $n^{-1/2}W_{n,\lfloor nt \rfloor} \Rightarrow \phi(L)(t)$. Hence, letting $(1-\rho)\sqrt{n} = \beta = 1$, we see that, for higher values of ρ, W_k^ρ should be of order $O(1/(1-\rho))$, while significant changes occur over time intervals of length $O(1/(1-\rho^2))$.

In contrast, with cyclic thinning, we have the alternative growth condition $(1-\rho)n \to \beta$ in (4.30), under which $W^c_{n,\lfloor nt \rfloor} \Rightarrow \phi(L)(t)$. Now we can let $(1-\rho_n)n = \beta = 1$ and obtain $n = 1/(1-\rho)$. With cyclic thinning, we have $W_k^{c,\rho}$ being of order $O(1)$, while significant changes occur over a time scale of $O(1/(1-\rho))$.

The time scaling by n with cyclic thinning can be better understood by considering the approximating $D/D/1$ model with the given traffic intensity. For example, suppose that the service times are all 1 and the interarrival times are all $n/(n-1)$ for some large positive integer n, corresponding to $(1-\rho_n)n = 1$. Suppose that the system is initialized by an arrival at time 0 who finds k customers in the queue and one more in service at the beginning of a service time. Because of the deterministic service times, that initial customer at time 0 has a waiting time of exactly $k+1$. In general, if there are k customers in queue upon arrival the waiting time is bounded below by k and bounded above by $k+1$. Thus, in the $D/D/1$ model, the waiting time is tightly linked to the queue length. For the specified initial conditions, it takes time n for the queue length to decrease by 1; the queue will first become empty, leaving one customer in service just beginning his service time, at time kn. Thus the waiting time of a new arrival at time jn will be about $k+1-j$. Thus we see that in the $D/D/1$ model with alternative initial conditions the waiting times change over time periods of order n. Theorem 4.2 is showing that remains true in the HT limit for the $G_n/G_n/1$ model. Finally, the extra variability in the $G_n/G_n/1$ model produces the stochastic limit without spatial scaling in Theorem 4.2. ∎

To highlight Remark 4.1, we state a corollary for first passage times, obtained by applying the continuous mapping theorem with the limit in (4.31). Let a and b be real numbers with $0 < a < b < \infty$. Consider $x \in D$ with $x(0) = c$ for c to be specified. Then define the first passage functions

$$T^+_{a,b}(x) \equiv \inf\{t \geq 0 : x(t) > b | x(0) = a\} \quad \text{and} \quad T^-_{b,a}(x) \equiv \inf\{t \geq 0 : x(t) < a | x(0) = b\}. \tag{4.39}$$

We will be interested in

$$\begin{aligned} T^+_{a,b}(\mathbf{W}^c_n) &= \frac{\min\{k \geq 1 : W^c_{n,k} > b | W^c_{n,0} = a\}}{n} \quad \text{and} \\ T^-_{b,a}(\mathbf{W}^c_n) &= \frac{\min\{k \geq 1 : W^c_{n,k} < a | W^c_{n,0} = b\}}{n}. \end{aligned} \tag{4.40}$$

For RBM and the $M/M/1$ queue, the distributions of the first passage times $T_{0,x}$ and $T_{x,0}$ are described in §5.7.5 of [14]. There the limits are described in terms of canonical BM and RBM with negative drift, having drift rate -1 and diffusion coefficient 1, which is convenient, because there are no parameters. It is easy to transform BM and RBM with general parameters to and from the canonical versions; see p. 174 of [14].

Corollary 4.2. (HT limit for first passage times in the $G_n/G_n/1$ models) *Under the assumptions of Theorem* 4.2*, including condition* (3.16) *of Theorem* 3.2*,*

$$\begin{aligned} \frac{\min\{k \geq 1 : W^c_{n,k} > b | W^c_{n,0} = a\}}{n} &\Rightarrow T^+_{a,b}(\phi(\sigma B - \beta e)) \stackrel{\mathrm{d}}{=} (\sigma^2/\beta^2) T^+_{\beta a/\sigma^2, \beta b/\sigma^2}(\phi(B - e)), \\ \frac{\min\{k \geq 1 : W^c_{n,k} < a | W^c_{n,0} = b\}}{n} &\Rightarrow T^-_{b,a}(\phi(\sigma B - \beta e)) \stackrel{\mathrm{d}}{=} T^-_{b,a}(\sigma B - \beta e) \\ &\qquad \stackrel{\mathrm{d}}{=} (\sigma^2/\beta^2) T^-_{\beta b/\sigma^2, \beta a/\sigma^2}(B - e). \end{aligned} \tag{4.41}$$

For example, since $E[T^-_{b,a}(B - e)] = Var(T^-_{b,a}(B - e) = b - a$, we have

$$E[(\sigma^2/\beta^2) T^-_{\beta b/\sigma^2, \beta a/\sigma^2}(B - e)] = \frac{(b-a)}{\beta}, \quad Var(\sigma^2/\beta^2) T^-_{\beta b/\sigma^2, \beta a/\sigma^2}(B - e) = \frac{\sigma^2(b-a)}{\beta^3}. \tag{4.42}$$

Remark 4.2. (*cyclic thinning applied to the arrivals alone*) The story is quite different if we apply cyclic thinning to either the arrival process alone or the service process alone, i.e., if we consider the limit for the $G_n/G/1$ model or the $G/G_n/1$ model, where the G component is not deterministic D. Then the entire queueing model is no longer nearly deterministic for large n (assuming that the G component is not really itself D).

The component without cyclic thinning forces us to scale in the conventional manner in §3 in order to get a proper limit. That additional spatial scaling (there is none in (4.28)) then knocks out the term with the cyclic scaling. The resulting HT limits are just as for the $D/G/1$ and $G/D/1$ models, already covered as a special case of §3. For example, with the $GI/GI/1$ base model, in the corresponding $GI_n/GI/1$ model, the limits L^u and L are changed from $\sigma_u B_u + e$ and $\sigma B - e$ to e and $\sigma_v B_v - e$, respectively; for the $GI/GI_n/1$ model, the the limits L^v and L are changed from $\sigma_v B_v$ and $\sigma B - e$ to $0e$ and $\sigma_u B_u - e$, respectively.

4.3. Beyond cyclic thinning

It is significant that the HT limits for waiting times we have established for $G_n/G_n/1$ models are not limited to the $G_n/G_n/1$ models with cyclic thinning of order n as $n \to \infty$ that we have considered. It suffices to have a sequence of general $G/G/1$ models, where the sequences of interarrival times and service times have the asymptotic properties derived for the interarrival and service times in the $G_n/G_n/1$ models; i.e., the results hold for more general nearly deterministic queues. That explains the title of the paper.

In particular, to have the limit $(\tilde{W}_n^c(0), \tilde{S}_n^{c,u}, \tilde{S}_n^{c,v}, \tilde{S}_n^c, \tilde{W}_n^c) \Rightarrow (\tilde{W}(0), \tilde{L}^u, \tilde{L}^u, \tilde{L}, \tilde{W})$ in $\mathbb{R} \times (\mathbb{R}^\infty)^3$ in (4.25) of Theorem 4.1, it suffices to directly have the convergence of the first three components. To achieve generality, we assume that we have a sequence of general $G/G/1$ models in which the partial sums of the interarrival times and service times in model n are denoted by $S_{n,k}^{c,u}$ and $S_{n,k}^{c,v}$ as before, but where these are not derived from cyclic thinning (the superscript c has no meaning). We then introduce the scaling in (4.23) and we directly *assume* that $(\tilde{S}_n^{c,u}, \tilde{S}_n^{c,v}) \Rightarrow (\tilde{L}^u, \tilde{L}^v)$ in $(\mathbb{R}^\infty)^2$ as $n \to \infty$.

In the same spirit, to have the limit $(\mathbf{W}_n^c(0), \mathbf{S}_n^{c,u}, \mathbf{S}_n^{c,v}, \mathbf{S}_n^c, \mathbf{W}_n^c) \Rightarrow (\mathbf{W}^c(0), \hat{L}^u + \beta e, \hat{L}^v, L, \phi(L))$ in $\mathbb{R} \times D^4$ as $n \to \infty$ in Theorem 4.2, it suffices to directly have the convergence of the first three components. As before, to achieve generality, we assume that we have a sequence of general $G/G/1$ models in which the partial sums of the interarrival times and service times in model n are denoted by $S_{n,k}^{c,u}$ and $S_{n,k}^{c,v}$ as before, but where these are not derived from cyclic thinning (again the superscript c has no meaning). We then introduce the scaling in (4.28) and we directly *assume* that $(\mathbf{S}_n^{c,u}, \mathbf{S}_n^{c,v}) \Rightarrow (\hat{L}^u + \beta e, \hat{L}^v)$ in D^2 as $n \to \infty$.

5. Associated counting processes

Theorem 4.2 shows that the individual partial sums $S^{c,u}_{n,k}$ and $S^{c,v}_{n,k}$ in (4.20), associated with the arrival process and service processes obtained from cyclic thinning, satisfy FCLT's with appropriate scaling, because the sequence of random elements $\{(\mathbf{S}^{c,u}_n, \mathbf{S}^{c,v}_n) : n \geq 1\}$ in (4.28) is asymptotically equivalent to the associated sequence of random elements $\{(\hat{\mathbf{S}}^u_{n^2} + \beta\bar{\mathbf{S}}^u_{n^2}, \hat{\mathbf{S}}^v_{n^2}) : n \geq 1\}$ in (3.11) and (4.35) with the scaling in (3.12) as $n \to \infty$, i.e., because of (4.36) and (4.38).

We would thus naturally expect that FCLT's would hold for the associated counting processes, by virtue of the continuous mapping theorem applied with the inverse function, as in §§13.6-13.8 of [14]. However, that is *not* so. Upon closer examination, we find that the asymptotically negligible differences between $(\mathbf{S}^{c,u}_n, \mathbf{S}^{c,v}_n)$ and $(\hat{\mathbf{S}}^u_{n^2} + \beta\bar{\mathbf{S}}^u_{n^2}, \hat{\mathbf{S}}^v_{n^2})$ significantly affect the associated counting processes. Nevertheless, we do establish FCLT's for the counting processes with different scaling.

To treat the associated counting processes, let

$$\begin{aligned} N^u_n(t) &\equiv \max\{k \geq 0 : S^u_{n,k} \leq t\}, \quad N^v_n(t) \equiv \max\{k \geq 0 : S^v_{n,k} \leq t\}, \\ N^{c,u}_n(t) &\equiv \max\{k \geq 0 : S^{c,u}_{n,k} \leq t\}, \quad N^{c,v}_n(t) \equiv \max\{k \geq 0 : S^{c,v}_{n,k} \leq t\} \quad (5.43) \end{aligned}$$

for $t \geq 0$. For simplicity, now assume in addition that $P(U_{n,k} > 0) = 1$ and $P(V_{n,k} > 0) = 1$ for all n and k, so that all these counting processes increase by unit jumps. By our initial conditions in §3.1, we have $N^u_n(0) = 1$ and $N^v_n(0) = 0$.

5.1. No time scaling

Before observing the technical problems with the counting processes with the scaling in Theorem 4.2, we observe that the time scaling there by n plays a critical role in obtaining interesting nondegenerate stochastic limits. In particular, we now show that, if we do not scale time by n in the $G_n/G_n/1$ models as $n \to \infty$ under the conditions of Theorem 4.2, we simply get convergence of all these queueing processes to the associated deterministic processes in the trivial $D/D/1$ model with traffic intensity 1 (and the specified initial conditions, having a 0^{th} customer arrive at time 0 to find an empty system). (The results in this subsection hold for both $(1-\rho_n)n \to \beta$ as in (4.30) or $(1-\rho_n)\sqrt{n} \to \beta$ as in (3.14).)

We start by considering the counting processes. For the counting processes, when we combine (5.43) with the basic definition in (4.20) and the initial conditions in §3.1, we obtain the important relations

$$N_n^u(nt) \;=\; 1 + n(N_n^{c,u}(t) - 1) + J_n^{c,u}(t) \quad \text{and} \quad N_n^v(nt) = nN_n^{c,v}(t) + J_n^{c,v}(t), \tag{5.44}$$

where $J_n^{c,u}(t)$ counts the number of interarrival time phases completed in the interarrival time in progress at time t, while $J_n^{c,v}(t)$ counts the number of service time phases completed in the service time in progress at time t. By our assumed initial conditions, $J_n^{c,u}(0) = J_n^{c,v}(0) = 0$ for all $n \geq 1$. Clearly, $0 \leq J_n^{c,u}(t) < n$ and $0 \leq J_n^{c,v}(t) < n$ for all $t \geq 0$ and $n \geq 1$. We can then rewrite the relations in (5.44) as

$$\begin{aligned} N_n^{c,u}(t) &= 1 + \frac{N_n^u(nt) - 1}{n} - \frac{J_n^{c,u}(t)}{n} = 1 + \lfloor (N_n^u(nt) - 1)/n \rfloor \quad \text{and} \\ N_n^{c,v}(t) &= \frac{N_n^v(nt)}{n} - \frac{J_n^{c,v}(t)}{n} = \lfloor N_n^v(nt)/n \rfloor, \end{aligned} \tag{5.45}$$

where $\lfloor t \rfloor$ is again the floor function, which is right continuous and thus an element of D.

Given the spatial scaling on the right in (5.45), we can obtain a FWLLN for $(N_n^{c,u}, N_n^{c,v})$, but only in the product space D^2. We cannot obtain convergence in $D([0,\infty), \mathbb{R}^2)$ with the usual J_1 topology, because the limit functions have common discontinuity points, and we require inconsistent time transformations in the two components. Convergence of the components separately is easy, because all processes are integer valued. Hence convergence in D is equivalent to convergence of the finite dimensional distributions. To state the results, define the following random elements in D:

$$\bar{\mathbf{N}}_n^u(t) \;\equiv\; \frac{N_n^u(nt)}{n}, \quad \bar{\mathbf{N}}_n^v(t) \equiv \frac{N_n^v(nt)}{n}, \quad \bar{\mathbf{J}}_n^{c,u} \equiv \frac{J_n^{c,u}(t)}{n}, \quad \bar{\mathbf{J}}_n^{c,v} \equiv \frac{J_n^{c,v}(t)}{n} \tag{5.46}$$

In (5.46) and in Theorem 5.1 below, we have no time scaling by n for the processes associated with the $G_n/G_n/1$ model, whereas we do for the associated $G/G/1$ base model.

Theorem 5.1. (FWLLN for the counting processes with cyclic thinning) *Consider a sequence of $G_n/G_n/1$ models associated with a single base $G/G/1$ model satisfying*

$$(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v) \Rightarrow (\hat{L}^u, \hat{L}^v) \quad \textit{in} \quad D^2 \tag{5.47}$$

for $(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v)$ *in* (3.11)*, where* $P((\hat{L}^u, \hat{L}^v) \in C^2) = 1$ *If either* (4.30) *or* (3.14) *holds, then*

$$(\bar{\mathbf{N}}_n^u, \bar{\mathbf{J}}_n^{c,u}, N_n^{c,u}, \bar{\mathbf{N}}_n^v, \bar{\mathbf{J}}_n^{c,v}, N_n^{c,v}) \Rightarrow (e, J, 1 + \lfloor e \rfloor, e, J, \lfloor e \rfloor) \quad \textit{in} \quad D^6 \quad \textit{as} \quad n \to \infty, \tag{5.48}$$

where e *is the identity map in* D*,* $\lfloor e \rfloor(t) \equiv \lfloor t \rfloor$ *and* $J = e - \lfloor e \rfloor$*.*

Proof. Assume (4.30); the reasoning is similar with (3.14). From the basic FCLT equivalence for partial sums and counting processes expressed in Theorem 7.3.2 of [14], we get

$$(\hat{\mathbf{N}}_n^u, \hat{\mathbf{N}}_n^v) \Rightarrow (-\hat{L}^u, -\hat{L}^v) \quad \text{in} \quad D^2 \quad \text{as} \quad n \to \infty, \tag{5.49}$$

where

$$\hat{\mathbf{N}}_n^u(t) \equiv \frac{N^u(nt) - nt}{\sqrt{n}} \quad \text{and} \quad \hat{\mathbf{N}}_n^v(t) \equiv \frac{N^v(nt) - nt}{\sqrt{n}}, \quad t \geq 0. \tag{5.50}$$

As a consequence, we get $(\bar{\mathbf{N}}_n^u, \bar{\mathbf{N}}_n^v) \Rightarrow (e, e)$ in D^2 as $n \to \infty$. Moreover, by (5.45), we can write

$$N_n^{c,v} = \lfloor \bar{\mathbf{N}}_n^v \rfloor \quad \text{for all} \quad n \quad \text{and} \quad d_t(N_n^{c,u}, 1 + \lfloor \bar{\mathbf{N}}_n^u \rfloor) \Rightarrow 0 \tag{5.51}$$

as $n \to \infty$ for all non-integer $t > 0$, where d_t is a metric inducing the J_1 topology on $D([0,t])$. Hence, we also have $(N_n^{c,u}, N_n^{c,v}) \Rightarrow (1 + \lfloor e \rfloor, \lfloor e \rfloor)$ in D^2. Finally, we also have $(\bar{\mathbf{J}}_n^{c,u}, \bar{\mathbf{J}}_n^{c,v}) \Rightarrow (J, J)$ in D^2 as $n \to \infty$ from the above and (5.45). That completes the proof.

5.2. Time-scaled counting processes

However, the story is different if we scale time by n, as we have already seen in Theorem 4.2. In particular, the queue-length and workload processes in the $G_n/G_n/1$ model for large n will fluctuate randomly over time intervals of length $O(n)$. We first consider the FCLT refinement of the FWLLN in Theorem 5.1. We get the following result, which is partly positive and partly negative. To state the result, in addition to the processes in (5.50), we introduce the following random elements for each $n \geq 1$:

$$\begin{aligned} \mathbf{N}_n^{c,u} &\equiv N_n^{c,u}(nt) - nt, \quad \mathbf{N}_n^{c,v} \equiv N_n^{c,v}(nt) - nt, \\ \mathbf{J}_n^{c,u} &\equiv \frac{J_n^{c,u}(nt)}{n} \quad \text{and} \quad \mathbf{J}_n^{c,v} \equiv \frac{J_n^{c,v}(nt)}{n}, \quad t \geq 0. \end{aligned} \tag{5.52}$$

Paralleling (3.6) and (4.28), in (5.52) the processes in $(\mathbf{N}_n^u, \mathbf{N}_n^u, \mathbf{J}_n^{c,u}, \mathbf{J}_n^{c,v})$ have spatial scaling, but the processes in $(\mathbf{N}_n^{c,u}, \mathbf{N}_n^{c,u})$ do not. To state the result, let $\leq_{st}$ denote

ordinary stochastic order for real-valued random variables; i.e., we write $X_1 \leq_{st} X_2$ for real-valued random variables if $P(X_1 > t) \leq P(X_2 > t)$ for all t.

Theorem 5.2. (FCLT for the counting processes with cyclic thinning) *Under the assumptions of Theorem* 5.1*,*

$$(\mathbf{N}_n^{c,u} + \mathbf{J}_n^{c,u}, \mathbf{N}_n^{c,v} + \mathbf{J}_n^{c,u}) \Rightarrow (1 - \hat{L}^u - \beta e, -\hat{L}^v) \quad \textit{in} \quad D^2 \quad \textit{as} \quad n \to \infty, \tag{5.53}$$

where $(\hat{L}^u, \hat{L}^v)$ *comes from* (3.13)*. However, the sequences* $\{\mathbf{J}_n^{c,u} : n \geq 1\}$*,* $\{\mathbf{J}_n^{c,v} : n \geq 1\}$*,* $\{\mathbf{N}_n^{c,u} : n \geq 1\}$ *and* $\{\mathbf{N}_n^{c,v}) : n \geq 1\}$ *in* D *are not tight and thus do not converge in* D*. Nevertheless, for each* $t \geq 0$*, the associated sequences of real-valued random variables* $\{\mathbf{J}_n^{c,u}(t) : n \geq 1\}$*,* $\{\mathbf{J}_n^{c,v}(t) : n \geq 1\}$*,* $\{\mathbf{N}_n^{c,u}(t) : n \geq 1\}$ *and* $\{\mathbf{N}_n^{c,v})(t) : n \geq 1\}$ *are tight. Moreover, for any convergent subsequence, with limits denoted by* $\mathbf{J}^{c,u}(t)$*,* $\mathbf{J}^{c,v}(t)$*,* $\mathbf{N}^{c,u}(t)$*, and* $\mathbf{N}^{c,v}(t)$*, we have the following bounds*

$$\begin{aligned} P(0 \leq \mathbf{J}^{c,u}(t) \leq 1) \quad &= \quad P(0 \leq \mathbf{J}^{c,v}(t) \leq 1) = 1 \\ -\hat{L}^u(t) - \beta t \quad &\leq_{st} \quad \mathbf{N}^{c,u}(t) \leq_{st} 1 - \hat{L}^u(t) - \beta t \\ -\hat{L}^v(t) - 1 \quad &\leq_{st} \quad \mathbf{N}^{c,v}(t) \leq_{st} -\hat{L}^v(t). \end{aligned} \tag{5.54}$$

Proof. By (5.45) and (5.52), we have

$$\mathbf{N}_n^{c,v} + \mathbf{J}_n^{c,v} = \mathbf{N}_{n^2}^v \quad \text{and} \quad \|(\mathbf{N}_n^{c,u} - 1 + \mathbf{J}_n^{c,u}) - \mathbf{N}_{n^2}^u\|_t = \frac{1}{n} \tag{5.55}$$

for all n and t, so that we can apply (5.49), which follows from the conditions, and the convergence-together theorem, Theorem 11.4.7 of [14], to obtain the positive result in (5.53). We use index n^2 instead of n.

We obtain the negative result by applying (5.55). We first observe that we have the established convergence of $(\mathbf{N}_{n^2}^u, \mathbf{N}_{n^2}^v)$ to a limit with continuous sample paths, as indicated in (5.49), so that

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} w_{\mathbf{N}_{n^2}^u}(\delta, t) = 0 \quad \text{and} \quad \lim_{\delta \downarrow 0} \limsup_{n \to \infty} w_{\mathbf{N}_{n^2}^v}(\delta, t) = 0 \quad \text{for all} \quad t > 0 \tag{5.56}$$

from the associated tightness. On the other hand, we have extra time scaling by n to go from $(\bar{\mathbf{J}}_n^{c,u}, \bar{\mathbf{J}}_n^{c,v})$ in (5.46) to $(\mathbf{J}_n^{c,u}, \mathbf{J}_n^{c,v})$ in (5.52). Since $(\bar{\mathbf{J}}_n^{c,u}, \bar{\mathbf{J}}_n^{c,v})$ converges to a nondegenerate limit with oscillations of size 1 over intervals of length 1 by Theorem 5.1, the oscillations will necessarily occur too rapidly as $n \to \infty$ when we add the extra time

scaling by n. Hence we cannot have tightness in D for either $\mathbf{J}_n^{c,u}$ or $\mathbf{J}_n^{c,v}$, and so also for $\mathbf{N}_{n^2}^u - \mathbf{J}_n^{c,u}$ or $\mathbf{N}_{n^2}^v - \mathbf{J}_n^{c,v}$, which is necessary for convergence. However, from (5.55), we have stochastic bounds for each n, which does imply tightness of the associated sequence of real-valued random variables. Moreover, these stochastic bounds must be preserved under passing to the limit along any subsequence.

Under additional regularity conditions, we will have $\mathbf{J}_n^{c,u}(t)$ and $\mathbf{J}_n^{c,v}(t)$ converge to random variables with the uniform distribution on the interval $[0,1]$. Then we will be able to replace the bounds in (5.54) by convergence in distribution in $\mathbb{R}$. We obtain such stronger results in a stationary framework in [11].

We conclude this subsection by indicating a framework in which we can get a limit for $(\mathbf{J}_n^{c,u}, \mathbf{J}_n^{c,v})$, but it is an unconventional one, even going beyond the spaces E and F in Chapter 15 of [14]. We will make the underlying space a subset of the collection of compact subsets of $\mathbb{R}^2$. We put the sample paths of the stochastic processes in D in this space by (i) looking at the graphs of the sample paths in $\mathbb{R}^2$, including the left limits, and (ii) by considering the restrictions of the domain $[0,\infty)$ to $[0,t]$ for various $t > 0$. As in the conventional spaces C and D, convergence will be characterized in terms of restrictions to the bounded interval $[0,t]$ for a sequence of time points t converging to infinity. For background on this graph approach, see Chapters 12 and 15 of [14] and references therein.

Instead of $D([0,t],\mathbb{R})$, we use the space $\mathbf{C}_t$ of compact subsets of $[0,t] \times \mathbb{R}$ in $\mathbb{R}^2$ using the Hausdorff metric, denoted by $d_{\mathrm{H},t}(A,B)$ and defined by

$$d_{\mathrm{H},t}(A,B) \equiv \max\{ \sup_{x\in A} \inf_{y\in B} d(x,y), \sup_{y\in B} \inf_{x\in A} d(x,y) \}, \tag{5.57}$$

where $A, B \in \mathbf{C}_t$ and d is a metric on $\mathbb{R}^2$, for which it is convenient to take the maximum metric: $d(x,y) \equiv d((x_1,x_2),(y_1,y_2)) \equiv \max\{|x_1 - y_1, |x_2 - y_2|\}$; see p. 381 of [14] for background. For each t, $(\mathbf{C}_t, d_{\mathrm{H},t})$ is a compact metric space. Let $\mathbf{C}$ be the subset of $\mathbb{R}^2$ for which all restrictions to $[0,t] \times \mathbb{R}$ are compact, and let $\mathbf{C}^k$ be the k-fold product space with the product topology. Define convergence in $\mathbf{C}$ to mean convergence of the restrictions in $\mathbf{C}_t$ for all $t > 0$. In our context, the limit of $\mathbf{J}_n^{c,u}$ is the deterministic set $\Upsilon \equiv [0,\infty) \times [0,1]$ in $\mathbf{C}$, which we call the *unit blur*. With this framework, we can obtain the following result.

Theorem 5.3. (convergence to the unit blur) *Under the assumptions of Theorem* 5.1*,*

$$(\mathbf{J}_n^{c,u}, \mathbf{J}_n^{c,v}) \Rightarrow (\Upsilon^u, \Upsilon^v) \quad in \quad \boldsymbol{C}^2 \quad as \quad n \to \infty, \tag{5.58}$$

where Υ^u *and* Υ^v *are unit blurs, so that*

$$(\mathbf{N}_n^{c,u}, \mathbf{J}_n^{c,u}, \mathbf{N}_n^{c,v}, \mathbf{J}_n^{c,u}) \Rightarrow (-\hat{L}^u - \beta e + \Upsilon^u, \Upsilon^u, -\hat{L}^v - \Upsilon^v, \Upsilon^v) \quad in \quad \boldsymbol{C}^4 \quad as \quad n \to \infty. \tag{5.59}$$

Proof. The statement in (5.58) is equivalent to convergence of the components separately, since the limit is deterministic, by Theorem 11.4.5 of [14]. First, for (5.58), we can exploit the FWLLN, Theorem 5.1. By (5.48),

$$(\bar{\mathbf{J}}_n^{c,u}, \bar{\mathbf{J}}_n^{c,v}) \Rightarrow (J, J) \quad \text{in} \quad D^2 \quad \text{as} \quad n \to \infty, \tag{5.60}$$

where $J(t) \equiv t - \lfloor t \rfloor$, $t \geq 0$. For $x \in D$ and any constant $b > 0$, let $(x \circ be)(t) \equiv x(bt)$, $t \geq 0$. From (5.60), we get the associated limit

$$(\bar{\mathbf{J}}_n^{c,u} \circ be, \bar{\mathbf{J}}_n^{c,v} \circ be) \Rightarrow (J \circ be, J \circ be) \quad \text{in} \quad D^2 \quad \text{as} \quad n \to \infty \tag{5.61}$$

for any $b > 0$. We now carry out the remaining analysis only for the arrival process; the service process is treated in the same way.

First, we observe that, for any $\epsilon > 0$, we can choose b_ϵ such that

$$d_{\mathrm{H},t}(J \circ be, \Upsilon^u) < \epsilon \quad \text{for all} \quad b > b_\epsilon \quad w.p.1. \tag{5.62}$$

Next observe that

$$d_{\mathrm{H},t}((\mathbf{J}_n^{c,u}, \Upsilon^u) \leq \sup_{b>n} d_{\mathrm{H},t}(\bar{\mathbf{J}}_n^{c,u} \circ be, \Upsilon^u), \tag{5.63}$$

where

$$d_{\mathrm{H},t}(\bar{\mathbf{J}}_n^{c,u} \circ be, \Upsilon^u) \leq d_{\mathrm{H},t}(\bar{\mathbf{J}}_n^{c,u} \circ be, J \circ be) + d_{\mathrm{H},t}(J \circ be, \Upsilon^u). \tag{5.64}$$

For any given $\epsilon > 0$, by (5.62), we can make the second term on the right in (5.64) less than $\epsilon/2$ by choosing b large enough. For that b chosen, by (5.61), we can make the first term on the right in (5.64) less than $\epsilon/2$ by choosing n large enough.

Finally, to establish (5.59), we use the fact that the convergence of $(\mathbf{N}_n^{c,u} + \mathbf{J}_n^{c,u}, \mathbf{N}_n^{c,v} + \mathbf{J}_n^{c,u})$ established in D^2 implies the corresponding weaker convergence in $\mathbf{C}^2$ to the limit, which has single-valued projections at t for all t. Since one of the limits has a single-valued projection, subtraction is continuous. The first component on the right in (5.59) is initially $1 - \hat{L}^u - \beta e - \Upsilon$, but $1 - \Upsilon = \Upsilon$.

Remark 5.1. (*interpreting the blur limits*) We note that we do not obtain the information about the joint distribution of Υ^u and Υ^v. In particular, we cannot conclude that $(\Upsilon^u, \Upsilon^v) = [0,\infty) \times [0,1] \times [0,1]$, because we do not know when each function assumes the values in $[0,1]$. In the framework above, $\Upsilon^u(t) = \Upsilon^v(t) = [0,1]$ for all $t \geq 0$, where $\Upsilon(t)$ is the projection of Υ on t. However, if we make additional stationarity and independence assumptions, we will have

$$(\mathbf{J}_n^{c,u}(t), \mathbf{J}_n^{c,v})(t)) \stackrel{\mathrm{d}}{=} (Y_n^u, Y_n^v) \Rightarrow (Y^u, Y^v) \quad \text{in} \quad \mathbb{R}^2, \tag{5.65}$$

where Y_n^u and Y_n^v are independent random variables uniformly distributed on the finite set$\{0, 1/n, \ldots, (n-1)/n\}$ while Y^u and Y^v are independent random variables uniformly distributed on $[0,1]$. Then we would have the limit at t for $\mathbf{J}_n^{c,u}(t)$ uniformly distributed over $[0,1]$, but our framework does not provide that extra level of detail. On the other hand, our framework has the virtue that it shows that, asymptotically as $n \to \infty$, for any $\epsilon > 0$, $\mathbf{J}_n^{c,u}(s)$ will be near *every* point in the interval $[0,1]$ for some $s \in (t-\epsilon, t+\epsilon)$ for all n suitably large. In other words, the unit blur captures the increasing rate of fluctuations. In the present context, that is more important than knowing the distribution for any one fixed t, because the value at t is not representative of the values for any time points near t, like white noise. Here the limit should be distributed something like uncountably many i.i.d. uniform random variables, which of course is not directly well defined. We should hope to obtain stronger results about integrals over finite intervals. Under additional regularity conditions, we should obtain the limit $\int_a^b \mathbf{J}_n^{c,u}(t)\, dt \Rightarrow (b-a)/2$ as $n \to \infty$.

5.3. FCLT's with stronger scaling

In this subsection we show that we can overcome the difficulties in Theorem 5.2 and obtain FCLT's for the counting processes, provided that we introduce a stronger scaling in the setting of §3.2. We will be brief and consider only the rate-1 processes with cyclic thinning, $S_k^{c,v}$ and $N^{c,v}(t)$. We introduce two new random elements in D^2, defined by

$$\begin{aligned} \tilde{\mathbf{S}}_n^{c,v}(t) &\equiv \frac{S^{c,v}_{n,\lfloor n^2 t \rfloor} - \lfloor n^2 t \rfloor}{\sqrt{n}} = \frac{S^v_{n\lfloor n^2 t \rfloor} - n\lfloor n^2 t \rfloor}{n^{3/2}} \approx \hat{\mathbf{S}}^v_{n^3}(t), \\ \tilde{\mathbf{N}}_n^{c,v}(t) &\equiv \frac{N_n^{c,v}(n^2 t) - n^2 t)}{\sqrt{n}} = \frac{N^v(n^3 t) - n^3 t - J_n^v(n^3 t)}{n^{3/2}} \approx \hat{\mathbf{N}}^v_{n^{3/2}}(t), \end{aligned} \tag{5.66}$$

for $t \ge 0$, using (5.45), where $\hat{\mathbf{S}}_n^v$ and $\hat{\mathbf{N}}_n^v$ are defined in (3.11) and (5.50), respectively. From the relations in (5.66), we obtain the following FCLT.

Theorem 5.4. (FCLT for counting processes with cyclic thinning and stronger scaling) *If* $\hat{\mathbf{S}}_n^v \Rightarrow \hat{L}^v$ *in* D *for* $\hat{\mathbf{S}}_n^v$ *in* (3.11), *where* $P(\hat{L}^v \in C) = 1$, *as assumed in Theorem* 5.1, *then*

$$(\tilde{\mathbf{S}}_n^{c,v}, \tilde{\mathbf{N}}_n^{c,v}) \Rightarrow (\hat{L}^v, -\hat{L}^v) \quad in \quad D^2 \quad as \quad n \to \infty, \tag{5.67}$$

for $(\tilde{\mathbf{S}}_n^{c,v}, \tilde{\mathbf{N}}_n^{c,v})$ *in* (5.66).

Proof. For $\tilde{\mathbf{S}}_n^{c,v}$, apply (5.66) plus the reasoning in the proof of Theorem 4.2. For $\tilde{\mathbf{N}}_n^{c,v}$, first note that the J_n^v term is asymptotically negligible, because $J_n^v(t)$ is bounded by n, but it is divided by $n^{3/2}$. Then apply (5.66) and then (5.49), which follows from (5.47), as noted in the proof of Theorem 5.1.

6. Queue length and workload processes

As in Chapter 9 of [14], we can define other queueing processes. In the setting of §3.1, in model n let $C_n(t)$ be the cumulative input of work, $X_n(t)$ the net input of work, $I_n(t)$ be the cumulative idle time, $B_n(t)$ be the cumulative busy time, $D_n(t)$ the cumulative number of departures, all over the interval $[0, t]$ for $t \ge 0$. Let $Q_n(t)$ the queue length (number in system) and $R_n(t)$ the remaining work in the system (the continuous-time workload), both at time t for $t \ge 0$. These can be defined by

$$\begin{aligned} C_n(t) &\equiv S^v_{n,N^u_n(t)}, \quad X_n(t) \equiv C_n(t) - t, \quad I_n(t) \equiv -\inf_{0\le s\le t}\{X_n(s)\}, \\ B_n(t) &\equiv t - I_n(t), \quad D_n(t) \equiv N^v_n(B_n(t)), \quad Q_n(t) \equiv N^u_n(t) - D_n(t), \\ R_n(t) &\equiv \psi(X_n)(t) = X_n(t) - \inf_{0\le s\le t}\{X_n(s)\} = X_n(t) + I_n(t), \quad t \ge 0. \end{aligned} \tag{6.68}$$

We add a superscript c to indicate when these processes are associated with cyclic thinning, i.e., when the n^{th} system is the $G_n/G_n/1$ model.

6.1. No time scaling

In the setting of §5.1, we can establish limits for almost all the processes in (6.68). In particular, we establish limits for all but the queue-length processes Q_n^c. For the queue-length process, we are only able to establish a strong bound. In particular, we

have the following result; we omit the proof for all but the queue length process, which follows easily from the basic definitions (4.20), (4.21), the assumption in (4.30) and the FWLLN following from (3.13).

Theorem 6.1. (HT limit in the $G_n/G_n/1$ models without time scaling) *Assume* (5.47), *again allowing either* (4.30) *or* (3.14). *Assume that* $W_{n,0}^c \Rightarrow W_0^c$. *the following limit for the processes without time scaling: As* $n \to \infty$,

$$\begin{aligned}&(S_n^{c,u}, S_n^{c,v}, S_n^c, W_n^c, N_n^{c,u}, N_n^{c,v}, C_n^c, X_n^c, I_n^c, B_n^c, R_n^c, D_n^c)\\ &\quad \Rightarrow (S^{c,u}, S^{c,v}, S^c, W^c, N^{c,u}, N^{c,v}, C^c, X^c, I^c, B^c, R^c, D^c) \qquad (6.69)\end{aligned}$$

in $(\mathbb{R}^\infty)^4 \times D^8$, *where*

$$\begin{aligned} S_k^{c,u} &\equiv k \equiv S_k^{c,v} \quad \textit{and} \quad S_k^c \equiv W_0^c \equiv W_k^c \quad k \geq 0, \quad N^{c,u}(t) \equiv 1 + \lfloor t \rfloor,\\ N^{c,v}(t) &\equiv \lfloor t \rfloor, C^c(t) \equiv W_0^c + 1 + \lfloor t \rfloor, \quad X^c(t) \equiv W_0^c + 1 + \lfloor t \rfloor - t,\\ I^c(t) &\equiv 0, \quad B^c(t) \equiv t, \quad R^c(t) \equiv X^c(t), \quad D^c(t) \equiv \lfloor t \rfloor. \qquad (6.70)\end{aligned}$$

In addition, if $W_0^c \equiv 0$, *then for all non-integer* t, $Q_n^c(t) \Rightarrow 1$ *as* $n \to \infty$ *and*

$$P(Q_n^c(s) \in \{0,1,2\} \quad \textit{for all} \quad s \in [0,t]) \to 1 \quad \textit{as} \quad n \to \infty \quad \textit{for all} \quad t > 0. \qquad (6.71)$$

Proof. We only have difficulty with Q_n^c. For it, we avoid specifying the number of customers in the system initially by assuming that $W_{n,0}^c \Rightarrow 0$. Still Q_n^c remains difficult because it involves subtraction of integer-valued functions with common discontinuity points. We have $Q_n^c(t) \equiv N_n^{c,u}(t) - D_n^c(t)$, where $D_n^c(t) = N_n^{c,v}(B_n^c(t)) \Rightarrow \lfloor e \rfloor$. The limit for non-integer t is thus $(1 + \lfloor e \rfloor)(t) - (\lfloor e \rfloor)(t) = 1$. But there can be problems at the integer points where both limits have discontinuities. But at each discontinuity point the error can be at most 1. Hence we have (6.71).

If instead of those initial conditions, we had the arrival at time 0 find $k-1$ others in the system at time 0, with one of those just beginning service, so that $Q^c(0) = k$, including the new arrival, we would have the alternative limits $Q^c(t) = k$ and $R^c(t) = k + \lfloor t \rfloor - t$, for all non-integer time points $t > 0$, so that $k-1 \leq R^c(t) \leq k$ and $k-1 \leq Q(t) \leq k+1$ for all $t \geq 0$. The practical implication is that the $G_n/G_n/1$ model for large n behaves like the $D/D/1$ model with $\rho = 1$ over short time intervals. (Note that, with deterministic unit service times, $Q(t) = \lceil W(t) \rceil$, where $\lceil t \rceil$ is the ceiling

function, giving the least integer greater than t.) Consequently, the queue-length and workload processes in the $G_n/G_n/1$ model for large n will change negligibly over short time intervals.

6.2. Time scaling by n

We now show the implications of the results in §5.2 for the other queueing processes introduced in (6.68). Define the following random elements of D:

$$\begin{aligned}
\mathbf{X}_n^c(t) &\equiv \mathbf{C}_n^c(t) \equiv C^c(nt) - nt = X^c(nt), \\
\mathbf{R}_n^c(t) &\equiv R_n^c(nt) = \phi(\mathbf{X}_n^c)(t), \\
\mathbf{I}_n^c(t) &\equiv I_n^c(nt) = nt - B_n^c(nt) \equiv -\mathbf{B}_n^c(t), \\
\mathbf{D}_n^c(t) &\equiv D_n^c(nt) - nt = N_n^{c,v}(B_n^c(nt)) - nt, \\
\mathbf{Q}_n^c(t) &\equiv Q_n^c(nt) = N_n^{c,u}(nt) - D_n^c(nt) = \mathbf{N}_n^{c,u}(t) - \mathbf{D}_n^c(t). \qquad (6.72)
\end{aligned}$$

Theorem 1. (HT limit for other processes in the $G_n/G_n/1$ model) *Under the assumptions of Theorem* 5.1*,*

$$(\mathbf{W}_n^c, \mathbf{X}_n^c, \mathbf{R}_n^c, \mathbf{B}_n^c, \mathbf{I}_n^c, \mathbf{D}_n^c, \mathbf{Q}_n^c) \Rightarrow (W^c, X^c, R^c, B^c, I^c, D^c, Q^c) \quad \textit{in} \quad \boldsymbol{C}^7, \qquad (6.73)$$

where

$$\begin{aligned}
W^c &\equiv \phi(L), \quad X^c \equiv L + \Upsilon^u, \quad R^c \equiv \phi(L) + \Upsilon^u, \quad B^c = -I^c \equiv L - \phi(L), \\
D^c &\equiv -\hat{L}^v - \Upsilon^v + L - \phi(L) \quad \textit{and} \quad Q^c \equiv \phi(L) + \Upsilon^u + \Upsilon^v, \qquad (6.74)
\end{aligned}$$

where $L \equiv \hat{L}^v - \hat{L}^u - \beta e$ *for* $(\hat{L}^u, \hat{L}^v)$ *in* (3.13)*,* ϕ *is the reflection map in* (3.7) *and* Υ^u *and* Υ^v *are the unit blurs associated with* $N_n^{c,u}$ *and* $N_n^{c,v}$*, respectively.*

Proof. We first obtain $\mathbf{W}_n^c \Rightarrow W^c \equiv \phi(L)$ in D and thus in $\mathbf{C}$ from Theorem 4.2. Jointly with that, we can obtain the following limits. Note that $\mathbf{X}_n^c = \mathbf{S}_n^{c,v} \circ \bar{\mathbf{N}}_n^{c,u} + \mathbf{N}_n^{c,u}$, where $(\mathbf{S}_n^{c,v}, \bar{\mathbf{N}}_n^{c,u}, \mathbf{N}_n^{c,u}) \Rightarrow (\hat{L}^v, e, -\hat{L}^u - \beta e + \Upsilon^u)$. Hence, $\mathbf{X}_n^c \Rightarrow L + \Upsilon^u$ as $n \to \infty$. Then $\mathbf{R}_n^c(t) \Rightarrow \phi(L + \Upsilon^u) = \phi(L) + \Upsilon^u$ and $\mathbf{I}_n^c = \mathbf{X}_n^c - \mathbf{R}_n^c \Rightarrow L - \phi(L)$. Next for $\mathbf{D}_n^c$, first note that $\bar{\mathbf{B}}_n^c \Rightarrow e$, where $\bar{\mathbf{B}}_n^c(t) \equiv B_n^c(nt)/n$, because $\mathbf{B}_n^c$ converges to a nondegenerate limit. Since $\mathbf{D}_n^c = \mathbf{N}_n^{c,v} \circ \bar{\mathbf{B}}_n^c + \mathbf{B}_n^c$, $\mathbf{D}_n^c \Rightarrow -\hat{L}^v - \Upsilon^v + L - \phi(L)$ and $\mathbf{Q}_n^c = \mathbf{N}_n^{c,u} - \mathbf{D}_n^c \Rightarrow -\hat{L}^u - \beta e + \Upsilon^u + \hat{L}^v + \Upsilon^v - L + \phi(L) = \phi(L) + \Upsilon^u + \Upsilon^v$.

Notice that as a special case of the limit in (6.73) above, we have the joint limit

$$(\mathbf{W}_n^c, \mathbf{R}_n^c, \mathbf{Q}_n^c) \Rightarrow (W^c, R^c, Q^c) \equiv (\phi(L), \phi(L) + \Upsilon^u, \phi(L) + \Upsilon^u + \Upsilon^v) \quad \text{in} \quad \mathbf{C}^3, \tag{6.75}$$

from which we see that the limits are ordered, i.e.,

$$W^c(t) \le R^c(t) \le W^c(t) + 1 \quad \text{and} \quad R^c(t) \le Q^c(t) \le R^c(t) + 1 \quad \text{for all} \quad t \ge 0, \tag{6.76}$$

with strict inequality holding for most t. We will develop explicit approximations for the associated steady-state quantities $(W_{n,\infty}^c, R_{n,\infty}^c, Q_{n,\infty}^c)$ in [11], which refine (6.75) but are consistent with it.

6.3. Time scaling by n^2

We now show the implications of Theorem 5.4 for the other queueing processes introduced in (6.68); we omit the elementary proof. Define the following new random elements of D by scaling time by n and dividing by $\sqrt{n}$ in the previous random elements in (6.72):

$$\begin{aligned}
\tilde{\mathbf{X}}_n^c(t) &\equiv \tilde{\mathbf{C}}_n^c(t) \equiv \mathbf{X}_n^c(nt)/\sqrt{n} = X^c(n^2t)/\sqrt{n}, \\
\tilde{\mathbf{R}}_n^c(t) &\equiv \mathbf{R}_n^c(nt)/\sqrt{n} = \phi(\tilde{\mathbf{X}}_n^c)(t), \quad \tilde{\mathbf{I}}_n^c(t) \equiv \mathbf{I}_n^c(nt)/\sqrt{n}, \\
\tilde{\mathbf{D}}_n^c(t) &\equiv \mathbf{D}_n^c(nt)/\sqrt{n}, \quad \tilde{\mathbf{Q}}_n^c(t) \equiv \mathbf{Q}_n^c(nt)/\sqrt{n} = \tilde{\mathbf{N}}_n^{c,u}(t) - \tilde{\mathbf{D}}_n^c(t).
\end{aligned} \tag{6.77}$$

Overall, time is now scaled by n^2, e.g., $\tilde{\mathbf{Q}}_n^c(t) = Q_n^c(n^2t)/\sqrt{n}$, $t \ge 0$.

Theorem 6.2. (HT limit for other processes in the $G_n/G_n/1$ model with stronger scaling) *Under the assumptions of Theorem* 5.1*,*

$$(\tilde{\mathbf{W}}_n^c, \tilde{\mathbf{X}}_n^c, \tilde{\mathbf{R}}_n^c, \tilde{\mathbf{B}}_n^c, \tilde{\mathbf{I}}_n^c, \tilde{\mathbf{D}}_n^c, \tilde{\mathbf{Q}}_n^c) \Rightarrow (\tilde{W}^c, \tilde{X}^c, \tilde{R}^c, \tilde{B}^c, \tilde{I}^c, \tilde{D}^c, \tilde{Q}^c) \quad \textit{in} \quad D^7, \tag{6.78}$$

where

$$\begin{aligned}
\tilde{W}^c &\equiv \phi(L), \quad \tilde{X}^c \equiv L, \quad \tilde{R}^c \equiv \phi(L), \quad \tilde{B}^c = -\tilde{I}^c \equiv L - \phi(L), \\
\tilde{D}^c &\equiv -\hat{L}^v + L - \phi(L) \quad \textit{and} \quad \tilde{Q}^c \equiv \phi(L),
\end{aligned} \tag{6.79}$$

where $L \equiv \hat{L}^v - \hat{L}^u - \beta e$ *for* $(\hat{L}^u, \hat{L}^v)$ *in* (3.13)*,* ϕ *is the reflection map in* (3.7)*.*

7. The motivating $G_n/D/\infty$ infinite-server queue

We now apply the FCLT just established in §5.3 to obtain a HT FCLT for the queue length (number of busy servers) in a $G_n/D/\infty$ model, addressing the motivating inventory problem from [12] mentioned in §2. We start with unit service times and a rate-1 arrival process; let the counting process be $\{N^u(t) : t \geq 0\}$. To achieve an associated rate-λ process, we scale time by λ, i.e., by using the process $\{N^u(\lambda t)\}$. Starting with a base $G/D/\infty$ queue with a rate-1 arrival process, we want to consider the associated $G_n/D/\infty$ queue with an arrival rate of n^2. We denote this arrival process as $\{N_n^{c,u}(t) : t \geq 0\}$. (Motivated by [12], the arrival rate should grow as the *square* of the cycle order n.)

Let the queue length at time t with arrival process $N_n^{c,u}$ and arrival rate n^2 be $Q_n^{\infty,c}(t)$. As in [5],

$$Q_n^{\infty,c}(t) = N_n^{c,u}(t) - N_n^{c,u}(t-1) \quad \text{for all} \quad t \geq 1. \tag{7.80}$$

Consequently, if the arrival process is time stationary, then $Q_n^{\infty,c}(t)$ has a fixed (stationary) distribution for all $t \geq 1$, regardless of the initial conditions. (It reaches steady state at time 1.) In general, we can obtain a HT FCLT for the scaled queue-length process

$$\mathbf{Q}_n^{\infty,c}(t) \equiv \frac{Q_n^{\infty,c}(t) - n^2}{\sqrt{n}}, \quad t \geq 0. \tag{7.81}$$

Theorem 7.1. *For the sequence of $G_n/D/\infty$ queues with arrival rate n^2 in model n (with cyclic thinning of order n) just defined, if the base stationary arrival counting process N^u satisfies a FCLT, i.e., if $\hat{\mathbf{N}}_n^u \Rightarrow \sigma_u B$ in D, where B is BM and $\mathbf{N}_n^u(t) \equiv (N^u(nt) - nt)/\sqrt{n}$, then*

$$\mathbf{Q}_n^{\infty,c} \Rightarrow \mathbf{Q}^{\infty,c} \quad \textit{in} \quad D([1,\infty),\mathbb{R}), \tag{7.82}$$

where $\mathbf{Q}_n^{\infty,c}$ is defined in (7.81) *above and $\mathbf{Q}^{\infty,c}(t) \equiv \sigma_u(B(t) - B(t-1)$, $t \geq 1$, so that*

$$\frac{Q_n^{\infty,c}(t) - n^2}{\sqrt{n}} \Rightarrow N(0,\sigma_u^2) \quad \textit{in} \quad \mathbb{R} \quad \textit{as} \quad n \to \infty \quad \textit{for all} \quad t \geq 1. \tag{7.83}$$

Proof. Essentially, we are applying Theorem 5.4. Directly, by (7.80) and the fact that $N_n^{u,c}$ has rate n^2,

$$\begin{aligned} Q_n^{\infty,c}(t) &\stackrel{\mathrm{d}}{=} N_n^{c,u}(t) - N_n^{c,u}((t-1)) \\ &= \frac{N^u(n^3 t) - J_n^{c,u}(n^3 t) - N^u(n^3(t-1)) + J_n^{c,u}(n^3(t-1))}{n}. \end{aligned} \tag{7.84}$$

Hence, when we divide by $\sqrt{n}$ in (7.84), as required for (7.81), the two $J_n^{c,u}$ terms on the right in (7.84) become asymptotically negligible. Ignoring them (using Theorems 11.4.5 and 11.4.7 of [14]), we have $\mathbf{Q}_n^{\infty,c}(t) \stackrel{\mathrm{d}}{=} \mathbf{N}_{n^3}^u(t) - \mathbf{N}_{n^3}^u((t-1))$. Hence the limit in (7.82) follows from the assumed FCLT for $\mathbf{N}_n^u$.

Because of the cyclic thinning, in model n the translation term in (7.81) is n^2, but the spatial scaling is only by $\sqrt{n}$. Nevertheless, this is consistent with [5] (even though it does not follow directly from [5]); e.g., for the $G_k/D/\infty$ model, with arrival rate n^2 but fixed cyclic thinning order k, from [5] we get the limit

$$\frac{Q_n^{\infty,c,k}(t) - n^2}{n} \Rightarrow N(0, \sigma_u^2/k) \stackrel{\mathrm{d}}{=} (1/\sqrt{k})N(0, \sigma_u^2) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty, \tag{7.85}$$

for each fixed k. If we multiply through by $\sqrt{k}$ in (7.85) and then (formally) let $n \to \infty$ with $k = n$, we obtain (7.83).

Acknowledgments

This research was supported by NSF grant CMMI 0948190. We thank doctoral students Yunan Liu for conducting supporting simulation experiments and Guodong Pang for helpful comments.

References

[1] Abate, J., Choudhury, G. L. and Whitt, W. (1993). Calculation of the GI/G/1 steady-state waiting-time distribution and its cumulants from Pollaczek's formula. *Int. J. Electronics and Communications* (*Archiv für Elektronik und Übertragungstechnik*, AEU) **47,** 311–321.

[2] Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, New York.

[3] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd edn. John Wiley, New York.

[4] Borovkov, A. A. (1984). *Asymptotic Methods in Queueing Theory*, John Wiley, New York.

[5] 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.

[6] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operat. Res.* **29,** 567–587.

[7] Jelenkovic, P., Mandelbaum, A. and Momcilovic, P. (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* **47,** 53–69.

[8] Kingman, J.F.C. (1961). The single-server queue in heavy traffic. *Proc. Camb. Phil. Soc.* **57,** 902–904.

[9] Liu, W. and Whitt, W. (2010). The heavily loaded many-server queue with abandonment and deterministic service times. IEOR Department, Columbia University, New York. http://www.columbia.edu/∼ww2040/allpapers.html

[10] Pang, G., Talreja, R. and Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4,** 193–267.

[11] Sigman, K. and Whitt, W. (2010). Heavy-traffic limits for nearly deterministic queues: stationary distributions. IEOR Department, Columbia University, New York. Submitted for publication. http://www.columbia.edu/∼ww2040/allpapers.html

[12] Song, J.-S. and Zipkin, P. H. (1996). The joint effect of leadtime variance and lot size in a parallel processing environment. *Management Sci.* 42, 1352–1363.

[13] Tijms, H. C. (1994). *Stochastic Models: An Algorithmic Approach*, John Wiley, New York.

[14] Whitt, W. (2002). *Stochastic-Process Limits*. Springer, New York.