Queueing Syst (2011) 69:145–173
DOI 10.1007/s11134-011-9253-y

Heavy-traffic limits for nearly deterministic queues: stationary distributions

Karl Sigman · Ward Whitt

Received: 18 April 2010 / Revised: 5 May 2011 / Published online: 30 July 2011

Abstract We establish heavy-traffic limits for stationary waiting times and other performance measures in $G_n/G_n/1$ queues, where G_n indicates that an original point process is modified by cyclic thinning of order n, i.e., the thinned process contains every nth point from the original point process. The classical example is the Erlang $E_n/E_n/1$ queue, where cyclic thinning of order n is applied to both the interarrival times and the service times, starting from a "base" $M/M/1$ model. The models $G_n/D/1$ and $D/G_n/1$ are special cases of $G_n/G_n/1$. 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 interarrival times are the sum of n consecutive interarrival times in the original model, the $G/D/n$ model is a special case as well. As $n \to \infty$, the $G_n/G_n/1$ models approach the deterministic $D/D/1$ model. We obtain revealing limits by letting $\rho_n \uparrow 1$ as $n \to \infty$, where ρ_n is the traffic intensity in model n.

Keywords Heavy traffic · Nearly deterministic queues · Cyclic thinning · Point processes · Stationary waiting times · Many-server queues · Deterministic service times · Gaussian random walk · Nearly deterministic queues · Limit interchange

Mathematics Subject Classification (2000) Primary 60K25 · Secondary 60F05 · 60G10

1 Introduction

This paper is a sequel to [21], in which we established heavy-traffic (HT) stochastic-process limits for $G_n/G_n/1$ queues, where G_n indicates that an original point process

K. Sigman · W. Whitt (✉)
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA
e-mail: ww2040@columbia.edu

K. Sigman
e-mail: ks20@columbia.edu

is modified by cyclic thinning of order n, i.e., the thinned process contains every nth point from the original point process. More precisely, with S_k denoting the kth point of a general (G) point process, for $k \geq 0$ with $S_0 \equiv 0$, S_{nk}/n is the kth point in the associated G_n point process. (We divide by n to keep the intensity unchanged.)

For large n, the $G_n/G_n/1$ queues are *nearly deterministic queues* because, under regularity conditions, the $G_n/G_n/1$ queues approach the deterministic $D/D/1$ queue as $n \to \infty$, by virtue of the law of large numbers. Consequently, for any fixed traffic intensity $\rho < 1$, the congestion is asymptotically negligible as $n \to \infty$. In [21], we obtained interesting nondegenerate limits as $n \to \infty$ by letting $\rho_n \uparrow 1$ as $n \to \infty$, where ρ_n is the traffic intensity as a function of n. In Sects. 1–2 of [21], we discussed motivation for considering these models.

In [21], we obtained HT limits with two different scalings:

$$(1 - \rho_n)\sqrt{n} \to \beta \quad \text{as } n \to \infty, \tag{1}$$

$$(1 - \rho_n)n \to \beta \quad \text{as } n \to \infty, \tag{2}$$

where β is a finite positive constant in each case. Let $W^c_{n,k}$ be the waiting time of arrival k in the $G_n/G_n/1$ queue, where the superscript c indicates that cyclic thinning of order n is applied to a base $G/G/1$ model. In case (1), we obtained a limit for the *spatially-scaled* waiting times $\sqrt{n}W^c_{n,k}$; in case (2), we obtained a limit for the *temporally-scaled* waiting times $W^c_{n,nk}$.

We now want to consider associated stationary waiting times. Under regularity conditions, the waiting times $W^c_{n,k}$ converge to stationary waiting times $W^c_{n,\infty}$ as $k \to \infty$ for each n. We can apply [21] to generate approximations for those stationary waiting times by considering the iterated limit in which first $n \to \infty$ and then $k \to \infty$. We now provide conditions under which the limit interchange is valid. In particular, we provide conditions under which $\sqrt{n}W^c_{n,\infty}$ and $W^c_{n,\infty}$ converge in distribution to proper limits as $n \to \infty$ in cases (1) and (2), respectively, and identify the limits with the iterated limits already established.

Special cases of the two limits were established previously. First, in Example 3.1 of [1], exploiting the known Laplace transform of the stationary waiting time, the authors showed in case (2) that $W^c_{n,\infty} \to W^c_\infty$ as $n \to \infty$, where W^c_∞ is an exponential random variable, when the $G_n/G_n/1$ model is $E_n/E_n/1$, i.e., when the base model is $M/M/1$. Second, in [13] the authors showed in case (1) that $\sqrt{n}W^c_{n,\infty} \to \tilde{W}^c_\infty$ as $n \to \infty$, where $\tilde{W}^c_\infty$ is the maximum of a Gaussian random walk with negative drift, when the $G_n/G_n/1$ model is $GI_n/D/1$. In [13], the authors actually considered the $G/D/n$ model, but they analyzed it by exploiting the fact that the waiting times are the same as in the associated $GI_n/D/1$ model. Our results here are extensions of those two results. We obtain some results for general $G_n/G_n/1$ models, but most of our results are for the special case $GI_n/GI_n/1$ in which the base model is $GI/GI/1$.

The two different scalings in (1) and (2) indicate that high-order cyclic thinning produces some interesting behavior. To a large extent, this phenomenon can be explained by the fact that two parts of the distribution of $W^c_{n,\infty}$ tend to have different asymptotic behavior. Paralleling the relatively well understood many-server queue (for example, see [25]), the delay probability $P(W^c_{n,\infty} > 0)$ and the conditional delay distribution $P(W^c_{n,\infty} > t | W^c_{n,\infty} > 0)$ behave differently. In case (1), the delay

probability $P(W^c_{n,\infty} > 0)$ has a nondegenerate limit α (with $0 < \alpha < 1$) as $n \to \infty$, without scaling, while $W^c_{n,\infty} \Rightarrow 0$, i.e., $P(W^c_{n,\infty} > t | W^c_{n,\infty} > 0) \to 0$ as $n \to \infty$ for each $t \geq 0$. On the other hand, in case (2), $P(W^c_{n,\infty} > 0) \to 1$, a degenerate limit, while $P(W^c_{n,\infty} > t | W^c_{n,\infty} > 0)$ has a nondegenerate limit as $n \to \infty$ for each t. This will be a unifying theme throughout the paper.

We first address a foundational issue. We show that stationary waiting times are well defined by placing the $G_n/G_n/1$ model in a stationary framework. To do so, we show that stationarity and ergodicity assumed for a point process are inherited by the new point process created by cyclic thinning. We postpone that discussion until Sect. 6, but we apply the conclusion in the rest of the paper. Afterwards, in Sect. 7 we combine Sect. 6 with Sect. 5 of [21] in order to establish limits for stochastic point processes modified by cyclic thinning. In Sect. 5 of [21] we showed that counting processes created by cyclic thinning do not have the same relatively simple asymptotic behavior as the associated partial sums. (The continuous mapping theorem with the inverse map discussed in Sect. 13 of [26] does not apply in the usual way.)

Here is how the rest of this paper is organized. In Sect. 2 we establish HT limits for stationary waiting times for general $G_n/G_n/1$ queues in case (1). In Sect. 3 we extract the consequences for the special case of $GI_n/GI_n/1$ queues, when all the models are $GI/GI/1$ queues. Sections 2 and 3 extend [13] and use similar reasoning.

In Sect. 4 we establish HT limits for the stationary waiting times in $GI_n/GI_n/1$ models in case (2). In Sect. 5 we examine the quality of the approximations for steady-state distributions in the $GI_n/GI_n/1$ model. In Sect. 6 we show that stationarity and ergodicity assumed for a point process are inherited by the new point process created by cyclic thinning. Afterwards, in Sect. 7 we combine Sect. 6 with Sect. 5 of [21] in order to establish limits for stochastic point processes modified by cyclic thinning.

In Sect. 8 we obtain HT limits for stationary queue lengths in $GI_n/GI_n/1$ models. The renewal arrival process allows us to apply the distributional version of Little's law. In Sect. 9 we establish HT limits for the stationary queue lengths in the $GI/D/n$ model. In Sect. 10 we discuss implications for staffing in the $GI/D/n$ model, based on steady-state performance constraints, relating to [13]. We show that there is a case for staffing in case (2) instead of in case (1). This parallels previous conclusions in [15, 25].

2 Heavy traffic limit in case (1): $(1 - \rho)\sqrt{n} \to \beta$

In this section we supplement the stochastic-process HT limit for $G_n/G_n/1$ models in case (1) provided by Theorem 4.1 of [21] by establishing an HT limit for associated stationary waiting time processes, assuming additional stationarity for the base $G/G/1$ model. As in [21], we primarily work in a single-sequence framework, but we initially consider a more general double-sequence framework. In each case, we assume that a FCLT holds jointly for the arrival and service processes.

2.1 The double-sequence framework

We consider general single-server queues with unlimited waiting room and the FCFS service discipline. For each $n \geq 1$, there is a base $G/G/1$ model specified by a se-

quence $\{(U_{n,k}, V_{n,k-1}) : k \geq 1\}$, where $U_{n,k}$ represents the interarrival time between customers $k-1$ and k and $V_{n,k}$ represents the service time of customer k. We assume that a 0th 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

$$W_{n,k} \equiv [W_{n,k-1} + V_{n,k-1} - U_{n,k}]^+, \quad k \geq 1, \tag{3}$$

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 the initial waiting time $W_{n,0}$ and the partial sums via

$$W_{n,k} \equiv W_{n,0} + S_{n,k} - \min_{0 \leq j \leq k} \{(W_{n,0} + S_{n,j}) \wedge 0\}, \quad k \geq 0, \tag{4}$$

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

$$S_{n,k} \equiv X_{n,1} + \cdots + X_{n,k} \quad \text{for } X_{n,k} \equiv V_{n,k-1} - U_{n,k},\ k \geq 1, \tag{5}$$

with $S_{n,0} \equiv 0$, so that $S_{n,k} = S^v_{n,k} - S^u_{n,k}$ with $S^u_{n,k} \equiv U_{n,1} + \cdots + U_{n,k}$, $S^v_{n,k} \equiv V_{n,0} + \cdots + V_{n,k-1}$, $k \geq 1$, $S^v_{n,0} \equiv 0$ and $S^u_{n,0} \equiv 0$; see Sect. 9.2 of [26].

Formula (4) constitutes 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

$$\tilde{\phi}(x)(k) \equiv x_k - \min_{0 \leq j \leq k} \{x_j \wedge 0\}, \quad k \geq 0. \tag{6}$$

We now introduce an associated sequence of $G_n/G_n/1$ models by applying cyclic thinning of order n to both the arrival and service processes in model n. Specifically, 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

$$S^{c,u}_{n,k} \equiv S^u_{n,kn}/n \quad \text{and} \quad S^{c,v}_{n,k} \equiv S^v_{n,kn}/n \quad \text{for all } n \geq 1 \text{ and } k \geq 1. \tag{7}$$

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

$$U^c_{n,k} \equiv S^{c,u}_{n,k} - S^{c,u}_{n,k-1} \quad \text{and} \quad V^c_{n,k-1} \equiv S^{c,v}_{n,k} - S^{c,v}_{n,k-1}. \tag{8}$$

From (7) and (8), 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 now review Theorem 4.1 of [21]. For that purpose, let $D \equiv D([0, \infty), \mathbb{R})$ be the function space 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, 26]. Let $D^k \equiv D \times \cdots \times D$ be the k-fold product space of D with itself, endowed with the usual product topology; let C and C^k be the

subsets of continuous functions in D and D^k, respectively. Let $\Rightarrow$ denote convergence in distribution.

Let random elements associated with the sequence of base $G/G/1$ models be defined by

$$\mathbf{S}_n^u(t) \equiv \frac{S^u_{n,\lfloor nt\rfloor} - \lfloor nt\rfloor}{\sqrt{n}}, \qquad \mathbf{S}_n^v(t) \equiv \frac{S^v_{n,\lfloor nt\rfloor} - \lfloor nt\rfloor}{\sqrt{n}}, \quad t \ge 0, \tag{9}$$

where $\lfloor t\rfloor$ is the greatest integer less than or equal to t.

We now introduce associated random elements of the space $\mathbb{R}^\infty$ for the $G_n/G_n/1$ models constructed above. For $n \ge 1$, let

$$\begin{aligned} \tilde{S}_n^{c,u}(k) &\equiv \sqrt{n}\big(S_{n,k}^{c,u} - k\big), \qquad \tilde{S}_n^{c,v}(k) \equiv \sqrt{n}\big(S_{n,k}^{c,v} - k\big), \\ \tilde{S}_n^{c}(k) &\equiv \sqrt{n}S_{n,k}^{c}, \quad \text{and} \quad \tilde{W}_n^{c}(k) \equiv \sqrt{n}W_{n,k}^{c}, \quad k \ge 1,\ n \ge 1, \end{aligned} \tag{10}$$

with $\tilde{S}_n^c(0) \equiv \sqrt{n}S_{n,0} \equiv 0$, where $(S_{n,k}^{c,u}, S_{n,k}^{c,v})$ is defined in (7), $S_{n,k}^c \equiv S_{n,k}^{c,v} - S_{n,k}^{c,u}$ and $W_{n,k}^c$ is defined in terms of $\{S_{n,k}^c : k \ge 0\}$ as in (3).

Theorem 1 (HT stochastic-process limit from [21]) *Consider a sequence of $G_n/G_n/1$ models associated with a sequence of base $G/G/1$ models and initial waiting times $W_{n,0}^c$, where*

$$\big(\sqrt{n}W_{n,0}^c, \mathbf{S}_n^u, \mathbf{S}_n^v\big) \Rightarrow \big(\tilde{W}(0), L^u, L^v\big) \quad \text{in } \mathbb{R}\times D^2, \tag{11}$$

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

$$\big(\tilde{W}_n^c(0), \tilde{S}_n^{c,u}, \tilde{S}_n^{c,v}, \tilde{S}_n^c, \tilde{W}_n^c\big) \quad \Rightarrow \quad \big(\tilde{W}(0), \tilde{L}^u, \tilde{L}^v, \tilde{L}, \tilde{W}\big) \quad \text{in } \mathbb{R}\times\big(\mathbb{R}^\infty\big)^4, \tag{12}$$

where $\tilde{W} \equiv \tilde{\phi}(\tilde{W}(0) + \tilde{L})$ for $\tilde{\phi}$ defined in (6), $\tilde{L} = \tilde{L}^v - \tilde{L}^u$, $\tilde{L}^u(k) \equiv L^u(k)$ and $\tilde{L}^v(k) \equiv L^v(k)$, $k \ge 1$.

We now proceed to our new result for stationary waiting time processes in the $G_n/G_n/1$ models. First, assume that sequences $\{(U_{n,k}, V_{n,k-1}) : k \ge 1\}$, $n \ge 1$, in the base $G/G/1$ models are strictly stationary and ergodic for each n, with finite means satisfying $E[V_{n,k-1}] < E[U_{n,k}]$. By Sect. 6, this stationary (and ergodic) framework carries over to the basic sequences $\{(U_{n,k}^c, V_{n,k-1}^c) : k \ge 1\}$, $n \ge 1$, in the $G_n/G_n/1$ models with cyclic thinning. It is well known that stationary waiting time processes exist under these assumptions; see Chap. 6 of [20]. We will use the same notation to refer to the stationary processes.

As is customary, for example, as on p. 207 of [4], we consider associated two-sided stationary infinite sequences $\{(U_{n,k}^c, V_{n,k-1}^c) : -\infty < k < \infty\}$ for each n. (The extension to two-sided stationary sequences can always be constructed; see, for example, [5].) Let $X_{n,k}^c \equiv V_{n,k-1}^c - U_{n,k}^c$ and let the reverse-time partial sums be defined by $S_{n,j,k}^{c,r} \equiv X_{n,j-1}^c + \cdots + X_{n,j-k}^c$. By the stationarity, $S_{n,j,k}^{c,r}$ is distributed the same as $S_{n,j,k}^c \equiv X_{n,j+1}^c + \cdots + X_{n,j+k}^c$ for each n, j and k, but the finite-dimensional distributions as a function of j and k are in general different for each n.

With this reverse-time framework, the stationary waiting times can be expressed as simple maxima, i.e., for each $n \geq 1$,

$$W_{n,j}^c \equiv \max_{k\geq 0}\left\{S_{n,j,k}^{c,r}\right\}, \quad j \geq 1 \text{ and } n \geq 1. \tag{13}$$

For $GI/GI/1$ models, we can exploit the independence to obtain the same relation for the one-dimensional marginal distributions using the forward partial sums, but not more generally.

In this stationary framework, scale all the random variables as in (10), using the same notation. Let $\tilde{X}_n^c(k) \equiv \tilde{S}_n^c(k) - \tilde{S}_n^c(k-1)$ for $\tilde{S}_n^c(k)$ in (10). Let $\stackrel{\mathrm{d}}{=}$ mean equal in distribution.

The following result extends Theorem 1 of [13] and is proved the same way, by applying the model stability result on p. 207 of Borovkov [4].

Theorem 2 (HT limit for the scaled stationary waiting times) *Consider a sequence of $G_n/G_n/1$ models associated with a sequence of base $G/G/1$ models for which the sequences $\{(U_{n,k}, V_{n,k-1}) : -\infty < k < \infty\}$, $n \geq 1$, are strictly stationary and ergodic for each n, with finite means satisfying $E[V_{n,k-1}] < E[U_{n,k}]$. Let the conditions of Theorem 1 be satisfied. Then there exists a process $Y \equiv \{Y(j) : -\infty < j < \infty\}$ with $Y(j) \stackrel{\mathrm{d}}{=} L^v(1) - L^u(1)$ for each j, for (L^v, L^u) in Theorem 1, such that, as $n \to \infty$,*

$$\left\{\tilde{X}_n^c(k) : -\infty < k < \infty\right\} \quad \Rightarrow \quad \left\{Y(k) : -\infty < k < \infty\right\} \quad \text{in } \mathbb{R}^\infty. \tag{14}$$

Assume that the (necessarily stationary) sequence $\{Y(k) : -\infty < k < \infty\}$ is also ergodic with $E[Y(1)] < 0$. If, in addition,

$$E\left[\tilde{X}_n^c(1)1_{\{\tilde{X}_n^c(1)>0\}}\right] \to E\left[Y(1)1_{\{Y(1)>0\}}\right] < \infty \quad \text{as } n \to \infty, \tag{15}$$

then

$$\left\{\tilde{W}_n^c(j) : j \geq 0\right\} \quad \Rightarrow \quad \left\{\tilde{W}(j) : j \geq 0\right\} \quad \text{in } \mathbb{R}^\infty \quad \text{as } n \to \infty, \tag{16}$$

where

$$\tilde{W}(j) \equiv \max_{k\geq 0}\left\{Y(j-1) + \cdots + Y(j-k)\right\}, \quad j \geq 1. \tag{17}$$

Condition (15) is satisfied if $P(Y(1) > 0) > 0$, 0 is a continuity point of the cdf $P(Y(1) \leq x)$ and

$$E\left[\tilde{X}_n^c(1)\right] \to E\left[Y(1)\right] < \infty \quad \text{as } n \to \infty. \tag{18}$$

In turn, condition (18) is satisfied if

$$\sup_{n\geq 1}\left\{E\left[\tilde{X}_n^c(1)^2\right]\right\} < \infty. \tag{19}$$

Proof We apply the model continuity (or stability) result on p. 207 of [4], which has three conditions. Condition I there requires that the limit process $\{Y(k) : -\infty < k < \infty\}$ be stationary and ergodic with $E[Y(1)] < 0$. Stationarity follows

by the convergence discussed below; we have directly assumed the ergodicity. Condition II in [4] requires convergence of the finite dimensional distributions as stated in (14), which we now justify. We first apply Theorem 1 to get the limit in (12), which extends immediately to two-sided sequences. (Here we are only concerned with the partial sums; we are not concerned with the initial conditions.) That implies the required convergence of the finite-dimensional distributions in (14). The final technical condition III in [4] is equivalent to condition (15). In turn, condition (15) holds if condition (18) holds, $P(Y(1) > 0) > 0$ and

$$P\big(\tilde{X}_n^c(1) > 0\big) \to P\big(Y(1) > 0\big) > 0 \quad \text{as } n \to \infty. \tag{20}$$

However, we can apply the conditions of Theorem 1 to deduce that condition (20) is satisfied, provided that 0 is a continuity point of the cdf $P(Y(1) \le x)$. In particular, by Theorem 1,

$$P\big(\tilde{X}_n^c(1) > 0\big) = P(S_n/\sqrt{n} > 0) \to P\big(\big(L^v - L^u\big)(1)\big) = P\big(Y(1) > 0\big). \tag{21}$$

Finally, (19) plus the convergence in distribution in (14) implies uniform integrability, which in turn implies (18). In conclusion, we remark that in the i.i.d. case it suffices to apply Theorem X.6.1 in [2], which has an easier proof than the theorem in [4]. □

2.2 The single-sequence framework

We now simplify the setting somewhat and exhibit quite general conditions under which all the conditions of Theorem 2 are satisfied, with the limit process being a tractable Gaussian random walk, for which explicit expressions are available.

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 base queueing models starting from a single sequence of ordered pairs of random variables $\{(U_k, V_{k-1}) : k \ge 1\}$.

Let the associated sequences of partial sums be

$$S_k^u \equiv U_1 + \cdots + U_k, \quad \text{and} \quad S_k^v \equiv V_0 + \cdots + V_{k-1}, \quad k \ge 1, \tag{22}$$

$S_0^v \equiv 0$ and $S_0^u \equiv 0$. Introduce the usual sequence $(\hat{\mathbf{S}}^u, \hat{\mathbf{S}}^v) \equiv \{(\hat{\mathbf{S}}_k^u, \hat{\mathbf{S}}_k^v) : k \ge 0\}$ of random elements of D associated by

$$\hat{\mathbf{S}}_n^u(t) \equiv \frac{S_{\lfloor nt \rfloor}^u - \lfloor nt \rfloor}{\sqrt{n}}, \quad \text{and} \quad \hat{\mathbf{S}}_n^v(t) \equiv \frac{S_{\lfloor nt \rfloor}^v - \lfloor nt \rfloor}{\sqrt{n}}, \quad t \ge 0. \tag{23}$$

In this context, our basic assumption is that the sequence $\{(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v) : n \ge 1\}$ 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

$$V_{n,k} \equiv V_k \quad \text{and} \quad U_{n,k} \equiv \frac{U_k}{\rho_n} \quad \text{for all } n, k \ge 1, \tag{24}$$

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 Sect. 2.

Theorems 1 and 2 imply the following corollary. Let $N(m, \sigma^2)$ denote a normally distributed random variable with mean m and variance σ^2.

Corollary 1 (More detail in Theorems 1 and 2) *Consider a sequence of $G_n/G_n/1$ models constructed from a single base $G/G/1$ model as indicated above. Let e be the identity map on D, i.e., $(e(t) = t, t \geq 0)$.*

(a) *Suppose that*

$$\left(\sqrt{n}W_{n,0}, \hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v\right) \quad \Rightarrow \quad \left(\tilde{W}(0), \hat{L}^u, \hat{L}^v\right) \quad \text{in } \mathbb{R} \times D^2 \tag{25}$$

for $(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v)$ in (23), 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 (1), then the limit in (12) holds with $L^u = \hat{L}^u + \beta e$ and $L^v = \hat{L}^v$.

(b) *If, in addition,*

$$\left(\hat{L}^u, \hat{L}^v\right) = (\sigma_u B_u, \sigma_v B_v), \tag{26}$$

where $\tilde{W}(0)$, B_u and B_v are mutually independent, and B_u and B_v are standard Brownian motions, then $L \equiv L^v - L^u \stackrel{\mathrm{d}}{=} \sigma B - \beta e$, where B is a standard BM and $\sigma^2 \equiv \sigma_u^2 + \sigma_v^2$, so that $\tilde{W} \equiv \tilde{\phi}(\tilde{W}(0) + \tilde{L})$ becomes a reflected Gaussian random walk with i.i.d. steps distributed as $N(-\beta, \sigma^2)$, starting at the independent initial state $\tilde{W}(0)$, in particular,

$$\tilde{W} \equiv \left\{\tilde{W}(k) : k \geq 0\right\} = \left\{\tilde{\phi}\left(\tilde{W}(0) + \sigma B - \beta e\right)(k) : k \geq 1\right\} \quad \text{in } \mathbb{R}^\infty. \tag{27}$$

(c) *If, in addition, the sequence $\{(U_k, V_{k-1}) : k \geq 1\}$ in the base model is stationary and ergodic with $E[U_k] = E[V_k] = 1$, which is consistent with the other assumptions above, then all the conditions of Theorem 2 are satisfied, with the limit of the scaled stationary waiting time processes being as in (27) above, where the initial value $\tilde{W}(0)$ is distributed as the stationary value.*

Proof When we add the extra regularity in Corollary 1 we see that the conditions in Theorems 1 and 2 are satisfied. For Theorem 2, the limit process Y becomes a stationary Gaussian random walk, which is necessarily ergodic as well as stationary. In addition, it follows that $P(Y(1) > 0) > 0$ and 0 is a continuity point of the cdf $P(Y(1) \leq x)$. Thus, in order to establish the desired limit (16) given (a) and (b), it suffices to establish the convergence of means in (18). That holds under (c):

$$E\left[\tilde{X}_n(1)\right] = E\left[\frac{S_n^v}{\sqrt{n}} - \frac{S_n^u}{\rho_n\sqrt{n}}\right] = \frac{(1 - \rho_n)\sqrt{n}}{\rho_n} \to -\beta = E\left[Y(1)\right] \tag{28}$$

as $n \to \infty$. □

3 The special case of a $GI/GI/1$ base queue

We now restrict attention to $GI/GI/1$ queues in the single-sequence framework. The following follows quite directly from Theorems 1 and 2 and Corollary 1.

Theorem 3 ($GI/GI/1$ queues) *Consider a sequence of $GI_n/GI_n/1$ models constructed from a single base $GI/GI/1$ model as indicated above*; *i.e.*, *suppose that $\{U_k : k \geq 1\}$ and $\{V_k : k \geq 0\}$ are sequences of i.i.d. random variables with mean 1 and variances $\sigma_u^2 \equiv \mathrm{Var}(U_k) < \infty$, $\sigma_v^2 \equiv \mathrm{Var}(V_k) < \infty$. Suppose that $\{(U_k, V_{k-1}) : k \geq 1\}$ is independent of $\{W_{n,0} : n \geq 1\}$ and $\sqrt{n}W_{n,0} \Rightarrow \tilde{W}(0)$ in $\mathbb{R}$ as $n \to \infty$. If $(1-\rho_n)\sqrt{n} \to \beta$, $0 < \beta < \infty$ as $n \to \infty$, as in (1), then the conditions and conclusions of Theorem 1 and Corollary 1(a) and (b) are satisfied. If, instead, we focus on the stationary waiting time processes, which is achieved by changing the initial conditions, then the conditions and conclusions of Theorem 2 hold with the limiting sequence $\{\tilde{W}(j) : j \geq 0\}$ being a reflected stationary Gaussian random walk; i.e., with $\{Y(j)\}$ being a sequence of i.i.d. random variables with $Y(k) \stackrel{\mathrm{d}}{=} N(-\beta, \sigma^2)$. If, instead, $(1-\rho_n)\sqrt{n} \to \infty$, then the stationary waiting times are asymptotically negligible, i.e., $\tilde{W}_n^c(j) \Rightarrow 0$; if, instead, $(1-\rho_n)\sqrt{n} \to 0$, then $\tilde{W}_n^c(j) \Rightarrow \infty$.*

Proof First, the GI assumptions here directly imply that the FCLT for the sequence $\{(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v) : n \geq 1\}$ in (23), by virtue of Donsker's theorem [3, 26], which in turn implies the FCLT for the partial sums in (9). Hence, the conditions of Theorems 1 and 2 and Corollary 1 are satisfied for the associated double sequence. Finally, the last statements follow from the main result, because $\tilde{W}_n^c(j)$ is clearly stochastically increasing in ρ_n; for example, see [16]. □

Remark 1 Theorem 3 only uses the GI framework to justify the extra conditions of Corollary 1. As usual with HT limits, we could have the same FCLT in (25) for various dependent sequences; see Sect. 4.4 of [26] for examples. Even in the GI case, Theorem 3 does not quite imply Theorem 1 of [13] as stated, because that stated result is in the more general double-sequence framework of Sect. 2. However, an extra condition, such as the Lindeberg condition or the Lyapounov condition, is needed in [13] in order for the CLT used in the proof of Theorem 1 in [13] to be valid. Theorem 1 of [13] is fine if this condition is added. Theorem 3 does extend to the more general double-sequence framework of Sect. 2 if such a condition is added here too; then it implies the $GI_n/D/1$ result in Theorem 1 of [13].

Remark 2 The limit $\tilde{W}(j)$ in (27) and Theorem 3 is the maximum of a Gaussian random walk. For more on this limit, see [10, 11] and references therein. For refinements to the approximation provided in the $E_n/D/1$ (or $M/D/n$) case, see [12].

In applications, it is natural to use performance measures such as the probability of delay $P(W > 0)$ and the mean delay $E[W]$. It is significant that convergence of these quantities is not yet covered by the results above. First, observe that convergence in distribution $Z_n \Rightarrow Z$ as $n \to \infty$ for nonnegative random variables does not imply that $P(Z_n > 0) \to P(Z > 0)$ as $n \to \infty$; for example, let $P(Z_n = 1/n, n \geq 1, Z = 0) = 1$.

Second, for the mean values, we need uniform integrability as well as convergence in distribution. Nevertheless, we can establish these important additional results. For these results, we do use the *GI* property more critically. Our next two results extend Corollary 1 of [13] from $GI/D/n \equiv GI_n/D/1$ to $GI_n/GI_n/1$, and provide alternative proofs. Let the standard normal cdf be $\Phi(x) \equiv P(N(0,1) \le x)$ and let ϕ be the associated density.

Theorem 4 (Stationary delay probabilities in the $GI_n/GI_n/1$ models) *In the setting of Theorem 3,*

$$P\left(W^c_{n,\infty} > 0\right) = e^{-\sum_{k=1}^{\infty}(1/k)P(S^c_{n,k}>0)} \to e^{-\sum_{k=1}^{\infty}(1/k)\Phi(-\sqrt{k}\beta/\sigma)} \equiv \alpha \tag{29}$$

as $n \to \infty$ for $0 < \alpha < 1$ if and only if

$$(1-\rho_n)\sqrt{n} \to \beta, \quad 0 < \beta < \infty, \tag{30}$$

where $\alpha \equiv \alpha(\beta)$ is given on the right in (29).

Proof For most of the proof we can follow [13]. First, as noted in [13], it is easy to see that $0 < \alpha < 1$ in (29) using basic properties of the normal distribution. Next assume that (30) holds. Reasoning as in Theorem 1 and Corollary 1, for each fixed k, we can apply the CLT and LLN to get

$$\begin{aligned} P\left(S^c_{n,k} > 0\right) &= P\left(\frac{S^v_{nk}}{n} - \frac{S^u_{nk}}{n\rho_n} > 0\right) = P\left(\frac{S^v_{nk} - S^u_{nk}}{\sqrt{n}} > \frac{S^u_{nk}}{n}\left(\beta + o(1)\right)\right) \\ &\to P\left(N\left(0, k\sigma^2\right) > k\beta\right) = \Phi(-\sqrt{k}\beta/\sigma) \quad \text{as } n \to \infty. \end{aligned} \tag{31}$$

The key technical step is to show that the infinite series in k, as a function of n, on the left in (29), converges uniformly in n. (Such a step seems to be needed in [13]; as it stands, the proof of (2) in [13] seems incomplete; for example, see the example before the theorem.) In particular, we will show that it is possible to choose a constant c and an integer n_0 such that $P(S^c_{n,k} > 0) < c/k$ for all $k \ge 1$ and all $n \ge n_0$, which will imply that the terms in the series are bounded above by c/k^2 for $n \ge n_0$. Choose n_0 so that $1/2 \le \rho_n \le 1$ and $(1-\rho_n)\sqrt{n} \ge \beta/2$ for all $n \ge n_0$, which can be done because of (30). Then we can apply Chebychev's theorem to get:

$$\begin{aligned} P\left(S^c_{n,k} > 0\right) &= P\left(\frac{S^c_{n,k} - E[S^c_{n,k}]}{\sqrt{\mathrm{Var}(S^c_{n,k})}} > \frac{-E[S^c_{n,k}]}{\sqrt{\mathrm{Var}(S^c_{n,k})}}\right) \\ &\le \frac{\mathrm{Var}(S^c_{n,k})}{(E[S^c_{n,k}])^2} \le \frac{8\sigma^2}{k\beta^2} \quad \text{for all } n \ge n_0 \text{ and } k \ge 1, \end{aligned} \tag{32}$$

because $\mathrm{Var}(S^c_{n,k}) \le 2k\sigma^2/n$ for all $n \ge n_0$ and k, $(E[S^c_{n,k}])^2 = [k(\rho_n^{-1} - 1)]^2 \ge [(k/\sqrt{n})(1-\rho_n)\sqrt{n}]^2$ and $[(1-\rho_n)\sqrt{n}]^2 \ge \beta^2/4$ for all $n \ge n_0$. As a consequence, there is a constant c such that the terms of the sum on the left are bounded by c/k^2 for all n and k. Hence, the tails of the infinite series are uniformly negligible. Finally,

to obtain the "only if" part, we exploit the monotonicity of $P(W^c_{n,\infty} > 0)$ in ρ_n. We can use bounding arguments from the established "if" part for high and low β. Hence there cannot be a subsequence such that $(1-\rho_n)\sqrt{n}$ converges to either 0 or ∞. Hence, the sequence must have a convergent subsequence, with all limits β' satisfying $0 < \beta' < \infty$. However, the limit along any such subsequence is determined by (29). □

We now establish a limit for the mean waiting times.

Theorem 5 (Mean stationary waiting times in the $GI_n/GI_n/1$ models) *In the setting of Theorem 3, if (30) holds, $E[U^2] < \infty$ and $E[V^2] < \infty$, then*

$$E\big[\tilde{W}^c_{n,\infty}\big] = \sum_{k=1}^{\infty} \frac{\sqrt{n}E[(S^c_{n,k})^+]}{k} \to \sum_{k=1}^{\infty} \frac{E[N(-k\beta, k\sigma^2)^+]}{k} = E\big[\tilde{W}^c_{\infty}\big]. \tag{33}$$

The limit can be expressed as

$$E\big[\tilde{W}^c_{\infty}\big] = \sum_{k=1}^{\infty} \frac{\beta\eta(\beta\sqrt{k}/\sigma)}{k} < \infty, \quad \text{where } \eta(x) \equiv \frac{\phi(x)}{x} - \Phi(-x). \tag{34}$$

Proof First, the series expressions on the left and right in (33) are established representations for the expected value of the maximum of a random walk with negative drift; see Proposition VIII.4.5 of [2]. Formula (34) follows from writing

$$\begin{aligned} &E\big[N\big(-k\beta, k\sigma^2\big)^+\big] \\ &\quad = E\big[N\big(-k\beta, k\sigma^2\big)\big|N\big(-k\beta, k\sigma^2\big) > 0\big]P\big(N\big(-k\beta, k\sigma^2\big) > 0\big). \end{aligned}$$

The (known) formula for the conditional mean is given in (18.29) of [6]. The task is to prove convergence. The convergence would follow from the limit in (16) via Theorem 3 plus uniform integrability, but it remains to establish uniform integrability. We establish convergence of the moments directly (and consequently the uniform integrability) by showing that the series on the left in (33) converges to the series on the right as $n \to \infty$. We do so by exploiting the tail integral representation of the mean and the (Lebesgue) dominated convergence theorem. Since

$$E\big[\tilde{W}^c_{n,\infty}\big] = \sum_{k=1}^{\infty} \int_0^{\infty} \frac{P(\sqrt{n}S^c_{n,k} > x)\,dx}{k}, \tag{35}$$

and a similar expression exists for the limit, it suffices to show that

$$\sum_{k=1}^{\infty} \int_0^{\infty} \frac{P(\sqrt{n}S^c_{n,k} > x)\,dx}{k} \to \sum_{k=1}^{\infty} \frac{P(N(-k\beta, k\sigma^2) > x)\,dx}{k} \tag{36}$$

as $n \to \infty$. First, by the CLT, we have $P(\sqrt{n}S^c_{n,k} > x) \to P(N(-k\beta, k\sigma^2) > x)$ for all x, because of the representation

$$P\left(\sqrt{n}S^c_{n,k} > x\right) = P\left(\frac{S^v_{nk} - S^u_{nk}}{\sqrt{n}} > x + \frac{S^u_{nk}(1-\rho_n)\sqrt{n}}{n\rho_n}\right). \tag{37}$$

(The final term on the right in the probability converges to βk.)

The proof is completed by bounding $P(\sqrt{n}S^c_{n,k} > x)$ above by a quantity for which the iterated sum and integral is bounded uniformly in n, allowing us to apply the dominated convergence theorem. For this step, we apply an inequality for partial sums in Corollary 1.11 of [17]: For a sum S_n of n i.i.d. random variables distributed as X, having mean 0 and finite variance σ^2,

$$P(S_n > y) \le C(r)\left(\frac{n\sigma^2}{y^2}\right)^r + nP(X > y/r) \tag{38}$$

for all $y > 0$ and $r > 0$, where we use C for a generic constant independent of n, here depending on r. We will be applying (38) with $r > 1$. We consider two cases, first focusing on the integrals from 0 to k and then focusing on the integrals from k to ∞. As an alternative to (37), we write

$$P\left(\sqrt{n}S^c_{n,k} > x\right) = P\left(S_{n,nk} > x\rho_n\sqrt{n} + (1-\rho_n)nk\right), \tag{39}$$

where $S_{n,nk}$ is the partial sum of nk i.i.d. random variables distributed as $X_{n,j} = \rho_n V_j - U_j + 1 - \rho_n$. (There is still dependence on n here, but we will show how it can be controlled.) Applying inequality (38) with representation (39), we get

$$\begin{aligned}
&\int_0^k P\left(S_{n,nk} > x\rho_n\sqrt{n} + (1-\rho_n)nk\right)dx \\
&\quad\le \int_0^k P\left(S_{n,nk} > (1-\rho_n)nk\right)dx \\
&\quad\le kP\left(S_{n,nk} > (1-\rho_n)nk\right) \\
&\quad\le C(r)\left(\sqrt{n}(1-\rho_n)\right)^{-2r}k^{1-r} + nk^2P\left(X_{n,1} > (1-\rho_n)nk/r\right).
\end{aligned} \tag{40}$$

From (40), we get the bound

$$\begin{aligned}
&\sum_{k=1}^{\infty}\frac{1}{k}\int_0^k P\left(S_{n,nk} > x\rho_n\sqrt{n} + (1-\rho_n)nk\right)dx \\
&\qquad\le C_1(r,\beta)\sum_{k=1}^{\infty}k^{-r} + \sum_{k=1}^{\infty}nkP(X_{n,1} > \beta\sqrt{n}k/r).
\end{aligned} \tag{41}$$

The first sum in (41) is finite for all $r > 1$, independent of n, so it remains to treat the second sum, showing uniformity in n. To estimate the second sum, we use the simple

inequality

$$\int_a^b x\,P(X > x)\,dx \geq (b-a)a\,P(X > b) \quad \text{for all } 0 < a < b. \tag{42}$$

We apply (42) to write, for $k \geq 2$,

$$nk\,P(X_{n,1} > \beta\sqrt{n}k/r) \leq C_2(r,\beta) \int_{\sqrt{n}\beta(k-1)/r}^{\sqrt{n}\beta k/r} x\,P(X_{n,1} > x)\,dx, \tag{43}$$

replacing $k/(k-1)$ by its upper bound 2 in the constant $C(r,\beta)$. We treat the term for $k = 1$ directly by Markov's inequality:

$$n\,P(X_{n,1} > \beta\sqrt{n}/r) \leq \frac{E[X_{n,1}^2]}{(\beta\sqrt{n}/r)^2} \leq C_1 < \infty, \tag{44}$$

uniformly in n. Hence,

$$\begin{aligned} \sum_{k=1}^{\infty} nk\,P(X_{n,1} > \beta\sqrt{n}k/r) &\leq C_1 + C_2(r,\beta) \int_0^{\infty} x\,P(X_{n,1} > x)\,dx \\ &\leq C_2 < \infty, \end{aligned} \tag{45}$$

where C_2 is independent of n.

We now turn to the integrals over the tail intervals $[k,\infty)$. From (38) and (39) again, we have

$$\begin{aligned} &\int_k^{\infty} P\big(S_{n,nk} > x\rho_n\sqrt{n} + (1-\rho_n)nk\big)\,dx \\ &\quad \leq \int_k^{\infty} P(S_{n,nk} > x\rho_n\sqrt{n})\,dx \\ &\quad \leq C(r)\rho_n^{-2r} \int_k^{\infty} x^{-2r}\,dx + nk \int_k^{\infty} P(X_{n,1} > x\rho_n\sqrt{n}/r)\,dx. \end{aligned} \tag{46}$$

Hence,

$$\begin{aligned} &\sum_{k=1}^{\infty} \frac{1}{k} \int_k^{\infty} P\big(S_{n,nk} > x\rho_n\sqrt{n} + (1-\rho_n)nk\big)\,dx \\ &\quad \leq C(r) \sum_{k=1}^{\infty} k^{-(2r-1)} + n \sum_{k=1}^{\infty} \int_k^{\infty} P(X_{n,1} > \beta\sqrt{n}k/r)\,dx. \end{aligned} \tag{47}$$

It now only remains to treat the second term in (47). To do so, we apply Tonelli's theorem to write

$$n \sum_{k=1}^{\infty} \int_k^{\infty} P(X_{n,1} > \beta\sqrt{n}k/r)\,dx$$

$$\le n \int_1^\infty x P(X_{n,1} > \beta\sqrt{n}k/r)\, dx$$

$$\le \rho_n^{-2} r^2 \int_0^\infty x P(X_{n,1} \ge x)\, dx, \tag{48}$$

which is uniformly bounded in n by virtue of the second moment condition. □

Remark 3 The limit is the same as for the $GI_n/D/1$ model stated in Corollary 1 of [13]. The proof in [13] relies on uniform integrability, but contrary to the claim on p. 58 of [13], for uniform integrability, given convergence in distribution of non-negative random variables, it does not suffice to show that the means are bounded above; we need some higher moment bounded above. The convergence in distribution and the convergence of moments we have established directly do imply uniform integrability.

4 Heavy traffic limit in case (2): $(1-\rho)n \to \beta$

We first review the stochastic process limit that holds with condition (2). For that purpose, we introduce the following random elements of D terms of random elements of D. For that purpose, let

$$\begin{aligned}
\mathbf{S}_n^{c,u}(t) &\equiv S_{n,\lfloor nt\rfloor}^{c,u} - \lfloor nt\rfloor = \frac{S_{n\lfloor nt\rfloor}^u}{\rho_n n} - \lfloor nt\rfloor, \\
\mathbf{S}_n^{c,v}(t) &\equiv S_{n,\lfloor nt\rfloor}^{c,v} - \lfloor nt\rfloor = \frac{S_{n,n\lfloor nt\rfloor}^v - n\lfloor nt\rfloor}{n} = \frac{S_{n\lfloor nt\rfloor}^v - n\lfloor nt\rfloor}{n}, \\
\mathbf{S}_n^c(t) &\equiv S_{n,\lfloor nt\rfloor}^c = \frac{S_{n,n\lfloor nt\rfloor}^v}{n} - \frac{S_{n,n\lfloor nt\rfloor}^u}{\rho_n n} \\
&= \big(\mathbf{S}_n^{c,v} - \mathbf{S}_n^{c,u}\big)(t), \\
\mathbf{W}_n^c(t) &\equiv W_{n,\lfloor nt\rfloor}^c = \frac{W_{n,n\lfloor nt\rfloor}}{n} = \phi\big(\mathbf{W}_n^c(0) + \mathbf{S}_n^c\big)(t).
\end{aligned} \tag{49}$$

Let $\phi : D \to D$ be the one-dimensional reflection map, defined by

$$\phi(x)(t) \equiv x(t) - \inf_{0\le s\le t} \big\{x(s) \wedge 0\big\}, \quad t \ge 0; \tag{50}$$

see Sects. 3.5 and 13.5 of [26]. The following is Theorem 4.2 of [21].

Theorem 6 (HT stochastic-process limit from [21]) *Consider a sequence of $G_n/G_n/1$ models associated with a single base $G/G/1$ model satisfying*

$$\big(W_{n,0}, \hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v\big) \quad \Rightarrow \quad \big(\mathbf{W}^c(0), \hat{L}^u, \hat{L}^v\big) \quad \text{in } \mathbb{R}\times D^2 \tag{51}$$

for $(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v)$ in (23), where $P((\hat{L}^u, \hat{L}^v) \in C^2) = 1$. If

$$(1-\rho_n)n \to \beta, \quad 0 < \beta < \infty, \quad \text{as } n \to \infty, \tag{52}$$

as in case (2), *then, as* $n \to \infty$,

$$\left(\mathbf{S}_n^c(0), \mathbf{S}_n^{c,u}, \mathbf{S}_n^{c,v}, \mathbf{S}_n^c, \mathbf{W}_n^c\right) \quad \Rightarrow \quad \left(\mathbf{W}^c(0), \hat{L}^u + \beta e, \hat{L}^v, L, \mathbf{W}^c\right) \quad \text{in } \mathbb{R} \times D^4, \tag{53}$$

where $\mathbf{W}^c \equiv \phi(\mathbf{W}^c(0) + L)$, ϕ *is given in* (50) *and* $L \equiv \hat{L}^v - \hat{L}^u - \beta e$. *If, in addition,* (26) *holds, where* $\mathbf{W}^c(0)$, B_u *and* B_v *are mutually independent, and* B_u *and* B_v *are standard Brownian motions, then* $L \stackrel{\mathrm{d}}{=} \sigma B - \beta e$ *and* $\mathbf{W}^c(0)$ *is independent of* B *where* B *is a standard BM and* $\sigma^2 = \sigma_u^2 + \sigma_v^2$.

We now establish a corresponding limit for stationary waiting times limit in the $GI/GI/1$ setting. For the base $GI/GI/1$ model, HT limits for stationary waiting times are given in Sect. X.7 of [2]. The following result for the $GI_n/GI_n/1$ models extends the result for the special case of $E_n/E_n/1$ Erlang queues (having $M/M/1$ base model) in Sect. 3 of [1].

Theorem 7 (HT limit for unscaled stationary waiting times) *Consider a base* $GI/GI/1$ *queueing model with* $E[U_k] = E[V_k] = 1$, $\mathrm{Var}(U_k) \equiv \sigma_u^2$ *and* $\mathrm{Var}(V_k) \equiv \sigma_v^2$, *where* $0 < \sigma^2 \equiv \sigma_u^2 + \sigma_v^2 < \infty$. *Consider the associated sequence of* $GI_n/GI_n/1$ *models with scaling in* (24) *and* $(1 - \rho_n)n \to \beta$, $0 < \beta < \infty$, *constructed according to the cyclic thinning for each* n *as in* (7). *Let* $W_n^c(0)$ *be the independent initial wait in model* n *and assume that* $W_n^c(0) \Rightarrow \mathbf{W}^c(0)$ *Then the conditions of Theorem* 6 *are satisfied, including* (26), *so that* $\mathbf{W}_n^c \Rightarrow \phi(\mathbf{W}^c(0) + \sigma B - \beta e)$ *in* D *as* $n \to \infty$. *In addition, for each* $n \geq 1$, *there exists a limiting stationary waiting time* $W_{n,\infty}^c$ *in the* $GI_n/GI_n/1$ *model having finite mean* $E[W_{n,\infty}^c]$. *As* $n \to \infty$,

$$W_{n,\infty}^c \quad \Rightarrow \quad W_\infty^c, \tag{54}$$

where W_∞^c *is an exponential random variable with* $E[W_\infty^c] = \sigma^2/2\beta$ *and* $\phi(\mathbf{W}^c(0) + \sigma B - \beta e)(t) \Rightarrow W_\infty^c$ *as* $t \to \infty$. *Finally, the sequence* $\{W_{n,\infty}^c : n \geq 1\}$ *is uniformly integrable, and*

$$E\left[W_{n,\infty}^c\right] \to E\left[W_\infty^c\right] = \frac{\sigma^2}{2\beta} \quad \text{as } n \to \infty. \tag{55}$$

Proof For each $n \geq 1$, the $GI_n/GI_n/1$ model is itself a $GI/GI/1$ queue with finite mean service time and interarrival time, and traffic intensity $\rho_n < 1$. It is well known that a steady state waiting time exists and coincides with the overall maximum of the partial sums; see Chap. 10 of [2]. So the stationary waiting time random variable $W_{n,\infty}^c$ is well defined for each n. By Kingman's bound in Theorem 2 of [14],

$$\begin{aligned} E\left[W_{n,\infty}^c\right] &\leq \frac{\mathrm{Var}(S_n^u/n\rho_n) + \mathrm{Var}(S_n^v/n)}{2(1-\rho_n)} \\ &= \frac{(\sigma_u^2/n\rho_n^2) + (\sigma_v^2/n)}{2(1-\rho_n)} \to \frac{(\sigma_u^2 + \sigma_v^2)}{2\beta} \end{aligned} \tag{56}$$

as $n \to \infty$ Hence the sequence of stationary waiting times $\{W_{n,\infty}^c : n \geq 1\}$ is stochastically bounded or tight. Consequently, there is a converging subsequence. Now con-

sider the associated sequence of stationary waiting time processes, obtained by letting the initial waiting time $W_{n,0}^c$ be distributed as $W_{n,\infty}^c$ for each n with indices in the convergent subsequence. We obtain a stationary sequence by that simple initialization because the sequence of waiting times in the $GI_n/GI_n/1$ model is a Markov chain. Since these initial distributions converge as $n \to \infty$, the full processes satisfy the HT limit theorem in Theorem 6. Since each of these converging processes is stationary, the limiting reflecting Brownian motion must also be stationary, but it has the unique stationary exponential distribution of W_∞^c. Hence the limit of the convergent subsequence must in fact be W_∞^c. Since every convergent subsequence must have this same limit, the entire sequence must converge to that same limit; i.e., we have shown (54). By Fatou's lemma, $E[W_\infty^c] \le \liminf_{n\to\infty} E[W_{n,\infty}^c]$. However, by (56), $\limsup_{n\to\infty} E[W_{n,\infty}^c] \le E[W_\infty^c]$. Hence, $E[W_{n,\infty}^c] \to E[W_\infty^c]$ as $n \to \infty$. Since these are nonnegative random variables, that implies that the sequence $\{W_{n,\infty} : n \ge 1\}$ is uniformly integrable. □

Without the *GI* independence conditions, HT limits for stationary waiting times are much more difficult to prove. For a single $G/G/1$ model, HT limits were established in [22, 23].

5 Evaluating the approximations for stationary waiting times

We suggest using the exact results for the stationary distributions of the $GI_n/GI_n/1$ model in Theorems 4, 5 and 7 as the basis for approximations for $G_n/G_n/1$ models more generally, provided that these limits hold in the Brownian case, for example, the conditions of Corollary 1(b) hold. That is supported by the process limits in Theorems 1 and 6, which in fact hold more generally than *GI*. We would then use the same approximations derived for $GI_n/GI_n/1$, with the exception that the parameters σ_u^2 and σ_v^2 would be obtained from the FCLT, and would not necessarily be the variances of the individual interarrival times and service times.

However, even for $GI_n/GI_n/1$ models with high ρ_n and n, we should not expect too much from these approximations, because these $GI_n/GI_n/1$ models are highly sensitive to small perturbations in the model. Small changes in the interarrival times or service times significantly alter the performance. For example, an increase of the traffic intensity by only $1 - \rho_n$ will make the model unstable.

The problematic nature of the approximations is well illustrated by the $E_{100}/D/1$ model with traffic intensity $\rho = 0.99$. Sample paths from simulation runs for this model were shown in Fig. 1 of [21]. This model can be regarded as the nth term in the sequence of $E_n/D/1$ models in Theorems 6 and 7 with $n = 100$ and $(1 - \rho_n)n = \beta = 1$. Those theorems suggest approximating the steady-state waiting-time distribution by an exponential random variable with mean $1/2$.

Interestingly, that is the identical approximation we obtain from the conventional HT limit, assuming that n is fixed and $\rho \uparrow 1$. The conventional HT approximation is again exponential with the mean

$$E[W] \approx \frac{E[V](c_u^2 + c_v^2)}{2(1-\rho)} = \frac{(\sigma_u^2 + \sigma_v^2)}{2(1-\rho)} = \frac{(0.01)}{2(0.01)} = \frac{1}{2}. \tag{57}$$

Table 1 A comparison of the approximations with exact numerical values computed using the numerical algorithm from [1] for three steady-state performance measures in the Erlang $E_k/E_l/1$ model for $k = 10^j$, $1 \le j \le 4$, in case (1) (approx($\sqrt{n}$)) and case (2) (approx(n))

$E_k/E_k/1$ queue with mean service time 1 and $\rho_k \equiv 1-(1/k)$				
	$k=10$	$k=100$	$k=1000$	$k=10{,}000$
$P(W>0)$ exact	0.7102	0.9036	0.9688	0.9900
Approx ($\sqrt{n}$)	0.6279	0.8666	0.9561	0.9843
Approx (n)	1.0000	1.0000	1.0000	1.0000
$E[W\|W>0]$ exact	1.054	1.0175	1.0055	1.0018
Approx ($\sqrt{n}$)	1.216	1.0265	1.0189	1.0076
Approx (n)	1.000	1.0000	1.0000	1.0000
$E[W]$ exact	0.7484	0.9195	0.97417	0.99018
Approx ($\sqrt{n}$)	0.7635	0.9201	0.9742	0.99018
Approx (n)	1.0000	1.0000	1.0000	1.00000

It is common to refine the approximation in (57) by adding ρ in the numerator, as in (44) of [24], but in this case that only changes the approximate mean to 0.495. The cancelation of the two 0.01 terms in the numerator and denominator of (57) is the basis for the high sensitivity to model perturbations.

We employed simulation to evaluate the accuracy of the approximations. A simulation run for a time interval of length 500,000, divided into 10 batches, yields an estimated 95% confidence interval for the mean of 0.4354 ± 0.0226.

The Kraemer-and-Langenbach-Belz (KL) approximation for the mean waiting time in a $GI/GI/1$ queue, given in (45) of [24], is known to perform remarkably well, but it too breaks down in this challenging $E_{100}/D/1$ model. It yields the approximate value of $E[W] \approx 0.256$. The basic HT approximation 0.500, the refinement 0.495 and the KL approximation 0.256 thus have relative errors ($|approx - exact|/exact$) of 14.8%, 13.7% and 41.2%, respectively, but at least all are of the right order. A simple $M/M/1$ approximation, obtained by using exponential random variables with the same means as the interarrival time and service time random variables, would be $E[W] \approx 99.0$, which overestimates by a factor of 227 (off by two orders of magnitude).

In fact, we have *two* candidate approximations for the stationary waiting time distribution provided by the HT limits in the two cases (1) and (2). We investigated the approximations for the mean and the probability of delay by comparing the HT approximations to exact numerical results from [1] and from simulation estimates for the $E_k/E_l/1$ high-order Erlang model.

Table 1 compares the approximations for the $E_k/E_k/1$ model with $\rho_k = 1-(1/k)$ to the exact numerical results in Table 1 of [1]. The scaling here is naturally in case (2), because $(1-\rho_k)k = 1$ for all k, but we considered both cases (1) and (2). In case (1) we evaluated formulas (29) for the delay probability and (34) for the mean using Matlab. In case (2), the approximate delay probability is 1, while the mean wait is given in (55).

Table 2 A comparison of approximations with simulation for high-order Erlang $E_k/E_l/1$ queues. The case D is represented as k or l being ∞; $\beta(n^p) \equiv (1-\rho)n^p$ and $ap(n^p)$ is the approximation based on scaling by n^p, with $p = 0.5$ in (1) and $p = 1$ in (2)

Steady-state performance in the $E_k/E_l/1$ queue with mean service time 1									
Parameters					$E[W\|W>0]$			$P(W>0)$	
k	l	ρ	$\beta(n)$ $(1-\rho)n$	$\beta(\sqrt{n})$ $(1-\rho)\sqrt{n}$	SIM	ap($\sqrt{n}$) (55)	ap(n)	SIM	ap($\sqrt{n}$) (29)
100	100	0.99	1.0	0.1	0.9991 ±.055	1.012	1.000	0.9038 ±.0048	0.9037
		0.90	10.0	1.0	0.1199 ±.0024	0.1187	0.1000	0.3119 ±.0025	0.3345
100	∞	0.99	1.0	0.1	0.5039 ±.026	0.513	0.500	0.8688 ±.0060	0.866
		0.90	10.0	1.0	0.0632 ±.0010	0.0632	0.050	0.1945 ±.0010	0.1995
∞	100	0.99	1.0	0.1	0.4998 ±.022	0.513	0.500	0.8597 ±.0037	0.866
		0.90	10.0	1.0	0.00655 ±.0013	0.0632	0.050	0.1658 ±.0013	0.1995

Even though the scaling puts these examples naturally in the domain of case (2), we find that the approximations based on case (1), referred to as approx ($\sqrt{n}$), consistently perform better than the approximations based on case (2), referred to as approx (n). Case (2) becomes competitive and even preferred to case (1) when we focus on the expected conditional delay, given that the wait is positive $E[W|W>0]$. For case (2), the approximation for the mean and the conditional mean agree, but they do not for case (1). Overall, we find these approximations remarkably effective, given the huge error using a simple $M/M/1$ approximation. For example, for $k = 100$, the $M/M/1$ approximation yields $P(W>0) = 0.99$ and $E[W] = 99$.

We made additional comparisons for the $E_k/E_l/1$ high-order Erlang model, where k and l are either 100 or ∞, with ∞ corresponding to deterministic (D). We considered the three models $E_{100}/E_{100}/1$, $E_{100}/D/1$ and $D/E_{100}/1$ with two different traffic intensities: $\rho = 0.99$ and $\rho = 0.90$. For $\rho = 0.99$, we have $(1-\rho)k = 1$, but $(1-\rho)\sqrt{k} = 0.1$; for $\rho = 0.90$, we have $(1-\rho)k = 10$, but $(1-\rho)\sqrt{k} = 1.0$. Thus, it is natural to regard $\rho = 0.99$ as case (2) and $\rho = 0.90$ as case (1). For this experiment we used simulation to estimate the exact values. We based our simulation estimates on 10 independent replications of 50,000 arrivals. The results are shown in Table 2.

The case $k = l = 100$ and $\rho = 0.99$ in Table 2 repeats the case $k = 100$ in Table 1, so that the simulation and numerical algorithm validate each other. In Table 2 we see that each approximation method works better for the conditional wait $E[W|W>0]$ in the expected case with $\beta = 1.0$. Overall, the approximations are impressively accurate.

Table 3 contains additional comparisons with simulations. Two cases with $k = 400$ and $k = 1600$ are chosen to be in case (1) with $(1-\rho)\sqrt{k} = \beta = 1$ but have larger

Table 3 A comparison of the approximations with exact numerical values computed using the numerical algorithm from [1] for three steady-state performance measures in the Erlang $E_k/E_l/1$ model for $k = 10^j$, $1 \le j \le 4$, in case (1) (approx($\sqrt{n}$)) and case (2) (approx(n))

$E_k/D/1$ queue with mean service time 1			
	$k = 400$	$k = 1600$	$k = 100$
	$\rho = 0.95$	$\rho = 0.975$	$\rho = 0.995$
$P(W > 0)$ simul.	0.19652	0.19927	0.93406
	±4.92e–06	±4.29e–06	±2.85e–05
Approx ($\sqrt{n}$)	0.1995	0.1995	0.9313
Approx (n)	1.0000	1.0000	1.0000
$E[W\|W > 0]$ simul.	0.0316	0.0158	0.9659
	±1.0e–04	±5.3e–05	±0.06
Approx ($\sqrt{n}$)	0.03168	0.01583	1.0872
Approx (n)	0.025000	0.0125	1.0000
$E[W]$ simul.	0.00621	0.00316	0.90217
	±9.7e–05	±5.3e–05	±0.06
Approx ($\sqrt{n}$)	0.00632	0.00316	1.0125
Approx (n)	0.0250	0.0125	1.0000

sample size. As expected, the Gaussian random walk approximations in case (1) are clearly superior. The other case with $k = 100$ is chosen to have even higher traffic intensity ρ, so that it is even more clearly in case (2). In these more extreme cases, the approximation we expect to perform better clearly does so for the conditional mean wait and even for the unconditional mean wait. The performance of the approximation for the probability of delay in case (1) is consistently good.

In conclusion, we remark that in our numerical results we found that the approximate value for the mean wait in (34) in case (1) consistently was an upper bound for the exact mean waiting time. Thus, we conjecture that to be valid.

6 Stationary point processes with cyclic thinning

We now establish stationarity properties of a single point process modified by cyclic thinning, using the notation in [20]; see [20] for background.

6.1 Preliminaries

Suppose that $\psi \equiv \{t_i : i \ge 0\}$ is a point-stationary ergodic simple point process with counting process $N(t) \equiv \max\{n \ge 1 : t_n \le t\}$, $t \ge 0$, $N(0) \equiv 0$. This means that the interarrival time sequence $T_i \equiv t_i - t_{i-1}$, $i \ge 1$ is a stationary ergodic sequence and $P(t_0 = 0) = 1$. We assume that $0 < 1/\lambda \equiv E(T_i) < \infty$; the arrival rate is λ.

It follows that for any $q \ge 2$ an integer, the thinned point process $\psi_q \equiv \{t_i(q) : i \ge 0\} = \{t_{iq} : i \ge 0\}$ is point-stationary too (with respect to its own points), since

its interarrival times $T_i(q) = \sum_{m=(i-1)q+1}^{iq} T_m$, $i \geq 1$, form a stationary sequence, and $t_0(q) = t_0 = 0$, w.p.1. Let $N_q(t) \equiv \max\{k \geq 1 : t_{kq} \leq t\}$ be the counting process for ψ_q. Let $\theta_s\psi$ denote the shifted point process from time s onwards with s relabeled as the origin, i.e., its points are given by $t_k(s) \equiv t_{N(s)+k} - s$, $k \geq 1$). Let $\theta(i) \equiv \theta_{t_i(q)}$, the shifts using $s = t_i(q)$. Then, the pair (ψ, ψ_q) is *jointly* invariant in distribution under these shifts; $(\theta(i)\psi, \theta(i)\psi_q)$ jointly has the same distribution for all $i \geq 0$. (However, they are *not* jointly invariant under the shifts using the single points $s = t_i$, because ψ_q is not invariant under them.)

We refer to ψ_q as a *q-cyclic thinning* of ψ. Given one of its interarrival times $T_i(q)$, we refer to the q variables T_m being added as its *q phases*. This is in keeping in the spirit of the special case when ψ is a Poisson process at rate λ and ψ_q is thus a renewal process with an Erlang(q, λ) interarrival time distribution with q exponential phases each at rate λ. The arrival rate of ψ_q is λ/q. Let $J(t) \equiv N(t) \bmod q$ be the number of completed phases at time $t \geq 0$ for the current interarrival time of ψ_q; $J(t) \in \{0, 1, 2, \ldots, q-1\}$. For $t \in [t_{(i-1)q}, t_{iq})$, during the ith interarrival time, $T_i(q)$, we have $J(t) = 0$ for $t_{(i-1)q} \leq t < t_{(i-1)q+1}$; $J(t) = 1$ for $t_{(i-1)q+1} \leq t < t_{(i-1)q+2}$; $\ldots$, $J(t) = q-1$ for $t_{iq-1} \leq t < t_{iq}$.

6.2 Marked point process approach

Given the framework of the previous section, we wish to proceed to jointly construct time-stationary versions of (ψ, ψ_q) and $\{J(t) : t \geq 0\}$. To do so, we will use a marked point process approach. Starting with our point-stationary $\psi = \{t_i : i \geq 0\}$, we add marks $\{M_i : i \geq 0\}$ to obtain a random marked point process (rmpp), $\psi_M = \{(t_i, M_i) : i \geq 0\}$ where $M_i = 1$ iff $i = 0 \bmod q$; 0 otherwise. Thus (starting with $t_0 = 0$), every qth point has a mark of $M = 1$, all others a mark of $M = 0$. Thus, ψ_q is the thinning of the points of ψ_M obtained by choosing only points for which $M_i = 1$. Let

$$\begin{aligned} t_0(q) &\equiv \min\{t_k : M_k = 1\}, \\ t_i(q) &\equiv \min\big\{t_k : t_k > t_{i-1}(q), M_k = 1\big\}, \quad i \geq 1. \end{aligned} \tag{58}$$

(Note that here, since $t_0 = 0$, we have $t_0(q) = 0$.) Then $N_q(t)$ is the counting process of the number of points of ψ_M that have marks of size 1, i.e.,

$$N_q(t) = \sum_{i=1}^{N(t)} I\{M_i = 1\}, \quad t \geq 0. \tag{59}$$

Letting $\theta_t\psi_M$ denote the marked point process shifted by t, and $\{M_i(t) : i \geq 1\}$ its associated marks ($M_i(t) =$ the mark of the ith point to the right of t), we now have

$$J(t) = q - \min\big\{i \geq 1 : M_i(t) = 1\big\}. \tag{60}$$

We let $J \equiv \{J(t) : t \geq 0\}$. Note that the pair (ψ_q, J) jointly is a function of ψ_M.

6.3 Stationary versions

Now observe that the rmpp ψ_M is neither point nor time stationary, but it is invariant under the $\theta(i) \equiv \theta_{t_i(q)}$ shifts; $\theta(i)\psi_M$ has the same distribution for all $i \geq 0$. Moreover, the cycle-length sequence $\{T_i(q) : i \geq 1\}$ is a stationary and ergodic sequence by the assumption that ψ is a point-stationary ergodic point process. Thus the rmpp ψ_M has the structure of a positive recurrent *synchronous process*, as in the Appendix in [7]. Consequently, ψ_M is *time asymptotically stationary (TAS)* with a time-stationary ergodic version, denoted by ψ_M^*:

$$P(\psi_M^* \in \cdot) \equiv \lim_{t\to\infty} \frac{1}{t} \int_0^t P(\theta_s \psi_M \in \cdot)\, ds = \frac{E\{\int_0^{T_1(q)} I\{\theta_s \psi_M \in \cdot\}\, ds\}}{E(T_1(q))}. \tag{61}$$

Intuitively, this time-stationary version is obtained by choosing the origin to be a randomly chosen time t way out in the infinite future. The rmpp $\theta_t \psi_M^*$ has the same distribution for all $t \geq 0$.

It thus also follows that ψ_M is *event asymptotically stationary* (*EAS*) (see [20]); there exists a point-stationary and ergodic version of ψ_M, denoted by ψ_M^0, such that $\theta_{t_i} \psi_M^0$ has the same distribution for all $i \geq 0$. Intuitively, a point-stationary version is obtained by choosing the origin to be a randomly chosen arrival point t_i way out in the infinite future.

We let ψ_q^*, $N_q^*(t)$, and J^* denote the corresponding processes constructed from ψ_M^*, and we let ψ_q^0, $N_q^0(t)$, and J^0 denote the corresponding processes constructed from ψ_M^0. (We explicitly use (58)–(60) in our constructions.)

From the construction (58)–(60) and the fact that $\theta_t \psi_M^*$ has the same distribution for all $t \geq 0$, we deduce that $\theta_t \psi_q^*$ and $\theta_t J^*$ also have the same distribution for all t jointly along with $\theta_t \psi_M^*$. (Recall that the pair (ψ_q^*, J^*) jointly is a function of ψ_M^*.) Thus the triple $(\psi_M^*, \psi_q^*, J^*)$ is jointly time-stationary. Consequently, $\psi^* = \{t_i^*\}$, the points without marks, is also time stationary; the 4-tuple $(\psi_M^*, \psi_q^*, J^*, \psi^*)$ is jointly time-stationary.

Clearly, $J^*(0)$ has the discrete uniform distribution on $\{0, 1, 2, \ldots, q-1\}$, since $P(J^*(0) = j)$ is the long-run proportion of time that phase $j+1$ of a "cycle" (e.g., an interarrival time) is in progress; $P(J^*(0) = j) = E(T_{j+1})/E(T_1(q)) = 1/q,\ j \in \{0, 1, 2, \ldots, q-1\}$. (Formally, we are using (61).)

Note that if we only were to observe the points of ψ^* themselves, we would not be able to figure out where the marks are, because we have no idea what phase we are in. The following theorem makes this precise, and summarizes what we have observed above.

Theorem 8 *The four-tuple* $(\psi_M^*, \psi_q^*, J^*, \psi^*)$ *is a jointly time stationary process. In addition,* $J^*(0)$ *has the discrete uniform distribution on* $\{0, 1, 2, \ldots, q-1\}$, *and* $J^*(t) = (J^*(0) + N^*(t)) \bmod q,\ t \geq 0$. *Furthermore* $\psi^* = \{t_n^* : n \geq 1\}$ *is independent of* $J^*(0)$.

Proof Only the last statement needs clarification. We will use (61). We must show that $P(J^*(0) = j, \psi^* \in \cdot) = (1/q)P(\psi^* \in \cdot)$, $j \in \{0, 1, 2, \ldots, q-1\}$. The amount of

time during $T_1(q)$ that $J(s) = j$ and that $\psi_s \in \cdot$ is given by the integral

$$I = \int_0^{T_1(q)} I\{J(s) = j,\ \psi_s \in \cdot\}\, ds = \int_{t_j}^{t_{j+1}} I\{\psi_s \in \cdot\}\, ds. \tag{62}$$

But ψ is also invariant under its own point shifts t_i by the assumption that ψ is point-stationary. Thus, by the Palm inverse formula for ψ, we already know that $P(\psi^* \in \cdot) = \lambda E(I)$. Since $E(T_1(q)) = q/\lambda$ we conclude that $E(I)/E(T_1(q)) = (1/q)P(\psi^* \in \cdot)$ as was to be shown. □

We also can consider the point-stationary analog. (Note that ψ^0 by itself has the same point-stationary distribution as the original ψ.) The following (and its proof) is analogous to Theorem 8, we omit the proof.

Theorem 9 *The four-tuple* $(\psi_M^0, \psi_q^0, J^0, \psi^0)$ *is a jointly point stationary process. In addition,* $J^0(0)$ *has the discrete uniform distribution on* $\{0, 1, 2, \ldots, q-1\}$, *and* $J^0(t) = (J^0(0) + N^0(t))$ *mod* q, $t \geq 0$. *Furthermore,* $\psi^0 = \{t_n^0 : n \geq 1\}$ *is independent of* $J^0(0)$.

7 Limits for a stationary point process with cyclic thinning

We now exploit the stationarity to extend Theorem 5.1 of [21] to include a FWLLN for the time stationary versions of the counting processes with cyclic thinning. For this, we return to the notation of Sect. 2. As in Sect. 5 of [21], let the counting processes be defined as

$$\begin{aligned} N_n^u(t) &\equiv \max\{k \geq 0 : S_{n,k}^u \leq t\}, \qquad N_n^v(t) \equiv \max\{k \geq 0 : S_{n,k}^v \leq t\}, \\ N_n^{c,u}(t) &\equiv \max\{k \geq 0 : S_{n,k}^{c,u} \leq t\} \quad \text{and} \quad N_n^{c,v}(t) \equiv \max\{k \geq 0 : S_{n,k}^{c,v} \leq t\} \end{aligned} \tag{63}$$

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 Sect. 2, we have $N_n^u(0) = 1$ and $N_n^v(0) = 0$. Recall the key relations given in (5.44) of [21], namely,

$$\begin{aligned} N_n^u(nt) &= 1 + n\big(N_n^{c,u}(t) - 1\big) + J_n^{c,u}(t) \quad \text{and} \\ N_n^v(nt) &= nN_n^{c,v}(t) + J_n^{c,v}(t), \end{aligned} \tag{64}$$

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 (2) as

$$\begin{aligned} N_n^{c,u}(t) &= 1 + \frac{N_n^u(nt) - 1}{n} - \frac{J_n^{c,u}(t)}{n} = 1 + \big\lfloor \big(N_n^u(nt) - 1\big)/n \big\rfloor \quad \text{and} \\ N_n^{c,v}(t) &= \frac{N_n^v(nt)}{n} - \frac{J_n^{c,v}(t)}{n} = \big\lfloor N_n^v(nt)/n \big\rfloor, \end{aligned} \tag{65}$$

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

We next review Theorem 5.1 of [21]. To state the results, define the following random elements in D:

$$\begin{aligned} \bar{\mathbf{N}}_n^u(t) &\equiv \frac{N_n^u(nt)}{n}, \qquad \bar{\mathbf{N}}_n^v(t) \equiv \frac{N_n^v(nt)}{n}, \\ \bar{\mathbf{J}}_n^{c,u} &\equiv \frac{J_n^{c,u}(t)}{n}, \qquad \bar{\mathbf{J}}_n^{c,v} \equiv \frac{J_n^{c,v}(t)}{n}. \end{aligned} \tag{66}$$

Theorem 10 (FWLLN for the counting processes from [21]) *Consider a sequence of $G_n/G_n/1$ models associated with a single base $G/G/1$ model satisfying*

$$\left(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v\right) \quad \Rightarrow \quad \left(\hat{L}^u, \hat{L}^v\right) \quad \text{in } D^2 \tag{67}$$

for $(\hat{\mathbf{S}}_n^u, \hat{\mathbf{S}}_n^v)$ in (23), where $P((\hat{L}^u, \hat{L}^v) \in C^2) = 1$, as in Corollary 1(a). If either (1) or (2) holds, then

$$\begin{aligned} &\left(\bar{\mathbf{N}}_n^u, \bar{\mathbf{J}}_n^{c,u}, N_n^{c,u}, \bar{\mathbf{N}}_n^v, \bar{\mathbf{J}}_n^{c,v}, N_n^{c,v}\right) \quad \Rightarrow \quad \left(e, J, 1 + \lfloor e \rfloor, e, J, \lfloor e \rfloor\right) \\ &\quad \text{in } D^6 \quad \text{as } n \to \infty, \end{aligned} \tag{68}$$

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

Now we introduce the stationarity assumption: Assume that $N^{u,*}$ is the arrival counting process associated with a time stationary point process, and let $\bar{\mathbf{N}}_n^{u,*}$ be the scaled version, defined as in (66). Let $N_n^{c,u,*}$ be the associated (unscaled) arrival counting process of the time stationary point process after cyclic thinning of order n to N^u. We can combine Sect. 6 with Theorem 10 to obtain

Theorem 11 *Under the conditions of Theorem 10, allowing either (1) or (2), if the base sequence $\{U_k : k \geq 1\}$ is stationary, then there are time stationary versions of the sequences $\{U_{n,k}^c : k \geq 1\}$ for each $n \geq 1$ and*

$$\left(\bar{\mathbf{N}}_n^{u,*}, N_n^{c,u,*}\right) \quad \Rightarrow \quad \left(e, \lfloor Y + e \rfloor\right) \quad \text{in } D^2 \quad \text{as } n \to \infty, \tag{69}$$

where Y is a random variable uniformly distributed on the interval $[0, 1]$. As a consequence,

$$\left(\bar{\mathbf{N}}_n^{u,*}(t), N_n^{c,u,*}(t)\right) \quad \Rightarrow \quad \left(t, \lfloor t + Y \rfloor\right) \quad \text{in } \mathbb{R}^2 \quad \text{as } n \to \infty \tag{70}$$

for each t, with $\lfloor t + Y \rfloor$ being a random variable with the two-point distribution

$$P\left(\lfloor t + Y \rfloor = \lfloor t \rfloor\right) = 1 - P\left(\lfloor t + Y \rfloor = \lfloor t \rfloor + 1\right) = 1 + \lfloor t \rfloor - t. \tag{71}$$

8 Other processes

Let $Q_n^c(\infty)$ the stationary queue length (number in system) and let $R_n^c(\infty)$ the stationary remaining work in the system (the continuous-time workload), both at an arbitrary continuous time in the nth $GI_n/GI_n/1$ model. By Little's law, we know that

$$P\big(Q_n^c(\infty)=0\big)=P\big(R_n^c(\infty)=0\big)=1-\rho_n \quad \text{for all } n\ge 1. \tag{72}$$

For the $GI_n/GI_n/1$ model, we have the explicit formulas for the distributions of $R_n^c(\infty)$ and $Q_n^c(\infty)$ in Theorems X.3.4 and X.4.2 in [2]. By Theorem X.3.4 in [2], for the $GI_n/GI_n/1$ model, we have

$$P\big(R_n^c(\infty)\le x\big)=1-\rho_n+\rho_n P\big(W_{n,\infty}^c+V_{n,e}^c\le x\big), \tag{73}$$

where $V_{n,e}^c$ is a random variable with the stationary excess (or residual lifetime) distribution associated with $V_{n,1}^c$, i.e.,

$$P\big(V_{n,e}^c\le x\big)\equiv\frac{1}{E[V_{n,1}^c]}\int_0^x P\big(V_{n,1}^c>y\big)\,dy, \quad x\ge 0. \tag{74}$$

Since $V_{n,1}^c\Rightarrow D$ as $n\to\infty$, with $E[V_{n,1}^c]=1$ for all n, we have $V_{n,e}^c\Rightarrow Y$ as $n\to\infty$, where Y is uniformly distributed on $[0,1]$.

By Theorem X.4.2 in [2] (the distributional version of Little's law, [8]), for the $GI_n/GI_n/1$ model, we have

$$Q_n^c(\infty)\overset{\mathrm{d}}{=}N_n^{c,u,*}\big(W_{n,\infty}^c+V_{n,1}^c\big) \quad \text{for all } n\ge 1, \tag{75}$$

where $\{N_n^{c,u,*}(t):t\ge 0\}$ is a time stationary version of the arrival counting process independent of $(W_{n,\infty}^c,V_{n,1}^c)$ and the random variables $W_{n,\infty}^c$ and $V_{n,1}^c$ are independent.

We can thus combine Theorems 11, 2 and 7 to obtain the following result.

Corollary 2 (HT limit for the stationary workloads and queue length in $GI_n/GI_n/1$ models) *Consider a sequence of $GI_n/GI_n/1$ models obtained from a base $GI/GI/1$ model.*

(a) *If $(1-\rho_n)\sqrt{n}\to\beta$ as $n\to\infty$ for $0<\beta<\infty$, then $W_n^c(\infty)\Rightarrow 0$ and*

$$\big(R_n^c(\infty),Q_n^c(\infty)\big)\ \Rightarrow\ (Y,1)\quad \text{in } \mathbb{R}^2\quad \text{as } n\to\infty, \tag{76}$$

where Y has the uniform distribution on $[0,1]$.

(b) *If $(1-\rho_n)n\to\beta$ as $n\to\infty$ for $0<\beta<\infty$, then $W_{n,\infty}^c\Rightarrow W_\infty^c$, where W_∞^c is an exponential random variable with mean $\sigma^2/2\beta$,*

$$R_n^c(\infty)\ \Rightarrow\ Y+W_\infty^c \quad \text{and} \quad Q_n^c(\infty)\ \Rightarrow\ \lfloor Y+W_\infty^c+1\rfloor \quad \text{as } n\to\infty, \tag{77}$$

where Y has the uniform distribution on $[0,1]$ and is independent of W_∞^c.

Notice that, with Y uniformly distributed on$[0, 1]$, $Y \le \lfloor Y + 1 \rfloor = 1 \le Y + 1$ in (76) and

$$Y + W_\infty^c \le \left\lfloor Y + W_\infty^c + 1 \right\rfloor \le Y + W_\infty^c + 1$$

in (77), consistent with (6.75) and (6.76) of [21]. Consistent with (72), Corollary 2 implies that $P(R_n^c(\infty) = 0) \to 0$ and $P(Q_n^c(\infty) = 0) \to 0$ as $n \to \infty$ in both cases (a) and (b).

9 Heavy-traffic limits for stationary queue lengths in $GI/D/n$

We now apply the results in previous sections to obtain HT limits for the stationary queue length (number in system) in the $GI/D/n$ model. As in Sect. 8, we do so by applying the distributional version of Little's law. Here this approach follows Corollary 2 in [13]. As noted in [13], for the queue length we cannot directly exploit the reduction to $G_n/D/1$.

Consider a sequence of $GI/D/n$ models with unit service times. Let the arrival processes be defined as in Sect. 3 in terms of a rate-1 renewal process by scaling time. Let the base unit mean interarrival times have finite variance σ_u^2. Let the time-stationary arrival counting process in system n be $N_n^{u,*}$. These counting process obey an FCLT, i.e., $\mathbf{N}_n^{u,*} \Rightarrow \sigma_u B$ in D as $n \to \infty$, where B is Brownian motion and

$$\mathbf{N}_n^{u,*}(t) \equiv \frac{N_n^{u,*}(t) - \rho_n n t}{\sqrt{n}}, \quad t \ge 0. \tag{78}$$

(If either a point stationary or time stationary point process satisfies an FCLT, then they both do, with common limit; see [18]. This is elementary for a renewal process.) Let $N(m, \sigma^2)$ denote a normal random variable with mean m and variance σ^2. Let $W_{n,\infty}$ denote the stationary waiting time in the $GI/D/n$ model, which coincides with $W_n^c(\infty)$ in the $GI_n/D/1$ model.

Theorem 12 (HT limit for the stationary queue lengths in $GI/D/n$ models) *Consider the sequence of $GI/D/n$ queueing models above, where $\mathbf{N}_n^{u,*} \Rightarrow \sigma_u B$ for $\mathbf{N}_n^{u,*}$ in (78).*

(a) *(From [13]) If $(1 - \rho_n)\sqrt{n} \to \beta$ as $n \to \infty$ for $0 < \beta < \infty$, then the assumptions of Theorems 1 and 2 are satisfied, so that $\sqrt{n} W_{n,\infty} \Rightarrow \tilde{W}_\infty$, where $\tilde{W}_\infty \stackrel{\mathrm{d}}{=} \max\{(\sigma_u B - \beta e)(k) : k \ge 0\}$ and*

$$\frac{Q_n(\infty) - n}{\sqrt{n}} \quad \Rightarrow \quad \tilde{Q}(\infty) \equiv N\left(-\beta, \sigma_u^2\right) + \tilde{W}_\infty \quad \text{in } \mathbb{R} \quad \text{as } n \to \infty, \tag{79}$$

where the two limiting random variables on the right in (79) are independent. In addition, $\tilde{Q}(\infty) \stackrel{\mathrm{d}}{=} \max\{\tilde{L}(k) : k \ge 1\}$ and $\tilde{W}_\infty \stackrel{\mathrm{d}}{=} \tilde{Q}(\infty)^+$.

(b) *If $(1 - \rho_n)n \to \beta \quad$ as $n \to \infty$ for $0 < \beta < \infty$, then the assumptions of Theorems 6 and 7 are satisfied, so that $W_{n,\infty} \Rightarrow W_\infty$, where W_∞ is an exponential random*

variable with $E[W_\infty] = \sigma_u^2/2\beta$, *and*

$$\frac{Q_n(\infty)}{n} \quad \Rightarrow \quad W_\infty + 1 \quad \text{in } \mathbb{R} \quad \text{as } n \to \infty. \tag{80}$$

Proof Given the assumed and established convergence, both (a) and (b) are proved by applications of the continuous mapping theorem with the composition map, in order to capture the random change. For both (a) and (b), we start with $Q_n(\infty) \stackrel{\mathrm{d}}{=} N_n^{u,*}(W_{n,\infty} + 1)$.

For (a), by the independence in the distributional version of Little's law, we can start with the joint convergence $(\mathbf{N}_n^{u,*}, \sqrt{n}W_{n,\infty}) \Rightarrow (\sigma_u B, \tilde{W}_\infty)$ in $D \times \mathbb{R}$ as $n \to \infty$, which is important for applying the continuous mapping with the composition map in order to treat a random time change. The established limit $\sqrt{n}W_{n,\infty} \Rightarrow \tilde{W}_\infty$ implies that $W_{n,\infty} + 1 \Rightarrow 1$. The limits $(1 - \rho_n)\sqrt{n} \to \beta$ and $\mathbf{N}_n^{u,*} \Rightarrow \sigma_u B$ imply that $\tilde{\mathbf{N}}_n^{u,*} \Rightarrow \sigma_u B - \beta e$, where $\tilde{\mathbf{N}}_n^{u,*}(t) \equiv (N_n^{u,*} - nt)/\sqrt{n}$, which has a different translation term. Then we have

$$\begin{aligned}
\frac{Q_n(\infty) - n}{\sqrt{n}} &\stackrel{\mathrm{d}}{=} \frac{N_n^{u,*}(W_{n,\infty} + 1) - n}{\sqrt{n}} \\
&= \frac{N_n^{u,*}(W_{n,\infty} + 1) - n(W_{n,\infty} + 1)}{\sqrt{n}} + \sqrt{n}\big((W_{n,\infty} + 1) - 1\big) \\
&= \tilde{\mathbf{N}}_n^{u,*}(W_{n,\infty} + 1) + \sqrt{n}W_{n,\infty} \\
\Rightarrow \quad & (\sigma_u B - \beta e)(1) + \tilde{W}_\infty.
\end{aligned} \tag{81}$$

For the concluding statement, since $\tilde{L}(1) \stackrel{\mathrm{d}}{=} N(-\beta, \sigma_u^2)$, we have $\tilde{W}_\infty \stackrel{\mathrm{d}}{=} \max\{\tilde{L}(k) : k \geq 0\}$, with $\tilde{L}(0) \equiv 0$ and $\tilde{Q}(\infty) \stackrel{\mathrm{d}}{=} \max\{\tilde{L}(k) : k \geq 1\}$, so that $\tilde{W}_\infty \stackrel{\mathrm{d}}{=} \tilde{Q}(\infty)^+$.

For (b), the limit $\mathbf{N}_n^{u,*} \Rightarrow \sigma_u B$ implies that $\bar{\mathbf{N}}_n^{u,*} \Rightarrow e$, where $\bar{\mathbf{N}}_n^{u,*}(t) \equiv N_n^{u,*}(t)/n$, $t \geq 0$. Then the established limit $W_{n,\infty} \Rightarrow W_\infty$ implies that

$$\frac{Q_n(\infty)}{n} \stackrel{\mathrm{d}}{=} \frac{N_n^{u,*}(W_{n,\infty} + 1)}{n} = \bar{\mathbf{N}}_n^{u,*}(W_{n,\infty} + 1) \quad \Rightarrow \quad W_\infty + 1. \tag{82}$$

□

Intuitively, Theorem 12 makes sense, because with n servers we expect to have $W_{n,\infty} \approx (Q_n(\infty) - n)^+/n$, so that in case (a) where $W_{n,\infty} = O(1/\sqrt{n})$, we expect to have $Q_n(\infty) - n = O(\sqrt{n})$, while in case (b) where $W_{n,\infty} = O(1)$, we expect to have $Q_n(\infty) - n = O(n)$.

10 Implications for staffing in the $GI/D/n$ system

In this section we discuss the implications of the waiting time limits in Sects. 2–4 for staffing in the $GI/D/n$ model. The queue-length limits in Sect. 9 provide another perspective, but here we focus on waiting times. As in the previous section, let

$W_{n,\infty}$ denote the stationary waiting time in the $GI/D/n$ model, which coincides with $W_n^c(\infty)$ in the $GI_n/D/1$ model.

By Corollary 1 of [13] and Theorem 4, for any α, $0 < \alpha < 1$,

$$P(W_{n,\infty} > 0) \to \alpha \quad \text{as } n \to \infty \tag{83}$$

if and only if

$$(1 - \rho_n)\sqrt{n} \to \beta, \quad 0 < \beta < \infty, \tag{84}$$

where $\beta \equiv \beta(\alpha)$ can be obtained by inverting the function $\alpha(\beta)$ in (29). This closely parallels Proposition 1 of [9] for the corresponding $M/M/n$ model.

Theorem 4 would seem to justify once again staffing by setting n so that $\rho_n \approx 1 - \beta/\sqrt{n}$ for appropriate β, and that is one possibility. However, a more common constraint in service-level agreements in practice is

$$P(W_{n,\infty} > \tau) \le \alpha \tag{85}$$

for some pair (τ, α) with $0 < \tau < \infty$ and $0 < \alpha < 1$. Thus we might seek asymptotics of the form $P(W_{n,\infty} > \tau) \to \alpha \quad \text{as } n \to \infty$. Indeed, the consequence of the alternative (customary) staffing constraint (85) in $M/M/n + GI$ models upon asymptotics is studied in [15]. However, the consequence is different in our nearly-deterministic setting.

One might suspect that we could actually achieve higher server utilization with the nearly deterministic model. In fact, we show that we can staff so that $\rho_n \approx 1 - \beta/n$ for appropriate β if our target is (85). The following is a consequence of Theorem 7.

Theorem 13 (Asymptotically nondegenerate tail probabilities) *Consider a sequence of $GI/D/n$ models indexed by n with unit deterministic service times and interarrival times having variances $\sigma_n^2 \to \sigma^2$, $0 < \sigma^2 < \infty$. Then, for any pair (τ, α) with $0 < \tau < \infty$ and $0 < \alpha < 1$,*

$$P(W_{n,\infty} > \tau) \to \alpha \quad \textit{and} \quad E[W_{n,\infty}] \to \frac{\tau}{-\log_e(\alpha)} \quad \textit{as } n \to \infty \tag{86}$$

if and only if

$$(1 - \rho_n)n \to \beta, \quad 0 < \beta < \infty, \tag{87}$$

where

$$\beta \equiv \beta(\tau, \alpha) \equiv \frac{-\log_e(\alpha)\sigma^2}{2\tau}. \tag{88}$$

As a consequence, in this setting,

$$W_{n,\infty} \quad \Rightarrow \quad 0 \quad \textit{and} \quad E[W_{n,\infty}] \to 0 \quad \textit{as } n \to \infty \tag{89}$$

if and only if $(1 - \rho_n)n \to \infty$.

Proof The implication (87) $\rightarrow$ (86) is implied by Theorem 7. The same will be true along any subsequence for which (87) holds. Hence, suppose that there is a subsequence for which $(1-\rho_n) \rightarrow \infty$. We can show that the limit in (86) cannot hold, because we can apply the result for very large finite β. Similarly, suppose that there is a subsequence for which $(1-\rho_n) \rightarrow 0$. We can again show that (86) cannot hold by comparing to the case in which (87) holds for very small β. The stochastic comparison is valid because in the general $G_n/G_n/1$ model, for fixed n and k, $W_{n,k}^c$ is increasing in ρ_n, under (24), (7) and (8). The final conclusion (89) is obtained along the way. □

We obtain the simple formula for the quality-of-service parameter β in (87) and (88) because the $W_{n,\infty}$ converges in distribution, without spatial scaling, to an exponential random variable with mean $\sigma^2/2\beta$ under condition (87). From (86), we see that the asymptotics for the mean are determined by the asymptotics for the tail probability: Given the pair (τ, α), the asymptotics for the mean are determined. On the other hand, given a target for the mean, there is a one parameter family of pairs (τ, α) yielding that same constraint on the mean. In other words, there is more freedom with the tail probability constraints. However, we can always change the tail probability constraint, without changing the staffing, provided that we keep the ratio $-\log_e(\alpha)/\tau$ fixed.

Note that we can simultaneously have both $P(W_{n,\infty} > 0) \rightarrow \alpha > 0$ and $W_{n,\infty} \Rightarrow 0$. In particular, Theorems 4 and 13 imply that these both occur if and only if $(1-\rho_n)\sqrt{n} \rightarrow \beta$, $0 < \beta < \infty$. We include the final (89) to contrast with Theorem 4. We regard Theorem 13 as providing strong support for staffing so that $1-\rho_n = O(1/n)$. However, if the model were not actually nearly deterministic, for example, if the model were $GI_n/GI/1$ where the GI service is neither D nor GI_n, so that the model would not be nearly deterministic, then the relevant HT regime would be the usual one, suggesting $1-\rho_n = O(1/\sqrt{n})$. The risk of service interruptions also cautions against trying to extract maximal economies of scale; see [19].

Acknowledgements 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. We thank a referee for showing how we can weaken previous moment conditions in Theorem 5.

References

1. Abate, J., Choudhury, G.L., Whitt, W.: Calculation of the $GI/G/1$ steady-state waiting-time distribution and its cumulants from Pollaczek's formula. AEÜ, Int. J. Electron. Commun. **47**, 311–321 (1993)
2. Asmussen, S.: Applied Probability and Queues, 2nd edn. Springer, New York (2003)
3. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley, New York (1999)
4. Borovkov, A.A.: Asymptotic Methods in Queueing Theory. Wiley, New York (1984)
5. Breiman, L.: Probability. Addison-Wesley, Reading (1968)
6. Browne, S., Whitt, W.: Piecewise-linear diffusion processes. In: Dshalalow, J. (ed.) Advances in Queueing, pp. 463–480. CRC Press, Boca Raton (1995)
7. Glynn, P.W., Sigman, K.: Uniform Cesaro limit theorems for synchronous processes with applications to queues. Stoch. Process. Appl. **40**, 29–44 (1992)
8. Haji, R., Newell, G.F.: A relation between stationary queue and waiting-time distributions. J. Appl. Probab. **8**, 617–620 (1971)

9. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**, 567–587 (1981)
10. Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: On Lerch's transcendent and the Gaussian random walk. Ann. Appl. Probab. **17**, 421–439 (2007)
11. Janssen, A.J.E.M., van Leeuwaarden, J.S.H.: Cumulants of the maximum of the Gaussian random walk. Stoch. Process. Appl. **117**, 1928–1959 (2007)
12. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, B.: Corrected diffusion approximations for a multi-server queue in the Halfin–Whitt regime. Queueing Syst. **58**, 261–301 (2008)
13. Jelenkovic, P., Mandelbaum, A., Momcilovic, P.: Heavy traffic limits for queues with many deterministic servers. Queueing Syst. **47**, 53–69 (2004)
14. Kingman, J.F.C.: Some inequalities for the queue $GI/G/1$. Biometrika **43**, 315–324 (1962)
15. Mandelbaum, A., Zeltyn, S.: Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Oper. Res. **57**, 1189–1205 (2009)
16. Müller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. Wiley, New York (2002)
17. Nagaev, S.V.: Large deviations of sums of independent random variables. Ann. Probab. **7**, 745–789 (1979)
18. Nieuwenhuis, G.: Equivalence of functional limit theorems for stationary point processes and their Palm distributions. Probab. Theory Relat. Fields **81**, 593–608 (1989)
19. Pang, G., Whitt, W.: Heavy-traffic limits for many-server queues with service interruptions. Queueing Syst. **61**, 167–202 (2009)
20. Sigman, K.: Stationary Marked Point Processes: An Intuitive Approach. Chapman-Hall/CRC, New York (1995)
21. Sigman, K., Whitt, W.: Heavy-traffic limits for nearly deterministic queues. J. Appl. Probab. **48** (2011, to appear)
22. Szczotka, W.: Exponential approximation of waiting time and queue size for queues in heavy traffic. Adv. Appl. Probab. **22**, 230–240 (1990)
23. Szczotka, W.: Tightness of the stationary waiting time in heavy traffic. Adv. Appl. Probab. **31**, 788–794 (1999)
24. Whitt, W.: The queueing network analyzer. Bell Syst. Tech. J. **62**, 2779–2815 (1983)
25. Whitt, W.: Understanding the efficiency of multi-server service systems. Manag. Sci. **38**, 708–723 (1992)
26. Whitt, W.: Stochastic-Process Limits. Springer, New York (2002)