Heavy Traffic Limits for Queues with Many Deterministic Servers

Predrag Jelenković[a] Avishai Mandelbaum[b,*] Petar Momčilović[a,†]

[a]Department of Electrical Engineering
Columbia University
New York, NY 10027
{predrag, petar}@ee.columbia.edu

[b]Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology
Haifa 32000, Israel
avim@tx.technion.ac.il

November 15, 2002, revised June 2003

Abstract

Consider a sequence of stationary GI/D/N queues indexed by $N \uparrow \infty$, with servers' utilization $1 - \beta/\sqrt{N}$, $\beta > 0$. For such queues we show that the scaled waiting times $\sqrt{N}W_N$ converge to the (finite) supremum of a Gaussian random walk with drift $-\beta$. This further implies a corresponding limit for the number of customers in the system, an easily computable non-degenerate limiting delay probability in terms of Spitzer's random-walk identities, and $\sqrt{N}$ rate of convergence for the latter limit. Our asymptotic regime is important for rational dimensioning of large-scale service systems, for example telephone- or internet-based, since it achieves, simultaneously, arbitrarily high service-quality and utilization-efficiency.

Keywords: Multi-server queue, GI/D/N, deterministic service time, heavy-traffic, Quality and Efficiency Driven (QED) or Halfin-Whitt regime, telephone call or contact centers, economies of scale, Gaussian random walk, Spitzer's identities

*Supported in part by the fund for the promotion of research at the Technion, by the Technion V.P.R. fund for the promotion of sponsored research, and by the Israel Science Foundation (grants 388/99 and 126/02). The research started while the author was visiting Columbia Business School, the hospitality of which is greatly appreciated.

†Supported in part by IBM PhD Fellowship.

1 Introduction

The dimensioning of a shared-resource facility necessarily balances between efficiency and quality, or more specifically between servers' capacity-utilization and customers' perceived service-quality: high utilization is typically achieved at the cost of frequent and long delays. It is thus commonly accepted that high efficiency and service-quality can not coexist. But here economies-of-scale come to the rescue. Indeed, large-scale service systems can operate in a regime, to which we refer as Quality & Efficiency Driven (QED), where *both* objectives are accomplished. The scaling that leads to the QED regime is of importance for dimensioning systems with high server costs where over provisioning is economically unacceptable. This is often the case for large telephone call centers in which the main operating cost is agents' (servers') salaries and wireless communication systems with inherently limited frequency spectrum.

Due to the desirable features of the QED regime, it has recently enjoyed considerable attention in the literature. But in fact, the importance of this regime was recognized as early as in Erlang's 1923 paper, that appeared in [12] and which addresses both Erlang-B (M/M/N/N) and Erlang-C (M/M/N) models. A precise characterization of the asymptotic expansion of the blocking probability, for Erlang-B in the QED regime, was given first in Jagerman [23]; see also [32]. However, the formal characterization of the QED regime, as one which accommodates both high operational efficiency (many heavily utilized servers) and high service level (a delay probability that is strictly between 0 and 1), was first recognized by Halfin and Whitt [20]. Specifically, they considered the GI/M/N queue in the QED regime, analyzing the scaled number of customers in both steady state and as a stochastic process. Convergence of this same scaled queueing process, in the more general GI/PH/N setting, was established in [28]. Application of QED queues to modelling and staffing of telephone call centers and communication networks, taking into account customers' impatience, can be found in [17] and [13], respectively. The optimality of the QED regime, under revenue maximization or constraint satisfaction, is discussed in [7, 26, 1, 2]. Readers are referred to Sections 4 and 5.1.4 of [16] for a survey of the QED regime, both practically and academically. Very recent references are [34, 35].

In this paper, we consider a sequence of stationary first-come-first-served GI/D/N queues, indexed by $N \uparrow \infty$. For the Nth system, the traffic intensity is $1 - \beta/\sqrt{N}$, $\beta > 0$, which is equivalent to a staffing level $N \approx R_N + \beta\sqrt{R_N}$; here R_N is its offered load, namely the arrival rate multiplied by mean service time. As will be shown, such scaling leads to the QED operational regime: traffic intensity increases to 1 (high efficiency) and simultaneously congestion levels diminish (high service level).

In the next section we introduce a decomposition property of the GI/D/N queue into N single-server queues. This yields convergence in distribution for the scaled waiting time and queue length. In each case, the limit has an explicit representation as the supremum of a Gaussian random walk with drift $-\beta$, and the two limits are related in a very simple way. We also show that the above scaling is equivalent to a non-degenerate limiting delay probability, and further establish the corresponding rate of convergence. The distribution

of the limiting waiting time is discussed in Section 4, specifically its asymptotics around the origin and the tail. The last section concludes with some possible generalizations, and comparisons of QED performance between GI/D/N and related systems.

2 Decomposition of GI/D/N into N single-server queues

A cyclic service discipline in an N-server queue is a scheduling policy under which every Nth customer is assigned to the same server. As will be shown, such policies achieve "perfect load balancing" among servers when the service requirements equal to a constant: the workloads of any two servers differ by at most a single service requirement.

Cyclic scheduling policies play a central role in establishing our main results. They have already been used for analyzing multi server queues. Indeed, in [36] they were utilized to establish stochastic upper bound for the performance of first-come-first-served multi server system, while in [31] to obtain an alternative proof of the existence of the stationary queue length and waiting time distributions. Recently, cyclic policies for the M/D/N queue enabled analysis in steady state [14] and in transience [15].

Consider an N-server queue under cyclic scheduling. Let $V_k^{(n)}$ be the workload of the kth server just before the nth arrival. It is assumed that the customers are assigned to servers upon their arrival. The following lemma, besides its load balancing property, states that the individual server workloads adhere to a cyclic permutation property.

Lemma 1. *Assume that all subscripts are* $\bmod N$. *For an arbitrary arrival sequence of customers, with unit service requirements (for convenience), if* $V_N^{(1)} - V_1^{(1)} \leq 1$ *and*

$$V_1^{(1)} \leq V_2^{(1)} \leq \cdots \leq V_N^{(1)},$$

then $V_{n-1}^{(n)} - V_n^{(n)} \leq 1$ *and*

$$V_n^{(n)} \leq \cdots \leq V_N^{(n)} \leq V_1^{(n)} \leq \cdots \leq V_{n-1}^{(n)}, \quad n \geq 1.$$

Proof. The proof is by induction; assume that the lemma holds for some $n > 1$. Let τ_{n+1} be the interarrival time between the nth and $(n+1)$st customers. After the arrival of the $(n+1)$st customer, by induction hypothesis, the smallest workload becomes the largest. Next, due to the cyclic policy we have $V_n^{(n+1)} = (V_n^{(n)} + 1 - \tau_{n+1})^+$, while the rest of the workloads $V_k^{(n+1)} = (V_k^{(n)} - \tau_{n+1})^+$, $k \neq n$. Then

$$\begin{aligned} V_n^{(n+1)} - V_{n+1}^{(n+1)} &\leq (V_n^{(n)} + 1 - \tau_{n+1})^+ - (V_{n+1}^{(n)} - \tau_{n+1})^+ \\ &\leq V_n^{(n)} + 1 - V_{n+1}^{(n)} \\ &\leq 1. \end{aligned}$$

□

The next lemma establishes equivalence between the first-come-first-served and cyclic multi server queue, assuming that service times are deterministic. It is an easy consequence

of the preceding lemma and, as such, it is a sample-paths result that does not require any probabilistic structure of the arrival sequence, e.g., being renewal. The relationship between the two scheduling policies was first documented in [22].

Lemma 2. *Consider two, initially empty, multi-server systems with first-come-first-served and cyclic service disciplines. Suppose that both cater to the same arrival sequence and provide a common fixed service time. Then both systems give rise to the same sequences of customer waiting times.*

An intuitive explanation of the lemma is as follows. Constant service times and first-come-first-served service discipline lead to the fact that customers depart in the order of their arrival. In other words, no customer can overtake any other customer. Therefore, upon arrival, a customer can be assigned to the same server that the Nth prior customer received service from (without causing any extra idleness relative to a work-conserving first-come-first-served discipline).

3 Main results

Consider a sequence of GI/D/N queues, indexed by N, with arrival rates $\lambda_N \to \infty$ as $N \to \infty$. For the Nth queue, the arrival process is a renewal process with interarrival times equal in distribution to τ_N, where $\mathbb{E}\tau_N = \lambda_N^{-1}$ and $\sigma_N^2 \triangleq \lambda_N^2 \mathrm{Var}(\tau_N) \to \sigma^2 < \infty$, as $N \to \infty$. Denote by $\tau_{N,n}$ the interarrival time between the $(n-1)$th and nth customers; $W_{N,n}$ stands for the waiting time of the nth customer. We assume that service requirements are constant and equal to $m > 0$*. Then the offered load is $R_N = \lambda_N m$.

Let the number of servers in the Nth system be $\lceil R_N + \beta\sqrt{R_N} \rceil$, for some $\beta > 0$. The traffic intensity $\rho_N \triangleq R_N/N$ then approaches 1 from below, as $N \to \infty$. (More precisely, $\sqrt{N}(1-\rho_N) = \beta + O(1/\sqrt{N})$, as $N \to \infty$.) Finally, we note that all stationary performance measures exist by the classical work of [24].

Denote by S_n the sum of n i.i.d. normal random variables with mean $-\beta$ and variance σ^2 equal to the asymptotic squared coefficient of variation of the interarrival times. Assume by convention that $S_0 = 0$. We use $\Longrightarrow$ for convergence in distribution. Our first main result concerns waiting time asymptotics.

Theorem 1. *The stationary waiting time W_N satisfies, as $N \to \infty$,*

$$\sqrt{N}\frac{W_N}{m} \Longrightarrow W \triangleq \sup_{n \geq 0} S_n.$$

Proof. By Lemma 2 it suffices to investigate the system under cyclic scheduling. Furthermore, by symmetry, one needs to consider only a single server queue. Then, the evolution

*All the results will be stated for a general $m > 0$, yet in the proofs, for notational simplicity and without loss of generality, we let $m = 1$.

of the waiting times in a single server is governed by Lindley's recursion

$$W_{N,N(n+1)} = \left(W_{N,Nn} + 1 - \sum_{i=Nn+1}^{N(n+1)} \tau_{N,i} \right)^+ \tag{1}$$

and, thus,

$$\sqrt{N} W_N \stackrel{d}{=} \sup_{n \geq 0} \left\{ n\sqrt{N} - \sum_{i=1}^{nN} \sqrt{N} \tau_{N,i} \right\},$$

where $\stackrel{d}{=}$ denotes equality in distribution. Next, easy algebraic steps and the CLT for triangular arrays (e.g., see [5, p. 359]) yield

$$\sqrt{N} - \sum_{i=1}^{N} \sqrt{N} \tau_{N,i} = \sqrt{\frac{N}{\lambda_N}} \frac{\lambda_N - \sum_{i=1}^{\lceil \lambda_N + \beta\sqrt{\lambda_N} \rceil} \lambda_N \tau_{N,i}}{\sqrt{\lambda_N}}$$
$$\Longrightarrow S_1, \qquad \text{as } N \to \infty.$$

Finally, the result follows from Theorem 1, p. 207 of [6]. □

Let $\Phi(\cdot)$ be the standard normal distribution function. In this paper, C denotes a sufficiently large positive constant; at different places, values of C are generally different as well, i.e., $C^2 = C$ or $C + C = C$. The following corollary relates the probability of delay and the number of servers.

Corollary 1. *The probability of delay has a nondegenerate limit*

$$\lim_{N \to \infty} \mathbb{P}[W_N > 0] = \alpha, \qquad 0 < \alpha < 1,$$

if and only if

$$\lim_{N \to \infty} (1 - \rho_N)\sqrt{N} = \beta, \qquad 0 < \beta < \infty,$$

in which case

$$\alpha \triangleq \alpha(\beta/\sigma) = 1 - e^{-\sum_{n=1}^{\infty} \frac{1}{n} \Phi(-\frac{\beta}{\sigma}\sqrt{n})}$$

and

$$\lim_{N \to \infty} \mathbb{E}\left[\sqrt{N} \frac{W_N}{m} \right] = \sum_{n=1}^{\infty} \left[\frac{\sigma}{\sqrt{2\pi n}} e^{-\frac{\beta^2 n}{2\sigma^2}} - \beta \Phi\left(-\frac{\beta\sqrt{n}}{\sigma} \right) \right].$$

Proof. Given that zero is not a point of continuity of the distribution function of W, we first establish $\mathbb{P}[W_N = 0] \to \mathbb{P}[W = 0]$, as $N \to \infty$. The fact that $P[W = 0] = 1 - \alpha$ follows from Spitzer's identity (see Section 8.5 of [11]). If $S_n(N) \triangleq n\sqrt{N} - \sum_{i=1}^{nN} \sqrt{N} \tau_{N,i}$ then the

expression for the supremum of the negative drift random walk (see [11, p. 291]), Fatou's lemma and the CLT yield

$$\begin{aligned}\varliminf_{N\to\infty} \mathbb{P}[W_N = 0] &= \varliminf_{N\to\infty} e^{-\sum_{n=1}^{\infty}\frac{1}{n}\mathbb{P}[S_n(N)>0]} \\ &\geq e^{-\sum_{n=1}^{\infty}\frac{1}{n}\mathbb{P}[S_n>0]} \\ &= e^{-\sum_{n=1}^{\infty}\frac{1}{n}\Phi(-\frac{\beta}{\sigma}\sqrt{n})} \\ &= \mathbb{P}[W = 0]. \end{aligned} \tag{2}$$

On the other hand, by the right-continuity of $\mathbb{P}[W \leq x]$ and Theorem 1, for any $x > 0$

$$\varlimsup_{N\to\infty} \mathbb{P}[W_N = 0] \leq \mathbb{P}[W \leq x]$$

and after $x \to 0$ one has

$$\varlimsup_{N\to\infty} \mathbb{P}[W_N = 0] \leq \mathbb{P}[W = 0]. \tag{3}$$

Now combining (2) and (3) implies the desired result.

To verify that the probability of delay α is in $(0, 1)$, note that the sum in its exponent must be finite and positive, which indeed follows from $\Phi(-x) = (\sqrt{2\pi}x)^{-1}e^{-x^2/2}(1+o(1))$ as $x \to \infty$. Alternatively, this can be seen from $\mathbb{P}[W > 0] = \mathbb{P}[W + S_1 > 0] = \mathbb{E}\Phi(\frac{W-\beta}{\sigma}) < 1$, where the first equality follows from $W \stackrel{d}{=} (W + S_1)^+$ and the latter strict inequality is a consequence of $W < \infty$, almost surely. To prove that there is convergence in L^1, given the convergence in distribution (Theorem 1), one must verify (see Theorem 4.5.2 in [11]) that

$$\sup_N \sqrt{N}\mathbb{E}W_N < \infty.$$

To this end, Lindley's recursion renders for W_N independent of $\tau_{N,i}$

$$W_N \stackrel{d}{=} \left(W_N + 1 - \sum_{i=1}^{N} \tau_{N,i}\right)^+,$$

which after raising to the square power, taking expectation on both sides and noting that by Theorem 2.1 of [3, p. 184] $\mathbb{E}W_N^2 < \infty$ yields

$$\mathbb{E}W_N \leq \frac{\mathbb{E}(1 - \sum_{i=1}^{N} \tau_{N,i})^2}{2(1 - \rho_N)} \leq \frac{C}{\sqrt{N}}.$$

Hence, $\mathbb{E}\sqrt{N}W_N \to \mathbb{E}W$; the first moment of W can be represented as [11, p. 287]

$$\mathbb{E}W = \sum_{n=1}^{\infty} \frac{1}{n}\mathbb{E}[S_n^+]$$

and the "if" statement now follows.

As far as the "only if" statement is concerned, we make use of the fact that for a given arrival sequence the probability of delay is non-increasing function in the number of servers. As in [20], if $(1-\rho_N)\sqrt{N} \to 0$ then for any $\beta > 0$

$$\lim_{N\to\infty} \mathbb{P}[W_N > 0] \leq \alpha(\beta/\sigma)$$

and, hence, the probability of delay converges to one. On the other hand if $(1-\rho_N)\sqrt{N} \to \infty$ then for any $\beta > 0$

$$\lim_{N\to\infty} \mathbb{P}[W_N > 0] \geq \alpha(\beta/\sigma)$$

and, hence, the probability of delay converges to zero. Finally, if $(1-\rho_N)\sqrt{N}$ fails to converge to any limit, then there exist two subsequences that converge to different limits and the above applies to each of the subsequences. Since $\alpha(\cdot)$ is strictly decreasing, the subsequences converge to different limits and the original sequence does not converge. □

Next, we turn our attention to Q_N, the number of customers in the system. The distribution of Q_N can not be derived by localizing to individual servers, as done with W_N. Yet, it equals to the number of arrivals during $W_N + m$ time periods, by the Distributional Little's Law [19], which enables one to deduce the asymptotics of Q_N from that of W_N.

Corollary 2. *For the stationary number of customers Q_N in the Nth system, we have*

$$\frac{Q_N - N}{\sqrt{N}} \Longrightarrow Q \triangleq \sup_{n\geq 1} S_n,$$

as $N \to \infty$, with the limiting variable satisfying $W \stackrel{d}{=} Q^+$.

Note that the quantity $(Q_N - N)/\sqrt{N}$ describes the scaled queue length if positive, and the scaled number of idle servers if negative. The latter converges to Q^-.

Proof. Let $\Lambda_N(\cdot)$ be the number of arrivals in $(0,\cdot)$ of a stationary renewal point process with interarrival times equal in distribution to τ_N (the first arrival time $\tau_{N,1}$ has the excess distribution of τ_N). Then, since customers depart in the order of arrival and the waiting time of a customer is independent of future arrivals, by the Distributional Little's Law [19] one has $Q_N \stackrel{d}{=} \Lambda_N(1+W_N)$, where W_N is independent of Λ_N. Therefore,

$$\begin{aligned}
\mathbb{P}\left[\frac{Q_N - N}{\sqrt{N}} > x\right] &= \mathbb{P}[\Lambda_N(1+W_N) \geq \lceil N + x\sqrt{N}\rceil] \\
&= \mathbb{P}\left[\sum_{i=1}^{\lceil N+x\sqrt{N}\rceil} \tau_{N,i} \leq 1 + W_N\right] \\
&= \mathbb{P}\left[\frac{\sum_{i=1}^{\lceil N+x\sqrt{N}\rceil} \lambda_N \tau_{N,i} - \lceil N + x\sqrt{N}\rceil}{\sqrt{N+x\sqrt{N}}} \leq \frac{\lambda_N W_N + \lambda_N - \lceil N + x\sqrt{N}\rceil}{\sqrt{N+x\sqrt{N}}}\right],
\end{aligned}$$

where the second equality follows from $\{\Lambda_N(t) \geq n\} = \{\sum_{i=1}^n \tau_{N,i} \leq t\}$. Next, the CLT for triangular arrays, Theorem 1 and the independence of W_N and $\{\tau_{N,i}\}$ (see [11, p. 92]) lead to

$$\lim_{N\to\infty} \mathbb{P}\left[\frac{Q_N - N}{\sqrt{N}} > x\right] = \mathbb{P}[S_1 + W \geq x] = \mathbb{P}[S_1 + W > x],$$

where S_1 and W are independent; the last equality holds since the distribution of $S_1 + W$ is absolutely continuous. The definition of W and the preceding relationship yield the statement of the corollary. □

The next result provides the rate of convergence for the probability of wait $\mathbb{P}[W_N > 0]$.

Theorem 2. *If* $\sup \lambda_N \mathbb{E}\tau_N^3 < \infty$ *and* $|\sigma - \sigma_N| \leq C/\sqrt{N}$*, then the relative error satisfies*

$$\frac{|\mathbb{P}[W_N = 0] - \mathbb{P}[W = 0]|}{\mathbb{P}[W = 0]} \leq \frac{C}{\sqrt{N}}.$$

Proof. Let $S_n(N) = n\sqrt{N} - \sum_{i=1}^{nN} \sqrt{N}\tau_{N,i}$. Using the expression for the supremum of the negative-drift random walk (see [11, p. 291]) one obtains

$$\begin{aligned}\frac{|\mathbb{P}[W_N = 0] - \mathbb{P}[W = 0]|}{\mathbb{P}[W = 0]} &= \Big|\prod_{n=1}^{\infty} e^{\frac{1}{n}(\mathbb{P}[S_n(N)>0]-\mathbb{P}[S_n>0])} - 1\Big| \\ &\leq \prod_{n=1}^{\infty} e^{\frac{1}{n}|\mathbb{P}[S_n(N)>0]-\mathbb{P}[S_n>0]|} - 1. \end{aligned} \tag{4}$$

Next, let $k_N = (N - \lambda_N)/\sqrt{N}$ and note that

$$\{S_n(N) > 0\} = \left\{\frac{nN - \sum_{i=1}^{nN} \lambda_N \tau_{N,i}}{\sqrt{nN}} > k_N\sqrt{n}\right\}.$$

The difference of probabilities in (4) can be represented in the following way:

$$\begin{aligned}\mathbb{P}[S_n(N) > 0] - \mathbb{P}[S_n > 0] &= \mathbb{P}\left[\frac{\sum_{i=1}^{nN}(1 - \lambda_N\tau_{N,i})}{\sigma_N\sqrt{nN}} > \frac{k_N\sqrt{n}}{\sigma_N}\right] - \Phi\left(-\frac{k_N\sqrt{n}}{\sigma_N}\right) \\ &\quad + \Phi\left(-\frac{k_N\sqrt{n}}{\sigma_N}\right) - \Phi\left(-\frac{\beta\sqrt{n}}{\sigma}\right) \\ &\triangleq f_1(n, N) + f_2(n, N). \end{aligned} \tag{5}$$

A bound on the absolute value of the first term in the preceding equation is due to the Berry-Esseen Theorem

$$\begin{aligned}|f_1(n, N)| &\leq \sup_{x\geq 0}\left|\mathbb{P}\left[\frac{\sum_{i=1}^{nN}(1 - \lambda_N\tau_{N,i})}{\sigma_N\sqrt{nN}} \leq x\right] - \Phi(x)\right| \\ &\leq \frac{C}{\sqrt{nN}}.\end{aligned}$$

The term f_2 is bounded as follows:

$$|f_2(n,N)| \leq \frac{1}{\sqrt{2\pi}} \int_{\frac{k_N\sqrt{n}}{\sigma_N}\wedge\frac{\beta\sqrt{n}}{\sigma}}^{\frac{k_N\sqrt{n}}{\sigma_N}\vee\frac{\beta\sqrt{n}}{\sigma}} e^{-\frac{x^2}{2}}\,dx$$
$$\leq \left|\frac{\beta}{\sigma} - \frac{k_N}{\sigma_N}\right| C\sqrt{n}\, e^{-\frac{n}{C}} \leq C\sqrt{\frac{n}{N}} e^{-\frac{n}{C}},$$

where the second inequality follows from $\sigma_N \to \sigma$ and $k_N \to \beta$, while the last inequality is due to

$$\left|\frac{\beta}{\sigma} - \frac{k_N}{\sigma_N}\right| \leq C|\beta\sigma_N - k_N\sigma|$$
$$\leq C|\sigma - \sigma_N| + C|k_N - \beta| \leq C/\sqrt{N},$$

by the assumption $|\sigma - \sigma_N| \leq C/\sqrt{N}$ and $|k_N - \beta| \leq C/\sqrt{N}$ since $N = \lceil \lambda + \beta\sqrt{\lambda}\rceil$. Substituting the bounds on f_1 and f_2 in (5) and (4) concludes the proof of the theorem, namely

$$\frac{|\mathbb{P}[W_N = 0] - \mathbb{P}[W = 0]|}{\mathbb{P}[W = 0]} \leq e^{\frac{C}{\sqrt{N}}\sum_{n=1}^{\infty}(n^{-3/2}+n^{-1/2}e^{-n/C})} - 1$$
$$\leq e^{C/\sqrt{N}} - 1 \leq C/\sqrt{N}. \qquad \square$$

4 On the distribution of W

The distribution of W is determined, in principle, by either the following Spitzer's identity [11, p. 286]

$$\mathbb{E}e^{itW} = \exp\left\{\sum_{n=1}^{\infty}\frac{1}{n}\left(\mathbb{E}\left[e^{itS_n^+}\right] - 1\right)\right\},$$

or the Wiener-Hopf (ladder heights) method that can be found in [3, Ch.VII]. In this section we are mainly exploring two aspects of this distribution: its tail and its atom at the origin. The later is important, being the characterizing performance measure of the QED regime. The tail turns out to be exponential, with parameter that coincides with that of the exponential distribution that arised in *conventional* heavy-traffic [3, Section VIII.6]. However, in contrast to conventional heavy-traffic, W is *not* exponentially distributed. Primary reference on Gaussian random walk is [10].

4.1 The tail of W

The distribution of W is stochastically bounded by an exponential distribution with rate $2\beta/\sigma^2$. To see that, denote by $B(t)$ a Brownian motion with drift $-\beta$, variance coefficient

σ^2 and $B(0) = 0$. Then the following bound prevails:

$$\begin{aligned}\mathbb{P}[W > x] &= \mathbb{P}\left[\sup_{n\geq 0} B(n) > x\right] \\ &\leq \mathbb{P}\left[\sup_{t\geq 0} B(t) > x\right] = e^{-\frac{2\beta}{\sigma^2}x},\end{aligned}$$

where the last equality follows from [21, p. 15]. In general, note that W is the limit of a Lindley process in discrete time [3, pp. 80-81]. Consequently, $W = (W + S_1)^+$ in distribution (W and S_1 are taken independent), which can be used to show that W can not be exponential. However, it is straightforward to show that $-\frac{1}{x}\log\mathbb{P}[W > x] \to 2\beta/\sigma^2$ as $x \to \infty$. Indeed, the upper bound given above is complemented with the lower bound

$$\begin{aligned}\mathbb{P}[W > x] &\geq \mathbb{P}[S_{\lfloor x/\beta\rfloor} > x] \\ &= 1 - \Phi\left(\frac{x + \beta\lfloor x/\beta\rfloor}{\sigma\sqrt{\lfloor x/\beta\rfloor}}\right) \\ &= \frac{\sigma}{2\sqrt{2\pi\beta x}} e^{-\frac{2\beta}{\sigma^2}\sqrt{\frac{x}{x-\beta}}x}(1 + o(1)), \quad \text{as } x \to \infty.\end{aligned}$$

More precise analysis yields the exact asymptotics. To this end, Wald's likelihood ratio identity results in

$$\mathbb{P}[W > x] = \xi(\beta/\sigma, x/\sigma)e^{-\frac{2\beta}{\sigma^2}x}$$

with the expression for $\xi(\cdot,\cdot)$ given in [30, p. 13]. Then, in [10] it is shown that $\xi(\beta/\sigma, x/\sigma)$ converges to $\nu(\beta/\sigma)$, as $x \to \infty$, exponentially fast over $0 \leq \beta/\sigma \leq 2\sqrt{\pi}$, where

$$\nu(\beta/\sigma) = \exp\left\{\frac{\beta}{\sigma}\sqrt{\frac{2}{\pi}}\sum_{n=0}^{\infty}\frac{\zeta(\frac{1}{2} - n)}{n!(2n+1)}\left(\frac{-\beta^2}{2\sigma^2}\right)^n\right\},$$

and $\zeta(\cdot)$ is the Riemann zeta function. In fact, for all $\beta/\sigma > 0$, the exact asymptotics is of the form $\kappa e^{-\frac{2\beta}{\sigma^2}x}$ (when $0 < \beta/\sigma \leq 2\sqrt{\pi}$ one has $\kappa = \nu(\beta/\sigma)$), where κ is a constant that depends on the ladder height distributions, e.g. see Theorem 5.3 in [3, Ch. XII].

Finally, we note that there exists a body of literature on Gaussian random walks that is only tangentially related to our work here. Readers are referred to the following and references therein. Using the Wiener-Hopf factorization method in [25], the author explores the excess distribution of boundary crossing for Gaussian random walks. The sample path difference (instead of the difference in steady-state considered here) between a Brownian motion and its embedded Gaussian random walk was studied in [4]. Correction terms for the diffusion approximation to one- and two-barrier crossing problems were examined in [29]. In the context of option pricing, a corresponding relationship between the continuous and discrete-time models was investigated in [8, 9].

4.2 Approximation of the delay probability α

As seen from Corollary 1, the probability of wait is expressed in terms of an infinite sum of Gaussian functions. While the sum converges fast for moderately large β/σ (say $\beta/\sigma > 1$), it has particularly slow rate of convergence for small values of β/σ (for small β/σ the first $O(\sigma/\beta)$ elements in the sum behave roughly as $1/n$). Hence, it is of interest to derive a simple approximation for small values of β/σ, which we now do. Consider the infinite sum of Gaussian functions

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{1}{n}\Phi\left(-\frac{\beta\sqrt{n}}{\sigma}\right) &= \sum_{n=1}^{k}\frac{1}{n}\Phi\left(-\frac{\beta\sqrt{n}}{\sigma}\right) + \int_k^{\infty}\frac{1}{u}\Phi\left(-\frac{\beta\sqrt{u}}{\sigma}\right)\,du + O(k^{-1})\\ &= \frac{1}{2}\sum_{n=1}^{k}\frac{1}{n} + O(\beta\sqrt{k}/\sigma) + 2\int_{\beta\sqrt{k}/\sigma}^{\infty} z^{-1}\Phi\left(-z\right)\,dz + O(k^{-1}). \qquad (6)\end{aligned}$$

Next, we evaluate the last integral using integration by parts

$$\begin{aligned}\int_x^{\infty} z^{-1}\Phi\left(-z\right)\,dz &= -\Phi(-x)\log x + \int_x^{\infty}\frac{\log z}{\sqrt{2\pi}}e^{-z^2/2}\,dz\\ &= -\Phi(-x)\log x - \frac{\gamma+\log 2}{4} + \int_0^{x}\frac{\log z}{\sqrt{2\pi}}e^{-z^2/2}\,dz\\ &= -\frac{\log x}{2} - \frac{\gamma+\log 2}{4} + o(1), \quad \text{as } x\downarrow 0, \qquad (7)\end{aligned}$$

where γ is Euler's constant, the second equality follows from equation (3.481) in [18, p. 387] and the last equality is due to $x\log x \to 0$, as $x \downarrow 0$. By replacing the approximation $(\log k + \gamma)$ for the harmonic number and (7) in (6), one obtains

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{1}{n}\Phi\left(-\frac{\beta\sqrt{n}}{\sigma}\right) &= \frac{\log k}{2} + \frac{\gamma}{2} + o(1) + O(\beta\sqrt{k}/\sigma) - \frac{\gamma}{2} - \frac{\log 2}{2} - \log\frac{\beta\sqrt{k}}{\sigma} + O(k^{-1})\\ &= -\log\sqrt{2}\frac{\beta}{\sigma} + o(1), \quad \text{as } \beta/\sigma\downarrow 0.\end{aligned}$$

Finally, substituting the preceding expression in Corollary 1 yields

$$\alpha(\beta/\sigma) = 1 - \sqrt{2}\frac{\beta}{\sigma}(1+o(1)), \quad \text{as } \beta/\sigma\downarrow 0.$$

However, one can further refine the approximation. Indeed, Chang and Peres in [10] obtained the following expansion for $0 < \beta/\sigma \le 2\sqrt{\pi}$ (see their Theorem 1.1 and (5))

$$\alpha(\beta/\sigma) = 1 - \sqrt{2}\frac{\beta}{\sigma}\exp\left\{\frac{\beta/\sigma}{\sqrt{2\pi}}\sum_{n=0}^{\infty}\frac{\zeta(\frac{1}{2}-n)}{n!(2n+1)}\left(-\frac{\beta/\sigma}{2}\right)^n\right\},$$

where $\zeta(\cdot)$ is the Riemann zeta function. This expansion easily yields, as $\beta/\sigma \downarrow 0$,

$$\alpha(\beta/\sigma) = 1 - \sqrt{2}\frac{\beta}{\sigma} - \frac{\zeta(\frac{1}{2})}{\sqrt{\pi}}\left(\frac{\beta}{\sigma}\right)^2 - \frac{\zeta^2(\frac{1}{2})}{2\sqrt{2}\pi}\left(\frac{\beta}{\sigma}\right)^3 - \left(-\frac{\zeta(-\frac{1}{2})}{6\sqrt{\pi}} + \frac{\zeta^3(\frac{1}{2})}{12\pi\sqrt{\pi}}\right)\left(\frac{\beta}{\sigma}\right)^4 - o((\beta/\sigma)^4)$$

with $\zeta(\frac{1}{2}) \approx -0.5826\sqrt{2\pi}$ and $\zeta(-\frac{1}{2}) \approx -0.0829\sqrt{2\pi}$. In Figure 1 we provide a graphical illustration of the approximation.

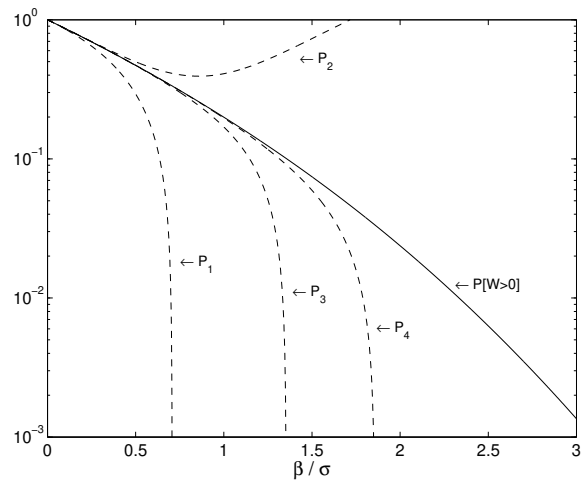

Figure 1: Approximations of the limiting probability of wait $\alpha(\beta/\sigma)$. The approximation P_k includes elements with up to the kth power of β/σ, i.e., $P_1 = 1 - \sqrt{2}\beta/\sigma$, etc.

5 Extensions and comparisons

In this concluding section we first discuss some possible extensions of our results and then compare them to others reported in the literature for related systems.

5.1 Interarrival times

Our primary result, Theorem 1, is easily generalizable. Indeed, in view of Theorem 1, p. 207 of [6], it applies to dependent (non-renewal) arrival sequences, as long as the process $\{nm\sqrt{N} - \sum_{i=-nN}^{1} \sqrt{N}\tau_{N,i}, n \geq 1\}$ converges in distribution, as $N \to \infty$, to an appropriate limit.

For Theorem 1 to hold, one needs the finiteness of the scaled second moment of the interarrival times, e.g., $\lambda_N \mathrm{Var}(\tau_N) \to \sigma^2 < \infty$. For infinite second moments, one may resort to Stable laws in order to obtain the appropriate scaling of the system. Due the higher variability in the arrival pattern, one needs more than "square-root" extra servers to get the probability of wait in $(0, 1)$. In particular, if for $\frac{1}{2} < H < 1$, $N \approx R_N + \beta(R_N)^H$, $\beta > 0$, and for a nondegenerate Y with a negative mean $-\beta$,

$$\frac{Nm - \sum_{i=1}^{N} N\tau_{N,i}}{mN^H} \Longrightarrow Y,$$

as $N \to \infty$, then the stationary waiting time W_N satisfies, as $N \to \infty$,

$$N^{1-H}\frac{W_N}{m} \Longrightarrow \sup_{n\geq 0}\sum_{i=1}^{n} Y_i,$$

where $\{Y_i\}$ are i.i.d. copies of Y. The proof closely follows the steps of the proof of Theorem 1. We omit the details and note that such dimensioning of the system, based on the infinite server approximation, is discussed in [33, Ch. 10].

5.2 Comparison of M/D/N and M/M/N

For a meaningful comparison, both systems serve customers with mean service requirements m arriving according to a Poisson process while operating in the QED regime, namely $\sqrt{N}(1-\rho_N) \to \beta$, as $N \to \infty$. We note that in the M/M/N case the waiting time satisfies [20]

$$\lim_{N\to\infty} \mathbb{P}[W_N^{\mathrm{M/M/N}} > 0] = \left(1 + \sqrt{2\pi}\beta\Phi(\beta)e^{\beta^2/2}\right)^{-1}. \tag{8}$$

Observe that the preceding relationship implies

$$\lim_{N\to\infty} \mathbb{P}[W_N > 0] = 1 - \sqrt{\pi/2}\beta + o(\beta),$$

as $\beta \downarrow 0$. Recalling the linear expansion for the M/D/N queue $1-\sqrt{2}\beta+o(\beta)$ (since $\sigma^2 = 1$), one concludes that in the latter the probability of delay decreases faster in the neighborhood of $\beta = 0$, as β increases. The probabilities of wait for the two systems, as a function of the parameter β, are shown in Figure 2. As seen, the two probabilities are quite close. To compare them in a more insightful way we define

$$\gamma(\beta) \triangleq \lim_{N\to\infty} \frac{\mathbb{P}[W_N^{\mathrm{M/M/N}} > 0]}{\mathbb{P}[W_N^{\mathrm{M/D/N}} > 0]},$$

numerically evaluate it by (8) and Corollary 1, and plot it in Figure 3. Observe that the ratio is bounded by 1.15 for all positive β. Using the well known approximation $\Phi(-x) = (\sqrt{2\pi}x)^{-1}e^{-x^2/2}(1+o(1))$, as $x \to \infty$, one can easily show that $\gamma(\beta) \to 1$ when $\beta \to \infty$. The same limit trivially holds for $\beta \to 0$.

To compare the expected waiting times we note that, in the M/M/N system, the wait, given that it is positive, is exponentially distributed with parameter $N(1-\rho_N)/m$, implying

$$\mathbb{E}\left[\sqrt{N}W_N^{\mathrm{M/M/N}} \middle| W_N^{\mathrm{M/M/N}} > 0\right] \to m\beta^{-1}, \quad \text{as } N \to \infty.$$

Similarly to the case of the probabilities of wait, we also consider the ratio of the two quantities given by

$$\eta(\beta) = \lim_{N\to\infty} \frac{\mathbb{E}[W_N^{\mathrm{M/M/N}} \mid W_N^{\mathrm{M/M/N}} > 0]}{\mathbb{E}[W_N^{\mathrm{M/D/N}} \mid W_N^{\mathrm{M/D/N}} > 0]}.$$

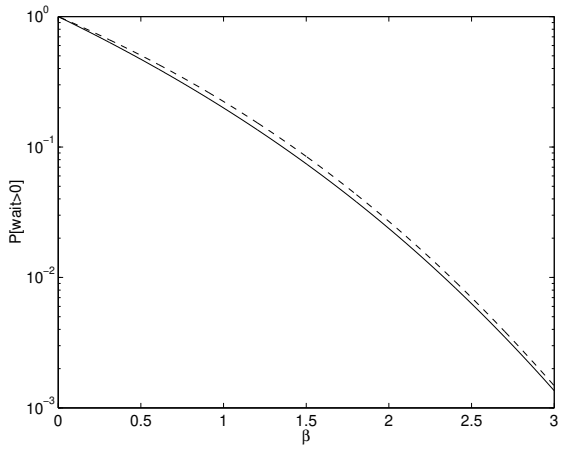

Figure 2: Limiting probability of wait in the corresponding M/D/N (solid line) and M/M/N (dashed line) queues, as a function of the parameter β.

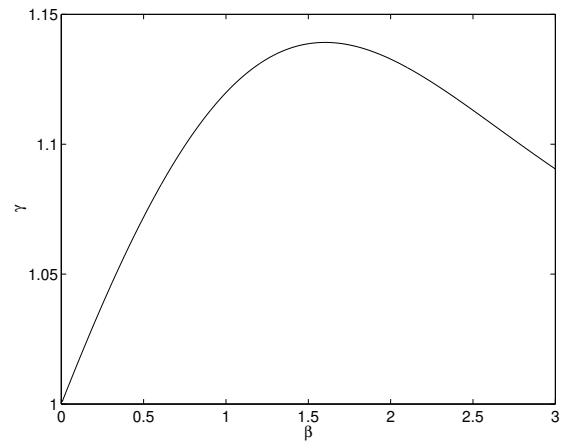

Figure 3: Ratio of limiting probabilities of wait in the corresponding M/M/N and M/D/N queues, as a function of the parameter β.

The numerical findings are shown in Figures 4 and 5; they are consistent with the simulation study [27]. By using the previously mentioned approximation for $\Phi(\cdot)$, one can get that the ratio of two expected delays tends to 1 as $\beta \to \infty$. Applying L'Hospital's rule twice and using elementary, but somewhat tedious, calculations one can also verify that

$$\lim_{\beta \downarrow 0} \eta(\beta) = \lim_{\beta \downarrow 0} \left(\sum_{n=1}^{\infty} \beta^3 \sqrt{\frac{n}{8\pi}} e^{-\frac{\beta^2 n}{2}} \right)^{-1} = 2.$$

5.3 Comparison of GI/D/N and GI/D/1

Finally, we compare $W_N^{\mathrm{GI/D/N}}$ with the waiting time of the GI/D/1 queue in conventional heavy traffic. Denote by $B(t)$ a Brownian motion with drift $-\beta$, variance coefficient σ^2 and $B(0) = 0$. When the capacity of the GI/D/1 queue is N (service duration equals $1/N$) and its utilization is $1 - \beta/\sqrt{N}$, the stationary waiting time $W_N^{\mathrm{GI/D/1}}$ satisfies [33, Ch. 9], as $N \to \infty$,

$$\sqrt{N} \frac{W_N^{\mathrm{GI/D/1}}}{m} \Longrightarrow \sup_{t \geq 0} B(t).$$

Therefore, the preceding limit and the first part of Section 4 imply that the waiting time in the multi server queue is stochastically smaller than the exponential one in the corresponding single server queue. However, one must keep in mind that the sojourn time in the GI/D/N system is $O(1)$ while that in the single server case is only $O(1/\sqrt{N})$.

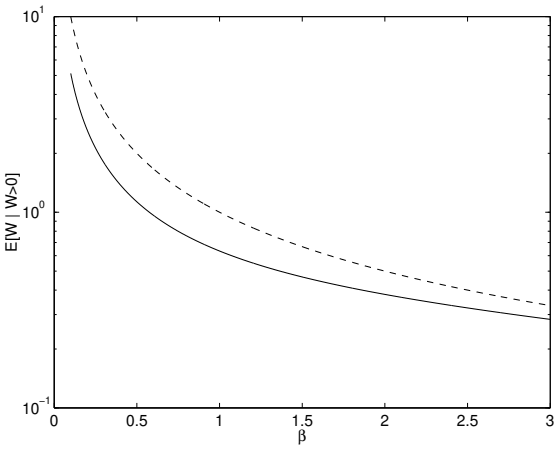

Figure 4: Limiting expected wait given wait in the corresponding M/D/N (solid line) and M/M/N (dashed line) queues, as a function of the parameter β.

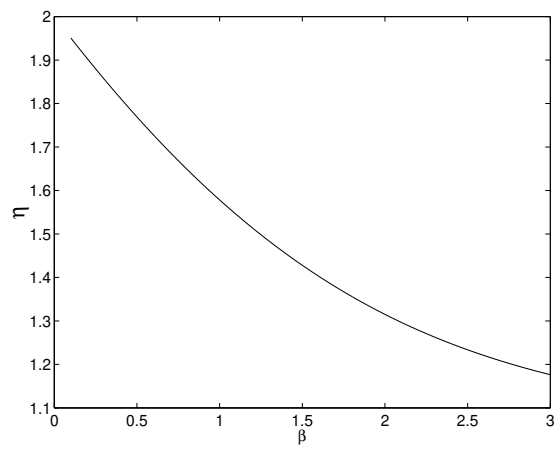

Figure 5: Ratio of limiting expected waits given wait in the corresponding M/M/N and M/D/N queues, as a function of the parameter β.

Acknowledgement

We thank Søren Asmussen, Paul Glasserman, Vladimir Lotov, Gennady Samorodnitsky and David Siegmund for providing us with some pointers to the literature. We are also grateful to an anonymous reviewer for a careful reading and helpful detailed comments.

References

[1] M. Armony and C. Maglaras. On customer contact centers with a call-back option: Customer decisions, sequencing rules, and system design. *Oper. Res.*, 52(2):271–292, 2004.

[2] M. Armony and C. Maglaras. Contact centers with a call-back option and real-time delay information. *Oper. Res.*, to appear.

[3] S. Asmussen. *Applied Probability and Queues.* Wiley, 1987.

[4] S. Asmussen, P. Glynn, and J. Pitman. Discretization error in simulation of one-dimensional reflecting Brownian motion. *Ann. Appl. Probab.*, 5(4):875–896, 1985.

[5] P. Billingsley. *Probability and Measure.* Wiley, 3rd edition, 1995.

[6] A. A. Borovkov. *Asymptotic Methods in Queueing Theory.* John Wiley & Sons, 1984.

[7] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning of large call centers. *Oper. Res.*, 52(1):17–34, 2004.

[8] M. Broadie, P. Glasserman, and S. Kou. A continuity correction for discrete barrier options. *Mathematical Finance*, 7(4):325–349, 1997.

[9] M. Broadie, P. Glasserman, and S. Kou. Connecting discrete and continuous path-dependent options. *Finance and Stochastics*, 3:55–82, 1999.

[10] J. Chang and Y. Peres. Ladder heights, Gaussian random walks, and the Riemann zeta function. *Ann. Probab.*, 25:787–802, 1997.

[11] K. L. Chung. *A Course in Probability Theory.* Academic Press, 2nd edition, 1974.

[12] A.K. Erlang. On the rational determination of the number of circuits. In E. Brockmeyer, H.L. Halstrom, and A. Jensen, editors, *The life and works of A.K. Erlang.* The Copenhagen Telephone Company, Copenhagen, 1948.

[13] P. Fleming, A. Stolyar, and B. Simon. Heavy traffic limit for a mobile phone system loss model. In *Proc. of 2nd Int'l Conf. on Telecomm. Syst. Mod. and Analysis*, Nashville, TN, 1994.

[14] G.J. Franx. A simple solution for the M/D/c waiting time distribution. *Oper. Res. Letters*, 29(5):221–229, 2001.

[15] G.J. Franx. The transient M/D/c queueing system. Preprint, available at http://www.cs.vu.nl/∼franx/, 2002.

[16] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5(2):79–141, 2003.

[17] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing and Service Operations Management*, 4(3):208–227, 2002.

[18] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals Series and Products.* Academic Press, 5th edition, 1994.

[19] R. Haji and G. Newell. A relationship between stationary queue and waiting time distributions. *J. Appl. Probab.*, 8:617–620, 1971.

[20] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29(3):567–588, 1981.

[21] J.M. Harrison. *Brownian Motion and Stochastic Flow Systems.* Wiley, 1985.

[22] V. Iversen. Decomposition of an M/D/rk queue with FIFO into k E$_k$/D/r queues with FIFO. *Oper. Res. Letters*, 2(1):20–21, 1983.

[23] D. Jagerman. Some properties of the Erlang loss function. *Bell System Techn. J.*, 53(3):525–551, 1974.

[24] J. Kiefer and J. Wolfowitz. On the theory of queues with many servers. *Trans. Amer. Math. Soc.*, 78:1–18, 1955.

[25] V. Lotov. On some boundary crossing problems for Gaussian random walks. *Ann. Probab.*, 24(4):2154–2171, 1996.

[26] C. Maglaras and A. Zeevi. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science*, 49(8):1018–1038, 2003.

[27] A. Mandelbaum and R. Schwartz. Simulation experiments with M/G/100 queues in the Halfin-Whitt (Q.E.D) regime. Technical report, Technion, 2002. http://iew3.technion.ac.il/serveng/References/references.html.

[28] A. Puhalskii and M. Reiman. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Probab.*, 32(3):564–595, 2000.

[29] D. Siegmund. Corrected diffusion approximations in certain random walk problem. *Adv. Appl. Probab.*, 11:701–719, 1979.

[30] D. Siegmund. *Sequential analysis: Tests and Confidence Intervals.* Springer, 1985.

[31] W. Whitt. Existence of limiting distributions in the GI/G/s queue. *Math. Oper. Res.*, 7(1):88–94, 1982.

[32] W. Whitt. Heavy traffic approximations for service systems with blocking. *AT&T Bell Lab. Tech. J.*, 63:689–708, 1984.

[33] W. Whitt. *Stochastic-Process Limits.* Springer, 2002.

[34] W. Whitt. A diffusion approximation for the G/GI/n/m queue. *Oper. Res.*, to appear.

[35] W. Whitt. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.*, to appear.

[36] R. Wolff. An upper bound for multi-channel queues. *J. Appl. Probab.*, 14:884–888, 1977.