Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Heavy-Traffic Asymptotics for the M/Gr/1 Quorum Queue: Exponential Limit and Second-Order Mean Correction

Lotfi Tadj

ltadj@alfaisal.edu

Alfaisal University

Research Article

Keywords: Quorum queue, bulk service, heavy traffic, second-order asymptotics, characteristic roots, probability-generating function

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11729742/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Heavy-Traffic Asymptotics for the $M/G^r/1$ Quorum Queue: Exponential Limit and Second-Order Mean Correction

Lotfi Tadj
College of Engineering and Advanced Computing, Alfaisal University
Riyadh, Saudi Arabia
ltadj@alfaisal.edu

Abstract

We study the heavy-traffic behavior of the single-server $M/G^r/1$ quorum queue, in which each service removes exactly r customers and the server waits for r customers to accumulate before beginning service. Assuming non-degenerate service times with finite second moment, we show that the stationary queue length, scaled by $1-\rho$, converges in distribution to an exponential random variable with rate $2/(1+rC_s^2)$, where C_s^2 is the squared coefficient of variation of the service time; the same limit holds at departure epochs and at arbitrary times. The leading term of the stationary mean, $(1+rC_s^2)/(2(1-\rho))$, is consistent with Kingman's heavy-traffic theorem for reflected random walks. Our main contribution is the next term. Starting from the exact root-based representation of the stationary mean at departure epochs, we prove that the roots of the characteristic equation inside the unit disk converge, as $\rho \uparrow 1$, to the interior roots of the critical-load characteristic equation, and we obtain, under the same second-moment assumption, the two-term expansion $L_\rho^+ = (1+rC_s^2)/(2(1-\rho))+c_0+O(1-\rho)$, with $c_0 = -rC_s^2 + \sum_{j<r} 1/(1-z_j(1))$ determined by those critical-load roots; the arbitrary-time mean differs by the constant $(r-1)/2$. For exponential service the correction is $c_0 = -(r+2)/3$, and for Erlang service it is determined by the exterior critical-load roots. Under a third-moment assumption the expansion extends to the Laplace transform of the scaled queue length at departure epochs, uniformly in the transform variable on compact sets, with a correction term whose coefficients are c_0 and a constant involving the third moment, and with a remainder of order $o(1-\rho)$, or $O((1-\rho)^2)$ under a fourth moment; the arbitrary-time version follows by a shift of the constants. The constant c_0 is not accessible from the weak limit: it depends on the full service-time distribution through the critical-load roots, not only on its first two moments, and it is precisely the quantity that the characteristic-root solution of the quorum queue makes computable.

Keywords: Quorum queue; bulk service; heavy traffic; second-order asymptotics; characteristic roots; probability-generating function.

Mathematics Subject Classification (2020): 60K25 (primary); 60F05, 90B22 (secondary).

1 Introduction

Bulk-service queues model systems in which customers are served in groups: shuttles and elevators, batch processing in manufacturing, group transmissions in communication networks. The quorum discipline is the simplest of these: a fixed number r of customers is served at a time, and if fewer than r are present when the server becomes free, it waits until the quorum is reached.

The resulting $M/G^r/1$ model is a special case of the general bulk-service queue $M/G^{a,b}/1$ with $a = b = r$; see Chaudhry and Templeton (1983) and Medhi (1984) for the classical theory and Chaudhry and Templeton (1981) for the $M/G^B/1$ system, in which the server takes up to B customers and so serves batches of variable size; the fixed-size quorum rule studied here is different, and none of the results we use are taken from that model.

The exact stationary analysis of the $M/G^r/1$ queue is in terms of the roots of the characteristic equation $z^r = \widetilde{S}(\lambda(1-z))$. Dshalalow and Tadj (1992) obtained the stationary probability-generating function (pgf), and Tadj (2003) gave explicit, transform-free expressions for the stationary probabilities and for the stationary means at departure epochs and at arbitrary times, using the roots inside and outside the unit disk. The mean at departure epochs takes the form

$$L_\rho^+ = \rho r + \frac{\lambda^2 m_2 - r(r-1)}{2r(1-\rho)} + \sum_{j=1}^{r-1} \frac{1}{1 - z_j(\rho)},$$

where $z_1(\rho), \ldots, z_{r-1}(\rho)$ are the non-unit characteristic roots in the unit disk. This representation is exact for every $\rho < 1$, but it does not by itself say how the queue behaves as $\rho \uparrow 1$: the first fraction diverges, and the behavior of the root sum near critical loading is not evident. The present paper carries out the heavy-traffic analysis in two steps, of which the second is the main contribution.

The first step is a distributional limit. We show (Theorem 3.1) that $(1-\rho)N_\rho$ converges in distribution to an exponential random variable with rate $2/(1+rC_s^2)$, where N_ρ is the stationary queue length at departure epochs and C_s^2 is the squared coefficient of variation of the service time, and (Corollary 4.1) that the queue length at an arbitrary time has the same limit. The proof is short and self-contained: it rests on the exact pgf and on a boundary identity $\sum_{j<r}(r-j)\pi_j = r(1-\rho)$ (Lemma 3.1). We note at once that this limit is not surprising. The embedded departure-epoch process is the reflected random walk $N_{n+1} = (N_n - r)^+ + A_n$, with A_n the number of arrivals during a service, whose increments have mean $-r(1-\rho)$ and variance tending to $r(1+rC_s^2)$; Kingman's (1961, 1962) heavy-traffic theorem for such walks predicts an exponential limit with rate equal to twice the drift divided by the variance, which is exactly $2/(1+rC_s^2)$. Theorem 3.1 therefore confirms that the quorum boundary, the states $0, \ldots, r-1$ in which the server idles, does not alter the first-order behavior. Its role in this paper is instrumental: it is obtained from the exact solution, and the same machinery can then be pushed one order further.

The second step, and the main contribution of the paper, is that further order. We show (Lemma 5.4) that the interior characteristic roots $z_j(\rho)$ remain bounded away from 1 as $\rho \uparrow 1$, and in fact converge to the interior roots $z_j(1)$ of the critical-load equation $z^r = \widetilde{S}(\frac{r}{m_1}(1-z))$, which has a double root at $z = 1$. This yields, with no assumption on the service-time distribution beyond $S > 0$ and $E[S^2] < \infty$, the two-term expansion (Theorem 6.1)

$$L_\rho^+ = \frac{1 + rC_s^2}{2(1-\rho)} + c_0 + O(1-\rho), \qquad c_0 = -rC_s^2 + \sum_{j=1}^{r-1} \frac{1}{1 - z_j(1)},$$

with the corresponding statement for the arbitrary-time mean. The leading term depends on the service-time distribution only through C_s^2, as in Kingman's theorem. The constant c_0 does not: it depends on the full distribution through the critical-load roots, and it cannot be obtained from the weak limit or from a diffusion approximation. It is not even of a fixed sign: $c_0 < 0$ for exponential service, so that the first-order approximation overestimates the mean, while $c_0 > 0$ for sufficiently regular (Erlang) service, where it underestimates it. For exponential

service the critical-load roots are the zeros of $1+2z+\cdots+rz^{r-1}$, and $c_0=-(r+2)/3$ in closed form (Corollary 6.1); for $r=1$ this reproduces the exact $M/M/1$ mean. The same expansion carries over to the Laplace transform of the scaled queue length: under a finite third moment, the transform of $(1-\rho)N_\rho$ equals $(1+\gamma u)^{-1}-(1-\rho)(c_0u+du^2)/(1+\gamma u)^2+o(1-\rho)$, with $\gamma=(1+rC_s^2)/2$ and an explicit constant d, the remainder improving to $O((1-\rho)^2)$ under a fourth moment (Theorem 7.1); the mean expansion is formally its derivative at $u=0$. The characteristic-root solution of Tadj (2003) is thus not only an exact finite-ρ tool but also the natural vehicle for second-order heavy-traffic asymptotics of bulk-service queues.

We are not aware of a heavy-traffic limit, let alone a second-order expansion, for the fixed-size quorum queue; in particular the monographs of Chaudhry and Templeton (1983) and Medhi (1984) treat the stationary analysis of bulk-service queues but not their critical-load asymptotics. Diffusion approximations for bulk queues have been proposed, notably by Chiamsiri and Leonard (1981). These replace the queue-length process by a diffusion with matched drift and variance and read off its stationary law; they do not establish convergence of the scaled stationary distribution, and by construction they carry no information beyond the first two moments of the increments, so they cannot produce the $O(1)$ term. General accounts of heavy-traffic limits are given by Whitt (2002) and Asmussen (2003). For single-service queues, second-order heavy-traffic corrections are classical. For the Lindley recursion, Siegmund (1979) showed that the $O(1)$ correction to the exponential approximation of the stationary waiting time is the mean overshoot of the ascending ladder height of the random walk, a renewal-theoretic quantity, and Blanchet and Glynn (2006) extended this to a complete expansion in powers of the drift; Knessl (1990) obtained higher-order terms for the $M/G/1$ and $GI/G/1$ workload by matched asymptotic expansions. The present paper obtains a second-order correction, with proof, for a bulk-service queue whose embedded chain is not a Lindley recursion, and identifies the characteristic roots at critical loading as the carriers of the second-order information. Whether c_0 admits a ladder-height interpretation of Siegmund's type for the walk with increments A_n-r and boundary set $\{0,\dots,r-1\}$ is an interesting question that we leave open; the root representation sidesteps the overshoot computation entirely.

The remainder of the paper is organized as follows. Section 2 describes the model and recalls the stationary pgf representation. Section 3 establishes the boundary identity and proves the departure-epoch heavy-traffic limit. Section 4 derives the arbitrary-time limit. Section 5 studies the characteristic roots near critical loading. Section 6 derives the two-term mean expansion, its exponential-service closed form, and its reduction to exterior roots for Erlang service, with numerical checks of the constant and of the accuracy of the approximation at moderate traffic. Section 7 extends the expansion to the Laplace transform, uniformly in the transform variable, and to a class of smooth functionals. Section 8 concludes.

2 Model and stationary pgf representation

Consider a single-server queue with Poisson arrivals at rate $\lambda>0$. Customers arrive individually. Service times are independent and identically distributed random variables with generic service time S, independent of the arrival process. We assume throughout that

$$P(S>0)>0, \qquad m_1=E[S]\in(0,\infty), \qquad m_2=E[S^2]<\infty.$$

The first condition only excludes the degenerate case $S\equiv 0$; it is what makes $|F_\rho(z)|<1$ for $|z|=1$, $z\neq 1$ (Lemmas 5.1 and 5.4), and an atom of S at 0 is harmless everywhere else. We

state results for "$S > 0$ almost surely" only where that phrase appears in the literature we compare with; $P(S > 0) > 0$ suffices throughout.

The server operates according to a fixed-size quorum discipline. Each service completion removes exactly r customers, where $r \geq 1$ is a fixed integer. If fewer than r customers remain after a service completion, the server waits until the queue contains at least r customers. A new service then begins.

Let N_n denote the number of customers immediately after the n-th service completion, and A_n the number of arrivals during the service time following the n-th completion. Since arrivals are Poisson and service times are independent of the arrival process, the conditional distribution of A_n given the service time is Poisson with parameter λS, and its pgf is

$$F_\rho(z) = E[z^{A_n}] = E\big[e^{-\lambda(1-z)S}\big] = \widetilde{S}(\lambda(1-z)),$$

where $\widetilde{S}$ denotes the Laplace–Stieltjes transform of the service-time distribution.

The embedded departure-epoch process satisfies

$$N_{n+1} = (N_n - r)^+ + A_n = N_n - \min(N_n, r) + A_n. \tag{2.1}$$

The traffic intensity is

$$\rho = \frac{\lambda m_1}{r}, \tag{2.2}$$

and we consider the stable case $0 < \rho < 1$.

For $N_n \geq r$ the conditional drift of the embedded chain is $E[N_{n+1} - N_n \mid N_n] = \lambda m_1 - r = -r(1-\rho) < 0$. The increments have finite first moment. Under the stated assumptions the embedded chain is irreducible and aperiodic, and the negative-drift condition outside the finite set $\{0, \ldots, r-1\}$ ensures positive recurrence; a stationary distribution therefore exists. Let $\pi_j = P(N_n = j)$, $j \geq 0$, denote its stationary probabilities, and define the stationary pgf and the boundary polynomial

$$P_\rho(z) = \sum_{j=0}^{\infty} \pi_j z^j, \qquad H_\rho(z) = \sum_{j=0}^{r-1} \pi_j z^j.$$

The stationary pgf satisfies

$$P_\rho(z) = \frac{F_\rho(z)\,[z^r H_\rho(1) - H_\rho(z)]}{z^r - F_\rho(z)}. \tag{2.3}$$

This is the established departure-epoch pgf representation for the fixed-size quorum queue; see Dshalalow and Tadj (1992) and Tadj (2003). We use it as the starting point for the heavy-traffic analysis.

Throughout the paper, write

$$C_s^2 = \frac{m_2 - m_1^2}{m_1^2} \tag{2.4}$$

for the squared coefficient of variation of the service-time distribution. Since $\lambda = r\rho/m_1$, the quantity appearing in the exact mean formula admits the exact polynomial expression

$$\lambda^2 m_2 - r(r-1) = r^2\rho^2(1 + C_s^2) - r(r-1), \tag{2.5}$$

which will be used repeatedly.

3 Departure-epoch heavy-traffic limit

We first establish an identity involving the stationary probabilities at the boundary states.

Lemma 3.1. *For $0 < \rho < 1$, the stationary probabilities satisfy*

$$\sum_{j=0}^{r-1}(r-j)\pi_j = r(1-\rho). \tag{3.1}$$

Proof. From (2.1), $N_{n+1}-N_n = A_n - \min(N_n, r)$, and $E|N_{n+1}-N_n| \le E[A_n]+r = \lambda m_1 + r < \infty$. Since the embedded chain is stationary, N_{n+1} and N_n have the same distribution. At this point $E[N_n]$ is not yet known to be finite (it will follow, a posteriori, from Theorem 6.1), so we cannot simply take expectations in the recursion; instead we use the following elementary equal-marginals argument, which only requires $E|N_{n+1} - N_n| < \infty$. If X and Y are identically distributed nonnegative random variables with $E|Y - X| < \infty$, put $\phi_M(x) = \min(x, M)$. Then $E[\phi_M(Y) - \phi_M(X)] = 0$ and $|\phi_M(Y) - \phi_M(X)| \le |Y - X|$, so dominated convergence as $M \to \infty$ gives $E[Y - X] = 0$.

Applying this to N_n and N_{n+1} yields $0 = E[A_n] - E[\min(N_n, r)]$, and since $E[A_n] = \lambda m_1 = r\rho$,

$$E[\min(N_n, r)] = r\rho.$$

On the other hand, using $\sum_{j\ge 0}\pi_j = 1$,

$$E[\min(N_n, r)] = \sum_{j=0}^{r-1} j\pi_j + r\sum_{j=r}^{\infty}\pi_j = r - \sum_{j=0}^{r-1}(r-j)\pi_j.$$

Equating the two expressions proves (3.1). □

Identity (3.1) is the same as the relation $H_\rho(1)\prod_{j<r}(1 - z_j(\rho)) = r(1-\rho)$ used in the exact solution of Tadj (2003), since both express $\frac{d}{dz}\left[z^r H_\rho(1) - H_\rho(z)\right]_{z=1}$. An immediate consequence is that, for every $j = 0, \dots, r-1$,

$$\pi_j = O(1-\rho), \qquad \rho \uparrow 1. \tag{3.2}$$

The proofs of this section and of Section 5 use Taylor expansions of the pgf F_ρ at $z = 1$, for complex z in the closed unit disk, under moment assumptions only. The following elementary lemma records what is needed.

Lemma 3.2. *Let $p \ge 1$ be an integer and suppose $E[S^p] < \infty$. Then, with $\delta = z - 1$,*

$$F_\rho(1+\delta) = \sum_{k=0}^{p}\frac{F_\rho^{(k)}(1)}{k!}\,\delta^k + o(|\delta|^p), \qquad |1+\delta| \le 1,\ \delta \to 0,$$

where $F_\rho^{(k)}(1) = E[A_n(A_n-1)\cdots(A_n-k+1)]$ are the factorial moments of A_n (finite, since $E[A_n^p] < \infty$), and the $o(|\delta|^p)$ is uniform in $\rho \in (0,1]$ and in the argument of δ, in the following quantitative sense: there is a nondecreasing function $\omega : (0,\infty) \to [0,\infty)$ with $\omega(t) \to 0$ as $t \downarrow 0$, depending on p and on the law of S but not on ρ, such that

$$\left|F_\rho(1+\delta) - \sum_{k=0}^{p}\frac{F_\rho^{(k)}(1)}{k!}\,\delta^k\right| \le |\delta|^p\,\omega(|\delta|), \qquad |1+\delta| \le 1,\ \rho \in (0,1]. \tag{3.3}$$

If moreover $E[S^{p+1}] < \infty$, the remainder is $O(|\delta|^{p+1})$, uniformly in $\rho \in (0,1]$ and in the argument of δ.

Proof. Write $F_\rho(z) = \sum_{n\ge 0} q_n(\rho) z^n$ with $q_n(\rho) = P(A_n = n)$. For $|z| \le 1$ and every n, two bounds hold for the Taylor remainder of z^n at 1:

$$\Big|z^n - \sum_{k=0}^{p} \binom{n}{k}\delta^k\Big| \le \binom{n}{p+1}|\delta|^{p+1}, \qquad \Big|z^n - \sum_{k=0}^{p} \binom{n}{k}\delta^k\Big| \le 2\binom{n}{p}|\delta|^{p},$$

the first because the $(p+1)$-st derivative of z^n is bounded by $n(n-1)\cdots(n-p)$ on the convex closed disk, the second by the same argument one order lower combined with $|\binom{n}{p}\delta^p| \le \binom{n}{p}|\delta|^p$. Summing against $q_n(\rho)$,

$$\Big|F_\rho(1+\delta) - \sum_{k=0}^{p} \frac{F_\rho^{(k)}(1)}{k!}\delta^k\Big| \le |\delta|^p \sum_{n\ge 0} q_n(\rho) \min\Big\{\binom{n}{p+1}|\delta|,\, 2\binom{n}{p}\Big\}.$$

The right-hand side depends on δ only through $|\delta|$, which gives the uniformity in the argument of δ: the segment from 1 to z lies in the closed disk by convexity, and the derivative bounds used above hold on the whole disk, so nothing in the estimate distinguishes real from complex δ.

We next show that the sum on the right tends to 0 as $\delta \to 0$, uniformly in $\rho \in (0,1]$. Two ingredients are needed. First, a monotonicity property: conditionally on S, A_n is Poisson with parameter $r\rho S/m_1$, and the Poisson family is stochastically increasing in its parameter; conditioning on S and mixing, there is a coupling of the laws of A_n at traffic intensities $\rho \le 1$ and 1 under which the former is pointwise no larger than the latter. Consequently, for every function h that is nondecreasing on the nonnegative integers,

$$\sum_{n\ge 0} h(n)\, q_n(\rho) \le \sum_{n\ge 0} h(n)\, q_n(1), \qquad \rho \in (0,1]. \tag{3.4}$$

Second, $\sum_n \binom{n}{p} q_n(1) < \infty$, because $E[A_n^p] \le c_p\big(1 + (r/m_1)^p E[S^p]\big) < \infty$ at $\rho = 1$.

Now fix $\eta > 0$. Apply (3.4) with $h(n) = \binom{n}{p}\mathbf{1}\{n > M\}$, which is nondecreasing since $\binom{n}{p}$ is nondecreasing in n: there is M such that $2\sum_{n>M}\binom{n}{p}q_n(\rho) \le 2\sum_{n>M}\binom{n}{p}q_n(1) < \eta/2$ for all $\rho \in (0,1]$ (the tail). For $n \le M$ use the first bound in the minimum: $\sum_{n\le M} q_n(\rho)\binom{n}{p+1}|\delta| \le \binom{M}{p+1}|\delta| < \eta/2$ once $|\delta| < \delta_0(\eta, M)$, for all ρ (the body). Hence the sum is less than η for $|\delta| < \delta_0$ and every $\rho \in (0,1]$. Define

$$\omega(t) := \sup_{\rho\in(0,1]} \sum_{n\ge 0} q_n(\rho) \min\Big\{\binom{n}{p+1} t,\; 2\binom{n}{p}\Big\}, \qquad t > 0.$$

Then ω is finite (each sum is at most $2\sum_n \binom{n}{p} q_n(1)$ by (3.4)), nondecreasing in t because each term of the sum is, and what was just shown is precisely that $\omega(t) \to 0$ as $t \downarrow 0$; with this ω the bound displayed above is (3.3). If $E[S^{p+1}] < \infty$, the first bound alone gives $|\delta|^{p+1}\sum_n \binom{n}{p+1} q_n(\rho) \le |\delta|^{p+1}\sum_n \binom{n}{p+1} q_n(1)$ by (3.4), which is $O(|\delta|^{p+1})$ uniformly in ρ. □

The lemma is a statement about the family of pgfs F_ρ alone; it uses nothing about the characteristic roots. The logical order of the root analysis in Section 5 is then linear: Lemma 3.2 gives the uniform expansion (3.8); this gives the root count below critical load (Lemma 5.1) and the separation Lemma 5.2; and these two are used in the critical-load Lemma 5.4.

Theorem 3.1. *Suppose that $S > 0$ almost surely and $E[S^2] < \infty$. Let N_ρ have the stationary departure-epoch distribution of the $M/G^r/1$ quorum queue. Then, as $\rho \uparrow 1$,*

$$(1-\rho)N_\rho \;\Rightarrow\; \mathrm{Exp}\left(\frac{2}{1+rC_s^2}\right), \tag{3.5}$$

where $\mathrm{Exp}(\gamma)$ *denotes the exponential distribution with rate* γ.

Proof. Put $\varepsilon = 1-\rho$ and $B_\rho(z) = z^r H_\rho(1) - H_\rho(z)$. Since $B_\rho(1) = 0$,

$$B_\rho'(1) = rH_\rho(1) - H_\rho'(1) = \sum_{j=0}^{r-1}(r-j)\pi_j = r\varepsilon$$

by Lemma 3.1. Because the polynomial B_ρ has fixed degree and all its coefficients are $O(\varepsilon)$ by (3.2), its second derivative is uniformly $O(\varepsilon)$ in a neighborhood of 1. Consequently, as $\delta \to 0$,

$$B_\rho(1+\delta) = r\varepsilon\delta + O(\varepsilon\delta^2). \tag{3.6}$$

Next, by Lemma 3.2 with $p = 2$, since $E[A_n^2] = \lambda m_1 + \lambda^2 m_2 < \infty$,

$$F_\rho(1+\delta) = 1 + \lambda m_1\delta + \frac{\lambda^2 m_2}{2}\delta^2 + o(|\delta|^2), \qquad |1+\delta| \le 1,\ \delta \to 0, \tag{3.7}$$

with the $o(|\delta|^2)$ uniform in $\rho \in (0,1]$. (In this section only real $\delta \le 0$ is needed; the complex version is used in Section 5.) Since $(1+\delta)^r = 1 + r\delta + \frac{r(r-1)}{2}\delta^2 + O(\delta^3)$, combining with (3.7) and $\lambda m_1 = r\rho$ yields

$$(1+\delta)^r - F_\rho(1+\delta) = r(1-\rho)\delta - \frac{\lambda^2 m_2 - r(r-1)}{2}\delta^2 + o(|\delta|^2). \tag{3.8}$$

Furthermore, by (2.5),

$$\lim_{\rho\uparrow 1}\left[\lambda^2 m_2 - r(r-1)\right] = r^2(1+C_s^2) - r(r-1) = r(1+rC_s^2). \tag{3.9}$$

Fix $u > 0$ and set $z_\rho = e^{-u\varepsilon}$, so that $\delta_\rho = z_\rho - 1 = -u\varepsilon + O(\varepsilon^2)$. Using (3.6),

$$B_\rho(z_\rho) = -ru\varepsilon^2 + o(\varepsilon^2), \tag{3.10}$$

and (3.8)–(3.9) give

$$z_\rho^r - F_\rho(z_\rho) = -r\Big[u + \frac{1+rC_s^2}{2}u^2\Big]\varepsilon^2 + o(\varepsilon^2). \tag{3.11}$$

Also $F_\rho(z_\rho) \to 1$. Substitution into (2.3) therefore gives

$$\lim_{\rho\uparrow 1} P_\rho\big(e^{-u(1-\rho)}\big) = \frac{u}{u + \frac{1+rC_s^2}{2}u^2} = \frac{1}{1 + \frac{1+rC_s^2}{2}u}. \tag{3.12}$$

The right-hand side is the Laplace transform of an exponential random variable with rate $2/(1+rC_s^2)$. Since the Laplace transforms converge to that of a proper probability distribution, the continuity theorem for Laplace transforms establishes (3.5). $\square$

Remark 3.1 (Consistency with Kingman's theorem). The embedded chain (2.1) is a random walk with increments $A_n - r$, reflected at a boundary layer of width r. Kingman's heavy-traffic theorem (Kingman, 1961, 1962; see Asmussen, 2003, Ch. X, Sect. 7, for a textbook statement) asserts that for the Lindley recursion $W_{n+1} = (W_n + U_n)^+$ driven by i.i.d. increments with mean $-\mu < 0$ and variance σ^2, the stationary distribution of μW converges, as $\mu \downarrow 0$ with σ^2 fixed, to an exponential law with rate $2/\sigma^2$. Here $E[A_n - r] = -r(1-\rho)$ and $\mathrm{Var}(A_n) = \lambda m_1 + \lambda^2 \mathrm{Var}(S) \to r + r^2 C_s^2 = r(1 + rC_s^2)$, so the rate predicted for $r(1-\rho)N_\rho$ is $2/(r(1+rC_s^2))$, i.e. rate $2/(1+rC_s^2)$ for $(1-\rho)N_\rho$, in agreement with (3.5). The theorem is not directly applicable, since (2.1) reflects at the set $\{0, \ldots, r-1\}$ rather than at 0, and the agreement is a consistency check rather than an alternative proof. Theorem 3.1 thus confirms that the quorum boundary does not affect the leading-order scaled limit (it does, of course, affect the lower-order terms studied in Sections 6 and 7); its independent proof from the exact pgf is what allows the second-order analysis of Section 6.

4 Arbitrary-time queue-length limit

Let Q_ρ denote the stationary number of customers in the system at an arbitrary time. The exact relationship between the stationary arbitrary-time and departure-epoch pgfs for this quorum queue is

$$\Pi_\rho(z) = \frac{1}{r}\big(1 + z + \cdots + z^{r-1}\big) P_\rho(z), \tag{4.1}$$

established in Dshalalow and Tadj (1992); see also Tadj (2003), eq. (4.2). Equivalently, in distribution,

$$Q_\rho \stackrel{d}{=} N_\rho + J, \tag{4.2}$$

where J is independent of N_ρ and uniformly distributed on $\{0, 1, \ldots, r-1\}$.

Corollary 4.1. *Under the assumptions of Theorem 3.1,*

$$(1-\rho)Q_\rho \;\Rightarrow\; \mathrm{Exp}\left(\frac{2}{1 + rC_s^2}\right). \tag{4.3}$$

Proof. By (4.2), $(1-\rho)Q_\rho \stackrel{d}{=} (1-\rho)N_\rho + (1-\rho)J$, and $0 \le (1-\rho)J \le (r-1)(1-\rho) \to 0$ since r is fixed. The result follows from Theorem 3.1 and Slutsky's theorem. □

Thus the departure-epoch and arbitrary-time queue lengths have the same heavy-traffic limit under the scaling $1 - \rho$.

5 The characteristic roots near critical loading

The distributional limit does not, by itself, imply convergence of the corresponding means, and it says nothing about the $O(1)$ term. To go further we use the exact mean formula of Tadj (2003), which requires control of the characteristic roots as $\rho \uparrow 1$.

Let $z_1(\rho), \ldots, z_{r-1}(\rho)$ denote the non-unit roots in the open unit disk of

$$\Phi_\rho(z) := z^r - F_\rho(z) = 0. \tag{5.1}$$

For every $\rho < 1$, equation (5.1) has exactly r roots in the closed unit disk, of which $z = 1$ is simple. This count is classical (Dshalalow and Tadj, 1992), but since everything below rests on

it, and since the usual quick arguments for it are not valid in the present generality, we include a proof.

Lemma 5.1 (Root count)**.** *Assume $P(S>0)>0$ and $E[S^2]<\infty$. For every $\rho\in(0,1)$, Φ_ρ has no zero on $\{|z|=1\}\setminus\{1\}$, the zero $z=1$ is simple, and Φ_ρ has exactly $r-1$ zeros, counted with multiplicity, in the open unit disk.*

Proof. Fix $\rho\in(0,1)$. For $|z|=1$, $z\neq 1$, put $c=\lambda\,\mathrm{Re}(1-z)=\frac{\lambda}{2}|z-1|^2>0$; then $|F_\rho(z)|\le E[e^{-cS}]<1=|z^r|$, the strict inequality because $P(S>0)>0$, so $\Phi_\rho(z)\neq 0$; by compactness the inequality $|F_\rho|<|z^r|$ is uniform on the circle away from any neighbourhood of 1. Also $\Phi_\rho'(1)=r-\lambda m_1=r(1-\rho)\neq 0$, so $z=1$ is simple.

For the count, we continue in ρ from $\rho\downarrow 0$. As $\rho\downarrow 0$ (that is, $\lambda\downarrow 0$), $F_\rho(z)=1-\lambda m_1(1-z)+O(\lambda^2)$ uniformly on the closed disk, so $\Phi_\rho(z)=z^r-1+\lambda m_1(1-z)+O(\lambda^2)$. At $\rho=0$ the zeros are the r-th roots of unity $\omega_k=e^{2\pi ik/r}$, all simple, with $\partial_z\Phi_0(\omega_k)=r\omega_k^{r-1}\neq 0$. By the implicit function theorem each ω_k continues to a zero $z_k(\rho)$ of Φ_ρ for small ρ, with $z_k'(0)=-\partial_\rho\Phi/\partial_z\Phi=-\frac{r}{m_1}m_1(1-\omega_k)/(r\omega_k^{r-1})=(\omega_k-1)/\omega_k^{r-1}$, where we used $\lambda=r\rho/m_1$. Hence

$$\begin{aligned}\frac{d}{d\rho}|z_k(\rho)|^2\Big|_{\rho=0}&=2\,\mathrm{Re}\left(z_k'(0)\overline{\omega_k}\right)=2\,\mathrm{Re}\left((\omega_k-1)\omega_k^{1-r}\overline{\omega_k}\right)\\&=2\,\mathrm{Re}\left(\omega_k^{1-r}-\omega_k^{-r}\right)=2\Big(\cos\frac{2\pi k}{r}-1\Big)<0\end{aligned}$$

for $k=1,\dots,r-1$, using $\omega_k^r=1$; while $z_0(\rho)=1$ for all ρ. So for small $\rho>0$ exactly $r-1$ zeros lie in the open disk and the remaining one is $z=1$; in particular, for $\rho\in(0,\rho_*]$ the count is $r-1$.

We now show that the zeros in the open disk stay away from the unit circle, locally uniformly in $\rho\in(0,1)$. Away from $z=1$ this is the first paragraph. Near $z=1$: if $\rho_n\to\rho_0\in(0,1)$ and $\Phi_{\rho_n}(1+\delta_n)=0$ with $\delta_n\to 0$, $\delta_n\neq 0$, then by (3.8), which is uniform in ρ on compact subsets of $(0,1]$ (Lemma 3.2, which needs $E[S^2]<\infty$), $0=r(1-\rho_n)\delta_n-\frac12 a_{\rho_n}\delta_n^2+o(|\delta_n|^2)$, and division by δ_n gives $r(1-\rho_0)=0$, a contradiction. Hence for every compact $J\subset(0,1)$ there is $R_J<1$ such that all zeros of Φ_ρ in the open disk lie in $\{|z|\le R_J\}$ for $\rho\in J$. On the circle $|z|=R_J'$, for any $R_J'\in(R_J,1)$, Φ_ρ is then nonvanishing and jointly continuous in $(\rho,z)\in J\times\{|z|=R_J'\}$, so by the argument principle the number of zeros of Φ_ρ inside that circle, which is the number in the open disk, is a continuous integer-valued function of $\rho\in J$, hence constant on J. Taking $J=[\rho_*,\rho]$ for any $\rho<1$ shows that the count is $r-1$ for every $\rho\in(0,1)$. □

Two remarks on this proof. First, the count is not a Rouché exercise on a fixed circle: on $|z|=1-\varepsilon_0$ one has $|F_\rho|\approx 1-r\rho\varepsilon_0$ and $|z^r|\approx 1-r\varepsilon_0$, so neither function dominates the other on the whole circle. Second, F_ρ itself may vanish inside the unit disk (for instance when $\widetilde{S}$ has a zero in $\{\mathrm{Re}\,s>0\}$, which happens for suitable discrete service laws), so the tempting argument "$\Phi_\rho=z^r-F_\rho$ has as many zeros as $-F_\rho$" is unsound; the continuation argument above does not use zero-freeness of F_ρ.

Since P_ρ in (2.3) is analytic in the open disk and continuous on its closure, the r roots of Lemma 5.1 are also zeros of the numerator polynomial $z^rH_\rho(1)-H_\rho(z)$, of degree r with leading coefficient $H_\rho(1)$, so the two zero sets coincide with multiplicities and

$$z^rH_\rho(1)-H_\rho(z)=H_\rho(1)\prod_{j=1}^{r}\big(z-z_j(\rho)\big),\qquad z_r(\rho)=1.\tag{5.2}$$

(If z^* is a zero of Φ_ρ of order m in the open disk, then $F_\rho(z^*) \neq 0$, since otherwise $z^{*r} = 0$, and P_ρ analytic at z^* forces the numerator to vanish to order at least m there.) Throughout, "roots" refers to the characteristic equation. Differentiating (5.2) at the simple zero $z = 1$ gives $H_\rho(1)\prod_{j<r}(1 - z_j(\rho)) = r(1-\rho)$ by Lemma 3.1, and differentiating twice gives

$$\frac{d^2}{dz^2}\big[z^r H_\rho(1) - H_\rho(z)\big]_{z=1} = 2r(1-\rho)\sum_{j=1}^{r-1}\frac{1}{1-z_j(\rho)}, \tag{5.3}$$

both valid with multiplicities. The stationary mean at departure epochs is (Tadj, 2003, eq. (3.22))

$$L_\rho^+ = \rho r + \frac{\lambda^2 m_2 - r(r-1)}{2r(1-\rho)} + \sum_{j=1}^{r-1}\frac{1}{1-z_j(\rho)}, \tag{5.4}$$

and the corresponding arbitrary-time mean is (Tadj, 2003, eq. (4.5))

$$L_\rho = L_\rho^+ + \frac{r-1}{2}. \tag{5.5}$$

Formula (5.4) is stated in Tadj (2003) under a simple-root assumption; it follows from (5.2)–(5.3) with multiplicities, so we use it in that form. For completeness: writing the numerator as $(z-1)\beta(z)$ and $\Phi_\rho(z) = (z-1)\varphi(z)$, one has $\beta(1) = r(1-\rho) = \varphi(1)$, $\beta'(1)$ equal to half of (5.3), and $\varphi'(1) = \frac{1}{2}\Phi_\rho''(1) = \frac{1}{2}[r(r-1) - \lambda^2 m_2]$, and the quotient rule applied to $P_\rho = F_\rho\beta/\varphi$ at $z = 1$ gives $L_\rho^+ = \lambda m_1 + \beta'(1)/\varphi(1) - \varphi'(1)/\varphi(1)$, which is (5.4). We also introduce the critical-load characteristic function

$$F_1(z) = \widetilde{S}\!\left(\frac{r}{m_1}(1-z)\right), \qquad \Phi_1(z) = z^r - F_1(z), \qquad |z| \le 1, \tag{5.6}$$

which is (5.1) with $\rho = 1$, that is, $\lambda = r/m_1$.

We record a regularity fact used below. Since $\widetilde{S}$ is analytic on $\{\mathrm{Re}\, s > 0\}$ and $\mathrm{Re}(1-z) > 0$ for every z in the closed unit disk other than $z = 1$, the function $(\rho, z) \mapsto \Phi_\rho(z)$ is jointly analytic on $(0,\infty) \times \big(\{|z| \le 1\} \setminus \{1\}\big)$, and indeed on $(0,\infty) \times U$ for the open set $U = \{z : \mathrm{Re}(1-z) > 0\} \supset \{|z| \le 1\} \setminus \{1\}$; moreover $\Phi_\rho \to \Phi_1$ locally uniformly on U as $\rho \to 1$, and uniformly on the closed unit disk by continuity of $\widetilde{S}$ on $\{\mathrm{Re}\, s \ge 0\}$.

Lemma 5.2 (Root separation)**.** *There exist constants $c > 0$ and $\rho_0 < 1$ such that*

$$|1 - z_j(\rho)| \ge c, \qquad j = 1, \dots, r-1,$$

for all $\rho_0 < \rho < 1$.

Proof. Suppose, to the contrary, that there exist $\rho_n \uparrow 1$ and roots z_n of Φ_{ρ_n} in the open unit disk such that $z_n \to 1$. Set $\varepsilon_n = 1 - \rho_n$ and $\delta_n = z_n - 1$, so that $\delta_n \to 0$, $\delta_n \neq 0$, and $\mathrm{Re}(\delta_n) < 0$ because $|1 + \delta_n| < 1$.

The expansion (3.8) is valid for complex δ with $|1 + \delta| \le 1$, uniformly in ρ, by Lemma 3.2. Applying it at $\delta = \delta_n$, $\rho = \rho_n$, the root condition $\Phi_{\rho_n}(1 + \delta_n) = 0$ becomes

$$0 = r\varepsilon_n\delta_n - \frac{a_{\rho_n}}{2}\delta_n^2 + o(|\delta_n|^2), \tag{5.7}$$

where, by (2.5) and (3.9),

$$a_\rho = \lambda^2 m_2 - r(r-1) = r^2\rho^2(1+C_s^2) - r(r-1) \longrightarrow r(1+rC_s^2) > 0.$$

Since $\delta_n \neq 0$, division of (5.7) by δ_n gives

$$r\varepsilon_n = \frac{a_{\rho_n}}{2}\delta_n + o(|\delta_n|). \tag{5.8}$$

Since $a_{\rho_n} \to r(1+rC_s^2) > 0$, there is $a_0 > 0$ with $a_{\rho_n} \geq a_0$ for large n; taking absolute values in (5.8) gives $\frac{a_0}{2}|\delta_n| \leq r\varepsilon_n + o(|\delta_n|)$, hence $|\delta_n| \leq (3r/a_0)\varepsilon_n$ for large n, i.e. $|\delta_n| = O(\varepsilon_n)$. Dividing (5.8) by ε_n and using $o(|\delta_n|) = o(\varepsilon_n)$,

$$\frac{\delta_n}{\varepsilon_n} \longrightarrow \frac{2r}{r(1+rC_s^2)} = \frac{2}{1+rC_s^2} > 0.$$

The limit is real and strictly positive, so $\mathrm{Re}(\delta_n/\varepsilon_n) > 0$ for all large n; since $\varepsilon_n > 0$, this means $\mathrm{Re}(\delta_n) > 0$, which contradicts $\mathrm{Re}(\delta_n) < 0$ established at the outset. Therefore no interior root approaches 1 as $\rho \uparrow 1$, and since the number of interior roots is fixed, the uniform separation follows. □

The proof of Lemma 5.2 shows that any root approaching 1 must do so from outside the unit disk, at the rate $2(1-\rho)/(1+rC_s^2)$. Whether such a root exists is a separate question, since F_ρ need not be defined beyond the closed disk. When it is, the root exists and is unique.

Lemma 5.3 (The exterior root)**.** *Suppose that $\widetilde{S}$ extends analytically to a neighbourhood of $s = 0$ (equivalently, $E[e^{\eta S}] < \infty$ for some $\eta > 0$; this holds in particular when $\widetilde{S}$ is rational). Then there exist $\rho_0 < 1$ and a neighbourhood V of $z = 1$ such that, for $\rho_0 < \rho < 1$, equation (5.1) has exactly two roots in V, namely $z = 1$ and a real root $z_{r+1}(\rho) > 1$, and $\rho \mapsto z_{r+1}(\rho)$ is real-analytic with*

$$z_{r+1}(\rho) = 1 + \frac{2(1-\rho)}{1+rC_s^2} + O\big((1-\rho)^2\big). \tag{5.9}$$

Proof. Under the hypothesis, $(\rho, z) \mapsto \Phi_\rho(z)$ is jointly analytic on a neighbourhood of $(1,1)$, and $\Phi_\rho(1) = 0$ for all ρ. Hence $\Psi(\rho,\delta) := \Phi_\rho(1+\delta)/\delta$ is jointly analytic near $(1,0)$ with $\Psi(\rho,0) = \Phi'_\rho(1) = r(1-\rho)$ and $\partial_\delta\Psi(1,0) = \frac{1}{2}\Phi''_1(1) = -r\gamma \neq 0$, where $\gamma = (1+rC_s^2)/2$ as in (3.9). By the analytic implicit function theorem there is a unique real-analytic $\delta(\rho)$ near $\rho = 1$ with $\delta(1) = 0$ and $\Psi(\rho,\delta(\rho)) = 0$, and all zeros of Ψ near $(1,0)$ are of this form. Differentiating $\Psi(\rho,\delta(\rho)) = 0$ at $\rho = 1$ gives $-r - r\gamma\,\delta'(1) = 0$, i.e. $\delta'(1) = -1/\gamma$, so $\delta(\rho) = (1-\rho)/\gamma + O((1-\rho)^2)$, which is (5.9); in particular $\delta(\rho) > 0$ for $\rho < 1$ close to 1, and $z_{r+1}(\rho) = 1 + \delta(\rho)$ is real and lies outside the unit disk. Since $\Phi_\rho(1+\delta) = \delta\,\Psi(\rho,\delta)$, the roots of (5.1) near 1 are exactly 1 and $z_{r+1}(\rho)$. □

For exponential service, (5.9) agrees with the explicit root of Tadj (2003), and $z_{r+1}(\rho)$ is the root governing the geometric decay rate $\theta_1 = 1/z_{r+1}(\rho)$ of the stationary distribution; its behavior $1 - \theta_1 \sim 2(1-\rho)/(1+rC_s^2)$ is the discrete counterpart of (3.5). Lemma 5.3 is not needed for the results below, which concern the interior roots only.

Lemma 5.4 (Roots at critical loading)**.** (i) *The critical-load function satisfies $\Phi_1(1) = \Phi'_1(1) = 0$ and $\Phi''_1(1) \neq 0$, the derivatives being taken within the closed disk (we refer to this as a double root at $z = 1$, although Φ_1 need not extend analytically across $z = 1$). Φ_1 has no other zero on the unit circle, and it has exactly $r-1$ zeros (with multiplicity) in the open unit disk; denote them $z_1(1), \ldots, z_{r-1}(1)$.*

(ii) *Let $R \in (0,1)$ be such that $|z_j(1)| < R$ for all j and such that Φ_1 has no zero on the circle $\{|z| = R\}$. There is an open interval $I \ni 1$ such that for every $\rho \in I$ the function Φ_ρ has exactly $r-1$ zeros in $\{|z| < R\}$, counted with multiplicity, and for $\rho \in I$, $\rho < 1$, these are precisely the interior roots $z_1(\rho), \dots, z_{r-1}(\rho)$. As $\rho \uparrow 1$ the multiset $\{z_1(\rho), \dots, z_{r-1}(\rho)\}$ converges to $\{z_1(1), \dots, z_{r-1}(1)\}$, and for every function f analytic on a neighbourhood of $\{|z| \le R\}$ the symmetric sum $\rho \mapsto \sum_{j=1}^{r-1} f(z_j(\rho))$ is a real-analytic function of $\rho \in I$; in particular*

$$\sum_{j=1}^{r-1} \frac{1}{1 - z_j(\rho)} = \sum_{j=1}^{r-1} \frac{1}{1 - z_j(1)} + O(1-\rho), \qquad \rho \uparrow 1. \tag{5.10}$$

(iii) *If in addition the roots $z_1(1), \dots, z_{r-1}(1)$ are simple, there exist $\rho_0 < 1$ and a labelling of the interior roots such that each $\rho \mapsto z_j(\rho)$ is real-analytic on $(\rho_0, 1]$ with $z_j(\rho) \to z_j(1)$.*

Proof. (i) Since F_ρ is the pgf of a random variable with mean $r\rho$ and finite second moment, $\Phi_\rho(1) = 0$ and $\Phi'_\rho(1) = r(1-\rho)$, the derivatives at $z = 1$ being taken within the closed disk. Hence $\Phi'_1(1) = 0$, while by (3.9) $\Phi''_1(1) = r(r-1) - r^2 m_2/m_1^2 = -r(1 + rC_s^2) \neq 0$; this is what we mean by a root of multiplicity exactly two at $z = 1$. (Note that Φ_1 need not extend analytically across $z = 1$, since no exponential moment of S is assumed; the statement is about the expansion (3.8) at $\rho = 1$, which reads as follows.)

$$\Phi_1(1+\delta) = -\frac{r(1 + rC_s^2)}{2}\,\delta^2 + o(|\delta|^2), \qquad |1+\delta| \le 1,\ \delta \to 0. \tag{5.11}$$

For $|z| = 1$, $z \neq 1$, one has $\operatorname{Re}\big(\frac{r}{m_1}(1-z)\big) > 0$ and $S > 0$ a.s., whence $|F_1(z)| = \big|E[e^{-\frac{r}{m_1}(1-z)S}]\big| < 1 = |z^r|$; thus Φ_1 has no other root on the circle.

For the count, (5.11) shows that there is an $\eta > 0$ such that $\Phi_1(z) \neq 0$ for all z in the closed unit disk with $0 < |z-1| \le \eta$. The set $K = \{|z| \le 1,\ |z-1| \ge \eta\}$ is compact, Φ_1 is continuous on it, and Φ_1 does not vanish on $K \cap \{|z| = 1\}$ by the previous paragraph; by continuity there is therefore a $\kappa > 0$ such that $\Phi_1 \neq 0$ on $K \cap \{|z| \ge 1 - \kappa\}$. Consequently every zero of Φ_1 in the open unit disk lies in the closed disk $\{|z| \le R_0\}$, $R_0 := 1 - \kappa < 1$, on which Φ_1 is analytic and not identically zero (indeed $\Phi_1(0) = -E[e^{-rS/m_1}] \neq 0$). Hence Φ_1 has finitely many zeros in the open unit disk; let m be their number, counted with multiplicity. That $m = r - 1$ is proved together with (ii) below, by continuity from $\rho < 1$, where the count of interior roots is Lemma 5.1, using the separation Lemma 5.2 to ensure that no interior root escapes to $z = 1$ in the limit.

(ii) Choose $R \in (R_0, 1)$; then $|z_j(1)| \le R_0 < R$ for every zero of Φ_1 in the open disk and Φ_1 has no zero on $C := \{|z| = R\}$, so R is as in the statement (any such R will do). Note that $\Phi_\rho(z) = z^r - \widetilde{S}(\frac{r\rho}{m_1}(1-z))$ is defined for all $\rho > 0$, including $\rho > 1$, on the closed unit disk, and that $\Phi_\rho \to \Phi_1$ uniformly on the closed unit disk as $\rho \to 1$, because $\widetilde{S}$ is uniformly continuous on compact subsets of $\{\operatorname{Re} s \ge 0\}$ and $\frac{r\rho}{m_1}(1-z)$ ranges over such a set. Since $\min_C |\Phi_1| > 0$, there is an open interval $I \ni 1$ such that $|\Phi_\rho - \Phi_1| < |\Phi_1|$ on C for all $\rho \in I$, and Rouché's theorem gives: for every $\rho \in I$, Φ_ρ has exactly m zeros in $\{|z| < R\}$, counted with multiplicity.

We now identify m. For $\rho \in I$, $\rho < 1$, the zeros of Φ_ρ in the closed unit disk are the simple zero $z = 1$ and the $r-1$ interior roots, and $z = 1$ lies outside C; hence $m \le r - 1$. Conversely, all interior roots lie inside C for ρ close to 1: otherwise there would be $\rho_n \uparrow 1$ and interior roots w_n of Φ_{ρ_n} with $|w_n| \ge R$; passing to a subsequence, $w_n \to w$ with $R \le |w| \le 1$, and $\Phi_1(w) = \lim \Phi_{\rho_n}(w_n) = 0$ by uniform convergence and continuity, so w would be a zero of Φ_1

in the closed disk with $|w| > R_0$, which forces $w = 1$, contradicting Lemma 5.2. Hence, after shrinking I, the m zeros of Φ_ρ inside C are exactly the $r-1$ interior roots for $\rho \in I$, $\rho < 1$, so $m = r-1$. This completes the proof of (i) and of the counting statement in (ii); the convergence of the multiset of interior roots to the multiset of zeros of Φ_1 inside C is Hurwitz's theorem (or follows from (5.12) below with $f(z) = z^k$, $k \le r-1$, and Newton's identities).

By the argument principle, for every f analytic on a neighbourhood of $\{|z| \le R\}$,

$$\sum_{j=1}^{r-1} f\big(z_j(\rho)\big) = \frac{1}{2\pi i} \oint_C f(z)\, \frac{\partial_z \Phi_\rho(z)}{\Phi_\rho(z)}\, dz, \qquad \rho \in I, \tag{5.12}$$

with the roots counted according to multiplicity; here f is defined at the roots because they lie inside C. The integrand is jointly analytic in (ρ, z) on a complex neighbourhood of $I \times C$, because $(\rho, z) \mapsto \Phi_\rho(z)$ is jointly analytic on $\{\rho \in \mathbb{C} : \operatorname{Re}\rho > 0\} \times \{|z| < 1\}$ (the argument $\frac{r\rho}{m_1}(1-z)$ of $\widetilde{S}$ then has positive real part) and Φ_ρ does not vanish on C for ρ in a complex neighbourhood of I. Hence the right-hand side of (5.12) is an analytic function of ρ on that complex neighbourhood, and in particular a real-analytic function of the real parameter $\rho \in I$. Taking $f(z) = 1/(1-z)$, which is analytic on the open unit disk and hence on a neighbourhood of $\{|z| \le R\}$, gives (5.10).

(iii) At a simple root $z_j(1)$ one has $\partial_z \Phi_1(z_j(1)) \neq 0$, so by the analytic implicit function theorem the root continues as a real-analytic function $z_j(\rho)$ on a neighbourhood of $\rho = 1$, distinct roots having disjoint continuations; since (ii) gives exactly $r-1$ roots inside C for ρ near 1, these continuations are all of them. When critical-load roots collide the individual roots need not depend analytically on ρ (Remark 6.1(d)), which is why (ii) is formulated for symmetric sums. □

Lemma 5.4 sharpens Lemma 5.2 from a bound to a limit, and it is the limit that the second-order expansion requires. Part (ii) is what the expansion uses, and it needs no assumption on the multiplicity of the critical-load roots: the quantity that enters the mean is a symmetric function of the roots, and symmetric functions of the zeros of an analytic family vary analytically with the parameter even when individual zeros collide. The contour radius R is a device of the proof only; the conclusion for $f(z) = 1/(1-z)$ does not depend on it. Part (iii) is included because the individual roots are of independent interest and because it holds in all the examples below (Lemma 6.1). Relation (5.12) is also a convenient way to *compute* the root sum without locating the roots, and it is used for the non-rational examples of Section 6.2.

6 Two-term expansion of the stationary mean

Theorem 6.1. *Suppose $S > 0$ almost surely and $E[S^2] < \infty$. Then, as $\rho \uparrow 1$,*

$$L_\rho^+ = \frac{1 + rC_s^2}{2(1-\rho)} + c_0 + O(1-\rho), \qquad c_0 = -\,rC_s^2 + \sum_{j=1}^{r-1} \frac{1}{1 - z_j(1)}, \tag{6.1}$$

and

$$L_\rho = \frac{1 + rC_s^2}{2(1-\rho)} + c_0 + \frac{r-1}{2} + O(1-\rho). \tag{6.2}$$

Proof. Put $\varepsilon = 1 - \rho$. By the exact identity (2.5), with $\rho^2 = 1 - 2\varepsilon + \varepsilon^2$,

$$\lambda^2 m_2 - r(r-1) = r(1 + rC_s^2) - 2r^2(1 + C_s^2)\varepsilon + r^2(1 + C_s^2)\varepsilon^2,$$

so that

$$\frac{\lambda^2 m_2 - r(r-1)}{2r(1-\rho)} = \frac{1 + rC_s^2}{2\varepsilon} - r(1 + C_s^2) + O(\varepsilon). \tag{6.3}$$

Also $\rho r = r - r\varepsilon$. Substituting (6.3) and (5.10) into (5.4),

$$L_\rho^+ = \frac{1 + rC_s^2}{2\varepsilon} + \big[r - r(1 + C_s^2)\big] + \sum_{j=1}^{r-1} \frac{1}{1 - z_j(1)} + O(\varepsilon),$$

and $r - r(1 + C_s^2) = -rC_s^2$, which is (6.1). Relation (6.2) follows from (5.5). □

Remark 6.1. (a) The leading term of (6.1) is the mean of the limiting exponential law in Theorem 3.1, and by Remark 3.1 it is what Kingman's theorem predicts; it depends on the service-time distribution only through C_s^2. The constant c_0, by contrast, depends on the full distribution through the critical-load roots $z_j(1)$ and is not accessible from the weak limit or from a diffusion approximation. Its computation requires exactly the root information on which the exact solution of Tadj (2003) is built.

(b) The remainder in (6.1) is $O(1-\rho)$, not merely $o(1)$. This relies on the fact that $\lambda^2 m_2 - r(r-1)$ is an exact quadratic polynomial in ρ, (2.5): the weaker statement $\lambda^2 m_2 - r(r-1) = r(1 + rC_s^2) + o(1)$ would only control the first fraction in (5.4) to within $o((1-\rho)^{-1})$, which is not enough even to assert that the correction is bounded. The $O(1-\rho)$ remainder is sharp in general: the next coefficient is nonzero in every example of Section 6.2 (the approach to c_0 in Table 2 and Figure 1 is linear in $1-\rho$, with slope about -0.14 for exponential service and $r = 3$, and larger in modulus for Erlang service), and for exponential service it can be read off the exact geometric solution.

(c) Theorem 6.1 contains, as its leading term, the first-moment asymptotics $L_\rho^+ \sim L_\rho \sim (1 + rC_s^2)/(2(1-\rho))$, established here through the exact representation rather than inferred from weak convergence, which does not imply convergence of means.

(d) No assumption on the multiplicity of the characteristic roots is needed, at critical loading or below it. If some $z_j(1)$ is a repeated root, the individual roots $z_j(\rho)$ typically behave like $z_j(1) + c(1-\rho)^{1/m}$ and are not analytic in ρ, but their sum in (5.4) is, by Lemma 5.4(ii); the expansion and its $O(1-\rho)$ remainder are unaffected. The simple-root assumption of Tadj (2003) is thus a convenience of exposition rather than a restriction on the service-time distribution.

Corollary 6.1 (Exponential service)**.** *For exponential service ($C_s^2 = 1$),*

$$\sum_{j=1}^{r-1} \frac{1}{1 - z_j(1)} = \frac{2(r-1)}{3}, \qquad c_0 = -\frac{r+2}{3}, \tag{6.4}$$

so that

$$L_\rho^+ = \frac{1+r}{2(1-\rho)} - \frac{r+2}{3} + O(1-\rho), \qquad L_\rho = \frac{1+r}{2(1-\rho)} + \frac{r-7}{6} + O(1-\rho). \tag{6.5}$$

Proof. For exponential service $F_\rho(z) = 1/(1+\rho r(1-z))$, and (Tadj, 2003)

$$z^r - F_\rho(z) = \frac{(1-z)\big(\rho r z^r - z^{r-1} - \cdots - z - 1\big)}{1+\rho r(1-z)}.$$

At $\rho = 1$ the polynomial factor vanishes at $z = 1$ and one checks directly that

$$rz^r - z^{r-1} - \cdots - z - 1 = (z-1)Q(z), \qquad Q(z) = \sum_{k=0}^{r-1}(k+1)z^k.$$

Thus $z_1(1), \ldots, z_{r-1}(1)$ are precisely the zeros of Q. Since $Q(z) = r\prod_{j<r}(z - z_j(1))$, logarithmic differentiation at $z = 1$ gives

$$\sum_{j=1}^{r-1}\frac{1}{1-z_j(1)} = \frac{Q'(1)}{Q(1)}, \quad Q(1) = \sum_{k=0}^{r-1}(k{+}1) = \frac{r(r+1)}{2}, \quad Q'(1) = \sum_{k=1}^{r-1}k(k{+}1) = \frac{(r-1)r(r+1)}{3},$$

so that the sum equals $2(r-1)/3$. It remains to check that the zeros of Q are simple. Multiplying out, $(z-1)^2Q(z) = R(z) := rz^{r+1} - (r+1)z^r + 1$, so a zero $w \neq 1$ of Q is repeated if and only if $R'(w) = 0$. But $R'(z) = r(r+1)z^{r-1}(z-1)$ vanishes only at 0 and 1, and $R(0) = 1 \neq 0$. Hence all zeros of Q are simple. (That they lie in the open unit disk is also seen directly: the Eneström–Kakeya theorem places the zeros of a polynomial with increasing positive coefficients $1 < 2 < \cdots < r$ in the closed disk, and Lemma 5.4(i) excludes the circle.) Theorem 6.1 therefore applies with $C_s^2 = 1$, which gives (6.4) and (6.5). □

For $r = 1$, (6.5) reduces to $L_\rho = (1-\rho)^{-1} - 1 = \rho/(1-\rho)$, the exact $M/M/1$ mean, so the expansion is sharp in that case.

6.1 Erlang service

For Erlang-k service with mean m_1 (so $C_s^2 = 1/k$), $F_\rho(z) = \big[k/(k + r\rho(1-z))\big]^k$ and the characteristic equation (5.1) is equivalent to the polynomial equation

$$P_\rho(z) := z^r\big(k + r\rho(1-z)\big)^k - k^k = 0, \tag{6.6}$$

of degree $r + k$; see Tadj (2003), eq. (3.12). For $\rho < 1$ it has $r - 1$ roots inside the unit disk, the simple root $z = 1$, and k roots outside. At $\rho = 1$ the outside root $z_{r+1}(\rho)$ of (5.9) merges with 1, and by Lemma 5.4(i) the remaining roots are $r - 1$ inside and $k - 1$ outside. We write $P = P_1$ and denote the outside roots at critical loading by $\zeta_1, \ldots, \zeta_{k-1}$.

Lemma 6.1. *For Erlang-k service and every $\rho \in (0, 1]$, all roots of (6.6) in the closed unit disk are simple, except $z = 1$ at $\rho = 1$. In particular the individual-root continuation of Lemma 5.4(iii) is available for all $k \geq 1$ and $r \geq 1$.*

Proof. Differentiating,

$$P'_\rho(z) = rz^{r-1}\big(k + r\rho(1-z)\big)^{k-1}\big[k + r\rho - \rho(r+k)z\big],$$

whose zeros are $z = 0$ (if $r \geq 2$), $z = (k + r\rho)/(r\rho)$ (if $k \geq 2$) and $z^* = (k + r\rho)/(\rho(r+k))$; when $r = 1$ or $k = 1$ the corresponding factor is absent and the argument below is unchanged. The first two are not roots of P_ρ, since P_ρ equals $-k^k$ at both. The third satisfies $z^* \geq 1$, with equality if and only if $\rho = 1$; thus for $\rho < 1$ no root in the closed disk is repeated, and for $\rho = 1$ the only candidate is $z = 1$, where the multiplicity is two by Lemma 5.4(i). □

Lemma 6.1 also shows that the only possible repeated root for $\rho < 1$ is the real root $z^* > 1$, which does not affect the interior sum in (5.4).

Corollary 6.2 (Erlang service)**.** *For Erlang-k service,*

$$\sum_{j=1}^{r-1} \frac{1}{1 - z_j(1)} = \frac{2(r-k)}{3k} + \sum_{j=1}^{k-1} \frac{1}{\zeta_j - 1}, \qquad c_0 = -\frac{r+2k}{3k} + \sum_{j=1}^{k-1} \frac{1}{\zeta_j - 1}, \tag{6.7}$$

where $\zeta_1, \dots, \zeta_{k-1}$ are the roots of $z^r\big(k + r(1-z)\big)^k = k^k$ outside the unit disk. For $k = 1$ this is Corollary 6.1.

Proof. By Lemma 5.4(i) and Lemma 6.1, $P(z) = (z-1)^2 G(z)$ with G a polynomial of degree $r + k - 2$ whose zeros are exactly $z_1(1), \dots, z_{r-1}(1)$ and $\zeta_1, \dots, \zeta_{k-1}$, all simple. Then $P''(1) = 2G(1)$ and $P'''(1) = 6G'(1)$, so that

$$\sum_{j=1}^{r-1} \frac{1}{1 - z_j(1)} + \sum_{j=1}^{k-1} \frac{1}{1 - \zeta_j} = \frac{G'(1)}{G(1)} = \frac{P'''(1)}{3P''(1)}.$$

Put $D(z) = \big(k + r(1-z)\big)^k$, so that $D(1) = k^k$, $D'(1) = -rk^k$, $D''(1) = r^2(k-1)k^{k-1}$ and $D'''(1) = -r^3(k-1)(k-2)k^{k-2}$. Leibniz's rule applied to $P = z^r D - k^k$ gives

$$P''(1) = r(r-1)D(1) + 2rD'(1) + D''(1) = -r(k+r)k^{k-1},$$

and

$$\begin{aligned} P'''(1) &= r(r-1)(r-2)D(1) + 3r(r-1)D'(1) + 3rD''(1) + D'''(1) \\ &= k^{k-2}\big[-2r(r^2-1)k^2 + 2r^3(k^2-1)\big] = 2r(k-r)(k+r)k^{k-2}. \end{aligned}$$

Hence $P'''(1)/(3P''(1)) = 2(r-k)/(3k)$, which gives the first relation in (6.7); the second follows from Theorem 6.1 with $C_s^2 = 1/k$, since $-r/k + 2(r-k)/(3k) = -(r+2k)/(3k)$. □

Remark 6.2. (a) Corollary 6.2 trades the $r-1$ interior roots for the $k-1$ exterior roots at critical loading. For Erlang-2 service only one root is needed, whatever the quorum size: ζ_1 is the unique real root greater than 1 of $z^{r/2}\big(2 + r(1-z)\big) = -2$, and $c_0 = -(r+4)/6 + 1/(\zeta_1 - 1)$. For $r = 2$, $\zeta_1 = 1 + \sqrt{2}$ and $c_0 = 1/\sqrt{2} - 1$, in agreement with the explicit roots of Tadj (2003).

(b) When $k = r$, equation (6.6) at $\rho = 1$ reads $\big(z(2-z)\big)^r = 1$, so its roots are $z = 1 \pm \sqrt{1-\omega}$ with $\omega^r = 1$, and the map $z \mapsto 2 - z$ exchanges the interior and exterior roots. Consequently the two sums in (6.7) coincide and, with the principal branch of the square root,

$$c_0 = -1 + \sum_{\omega^r = 1,\ \omega \neq 1} \frac{1}{\sqrt{1-\omega}} \qquad (k = r),$$

which gives $1/\sqrt{2} - 1$, 0.4679, 1.2609 and 2.0759 for $r = 2, 3, 4, 5$.

(c) Table 1 lists c_0 from (6.7) for small k and r. In every case shown the exterior-root sum is positive and c_0 increases with k for fixed r, i.e. the correction grows as service-time variability decreases; we have not attempted to prove either property in general.

$k\backslash r$	2	3	4	5
1	−1.3333	−1.6667	−2.0000	−2.3333
2	−0.2929	−0.0770	0.1424	0.3634
3	0.0598	0.4679	0.8823	1.3000
4	0.2378	0.7451	1.2609	1.7813

Table 1: The constant c_0 of Theorem 6.1 for Erlang-k service, from (6.7).

6.2 Numerical check

Table 2 compares c_0 computed from the critical-load roots with $L_\rho^+ - (1+rC_s^2)/(2(1-\rho))$, where L_ρ^+ is the stationary mean of the embedded chain (2.1), obtained by direct numerical solution of its balance equations on $\{0,\dots,N\}$ (the number of arrivals per service is negative binomial for Erlang service). The chain is truncated by lumping the transition mass beyond N into state N, with $N = \lceil 30\gamma/(1-\rho)\rceil + 200$; since the stationary distribution decays geometrically at rate $\theta_1 = 1-(1-\rho)/\gamma + O((1-\rho)^2)$ by Lemma 5.3, the neglected tail mass is of order $\theta_1^N \approx e^{-30}$, far below the precision reported. The discrepancy halves whenever $1-\rho$ halves (between $\rho = 0.98$, 0.99 and 0.995; from $\rho = 0.95$ to 0.98 the factor is 2.5), in agreement with the $O(1-\rho)$ remainder in (6.1). The truncation was validated by recomputing the means with $2N$ and $3N$ states: the changes are below 10^{-4} in every case (below 10^{-5} for Erlang-3 and deterministic service), and they grow linearly rather than decaying with N, which identifies them as rounding in the long banded solve rather than truncation; the lumped tail mass contributes at most $N\theta_1^N \approx Ne^{-30}$ to the mean. Entries are therefore reliable to the four decimals shown.

The lower block of Table 2 goes beyond the rational case. For deterministic service ($F_\rho(z) = e^{-r\rho(1-z)}$, $C_s^2 = 0$, so $\gamma = \frac{1}{2}$ and $c_0 = \sum_{j<r} 1/(1-z_j(1))$) and for a two-phase hyperexponential service with $C_s^2 = 2$ (that is, S is exponential with rate μ_1 with probability p and with rate μ_2 with probability $1-p$, with balanced means $p/\mu_1 = (1-p)/\mu_2 = \frac{1}{2}$, which forces $p(1-p) = 1/(2(C_s^2+1)) = 1/6$; so $p = (1-1/\sqrt{3})/2 \approx 0.2113$, $\mu_1 = 2p \approx 0.4226$, $\mu_2 = 2(1-p) \approx 1.5774$), the root sum was evaluated directly from the contour integral (5.12), with $f(z) = 1/(1-z)$ and $C = \{|z| = 0.97\}$, by the trapezoidal rule with 4×10^4 nodes (the same integral with $f \equiv 1$ returns $r-1$ to twelve digits, confirming that C encloses all interior roots and nothing else; doubling the number of nodes or moving the radius to 0.95 changes the computed c_0 by less than 10^{-10}). For deterministic service the characteristic equation $z^r = e^{-r(1-z)}$ is transcendental and has infinitely many roots in the plane, but by Lemma 5.4 exactly $r-1$ of them lie in the open unit disk, and these are the only ones that enter the stationary mean; the contour integral sees no others. For $r = 2$ the interior root is explicit: writing $z = -x$ with $x > 0$ in $z^2 = e^{-2(1-z)}$ gives $xe^x = e^{-1}$, so $z_1(1) = -W(e^{-1})$ with W the Lambert function, and

$$c_0 = \frac{1}{1+W(e^{-1})} = 0.78218829\ldots \qquad \text{(deterministic service, } r=2\text{)},$$

in agreement with the table. The number of arrivals per service is Poisson, respectively a mixture of two geometric laws. The exponential case $r = 10$ ($c_0 = -4$) is included to show that nothing deteriorates for larger quorum sizes. The scripts that produce all tables and the figure (chain construction, contour evaluation of the root sum, truncation and quadrature checks) are available from the author. Figure 1 displays the convergence for four service laws over the whole range $0.8 \le \rho \le 0.99$; the approach to c_0 is visibly linear in $1-\rho$, and the sign of c_0 switches

from negative for the two high-variability laws to positive for the two low-variability ones.

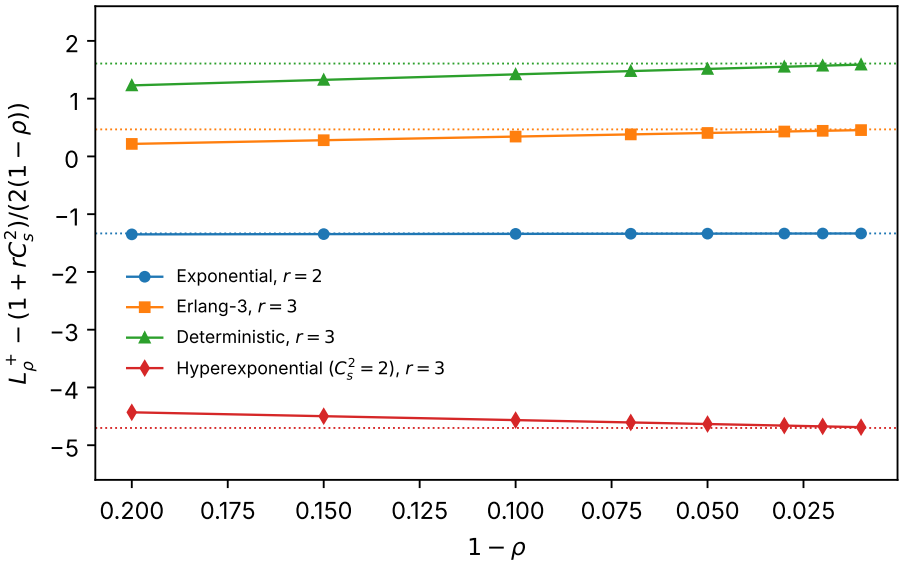

Figure 1: $L_\rho^+ - (1 + rC_s^2)/(2(1-\rho))$ as a function of $1-\rho$ (decreasing to the right) for four service-time distributions, with the limits c_0 of Theorem 6.1 as dotted lines.

6.3 Accuracy at moderate traffic

Table 3 shows how far down in ρ the expansion remains useful, by comparing the leading term $(1 + rC_s^2)/(2(1-\rho))$ and the two-term approximation $(1 + rC_s^2)/(2(1-\rho)) + c_0$ with the exact stationary mean. For exponential service the two-term approximation is remarkably good: its relative error is below 5% at $\rho = 0.5$ and below 1% from $\rho = 0.7$ on, whereas the leading term alone overestimates the mean by 70–85% at $\rho = 0.5$ and still by about 10% at $\rho = 0.9$. The explanation is that for $M/M^r/1$ the $O(1-\rho)$ remainder in (6.1) has a small coefficient (about 0.1 in the cases shown).

For Erlang service the picture is more sober. The coefficient of the $O(1-\rho)$ remainder is several times larger (about 0.5–0.9), so the second term only improves matters appreciably for $\rho \geq 0.9$. For Erlang-3 service the leading term is, by accident, the better approximation at moderate traffic: $c_0 > 0$ there while the next correction is negative, and the two partially cancel in the exact mean. This is not a defect of the expansion, whose remainder is $O(1-\rho)$ exactly as claimed, but it does show that the two-term formula should be used as an asymptotic statement about $\rho \uparrow 1$ and not as a general-purpose approximation, except in the exponential case.

7 Second-order expansion of the Laplace transform

The constant c_0 of Theorem 6.1 is the second-order term of the mean. We now show that the same roots govern the second-order term of the Laplace transform of the scaled queue length: the transform in Theorem 3.1 admits an expansion to order $1-\rho$, uniform in the transform variable on compact sets, whose coefficient is a rational function of u built from c_0, from the

Service	r	$\rho = 0.95$	$\rho = 0.98$	$\rho = 0.99$	$\rho = 0.995$	c_0
Exponential	2	-1.3371	-1.3348	-1.3341	-1.3337	$-4/3$
Exponential	3	-1.6738	-1.6695	-1.6681	-1.6674	$-5/3$
Erlang-2	2	-0.3232	-0.3050	-0.2989	-0.2959	$1/\sqrt{2} - 1 = -0.2929$
Erlang-2	3	-0.1247	-0.0960	-0.0865	-0.0817	-0.0770
Erlang-3	3	$+0.4059$	$+0.4432$	$+0.4555$	$+0.4617$	$+0.4679$
Exponential	10	-4.0280	-4.0110	-4.0055	-4.0027	-4
Deterministic	2	$+0.7235$	$+0.7588$	$+0.7705$	$+0.7763$	$+0.7822$
Deterministic	3	$+1.5145$	$+1.5706$	$+1.5893$	$+1.5986$	$+1.6079$
Deterministic	5	$+3.1612$	$+3.2599$	$+3.2926$	$+3.3090$	$+3.3253$
Hyperexp. ($C_s^2 = 2$)	2	-3.3045	-3.3324	-3.3417	-3.3463	-3.3510
Hyperexp. ($C_s^2 = 2$)	3	-4.6331	-4.6740	-4.6877	-4.6944	-4.7014

Table 2: $L_\rho^+ - (1 + rC_s^2)/(2(1-\rho))$ from the exact stationary mean, against the constant c_0 of Theorem 6.1.

Service	r	$\rho = 0.5$			$\rho = 0.7$			$\rho = 0.9$			$\rho = 0.95$		
		L_ρ^+	1-t.	2-t.	L_ρ^+	1-t.	2-t.	L_ρ^+	1-t.	2-t.	L_ρ^+	1-t.	2-t.
Exponential	2	1.62	+85.4	+3.0	3.64	+37.3	+0.7	13.66	+9.8	+0.1	28.66	+4.7	+0.0
Exponential	3	2.24	+78.5	+4.1	4.95	+34.6	+1.0	18.32	+9.2	+0.1	38.33	+4.4	+0.0
Exponential	5	3.49	+71.9	+5.0	7.57	+32.0	+1.2	27.64	+8.5	+0.1	57.65	+4.1	+0.0
Erlang-2	3	1.92	+30.4	+26.4	3.79	+9.8	+7.8	12.33	+1.4	+0.8	24.88	+0.5	+0.2
Erlang-3	3	1.82	+10.2	+36.0	3.42	−2.5	+11.2	10.34	−3.3	+1.2	20.41	−2.0	+0.3

Table 3: Exact departure-epoch mean L_ρ^+ (from the embedded chain, as in Table 2) and the relative errors, in percent, of the one-term approximation $(1 + rC_s^2)/(2(1-\rho))$ ("1-t.") and of the two-term approximation $(1 + rC_s^2)/(2(1-\rho)) + c_0$ ("2-t.").

third moment of S, and from the root sum S_1. One more moment of S is needed. The result is a statement about the transform; what it implies for the distribution is made precise in Corollary 7.1 and discussed in Remark 7.1(d).

Throughout this section write

$$\gamma = \frac{1 + rC_s^2}{2}, \qquad S_1 = \sum_{j=1}^{r-1} \frac{1}{1 - z_j(1)} = c_0 + rC_s^2, \qquad \kappa_3 = \frac{m_3}{m_1^3}, \quad m_3 = E[S^3], \tag{7.1}$$

so that $1/\gamma$ is the rate of the limit law in (3.5), and define

$$d = \gamma\Big(r - \frac{1}{2} + S_1\Big) + \frac{(r-1)(r-2) - r^2\kappa_3}{6}. \tag{7.2}$$

Theorem 7.1. *Suppose $S > 0$ almost surely and $E[S^3] < \infty$. Then, as $\rho \uparrow 1$,*

$$E\big[e^{-u(1-\rho)N_\rho}\big] = \frac{1}{1+\gamma u} - (1-\rho)\,\frac{c_0 u + d\,u^2}{(1+\gamma u)^2} + R_\rho(u), \tag{7.3}$$

where $\sup_{0\le u\le U} |R_\rho(u)| = o(1-\rho)$ for every $U < \infty$. If moreover $E[S^4] < \infty$, then $\sup_{0\le u\le U} |R_\rho(u)| = O\big((1-\rho)^2\big)$ for every $U < \infty$.

Proof. Put $\varepsilon = 1-\rho$, $z = e^{-u\varepsilon}$ and $\delta = z-1 = -u\varepsilon + \frac{1}{2}u^2\varepsilon^2 + O(\varepsilon^3)$. We expand the three factors of (2.3) to the order required.

The boundary factor. By the factorization $z^r H_\rho(1) - H_\rho(z) = H_\rho(1)\prod_{j=1}^{r}(z - z_j(\rho))$, $z_r(\rho) = 1$, of Tadj (2003), together with $H_\rho(1)\prod_{j<r}(1-z_j(\rho)) = r\varepsilon$ (the form of (3.1) noted after Lemma 3.1),

$$B_\rho(z) = z^r H_\rho(1) - H_\rho(z) = r\varepsilon\,(z-1)\prod_{j=1}^{r-1}\frac{z - z_j(\rho)}{1 - z_j(\rho)} = r\varepsilon\,\delta\prod_{j=1}^{r-1}\Big(1 + \frac{\delta}{1-z_j(\rho)}\Big).$$

Expanding the product, $\prod_{j<r}(1+\delta w_j) = 1 + \delta S_1(\rho) + \delta^2 e_2(\rho) + \cdots$, where $w_j = 1/(1-z_j(\rho))$, $S_1(\rho) = \sum_j w_j$ and $e_2(\rho)$ is the second elementary symmetric function of the w_j. By Lemma 5.2 the w_j are bounded for ρ near 1, so all elementary symmetric functions are $O(1)$, and since $\delta = O(\varepsilon)$ the product equals $1 + \delta\, S_1(\rho) + O(\varepsilon^2)$. Moreover $S_1(\rho) = S_1 + O(\varepsilon)$ by (5.10), and replacing $S_1(\rho)$ by S_1 costs $r\varepsilon\cdot\delta^2\cdot O(\varepsilon) = O(\varepsilon^4)$. Hence

$$B_\rho(z) = r\varepsilon\big(\delta + \delta^2 S_1\big) + O(\varepsilon^4) = -ru\,\varepsilon^2 + ru^2\Big(\frac{1}{2} + S_1\Big)\varepsilon^3 + O(\varepsilon^4). \tag{7.4}$$

The denominator. By Lemma 3.2 with $p = 3$, since $E[S^3] < \infty$,

$$\Phi_\rho(1+\delta) = r\varepsilon\,\delta - \frac{a_\rho}{2}\delta^2 + \frac{b_\rho}{6}\delta^3 + \tau_\rho(\delta), \qquad a_\rho = \lambda^2 m_2 - r(r-1), \quad b_\rho = r(r-1)(r-2) - \lambda^3 m_3,$$

where the remainder $\tau_\rho(\delta)$ is $o(|\delta|^3)$ uniformly in ρ, and is $O(|\delta|^4)$ uniformly in ρ if $E[S^4] < \infty$. Both coefficients are exact polynomials in ρ: by (2.5), $a_\rho = 2r\gamma - 2r^2(1+C_s^2)\varepsilon + r^2(1+C_s^2)\varepsilon^2$, and $b_\rho = r(r-1)(r-2) - r^3\rho^3\kappa_3 = b + 3r^3\kappa_3\varepsilon + O(\varepsilon^2)$ with $b = r(r-1)(r-2) - r^3\kappa_3$. Substituting $\delta = -u\varepsilon + \frac{1}{2}u^2\varepsilon^2 + O(\varepsilon^3)$ and collecting powers of ε,

$$\Phi_\rho(z) = -ru(1+\gamma u)\,\varepsilon^2 + \Big[\frac{ru^2}{2} + r\gamma u^3 + r^2(1+C_s^2)u^2 - \frac{b}{6}u^3\Big]\varepsilon^3 + \tau_\rho(\delta) + O(\varepsilon^4), \tag{7.5}$$

where the $O(\varepsilon^4)$ collects the polynomial terms of order four and higher (which involve only a_ρ, b_ρ and the expansion of δ) and $\tau_\rho(\delta) = o(\varepsilon^3)$, respectively $O(\varepsilon^4)$ under the fourth-moment assumption, since $|\delta| \le u\varepsilon$.

The numerator pgf. By Lemma 3.2 with $p = 1$ and $E[S^2] < \infty$, $F_\rho(z) = 1 + \lambda m_1\delta + O(\varepsilon^2) = 1 - ru\varepsilon + O(\varepsilon^2)$.

Assembly. Write (7.4) as $\varepsilon^2(B_2 + B_3\varepsilon + O(\varepsilon^2))$ and (7.5) as $\varepsilon^2(\Phi_2 + \Phi_3\varepsilon + \varepsilon^{-2}\tau_\rho(\delta) + O(\varepsilon^2))$, with $\Phi_2 = -ru(1+\gamma u) \neq 0$. Then

$$\frac{B_\rho(z)}{\Phi_\rho(z)} = \frac{B_2}{\Phi_2}\Big[1 + \Big(\frac{B_3}{B_2} - \frac{\Phi_3}{\Phi_2}\Big)\varepsilon\Big] + R_\rho(u), \qquad R_\rho(u) = O\big(\varepsilon^{-2}|\tau_\rho(\delta)|\big) + O(\varepsilon^2),$$

so that $R_\rho(u) = o(\varepsilon)$ in general and $R_\rho(u) = O(\varepsilon^2)$ under the fourth-moment assumption; this is the only place where the quality of the Taylor remainder enters. The uniformity in $u \in [0, U]$ is read off from the same bookkeeping: every term in (7.4) and (7.5) carries at least the power of u shown, so $B_\rho(z)$ and $\Phi_\rho(z)$ are both $-ru\varepsilon^2$ times $1 + O(u\varepsilon)$, uniformly for $u \in [0, U]$, with $\Phi_2/(-ru) = 1 + \gamma u$ bounded away from 0; and $|\delta| \le u\varepsilon$ gives $|\tau_\rho(\delta)| \le (u\varepsilon)^3\omega(u\varepsilon)$ with ω from Lemma 3.2, so $\varepsilon^{-2}|\tau_\rho(\delta)|/(ru) \le u^2\varepsilon\,\omega(U\varepsilon) = o(\varepsilon)$ uniformly on $[0, U]$ (and $O(\varepsilon^2)$ under four

moments). At $u = 0$ both sides of (7.3) equal 1. We now collect the terms explicitly. With $B_2 = -ru$, $B_3 = ru^2(\frac{1}{2} + S_1)$, $\Phi_2 = -ru(1+\gamma u)$ and $\Phi_3 = \frac{ru^2}{2} + r\gamma u^3 + r^2(1+C_s^2)u^2 - \frac{b}{6}u^3$,

$$\frac{B_3}{B_2} = -u\Big(\frac{1}{2} + S_1\Big), \qquad \frac{\Phi_3}{\Phi_2} = -\frac{\frac{u}{2} + r(1+C_s^2)u + u^2\big(\gamma - \frac{b}{6r}\big)}{1+\gamma u},$$

so that, before multiplication by $F_\rho(z)$,

$$\frac{B_\rho(z)}{\Phi_\rho(z)} = \frac{1}{1+\gamma u} + \varepsilon\,\frac{u\big(r(1+C_s^2) - S_1\big) + u^2\big(\gamma(\frac{1}{2} - S_1) - \frac{b}{6r}\big)}{(1+\gamma u)^2} + R_\rho(u), \tag{7.6}$$

where the numerator was obtained by writing $-u(\frac{1}{2}+S_1) = -u(\frac{1}{2}+S_1)(1+\gamma u)/(1+\gamma u)$ and adding. Multiplying (7.6) by $F_\rho(z) = 1 - ru\varepsilon + O(\varepsilon^2)$ adds $-ru\varepsilon/(1+\gamma u) = -\varepsilon(ru + r\gamma u^2)/(1+\gamma u)^2$ at order ε. The coefficient of u in the resulting numerator is therefore $r(1+C_s^2) - S_1 - r = -(S_1 - rC_s^2) = -c_0$, and the coefficient of u^2 is $\gamma(\frac{1}{2} - S_1) - \frac{b}{6r} - r\gamma = -\big[\gamma(r - \frac{1}{2} + S_1) + \frac{(r-1)(r-2) - r^2\kappa_3}{6}\big] = -d$, using $b/(6r) = \big((r-1)(r-2) - r^2\kappa_3\big)/6$. This is (7.3). □

The uniformity in u converts the transform expansion into an expansion for a class of smooth functionals of the scaled queue length.

Corollary 7.1. *Under the assumptions of Theorem 7.1, let $g(x) = \int_{[0,U]} e^{-ux}\,\nu(du)$ for a finite signed measure ν on a bounded interval $[0,U]$; examples are finite linear combinations of exponentials $e^{-u_k x}$, and $g(x) = (1-e^{-Ux})/x$, for which ν is Lebesgue measure on $[0,U]$. Then*

$$E\big[g\big((1-\rho)N_\rho\big)\big] = \int_{[0,U]} \Big[\frac{1}{1+\gamma u} - (1-\rho)\frac{c_0 u + du^2}{(1+\gamma u)^2}\Big]\nu(du) + o(1-\rho),$$

with $O((1-\rho)^2)$ in place of $o(1-\rho)$ if $E[S^4] < \infty$.

Proof. By Fubini's theorem, $E[g(\varepsilon N_\rho)] = \int E[e^{-u\varepsilon N_\rho}]\,\nu(du)$, and the error term is bounded by $|\nu|([0,U])$ times $\sup_{0\le u\le U}|R_\rho(u)|$. □

The class of functionals is narrow, and deliberately so: for indicator functions, which would give the distribution function itself, no such expansion can hold pointwise, as Remark 7.1(d) shows.

Table 4 checks the constants of Theorem 7.1 numerically for Erlang service, where neither the exact $M/M/1$ transform nor the geometric solution of Remark 7.1(c) applies. The transform $E[e^{-u\varepsilon N_\rho}]$ is computed from the stationary distribution of the truncated embedded chain, as in Section 6.2, and the quantity $\big(E[e^{-u\varepsilon N_\rho}] - (1+\gamma u)^{-1}\big)/\varepsilon$ is compared with the predicted $-(c_0 u + du^2)/(1+\gamma u)^2$. The discrepancy halves when ε halves, as it must if the remainder is $O(\varepsilon^2)$ (Erlang service has all moments).

Remark 7.1. (a) The u-linear coefficient of the correction is c_0, so formal differentiation of (7.3) at $u = 0$ returns $(1-\rho)L_\rho^+ = \gamma + (1-\rho)c_0 + \cdots$, which is Theorem 6.1. (The differentiation is only formal, since (7.3) is a statement for fixed $u > 0$; Theorem 6.1 is proved independently and needs only $E[S^2] < \infty$ for its $O(1-\rho)$ remainder, because it never expands the pgf.)

(a″) The correction in (7.3) depends on the roots only through the critical-load sum S_1, not through the rate $S_1'(1)$ at which the root sum approaches it: the difference $S_1(\rho) - S_1 = O(1-\rho)$ enters $B_\rho(z)$ only at order ε^4, as the proof shows. This is why the second-order constants of both Theorem 6.1 and Theorem 7.1 are determined by the critical-load equation alone.

Service	r	u	$\varepsilon = 0.04$	$\varepsilon = 0.02$	$\varepsilon = 0.01$	predicted
Exponential	3	1	0.3024	0.2993	0.2978	0.2963
Erlang-2	3	1	−0.0956	−0.1125	−0.1213	−0.1304
Erlang-2	3	3	−0.1878	−0.2319	−0.2567	−0.2838
Erlang-3	2	1	−0.0988	−0.1136	−0.1212	−0.1290
Erlang-3	2	3	−0.2025	−0.2435	−0.2657	−0.2893

Table 4: $\big(E[e^{-u\varepsilon N_\rho}]-(1+\gamma u)^{-1}\big)/\varepsilon$ from the exact stationary distribution, against the coefficient $-(c_0u+du^2)/(1+\gamma u)^2$ of Theorem 7.1; $\varepsilon = 1-\rho$.

(a′) The remainder in Theorem 7.1 is $o(1-\rho)$ under a third moment and $O((1-\rho)^2)$ under a fourth. This is the usual price in refined heavy-traffic expansions, where each further order in $1-\rho$ costs one further moment, cf. Siegmund (1979) and Blanchet and Glynn (2006) for the Lindley recursion: the coefficient d is determined by the third moment, but the rate at which the next term is controlled is governed by the Taylor remainder of F_ρ, which is $o(|\delta|^3)$ with three moments and $O(|\delta|^4)$ with four. The mean expansion of Theorem 6.1 escapes this because it starts from the exact root representation rather than from the pgf.

(b) For $r=1$ the queue is $M/G/1$ and $c_0=-C_s^2$, $S_1=0$, $d=\frac{\gamma}{2}-\frac{\kappa_3}{6}$. For $M/M/1$ ($C_s^2=1$, $\kappa_3=6$) this gives $\gamma=1$, $c_0=-1$, $d=-\frac12$, and (7.3) reads $\frac{1}{1+u}+\varepsilon\frac{u(1+u/2)}{(1+u)^2}+O(\varepsilon^2)$ (all moments being finite), which is the exact expansion of $E[e^{-u\varepsilon N}]=\varepsilon/(1-(1-\varepsilon)e^{-u\varepsilon})$.

(c) For exponential service in general ($C_s^2=1$, $\kappa_3=6$, $S_1=2(r-1)/3$, $\gamma=(r+1)/2$) the constant simplifies to

$$d=-\frac{r+1}{4}=-\frac{\gamma}{2},\qquad \text{so that, since } -(c_0u+du^2)=\tfrac{r+2}{3}u+\tfrac{r+1}{4}u^2=u\big(\tfrac{r+2}{3}+\tfrac{\gamma}{2}u\big),$$

$$E\big[e^{-u(1-\rho)N_\rho}\big]=\frac{1}{1+\gamma u}+(1-\rho)\,\frac{u\big(\frac{r+2}{3}+\frac{\gamma}{2}u\big)}{(1+\gamma u)^2}+O\big((1-\rho)^2\big),$$

the remainder being $O((1-\rho)^2)$ since the exponential law has moments of all orders; the same applies to Erlang service. This formula can be verified independently of the proof of Theorem 7.1, for every r, from the exact geometric solution $P(N_\rho=n)=(1-\theta)\theta^n$ of the $M/M^r/1$ queue (Tadj, 2003), where $\theta=1/z_{r+1}(\rho)$ solves $\theta+\theta^2+\cdots+\theta^r=r\rho$. Writing $\theta=1-a\varepsilon+b\varepsilon^2+O(\varepsilon^3)$ and expanding, one finds $a=2/(r+1)=1/\gamma$ and $b=-4(r-1)/(3(r+1)^2)$, so that $1-b\gamma^2=(r+2)/3$, and then

$$E\big[e^{-u\varepsilon N_\rho}\big]=\frac{1-\theta}{1-\theta e^{-u\varepsilon}}=\frac{1}{1+\gamma u}+\varepsilon\,\frac{u(1-b\gamma^2)+\frac{\gamma}{2}u^2}{(1+\gamma u)^2}+O(\varepsilon^2),$$

in agreement with the display above. Since this computation uses only the geometric form of the solution and the characteristic equation, it checks the constants c_0 and d of Theorem 7.1 for exponential service and all quorum sizes, not merely for $r=1$.

(d) *Formal inversion and the lattice effect.* Writing $u/(1+\gamma u)^2$ and $u^2/(1+\gamma u)^2$ as combinations of 1, $(1+\gamma u)^{-1}$ and $(1+\gamma u)^{-2}$, the ε-term of (7.3) is the Laplace transform of the signed measure

$$-\frac{d}{\gamma^2}\,\delta_0(dx)+\Big[\frac{2d-c_0\gamma}{\gamma^3}+\frac{c_0\gamma-d}{\gamma^4}\,x\Big]e^{-x/\gamma}\,dx,$$

of total mass zero, so that, *formally* (the relation is not asserted pointwise; see below),

$$P\big((1-\rho)N_\rho > x\big) \text{ “=” } e^{-x/\gamma}\Big[1+(1-\rho)\Big(\frac{d}{\gamma^2}+\frac{c_0\gamma-d}{\gamma^3}\,x\Big)\Big]+o(1-\rho), \qquad x>0. \tag{7.7}$$

The atom at 0 carries the $O(1-\rho)$ mass of the boundary states $\{0,\dots,r-1\}$, cf. (3.2). Relation (7.7) cannot hold pointwise, because N_ρ is lattice: already for $M/M/1$, $P(\varepsilon N>x)=\rho^{\lfloor x/\varepsilon\rfloor+1}=e^{-x}\big[1-\varepsilon\big(\frac{x}{2}+1-\{x/\varepsilon\}\big)\big]+O(\varepsilon^2)$, whose $O(\varepsilon)$ term oscillates with the fractional part $\{x/\varepsilon\}$ between $-e^{-x}\varepsilon(x/2+1)$ and $-e^{-x}\varepsilon x/2$, while (7.7) (with $\gamma=1$, $c_0=-1$, $d=-\frac12$) gives the average, $-e^{-x}\varepsilon(x/2+\frac12)$. Thus (7.7) describes the tail smoothed over the lattice, and the rigorous content of the second-order correction is Theorem 7.1 together with Corollary 7.1: the $O(1-\rho)$ term is a signed measure of total mass zero, and it is a correction to expectations of functionals in the class of Corollary 7.1, not to probabilities of events. The atom at 0 is the lattice analogue of the boundary-layer corrections that Knessl (1990) obtains for the continuous $M/G/1$ workload, where the higher-order terms take different forms near the origin and in the bulk.

Corollary 7.2. *Under the assumptions of Theorem 7.1, the arbitrary-time queue length satisfies* (7.3)*, with the same remainder $R_\rho(u)$, with c_0 and d replaced by*

$$c_0+\frac{r-1}{2} \qquad \textit{and} \qquad d+\frac{\gamma(r-1)}{2}.$$

Proof. By (4.2), $E[e^{-u\varepsilon Q_\rho}]=E[e^{-u\varepsilon N_\rho}]\,E[e^{-u\varepsilon J}]$ with $E[e^{-u\varepsilon J}]=\frac1r\sum_{j=0}^{r-1}e^{-ju\varepsilon}=1-\frac{r-1}{2}u\varepsilon+O(\varepsilon^2)$. Multiplying by (7.3),

$$E\big[e^{-u\varepsilon Q_\rho}\big]=\frac{1}{1+\gamma u}-\varepsilon\Big[\frac{c_0u+du^2}{(1+\gamma u)^2}+\frac{r-1}{2}\,\frac{u}{1+\gamma u}\Big]+R_\rho(u)+O(\varepsilon^2),$$

and since $\frac{u}{1+\gamma u}=\frac{u+\gamma u^2}{(1+\gamma u)^2}$ the bracket equals $\frac{(c_0+\frac{r-1}{2})u+(d+\frac{\gamma(r-1)}{2})u^2}{(1+\gamma u)^2}$. The shift of the u-linear coefficient by $(r-1)/2$ is consistent with (5.5), as formal differentiation at $u=0$ shows, and for $r=1$ both shifts vanish, as they must since then $Q_\rho=N_\rho$. □

8 Discussion and conclusion

The stationary queue length of the fixed-size $M/G^r/1$ quorum queue grows on the scale $1/(1-\rho)$ as the system approaches critical loading, and $(1-\rho)N_\rho$ converges to an exponential law with rate $2/(1+rC_s^2)$, the same at departure epochs and at arbitrary times. The rate separates two influences on congestion: service-time variability through C_s^2, and the quorum size through the factor r. This first-order behavior is what Kingman's theorem predicts for the underlying reflected random walk, and Theorem 3.1 confirms that the quorum boundary does not disturb it.

The paper's main result concerns the next order. The exact root-based representation of the stationary mean, combined with the convergence of the interior characteristic roots to their critical-load limits (Lemma 5.4), yields the two-term expansion of Theorem 6.1, with the bounded correction $c_0=-rC_s^2+\sum_{j<r}1/(1-z_j(1))$. Unlike the leading term, c_0 is a property of the full service-time distribution, carried by the characteristic roots at $\rho=1$; for exponential service it equals $-(r+2)/3$, and for Erlang-k service it is determined by the $k-1$ exterior roots

at critical loading (Corollary 6.2). The numerical evidence of Table 2 confirms both the constant and the $O(1-\rho)$ rate of the remainder.

The sign of c_0 deserves comment. For exponential service $c_0 < 0$, so the leading term $(1+r)/(2(1-\rho))$ overestimates the mean. In the cases computed, the correction increases with the regularity of the service time and changes sign (Tables 2 and 1): for Erlang-3 service with $2 \le r \le 5$, for Erlang-2 with $r = 4, 5$, and for deterministic service with $r = 2, 3, 5$, one has $c_0 > 0$, and the leading term *under*estimates the mean. We have not attempted to prove a general monotonicity statement. A first-order heavy-traffic approximation of Kingman type, which knows the service-time distribution only through C_s^2, therefore cannot even predict the direction of its own error; the characteristic roots at $\rho = 1$ do.

Section 7 shows that the same roots govern the second-order term of the Laplace transform of the scaled queue length, not only of the mean: the Laplace transform of $(1-\rho)N_\rho$ has an $O(1-\rho)$ correction whose coefficients are c_0 and a constant d involving the third moment of S, with a remainder $o(1-\rho)$ that becomes $O((1-\rho)^2)$ under a fourth moment, and the exact $M/M/1$ transform is recovered at $r = 1$. Because the queue length is lattice-valued, the pointwise tail inherits an $O(1-\rho)$ oscillation, and the inverted correction (7.7) is to be read as a lattice-smoothed statement.

A natural continuation is the general bulk-service queue $M/G^{a,b}/1$, in which the server starts when at least a customers are present and serves up to b at a time. Its characteristic equation is $z^b = F_\rho(z)$, with $b-1$ interior roots, so Lemmas 5.2 and 5.4 carry over with r replaced by b and $\rho = \lambda m_1/b$; the root analysis presents no new difficulty. What changes is the boundary: the idle states $\{0, \dots, a-1\}$ and the partially-filled states $\{a, \dots, b-1\}$ enter the boundary polynomial differently, so the identity of Lemma 3.1 and the exact mean formula must be rederived with the two thresholds, and the constant c_0 will depend on a through the boundary terms as well as on the roots. We expect the resulting expansion to have the same form, with the leading term depending on b and C_s^2 only.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

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

[2] Chaudhry, M. L. and Templeton, J. G. C. (1981). The queuing system $M/G^B/1$ and its ramifications. *European Journal of Operational Research*, **6**(1), 56–60.

[3] Chaudhry, M. L. and Templeton, J. G. C. (1983). *A First Course in Bulk Queues*. Wiley, New York.

[4] Blanchet, J. and Glynn, P. (2006). Complete corrected diffusion approximations for the maximum of a random walk. *Annals of Applied Probability*, **16**(2), 951–983.

[5] Chiamsiri, S. and Leonard, M. S. (1981). A diffusion approximation for bulk queues. *Management Science*, **27**(10), 1188–1199.

[6] Dshalalow, J. H. and Tadj, L. (1992). Queueing system with a fixed accumulation level, random server capacity, and capacity dependent service time. *International Journal of Mathematics and Mathematical Sciences*, **15**(1), 189–194.

[7] Kingman, J. F. C. (1961). The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, **57**(4), 902–904.

[8] Kingman, J. F. C. (1962). On queues in heavy traffic. *Journal of the Royal Statistical Society, Series B*, **24**(2), 383–392.

[9] Knessl, C. (1990). Refinements to heavy traffic limit theorems in queueing theory. *Operations Research*, **38**(5), 826–837.

[10] Medhi, J. (1984). *Recent Developments in Bulk Queueing Models*. Wiley Eastern, New Delhi.

[11] Siegmund, D. (1979). Corrected diffusion approximations in certain random walk problems. *Advances in Applied Probability*, **11**(4), 701–719.

[12] Tadj, L. (2003). Explicit solution of a quorum queueing system. *Stochastic Analysis and Applications*, **21**(3), 703–717.

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