Adv. Appl. Prob. **43,** 243–263 (2011)
Printed in Northern Ireland
© *Applied Probability Trust* 2011

THE ASYMPTOTIC VARIANCE OF DEPARTURES IN CRITICALLY LOADED QUEUES

A. AL HANBALI,[*] *University of Twente*
M. MANDJES,[**] *University of Amsterdam and CWI*
Y. NAZARATHY,[***] *EURANDOM and Eindoven University of Technology*
W. WHITT,[****] *Columbia University*

Abstract

We consider the asymptotic variance of the departure counting process $D(t)$ of the GI/G/1 queue; $D(t)$ denotes the number of departures up to time t. We focus on the case where the system load ϱ equals 1, and prove that the asymptotic variance rate satisfies $\lim_{t\to\infty} \operatorname{var} D(t)/t = \lambda(1-2/\pi)(c_a^2+c_s^2)$, where λ is the arrival rate, and c_a^2 and c_s^2 are squared coefficients of variation of the interarrival and service times, respectively. As a consequence, the departures variability has a remarkable singularity in the case in which ϱ equals 1, in line with the BRAVO (balancing reduces asymptotic variance of outputs) effect which was previously encountered in finite-capacity birth–death queues. Under certain technical conditions, our result generalizes to multiserver queues, as well as to queues with more general arrival and service patterns. For the M/M/1 queue, we present an explicit expression of the variance of $D(t)$ for any t.

Keywords: GI/G/1 queue; critically loaded system; uniform integrability; departure process; renewal theory; Brownian bridge; multiserver queue

2010 Mathematics Subject Classification: Primary 90B22
Secondary 60G55

1. Introduction

In the study of queueing systems, the analysis of departure processes has played an important role. Following Burke's theorem [5], which states that departures of a stationary M/M/1 queue form a Poisson process, many papers have dealt with properties of interdeparture times, departure counting processes, and approximations. A classic survey is by Daley [8], while other useful references in this area are [9] and [10, Chapter VII].

A key object in the analysis of departure processes is the variance of the number of departures between time 0 and t, in the sequel denoted by $D(t)$; see, e.g. [7]. From an application point of view, insight into $\operatorname{var} D(t)$ is of crucial importance in the performance analysis of supply chain and manufacturing networks; several recent studies have investigated approximations for departure processes in complex queueing systems; see, e.g. [20] and the references therein.

Received 8 May 2010; revision received 12 October 2010.

[*] Postal address: School of Management and Governance, University of Twente, Enschede, The Netherlands. Email address: a.m.alhanbali@utwente.nl
[**] Postal address: Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands. Email address: m.r.h.mandjes@uva.nl
[***] Current address: Faculty of Engineering and Industrial Sciences, Swinburne University of Technology, John Street, Hawthorn, Victoria, 3122 Australia. Email address: ynazarathy@swin.edu.au
[****] Postal address: Department of Industrial Engineering and Operations Research, Columbia University, S. W. Mudd Building, 500 West 120th Street, New York, NY 10027, USA. Email address: ww2040@columbia.edu

Related research deals with decoupling queueing networks into subsystems where the output of one or several queues is fed as an input to other queues; see, e.g. [24]. In such cases, it is of crucial importance to understand the structure of var $D(t)$.

1.1. Main result

In this paper we contribute to the analysis of var $D(t)$ by considering the *critically loaded* GI/G/1 queue. This critically loaded regime, in which the mean interarrival time equals the mean service time, is relevant from a practical standpoint (as in many real-life situations queues are saturated or close to saturation). Moreover, it is mathematically interesting since it leads to counterintuitive results in line with the BRAVO (balancing reduces asymptotic variance of outputs) effect observed previously in finite-capacity birth–death queues [17]; see also [16].

We now describe the contribution of our work and its relation to BRAVO in more detail. We let ζ_A represent a generic interarrival time and ζ_S represent a generic service time. We denote the system load by $\varrho := \lambda/\mu$, with $\lambda := 1/\mathrm{E}\,\zeta_A$ and $\mu := 1/\mathrm{E}\,\zeta_S$, and we let the squared coefficients of variations (ratio of variance and square of the mean) of ζ_A and ζ_S be c_a^2 and c_s^2, respectively. We study the asymptotic variance of the departure process, defined as

$$\sigma := \lim_{t\to\infty} \frac{\operatorname{var} D(t)}{t},$$

when the queue is critically loaded, that is, $\varrho = 1$. Under suitable regularity conditions, it is not hard to prove that

$$m := \lim_{t\to\infty} \frac{\mathrm{E}\,D(t)}{t} = \min\{\lambda, \mu\},$$

whereas $\sigma = \lambda c_a^2$ for $\varrho < 1$ and $\sigma = \mu c_s^2$ for $\varrho > 1$. However, there is evidently no explicit expression for σ in the case $\varrho = 1$ in the literature. We show that

$$\sigma = \lambda\left(1 - \frac{2}{\pi}\right)(c_a^2 + c_s^2), \qquad \varrho = 1. \tag{1.1}$$

It thus follows that the variability function $v(\varrho) := \sigma/m = \lim_{t\to\infty} \operatorname{var} D(t)/\mathrm{E}\,D(t)$ has a singular point at $\varrho = 1$, which can be regarded as a manifestation of the BRAVO phenomenon. More specifically, for $\varrho \neq 1$, $v(\varrho)$ is essentially determined by either the arrival or the service process; for $\varrho = 1$, $v(\varrho)$ is determined by both the arrival and service processes. Consider, for instance, the M/M/1 queue; then $v(\varrho) = 1$ for $\varrho \neq 1$, but it is *reduced* to $2(1 - 2/\pi) \approx 0.72$ at $\varrho = 1$.

In addition to the GI/G/1, (1.1) is a fundamental quantity which appears in a variety of critically loaded systems. We show that it holds for the GI/G/s queue (with $s \in \mathbb{N}$ servers), and generalizes to multichannel, multiserver queues with more general (nonrenewal) arrival and service patterns (see Theorems 6.1 and 6.2). We also demonstrate numerically that, when $\rho \approx 1$ (but is not necessarily equal to 1), the variance for finite t approximately follows (1.1); see Figure 1 in Section 5. This numerical experiment illustrates that the BRAVO phenomenon may also be observed in practice in systems that are nearly critically loaded.

Our starting point for obtaining (1.1) is a diffusion limit presented in [12, Section 4], where it was shown that, for critically loaded queues, the sequence of processes

$$\hat{D}_n(t) = \frac{D(nt) - \lambda nt}{\sqrt{n}}, \qquad n = 1, 2, \dots,$$

converges weakly to

$$\hat{D}(t) = \inf_{0 \le s \le t} \{c_a^2 B_1(s) + c_s^2 B_2(t-s)\},$$

where $B_1(\cdot)$ and $B_2(\cdot)$ are independent standard Brownian motions. It then turns out that $\sigma = \lambda \operatorname{var} \hat{D}(1)$, given suitable uniform integrability (UI) conditions.

1.2. Auxiliary technical results

On route to obtaining (1.1), we derive some important auxiliary results which are of independent interest. We first identify the distribution of $\hat{D}(1)$ (which, for brevity, we denote simply by $\hat{D}$) and show that $\operatorname{var} \hat{D} = (1 - 2/\pi)(c_a^2 + c_s^2)$.

We then continue to attack the required UI conditions. This is often a challenging task in applied probability. We remind the reader that a collection of random variables $\{Z_t\}$ is uniformly integrable if

$$\lim_{M \to \infty} \Bigl(\sup_t \mathrm{E}\, |Z_t|\, \mathbf{1}_{\{|Z_t| \ge M\}} \Bigr) = 0.$$

A well-known sufficient condition is to have

$$\sup_t \mathrm{E}(|Z_t|^{1+\varepsilon}) < \infty \quad \text{for some } \varepsilon > 0.$$

We denote by $Z_t \Rightarrow Z$ the fact that Z_t converges in distribution to Z. In case Z_t is uniformly integrable, this also implies that $\lim_{t\to\infty} \mathrm{E}\, Z_t = \mathrm{E}\, Z$; see [4].

In handling the UI conditions, our problem narrows down to proving that the sequence $\{Q(t)^2/t\}$ is uniformly integrable, where $Q(t)$ denotes the number of customers in the system in our critically loaded queue. Using a sequence of steps which rely on the reflection map for the queue length, stochastic ordering results, and renewal-theoretic results, we are able to establish UI for the GI/NWU/1 queue (where NWU stands for *new worse than used*—see Section 4.2 for a definition). We also find that a sufficient condition for the UI requirement is that

$$\mathrm{P}(B > x) \sim L(x) x^{-1/2},$$

where B is a generic busy period and $L(\cdot)$ is a slowly varying function (i.e. $L(ax)/L(x) \to 1$ as $x \to \infty$ for every $a > 0$) that is bounded by a constant.

We refer to Theorem 2.2 for an exact statement of our results. It should be noted that we believe that the complications when establishing the UI requirement are primarily of a technical nature, and that we in fact believe that (1.1) holds for a broader class of critically loaded GI/G/1 queues. This conjecture is formalized following Theorem 2.2. We are also able to handle the UI conditions for some GI/G/s queues (Theorem 6.2).

To complement our asymptotic results, we perform an explicit analysis for the departure process of the M/M/1 queue, and obtain $\operatorname{var} D(t)$ at all time points in terms of Bessel functions. This yields an alternative derivation of (1.1) for this case as well as other more refined properties.

1.3. Organization

This paper is organized as follows. In Section 2 we present the main result and formulate related conjectures. In Section 3 we derive the distribution of $\hat{D}$, and compute the explicit expression for $\operatorname{var} \hat{D}$. In Section 4 we find conditions under which the process $\{Q(t)^2/t\}$ is uniformly integrable. In Section 5 we find the variance curve of the M/M/1 queue. We conclude in Section 6 with a discussion on the extensions to the multiserver GI/G/s queue, as well as to queues with more general arrival and service patterns.

2. Main results

In this section we present our main results on the critically loaded GI/G/1 queue operating under the first-come–first-served (FCFS) discipline. Assume that $Q(0) = 0$, and denote by $A(t)$ the number of arrivals during $[0, t]$. In addition, assume that the first interarrival time is identically distributed to the generic interarrival time ζ_A. We further denote by $S(t)$ the renewal counting process induced by the service times. A key role is played by the process $\mathcal{Q}$, defined as

$$\mathcal{Q} = \left\{ \frac{Q(t)^2}{t}, \ t \geq t_0 \right\}$$

for some $t_0 > 0$.

We state our main result in Theorems 2.1 and 2.2 below. We first establish the following elementary lemma which is used throughout the paper.

Lemma 2.1. *For real x and y and $r \geq 1$,*

$$|x + y|^r \leq 2^{r-1}(|x|^r + |y|^r). \tag{2.1}$$

Proof. It is easy to check that, for $z \geq 0$, the function $f(z) = (1 + z)^r/(1 + z^r)$ is globally maximized when $z = 1$ and attains a maximum of 2^{r-1}. The result follows by taking $z = |y|/|x|$ and using the triangle inequality.

Theorem 2.1. *Consider the critically loaded GI/G/1 queue with $\mathrm{E}\,\zeta_A^2 < \infty$ and $\mathrm{E}\,\zeta_S^2 < \infty$. Assume that $\mathcal{Q}$ is uniformly integrable. Then*

$$\sigma = \lambda\left(1 - \frac{2}{\pi}\right)(c_a^2 + c_s^2). \tag{2.2}$$

Proof. From the heavy-traffic functional central limit theorem in [12, Theorem 4.1], upon applying the projection map (at the time $t = 1$) and the continuous mapping theorem, we have

$$\frac{D(t) - \lambda t}{\sqrt{\lambda t}} \Rightarrow \hat{D} \quad \text{as } t \to \infty. \tag{2.3}$$

Furthermore, using the continuous mapping theorem, we obtain

$$\frac{(D(t) - \lambda t)^2}{\lambda t} \Rightarrow \hat{D}^2 \quad \text{as } t \to \infty. \tag{2.4}$$

Under the UI conditions established below, we have, from (2.3) and (2.4),

$$\lim_{t\to\infty} \mathrm{E}\left(\left(\frac{D(t) - \lambda t}{\sqrt{\lambda t}}\right)^k\right) = \mathrm{E}(\hat{D}^k), \qquad k = 1, 2. \tag{2.5}$$

Observe that $\operatorname{var} D(t) = \mathrm{E}(D(t) - \lambda t)^2 - (\mathrm{E}\,D(t) - \lambda t)^2$, and combine this with (2.5) to obtain

$$\begin{aligned}
\frac{\sigma}{\lambda} &= \lim_{t\to\infty} \frac{\operatorname{var} D(t)}{\lambda t} \\
&= \lim_{t\to\infty} \frac{\mathrm{E}(D(t) - \lambda t)^2}{\lambda t} - \left(\lim_{t\to\infty} \frac{\mathrm{E}\,D(t) - \lambda t}{\sqrt{\lambda t}}\right)^2 \\
&= \operatorname{var} \hat{D},
\end{aligned}$$

which yields the desired result using Proposition 3.1 below.

It now remains to establish the convergence of the moments in (2.5). To do so, we establish that the sequences $\{[(D(t)-\lambda t)/\sqrt{\lambda t}]^k,\ t \geq t_0\}$, $k = 1, 2$, are uniformly integrable. First note that $D(t) = A(t) - Q(t)$. Combining this with (2.1) yields

$$\left|\frac{D(t)-\lambda t}{\sqrt{\lambda t}}\right| \leq \left|\frac{A(t)-\lambda t}{\sqrt{\lambda t}}\right| + \left|\frac{Q(t)}{\sqrt{\lambda t}}\right|, \qquad \left|\frac{D(t)-\lambda t}{\sqrt{\lambda t}}\right|^2 \leq 2\left(\left|\frac{A(t)-\lambda t}{\sqrt{\lambda t}}\right|^2 + \left|\frac{Q(t)}{\sqrt{\lambda t}}\right|^2\right).$$

It thus suffices to show that the sequences $\{(A(t)-\lambda t)^2/\lambda t,\ t \geq t_0\}$ and $\mathcal{Q}$ are uniformly integrable. UI of the first sequence is a standard result from renewal theory; cf. [11, p. 49]. UI of the second sequence is an assumption (which we partially prove in Theorem 2.2 below).

The above theorem is generalized in Section 6 for multichannel, multiserver queues with more general arrival and service processes. We are able to establish the UI of $\mathcal{Q}$ needed by Theorem 2.1 for different cases.

Theorem 2.2. *If* $\mathrm{E}\,\zeta_A^4 < \infty$ *and* $\mathrm{E}\,\zeta_S^4 < \infty$, *then* $\mathcal{Q}$ *is uniformly integrable in the following cases.*

(i) *Any critically loaded GI/G/1 queue with* $\mathrm{P}(B > x) \sim L(x)x^{-1/2}$, *where* $L(\cdot)$ *is a bounded, slowly varying function.*

(ii) *The critically loaded M/G/1 queue.*

(iii) *The critically loaded GI/NWU/1 queue.*

(iv) *The critically loaded D/G/1 queue with* $\mathrm{P}(\zeta_S > b) = 1$ *for some* $b > 0$.

The theorem is proved by a sequence of arguments in Section 4. A version of this theorem for the GI/G/s queue is given in Section 6.

We conjecture that our result also holds under milder conditions. To this end, we first remark that the condition in Theorem 2.2(i) has been shown to be true in [27] for the critically loaded M/G/1 queue with $\mathrm{E}\,\zeta_S^2 < \infty$. We conjecture that this also holds for the critically loaded GI/G/1.

Conjecture 2.1. *For the critically loaded GI/G/1 queue with* $\mathrm{E}\,\zeta_A^2 < \infty$ *and* $\mathrm{E}\,\zeta_S^2 < \infty$,

$$\mathrm{P}(B > x) \sim L(x)x^{-1/2},$$

where $L(\cdot)$ *is a bounded, slowly varying function.*

Conjecture 2.1, along with Theorem 2.2(i), implies UI for all GI/G/1 queues with finite fourth moments. We also conjecture that the fourth moment condition may be reduced to $2+\varepsilon$ moments for some strictly positive ε. Combining this with the multiserver result of Section 6, we conjecture the following.

Conjecture 2.2. *Consider the critically loaded GI/G/s multiserver queue. Assume that*

$$\mathrm{E}\,\zeta_A^{2+\varepsilon} < \infty \quad \textit{and} \quad \mathrm{E}\,\zeta_S^{2+\varepsilon} < \infty$$

for any $\varepsilon > 0$. *Then (2.2) holds.*

3. The distribution of $\hat{D}$

In this section we derive the distribution of the random variable $\hat{D}$, defined as

$$\inf_{0 \leq t \leq 1} \{c_1 B_1(t) + c_2 B_2(1-t)\},$$

where B_1 and B_2 are two independent standard Brownian motions. This answers an open question posed in [12]. As usual, $\Phi(x)$ is the distribution function of a standard normal random variable.

Theorem 3.1. *Let $c_1, c_2 \geq 0$. Then*

$$\begin{aligned}\mathrm{P}(\hat{D} \leq x) = {} & \Phi\Big(\frac{x}{c_1}\Big) + \Phi\Big(\frac{x}{c_2}\Big) - \Phi\Big(\frac{x}{c_1}\Big)\Phi\Big(\frac{x}{c_2}\Big) \\ & + \frac{1}{\sqrt{2\pi}} \int_0^\infty \mathrm{e}^{-L(u,x)} \Phi(-M(u,x)) \,\mathrm{d}u, \end{aligned} \tag{3.1}$$

where

$$L(u,x) := \frac{1}{2}\Big(\frac{u(c_1^2 - c_2^2)}{\check{c}^2} - \frac{x}{c_1}\Big)^2, \qquad M(u,x) := \frac{2uc_1c_2}{\check{c}^2} + \frac{x}{c_2}, \qquad \check{c} := \sqrt{c_1^2 + c_2^2}.$$

For the case where $c_1 = c_2 = c$, the last term on the right-hand side of (3.1) simplifies to

$$\frac{\mathrm{e}^{-x^2/(2c^2)}}{\sqrt{2\pi}}\Big(\frac{\mathrm{e}^{-x^2/(2c^2)}}{\sqrt{2\pi}} - \frac{x\Phi(-x/c)}{c}\Big).$$

Proof. Define the event

$$\mathscr{E}(b_1, b_2) := \{B_1(1) = b_1,\ B_2(1) = b_2\}$$

for arbitrary b_1 and b_2. Furthermore, denote by $B^{(b)}(t)$ a Brownian bridge process which starts at 0 at time 0 and ends at b at time 1 (i.e. $B^{(b)}(t) = B(t) - t(B(1) - b)$, where $B(\cdot)$ is a standard Brownian motion). Conditioning on $\mathscr{E}(b_1, b_2)$ we have

$$\mathrm{P}(\hat{D} \leq x \mid \mathscr{E}(b_1, b_2)) = \begin{cases} \mathrm{P}\Big(\inf\limits_{0\leq t\leq 1}\{b_2c_2 + \check{c}B^{(d)}(t)\} \leq x\Big), & x \leq \min(b_1c_1, b_2c_2), \\ 1, & x > \min(b_1c_1, b_2c_2), \end{cases}$$

where $d := (b_1c_1 - b_2c_2)/\check{c}$.

Manipulating the above probability of the Brownian bridge, we obtain

$$\begin{aligned}\mathrm{P}\Big(\inf_{0\leq t\leq 1}\{b_2c_2 + \check{c}B^{(d)}(t)\} \leq x\Big) &= \mathrm{P}\Big(\sup_{0\leq t\leq 1}\{-B^{(d)}(t)\} \geq \frac{b_2c_2 - x}{\check{c}}\Big) \\ &= \mathrm{P}\Big(\sup_{0\leq t\leq 1}\{B^{(-d)}(t)\} \geq \frac{b_2c_2 - x}{\check{c}}\Big).\end{aligned}$$

The first equality is trivial and the second step follows from the symmetry of the Brownian bridge. Now use (see [15, Chapter V])

$$\mathrm{P}\Big(\sup_{0\leq t\leq 1}\{B^{(b)}(t)\} > y\Big) = \mathrm{e}^{-2y(y-b)},$$

to arrive at

$$\mathrm{P}(\hat{D} \leq x \mid \mathscr{E}(b_1, b_2)) = \begin{cases} \exp\Big\{-\dfrac{2}{\check{c}^2}(x - b_1c_1)(x - b_2c_2)\Big\}, & x \leq \min(b_1c_1, b_2c_2), \\ 1, & x > \min(b_1c_1, b_2c_2). \end{cases}$$

By unconditioning we obtain

$$
\begin{aligned}
&\mathrm{P}(\hat{D} \le x) \\
&\quad = \frac{1}{2\pi} \int_{(b_1,b_2)\in\mathbb{R}^2} \mathrm{P}(\hat{D} \le x \mid \mathcal{E}(b_1,b_2)) \mathrm{e}^{-(b_1^2+b_2^2)/2} \,\mathrm{d}b_1\,\mathrm{d}b_2 \\
&\quad = \frac{1}{2\pi} \int_{\min(b_1c_1,b_2c_2)<x} \mathrm{e}^{-(b_1^2+b_2^2)/2} \,\mathrm{d}b_1\,\mathrm{d}b_2 \\
&\qquad + \frac{1}{2\pi} \int_{\min(b_1c_1,b_2c_2)\ge x} \exp\biggl\{-\biggl(\frac{2}{\check{c}^2}(x-b_1c_1)(x-b_2c_2) + \frac{1}{2}(b_1^2+b_2^2)\biggr)\biggr\} \,\mathrm{d}b_1\,\mathrm{d}b_2.
\end{aligned}
$$

The first integral in the above expression can be represented as

$$
\Phi\biggl(\frac{x}{c_1}\biggr) + \Phi\biggl(\frac{x}{c_2}\biggr) - \Phi\biggl(\frac{x}{c_1}\biggr)\Phi\biggl(\frac{x}{c_2}\biggr).
$$

For the integral on the right-hand side, we first change the region of integration to the positive quadrant, move the terms involving only b_1 out of the inner integral, and then complete the square:

$$
\begin{aligned}
&\frac{1}{2\pi}\int_0^\infty \int_0^\infty \exp\biggl\{-\biggl(\frac{2c_1c_2}{\check{c}^2}b_1b_2 + \frac{1}{2}\biggl(b_1+\frac{x}{c_1}\biggr)^2 + \frac{1}{2}\biggl(b_2+\frac{x}{c_2}\biggr)^2\biggr)\biggr\} \,\mathrm{d}b_1\,\mathrm{d}b_2 \\
&\qquad = \frac{1}{\sqrt{2\pi}} \int_0^\infty \mathrm{e}^{-L(u,x)} \Phi(-M(u,x)) \,\mathrm{d}u. \qquad (3.2)
\end{aligned}
$$

For the case where $c_1 = c_2 = c$, the remaining integral can be simplified to the desired expression by changing the order of integration.

We are now able to obtain an explicit expression for $\operatorname{var}\hat{D}$.

Proposition 3.1. *We have*

$$
\mathrm{E}\,\hat{D} = -\sqrt{\frac{2(c_1^2+c_2^2)}{\pi}}, \qquad \mathrm{E}\,\hat{D}^2 = c_1^2 + c_2^2, \qquad \operatorname{var}\hat{D} = (c_1^2+c_2^2)\biggl(1-\frac{2}{\pi}\biggr).
$$

Proof. We first determine the density of $\hat{D}$ by differentiating the distribution function, and then calculate the first and second moments in the standard manner. The part of the density obtained from $\Phi(x/c_1)+\Phi(x/c_2)-\Phi(x/c_1)\Phi(x/c_2)$, multiplied by x or x^2, can be integrated relatively easily. The part related to (3.2) should first be integrated over x (after multiplication by x or x^2). In both cases, this yields an integral over the positive quadrant of a function proportional to bivariate independent Gaussian distributions, which can therefore be simplified. Upon combining these terms, we obtain the result.

4. Uniform integrability

Our main result, Theorem 2.1, involves the assumption that $\mathcal{Q}$ is uniformly integrable. In this section we find sufficient conditions for this assumption to hold, thus establishing (i)–(iv) of Theorem 2.2. We apply several methods in the analysis. In Section 4.1 we use the reflection mapping for the queue to establish UI for the GI/M/1 queue. In Section 4.2 we construct couplings that involve the reflected queueing process, the actual queue process, and the count of the number of busy cycles. This allows us to establish the UI for the GI/NWU/1 queue, and for the GI/G/1 queue under an additional condition on the tail of the busy period. In Section 4.3 we show UI for the D/GI/1 case by using a different approach: we relate $Q(t)$ and W_n, the workload seen by the nth arrival, and then apply a UI result from [22].

4.1. Reflection mapping for queue length

In this subsection we prove UI for the GI/M/1 case. We do so by first introducing a process $\{Q'(t)\}$ (which is closely related to $\{Q(t)\}$), and prove UI for $\mathcal{Q}'$, defined as

$$\mathcal{Q}' = \left\{ \frac{Q'(t)^2}{t},\ t \geq t_0 \right\}$$

for some $t_0 > 0$. The following proposition plays a crucial role. Define $X(t) := A(t) - S(t)$, and let

$$Q'(t) = X(t) - \inf_{0 \leq s \leq t} X(s)$$

denote the associated reflected process. Note that, for the GI/M/1 case, it holds that $Q'(t)$ equals in distribution $Q(t)$ (see, e.g. [18, p. 68]); this does not hold for the GI/G/1 case. For the M/M/1 case, the reflected process is distributed as $\sup_{0 \leq s \leq t} X(s)$, but this is in general not true for GI/M/1; cf. [3, p. 98].

Proposition 4.1. *Assume that both*

$$\mathrm{E}\Big(\Big(\sup_{0 \leq s \leq t} \{|A(s) - \lambda s|\}\Big)^4\Big) \quad \textit{and} \quad \mathrm{E}\Big(\Big(\sup_{0 \leq s \leq t} \{|S(s) - \lambda s|\}\Big)^4\Big) \tag{4.1}$$

are $O(t^2)$. Then,

(i) $\mathrm{E}(Q'(t)^4) = O(t^2)$,

(ii) $\sup_{t \geq t_0} \mathrm{E}(Q'(t)^4/t^2) < \infty$,

(iii) *$\mathcal{Q}'$ is uniformly integrable.*

Proof. Use inequality (2.1), with $r = 4$, to obtain

$$Q'(t)^4 \leq 8\Big(X(t)^4 + \Big(\sup_{0 \leq s \leq t} -X(s)\Big)^4\Big).$$

We now deal with both terms separately. The first term is bounded as

$$X(t)^4 = ((A(t) - \lambda t) - (S(t) - \lambda t))^4 \leq 8(|A(t) - \lambda t|^4 + |S(t) - \lambda t|^4),$$

and, therefore, it follows from (4.1) that

$$\mathrm{E}(X(t)^4) = O(t^2). \tag{4.2}$$

We now consider the second term:

$$\begin{aligned}
\Big(\sup_{0 \leq s \leq t} -X(s)\Big)^4 &\leq \Big(\sup_{0 \leq s \leq t} |X(s)|\Big)^4 \\
&= \Big(\sup_{0 \leq s \leq t} |(S(s) - \lambda s) + (\lambda s - A(s))|\Big)^4 \\
&\leq \Big(\sup_{0 \leq s \leq t} \{|S(s) - \lambda s| + |A(s) - \lambda s|\}\Big)^4 \\
&\leq \Big(\sup_{0 \leq s \leq t} |S(s) - \lambda s| + \sup_{0 \leq s \leq t} |A(s) - \lambda s|\Big)^4 \\
&\leq 8\Big(\Big(\sup_{0 \leq s \leq t} |S(s) - \lambda s|\Big)^4 + \Big(\sup_{0 \leq s \leq t} |A(s) - \lambda s|\Big)^4\Big).
\end{aligned}$$

Again invoking (4.1) yields

$$\mathrm{E}\Big(\Big(\sup_{0\le s\le t} -X(s)\Big)^4\Big) = O(t^2). \tag{4.3}$$

Upon combining (4.2) and (4.3), we obtain (i). Result (ii) follows directly from (i), and (iii) follows from the sufficient condition of UI in (ii).

Recall that Doob's L_p maximum inequality for both continuous time and discrete time states that, for any $p > 1$ and martingale $\{M_t\}$,

$$\mathrm{E}\Big(\Big(\sup_{0\le s\le t} |M_s|\Big)^p\Big) \le \Big(\frac{p}{p-1}\Big)^p \mathrm{E}(|M_t|^p).$$

We may use Doob's inequality together with Proposition 4.1 to show that in the M/M/1 case, $\mathcal{Q}$ is uniformly integrable. This can be done by observing that $\{A(t) - \lambda t\}$ and $\{S(t) - \lambda t\}$ are martingales. Applying Doob's inequality shows that

$$\mathrm{E}\Big(\Big(\sup_{0\le s\le t} (A(s) - \lambda s)\Big)^4\Big) \le \Big(\frac{4}{3}\Big)^4 (3\lambda^2 t^2 + \lambda t) = O(t^2),$$

with the same result holding for $\{S(t) - \lambda t\}$.

For the GI/M/1 case, $\{A(t) - \lambda t\}$ is no longer a martingale, yet the following theorem is useful (and is of independent interest).

Theorem 4.1. *Let $\{\zeta_i,\ i \ge 0\}$ be a sequence of nonnegative, independent, and identically distributed random variables, and let $S_n := \sum_{i=1}^n \zeta_i$ be their partial sums. Denote the corresponding renewal counting process by $N(t) := \sup\{n\colon S_n \le t\}$. Define $\mathrm{E}\,\zeta_1 := \gamma^{-1}$, and assume that $\mathrm{E}\,\zeta_1^4 < \infty$. Then*

$$\mathrm{E}\Big(\Big(\sup_{0\le s\le t} \{|N(s) - \gamma s|\}\Big)^4\Big) = O(t^2).$$

Proof. Define $V(t) = \inf_n\{n\colon S_n \ge t\}$, so that $N(t) + 1 = V(t)$ and $S_{N(t)} \le t \le S_{V(t)}$. As a result of these inequalities, we have

$$\gamma s - N(s) \le \gamma S_{V(s)} - N(s) = \gamma S_{V(s)} - V(s) + 1 \le \sup_{0\le n\le V(s)} \{\gamma S_n - n\} + 1,$$

and, on the other hand,

$$\begin{aligned} N(s) - \gamma s &\le N(s) - \gamma S_{N(s)} \\ &\le \sup_{0\le n\le N(s)} \{n - \gamma S_n\} \\ &\le \sup_{0\le n\le V(s)} \{n - \gamma S_n\} \\ &\le \sup_{0\le n\le V(s)} |\gamma S_n - n| + 1. \end{aligned}$$

Combining these two inequalities, we obtain

$$|N(s) - \gamma s| \le \sup_{0\le n\le V(s)} |\gamma S_n - n| + 1.$$

Define $M_n := \sum_{i=1}^{n} \xi_i$, where $\xi_i := \gamma\zeta_i - n$ (which is a martingale). Taking the supremum over s between 0 and t yields

$$\sup_{0\leq s\leq t} |N(s) - \gamma s| \leq \sup_{0\leq n\leq V(t)} |\gamma S_n - n| + 1 = \sup_{0\leq n\leq V(t)} |M_n| + 1. \tag{4.4}$$

We are interested in the fourth moment of the quantity on the left-hand side of (4.4). Owing to (2.1), we have

$$\mathrm{E}\Big(\Big(\sup_{0\leq s\leq t} |N(s) - \gamma s|\Big)^4\Big) \leq 8\,\mathrm{E}\Big(\Big(\sup_{0\leq n\leq V(t)} |M_n|\Big)^4\Big) + 8.$$

Recalling that M_n is a martingale, observe that $V(t)$ is a stopping time with respect to the natural filtration of $\{M_n\}$ and, hence, $M_{n\wedge V(t)}$ is a martingale as well. Therefore, owing to Doob's maximum inequality, for $k = 0, 1, \dots,$

$$\mathrm{E}\Big(\Big(\sup_{0\leq n\leq k} |M_{n\wedge V(t)}|\Big)^4\Big) \leq \left(\frac{4}{3}\right)^4 \mathrm{E}((M_{k\wedge V(t)})^4). \tag{4.5}$$

Furthermore, observe that the sequence $\{\sup_{0\leq n\leq k} |M_{n\wedge V(t)}|\}$ is monotone increasing in k, and, almost surely,

$$\lim_{k\to\infty} \Big(\sup_{0\leq n\leq k} |M_{n\wedge V(t)}|\Big)^4 = \Big(\sup_{0\leq n} |M_{n\wedge V(t)}|\Big)^4 = \Big(\sup_{0\leq n\leq V(t)} |M_n|\Big)^4.$$

Applying the monotone convergence theorem, we obtain

$$\lim_{k\to\infty} \mathrm{E}\Big(\Big(\sup_{0\leq n\leq k} |M_{n\wedge V(t)}|\Big)^4\Big) = \mathrm{E}\Big(\Big(\sup_{0\leq n\leq V(t)} |M_n|\Big)^4\Big). \tag{4.6}$$

Furthermore, observe that, almost surely,

$$\lim_{k\to\infty} |M_{k\wedge V(t)}|^4 = |M_{V(t)}|^4.$$

Also, $\mathrm{E}\sup_k |M_{k\wedge V(t)}|^4 < \infty$, as follows from

$$(M_{k\wedge V(t)})^4 \leq 8\gamma^4 (S_{k\wedge V(t)})^4 + 8(k\wedge V(t))^4 \leq 8\gamma^4 (S_{V(t)})^4 + 8(V(t))^4$$

and the fact that, for fixed t, the right-hand side has finite mean; see, e.g. [11].

Now applying the dominated convergence theorem, we obtain

$$\lim_{k\to\infty} \mathrm{E}((M_{k\wedge V(t)})^4) = \mathrm{E}((M_{V(t)})^4). \tag{4.7}$$

Combining (4.5), (4.6), and (4.7), we obtain

$$\mathrm{E}\Big(\Big(\sup_{0\leq n\leq V(t)} |M_n|\Big)^4\Big) \leq \left(\frac{4}{3}\right)^4 \mathrm{E}(M_{V(t)}^4).$$

We now complete the proof by showing that the right-hand side of the previous display is $O(t^2)$. To this end, define $\mathrm{E}(\xi_i^\ell) = m_\ell$, and recall that it was assumed that $m_\ell < \infty$, $\ell = 1, 2, 3, 4$. Furthermore, let $\gamma(r)$ denote the cumulant generating function of ξ_i, i.e. $\gamma(r) = \log(\mathrm{E}(\mathrm{e}^{r\xi_i}))$, $\mathrm{Re}(r) \leq 0$. Let $\gamma^{(n)}(r)$ denote the nth derivative of $\gamma(r)$. Observe that $\gamma(0) = 0$,

$\gamma^{(1)}(0) = m_1 = 0$, and $\gamma^{(2)}(0) = \operatorname{var}(\xi_i) = m_2$, and that $\gamma^{(3)}(0)$ and $\gamma^{(4)}(0)$ can be expressed in terms of m_ℓ, $\ell = 2, 3, 4$. Since $V(t)$ is a stopping time, Wald's identity [21] yields

$$\mathrm{E}\exp\{rM_{V(t)} - V(t)\gamma(r)\} = 1.$$

Taking the second-order derivative and fourth-order derivative (with respect to r) of the latter equation at 0, we find that

$$\mathrm{E}((M_{V(t)})^2) = \mathrm{E}\,V(t)m_2, \tag{4.8}$$

$$\begin{aligned}\mathrm{E}((M_{V(t)})^4) &= \gamma^{(4)}(0)\,\mathrm{E}\,V(t) + 4\gamma^{(3)}(0)\,\mathrm{E}\,V(t)M_{V(t)} \\ &\quad - 3\,\mathrm{E}\,V(t)^2m_2^2 + 6m_2\,\mathrm{E}\,V(t)(M_{V(t)})^2.\end{aligned} \tag{4.9}$$

Then, note that the Cauchy–Schwarz inequality gives

$$\mathrm{E}\,V(t)M_{V(t)} \le \sqrt{\mathrm{E}\,V(t)^2\,\mathrm{E}(M_{V(t)})^2}, \tag{4.10}$$

$$\mathrm{E}\,V(t)(M_{V(t)})^2 \le \sqrt{\mathrm{E}\,V(t)^2\,\mathrm{E}((M_{V(t)})^2)^2} = \mathrm{E}(M_{V(t)})^2\sqrt{\mathrm{E}\,V(t)^2}. \tag{4.11}$$

Also, $\mathrm{E}\,V(t) = O(t)$ and $\mathrm{E}\,V(t)^2 = O(t^2)$; see, e.g. [3, Chapter V]. From (4.8) we deduce that $\mathrm{E}(M_{V(t)})^2 = O(t)$. Using (4.10) and (4.11), the latter equation gives $\mathrm{E}\,V(t)M_{V(t)} = O(t^{3/2})$ and $\mathrm{E}\,V(t)(M_{V(t)})^2 = O(t^2)$. Substituting these results into (4.9) yields

$$\mathrm{E}(M_{V(t)})^4 = O(t^2),$$

as desired.

Corollary 4.1. *For the critically loaded GI/M/1 queue, with* $\mathrm{E}\,\zeta_A^4 < \infty$, $\mathcal{Q}$ *is uniformly integrable.*

Proof. Theorem 4.1 gives (4.1), which completes the proof.

4.2. Coupling Q and Q'

In the previous subsection we were able to establish UI for the GI/M/1 queue using the fact that $Q'(t)$ is distributed the same as $Q(t)$. This property does not carry over to queues with nonexponential service times, but nevertheless we can obtain the desired UI from $Q'(t)$ for a large class of service times using the following result, which we prove using a coupling argument.

Recall that the distribution of a random variable X is *new worse than used* (NWU) if

$$\mathrm{P}(X > x) \le \frac{\mathrm{P}(X > t + x)}{\mathrm{P}(X > t)} \quad \text{for all } x, t \ge 0.$$

We denote by GI/NWU/1 the single-server queue with service times having an NWU service distribution; see [19] for more background.

Theorem 4.2. *Denote by* $C(t)$ *the number of busy cycles of the process* $\{Q(t)\}$ *during the time interval* $[0, t]$. *Let* $r \ge 1$. *Then,*

(i) *for GI/NWU/1,* $Q(t) \le_{\mathrm{st}} Q'(t)$, $t \ge 0$,

(ii) *for GI/G/1,* $\mathrm{E}\,Q(t)^r \le 2^{r-1}(\mathrm{E}\,Q'(t)^r + \mathrm{E}\,C(t)^r)$, $t \ge 0$.

Proof. We begin with (i). Let $\mathcal{L}(\cdot)$ denote the probability law of a stochastic process. We will construct a probability space supporting two coupled processes $\{\tilde{Q}(t)\}$ and $\{\tilde{Q}'(t)\}$ such that

$$\tilde{Q}(t) \le \tilde{Q}'(t), \qquad t \ge 0, \text{ with probability } 1, \tag{4.12}$$

where $\mathcal{L}(Q) = \mathcal{L}(\tilde{Q})$ and $\mathcal{L}(Q') = \mathcal{L}(\tilde{Q}')$. Establishing such a construction is equivalent to a stochastic order on the function space of sample paths (see [14]), from which (i) is an elementary consequence. We let $\tilde{Q} = Q$, so that it remains to produce (4.12) with $\mathcal{L}(Q') = \mathcal{L}(\tilde{Q}')$. We let both systems start empty and give both systems the given arrival process for Q. We redefine the service times of the upper bound system $\{\tilde{Q}'(t)\}$ every time an arrival enters an empty system. Otherwise, arrivals are assigned identical service times in both systems, which are taken from the given independent and identically distributed service times for Q. The construction is recursive over busy cycles of the process $\tilde{Q}$, i.e. we do mathematical induction over successive epochs at which an arrival finds the upper bound system empty. Clearly, the sample paths of the two systems are identical until the first time that an arrival in the upper bound system finds the system empty. Because of the reflection construction, the actual service time in the upper bound system is a residual service time, but, by the NWU assumption, the residual service time is stochastically larger than an ordinary service time. Given that stochastic order, we can construct a new service time for the upper bound process that is greater than or equal to the corresponding service time in the lower bound system with probability 1, and yet has its given probability law. Performing this construction maintains $\mathcal{L}(Q') = \mathcal{L}(\tilde{Q}')$. We repeat this construction each time an arrival at the upper bound system finds an idle server; necessarily, the corresponding arrival in the lower bound system finds the server idle too. By this special construction we make the service times of the upper bound process greater than or equal to the service times in the lower bound process with probability 1, while their distributions remain unchanged. It is known and not difficult to show that the queue length sample paths will be ordered with probability 1 if two systems differ only by service times that are all ordered; this is, e.g. the basis for Theorems 5 and 8 and the remark on page 216 of [23]. Hence, we achieve the sample-path order in (4.12) while keeping the relations $\mathcal{L}(Q) = \mathcal{L}(\tilde{Q})$ and $L(Q') = \mathcal{L}(\tilde{Q}')$. This sample path order holds over the successive finite time segments $[0, \tau_n)$, where τ_n is the time that the nth busy cycle begins. By mathematical induction, it thus holds over the entire positive halfline. We thus have (i).

We now turn to (ii). We will achieve the moment inequality by constructing a coupling of $Q(t)$, $C(t)$, and $Q'(t)$ on the same probability space. Again, we let $\tilde{Q} = Q$, so it remains to produce

$$Q(t) \le \tilde{Q}'(t) + C(t), \qquad t \ge 0, \text{ with probability } 1,$$

where $\mathcal{L}(Q') = \mathcal{L}(\tilde{Q}')$. We will do the construction by finding an intermediate system $\hat{Q}$ with

$$Q(t) \le \hat{Q}(t) \le \tilde{Q}'(t) + C(t), \qquad t \ge 0, \text{ with probability } 1, \tag{4.13}$$

where still $\mathcal{L}(Q') = \mathcal{L}(\tilde{Q}')$.

We let all three systems start empty and give them the specified arrival process for Q. We let all three systems be assigned the same service times from the sequence of independent and identically distributed random variables for Q.

The right-hand side of (4.13) indicates $Q'(t)$ with an additional customer added per busy period which is added whenever an arrival finds an empty system in Q. We let the service time of the extra arrival for $\tilde{Q}'$ match the service time of the arrival in Q, so that we can think of an extra initial customer with the residual service time, and otherwise the same arrivals having identical service times. We now construct the system $\hat{Q}$ from $\tilde{Q}'$ by combining the customer with the residual service time and the new customer into a single customer with the sum of the residual service time and the new service time. Hence, by this 'combining' of customers, at the start of every busy period, $\hat{Q}(t)$ is initially less than $\tilde{Q}'(t)+1$, and the inequality holds throughout the busy period. Again, using induction as in (i), we have the second inequality in (4.13).

With this construction, note that $\hat{Q}$ differs from Q only by having some customers with longer service times. In particular, whenever an arrival in Q finds an empty system, that customer has a shorter service time than the corresponding arrival in $\hat{Q}$. As a consequence, by the same reasoning as in part (i), we have the first inequality in (4.13). Now combining with (2.1) directly implies the final claimed moment inequality.

Note that the coupling in part (ii) of the above proof also implies that there exists a joint distribution between $Q'(t)$ and $C(t)$ such that $Q(t) \le Q'(t) + C(t)$ with probability 1. Also, note that in the above theorem we did not use the renewal structure of the arrival process and, thus, the result actually holds for queues with arbitrary arrival processes.

We now have UI of $\mathcal{Q}$ for the GI/NWU/1 queue.

Corollary 4.2. *For the critically loaded GI/NWU/1 queue with* $\mathrm{E}\,\zeta_A^4 < \infty$ *and* $\mathrm{E}\,\zeta_S^4 < \infty$, $\mathcal{Q}$ *is uniformly integrable.*

Proof. From Theorem 4.2(i) we deduce that $\mathrm{E}\,Q(t)^4 \le \mathrm{E}\,Q'(t)^4$. By Proposition 4.1(ii) we have $\mathrm{E}\,Q'(t)^4 = O(t^2)$. Thus, $\mathrm{E}\,Q(t)^4 = O(t^2)$, which completes the proof.

In order to use the stochastic order in Proposition 4.1(ii) for the UI of $\mathcal{Q}$ in the GI/G/1 queue, we need first to establish the order of growth of the moments of $C(t)$. The following theorem is attributed to A. Löpker (personal communication). To the best of the authors' knowledge, this general result about renewal processes has not appeared elsewhere.

Theorem 4.3. *Let* $N(t)$ *and* ζ_i *be defined as in Theorem 4.1. Suppose that* $\mathrm{P}(\zeta_i \ge x) = 1 - F(x) \sim L(x)x^{-\alpha}$ *with* $\alpha \in [0, 1)$ *and* $L(\cdot)$ *slowly varying. Then*

$$\mathrm{E}\,N(t)^m \sim t^{\alpha m} L(t)^{-m} \frac{\Gamma(1+m)}{\Gamma(1-\alpha)^m \Gamma(1+\alpha m)}, \qquad t \to \infty.$$

Proof. We have

$$\begin{aligned}\mathrm{E}\,N(t)^m &= \sum_{i=1}^{\infty} i^m \mathrm{P}(N(t) = i) \\ &= \sum_{i=1}^{\infty} i^m (F^{*i}(t) - F^{*i+1}(t)) \\ &= \sum_{i=1}^{\infty} i^m F^{*i}(t) - \sum_{i=2}^{\infty} (i-1)^m F^{*i}(t) \\ &= \sum_{i=1}^{\infty} a(i) F^{*i}(t),\end{aligned}$$

where $a(i) = i^m - (i-1)^m$. Clearly, $\sum_{i=1}^{n} a(i) = n^m$. Now using Omey's theorem [2, Theorem D] with $\rho = m$ and $L_1(x) = 1$ (where ρ and $L_1(\cdot)$ follow the notation of [2]), the result follows.

We are now in a position to relate the growth rate of $C(t)$ to the tail asymptotics of the busy period distribution.

Corollary 4.3. *For the critically loaded GI/G/1 queue with* $\mathrm{E}\,\zeta_A^4 < \infty$ *and* $\mathrm{E}\,\zeta_S^4 < \infty$*, if*

$$\mathrm{P}(B > x) \sim L(x)x^{-1/2} \tag{4.14}$$

with $L(x)$ *slowly varying and bounded, then* $\mathcal{Q}$ *is uniformly integrable.*

Proof. We apply Theorem 4.3 with $m = 4$ to $C(t)$ of Theorem 4.2, to obtain $\mathrm{E}\,C(t)^4 = O(t^2)$. Furthermore, observe that Theorem 4.1 applied to $A(t)$ and $S(t)$ implies condition (4.1), and, thus, by Proposition 4.1(i), we have $\mathrm{E}\,Q'(t)^4 = O(t^2)$. Since Theorem 4.2(i) implies that $\mathrm{E}\,Q(t)^4 \leq 8(\mathrm{E}\,Q'(t)^4 + \mathrm{E}\,C(t)^4)$, we have $\mathrm{E}\,Q(t)^4 = O(t^2)$, and, as a result,

$$\sup_{t\geq t_0} \mathrm{E}\biggl(\frac{Q'(t)^2}{t}\biggr)^2 < \infty.$$

Hence, we conclude that $\mathcal{Q}$ is uniformly integrable.

Corollary 4.4. *For the critically loaded M/G/1 queue with* $\mathrm{E}\,\zeta_S^4 < \infty$*,* $\mathcal{Q}$ *is uniformly integrable.*

Proof. The tail asymptotics for the busy period in (4.14) have been established for the critically loaded M/G/1 queue in [27, Theorem 4.1]. Consequently, the result follows from Theorem 4.3.

4.3. The D/GI/1 case

The approach we follow for the D/GI/1 queue differs substantially from the approach taken in the previous subsections. Here we simply relate the queue size to the workload and use a previous result of UI stated in [22].

Proposition 4.2. *For the critically loaded D/G/1 queue with* $\mathrm{E}\,\zeta_S^4 < \infty$ *and* $\mathrm{P}(\zeta_S > b) = 1$ *for some* $b > 0$*,* $\mathcal{Q}$ *is uniformly integrable.*

Proof. In the following we relate $Q(t)$ and W_n, the workload seen by the nth arrival. Note that in [22, Theorem 4.1] it was shown that if $\mathrm{E}\,\zeta_S^{2m} < \infty$ then $(W_n/\sqrt{n})^k$, $k \leq 2m$, is uniformly integrable. Moreover, it is well known that if we have the nonnegative sequences of random variables X_n, Y_n, and Z_n such that $Z_n < X_n + Y_n$, and X_n and Y_n are uniformly integrable, then so is Z_n.

We have $Q(t) \leq W(t)/b + 1$ and $W(t) = W_{A(t)} - (t - \tau_{A(t)}) \leq W_{A(t)}$, where $\tau_{A(t)}$ is the arriving time of the $A(t)$th arrival. Therefore, we see that, for $\lfloor \lambda t \rfloor > 0$,

$$\begin{aligned}
\frac{Q(t)}{\sqrt{t}} &\leq b^{-1}\frac{W(t) + b}{\sqrt{t}} && (4.15)\\
&\leq b^{-1}\frac{W_{A(t)} + b}{\sqrt{t}} && (4.16)\\
&\leq b^{-1}\sqrt{\lambda}\frac{W_{A(t)} + b}{\sqrt{A(t)}} \\
&= b^{-1}\sqrt{\lambda}\biggl(\frac{W_{\lfloor \lambda t \rfloor}}{\sqrt{\lfloor \lambda t \rfloor}} + \frac{b}{\sqrt{\lfloor \lambda t \rfloor}}\biggr), && (4.17)
\end{aligned}$$

where the third inequality and last line follow from $A(t) = \lfloor \lambda t \rfloor \leq \lambda t$ (at time 0 the queue is empty and an interarrival time is deterministic and equal to $1/\lambda$). Using (2.1) with $r = 4$, (4.17) then gives

$$\left(\frac{Q(t)}{\sqrt{t}}\right)^4 \le 8b^{-4}\lambda^2\left\{\left(\frac{W_{\lfloor \lambda t\rfloor}}{\sqrt{\lfloor \lambda t\rfloor}}\right)^4 + \left(\frac{b}{\sqrt{\lfloor \lambda t\rfloor}}\right)^4\right\}. \tag{4.18}$$

Note that $(b/\sqrt{\lfloor \lambda t\rfloor})^4$ is bounded from above by b^4, $t \ge t_0 > 0$, which implies that it is uniformly integrable. Moreover, under the assumption that $\mathrm{E}\,\zeta_S^4 < \infty$, $(W_{\lfloor \lambda t\rfloor}/\sqrt{\lfloor \lambda t\rfloor})^4$ is uniformly integrable; see [22, Theorem 4.1]. Hence, both terms on the right-hand side of (4.18) are uniformly integrable, and, hence, so is $\mathcal{Q}$.

5. The variance curve of the M/M/1 queue

In this section we consider the M/M/1 queue and obtain expressions for the first and second moments of $D(t)$ for any $t \ge 0$. We first consider arbitrary $\lambda, \mu > 0$, and obtain cumbersome yet computationally tractable expressions for $\mathrm{E}\,D(t)$ and $\operatorname{var} D(t)$ in terms of integrals of Bessel functions (Theorem 5.1). These expressions are useful for numerically illustrating the presence of the BRAVO effect for finite t and $\varrho \approx 1$ (see Figure 1). For the critically loaded case, some simplification occurs and these integrals evaluate to simpler explicit expressions, given in terms of Bessel functions (Corollary 5.1).

We are further able to perform asymptotic expansions for $\mathrm{E}\,D(t)$ and $\operatorname{var} D(t)$ for large t (Theorem 5.2). These expansions show that in the critically loaded case, the variance and expectation curves have a lower-order square root term that does not exist when $\lambda \ne \mu$. They also serve as an alternative proof to our main result in the specific case of M/M/1.

Notation. We denote the convolution operator by '$*$' and make use of the modified Bessel function of the first kind:

$$I_j(2t) = \sum_{n=0}^{\infty} \frac{t^{j+2n}}{(j+n)!\,n!}.$$

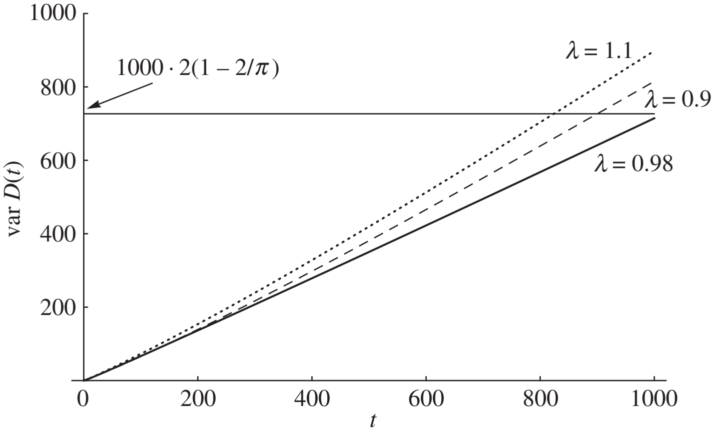

FIGURE 1: Demonstration of the BRAVO effect for $\lambda \approx \mu$ and finite t: $\operatorname{var} D(t)$ is plotted for M/M/1 systems with $\mu = 1$. The dashed line is for $\lambda = 0.9$, the solid line is for $\lambda = 0.98$, and the dotted line is for $\lambda = 1.1$. The horizontal line is at the height $1000 \cdot 2(1 - 2/\pi)$.

Theorem 5.1. *For the M/M/1 queue with* $Q(0) = 0$,

$$\mathrm{E}\, D(t) = \sqrt{\lambda\mu} \int_0^t (t-u) \frac{I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}}{u}\, \mathrm{d}u,$$

$$\begin{aligned}\operatorname{var} D(t) &= \mu t(\mu t + 2) - \sqrt{\lambda\mu} \int_0^t (t-u)(\mu(t-u)+2) \frac{I_1(2u\sqrt{\lambda\mu})}{u} \mathrm{e}^{-(\lambda+\mu)u}\, \mathrm{d}u \\ &\quad + 2\lambda\mu \int_0^t (t-u)^2 \frac{I_2(2u\sqrt{\lambda\mu})}{u} \mathrm{e}^{-(\lambda+\mu)u}\, \mathrm{d}u \\ &\quad + \mu \int_0^t (\mu(\lambda-\mu)(t-u)^2 - 4\mu(t-u) - 2) I_0(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}\, \mathrm{d}u \\ &\quad + \mu\sqrt{\lambda\mu} \int_0^t (t-u)((\mu-\lambda)(t-u)+2) I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}\, \mathrm{d}u \\ &\quad + \sqrt{\lambda\mu} \int_0^t (t-u) \frac{I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}}{u}\, \mathrm{d}u \\ &\quad - \lambda\mu \biggl(\int_0^t (t-u) \frac{I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}}{u}\, \mathrm{d}u \biggr)^2 .\end{aligned}$$

Proof. Let X_α be an exponential random variable with mean $1/\alpha$. Let $\phi_\alpha(z)$ denote the probability generating function (PGF) of the number of departures at the random time X_α. We have

$$\phi_\alpha(z) = \mathrm{E}_{X_\alpha} \mathrm{E}(z^{D(X_\alpha)} \mid X_\alpha) = \int_0^\infty \alpha \mathrm{e}^{-\alpha t}\, \mathrm{E}(z^{D(t)})\, \mathrm{d}t.$$

Note that $\phi_\alpha(z)/\alpha$ can be interpreted as the Laplace transform of $\mathrm{E}\, z^{D(t)}$. Define $\phi_\alpha^1 := \phi_\alpha'(1)/\alpha$ and $\phi_\alpha^2 := \phi_\alpha''(1)/\alpha$, and denote by $\mathcal{L}^{-1}(\cdot)$ the inverse Laplace transform. Thus, it is readily seen that

$$\mathrm{E}\, D(t) = \mathcal{L}^{-1}(\phi_\alpha^1), \qquad \mathrm{E}\, D(t)^2 = \mathcal{L}^{-1}(\phi_\alpha^2 + \phi_\alpha^1). \tag{5.1}$$

From [6, Equation (2.71), p. 199,], inserting $k = 0$, $\rho = \alpha$, $q = z$, and $x_2(q) = r(z, \alpha)$ (see also [1, Equation (25)]), we have the following simple expression:

$$\frac{\phi_\alpha(z)}{\alpha} = \frac{z}{\mu(1-z)+\alpha} \frac{1 - r(z,\alpha)}{z - r(z,\alpha)}. \tag{5.2}$$

Here

$$r(z,\alpha) = \frac{\lambda + \mu + \alpha - \sqrt{(\lambda+\mu+\alpha)^2 - 4\lambda\mu z}}{2\lambda}.$$

Furthermore, let

$$s(z,\alpha) = \frac{\lambda + \mu + \alpha + \sqrt{(\lambda+\mu+\alpha)^2 - 4\lambda\mu z}}{2\lambda} = \frac{\mu z}{\lambda r(z,\alpha)}.$$

Differentiating (5.2) with respect to z at the point $z = 1$ yields

$$\phi_\alpha^1 = \frac{\lambda}{\alpha^2} r(1,\alpha), \tag{5.3}$$

$$\begin{aligned}\phi_\alpha^2 &= 2\mu \frac{\mu+\alpha}{\alpha^3} - 2\lambda \frac{\mu+\alpha}{\alpha^3} r(1,\alpha) + \frac{2\lambda^2}{\alpha^3} r(1,\alpha)^2 \\ &\quad + \frac{2\mu}{\alpha^3}\biggl(\mu - \frac{(\mu+\alpha)^2}{\lambda}\biggr) \frac{1}{s(1,\alpha) - r(1,\alpha)} + \frac{2\mu}{\alpha^3}(\mu+\alpha-\lambda) \frac{r(1,\alpha)}{s(1,\alpha)-r(1,\alpha)}.\end{aligned}$$

Now using an explicit inversion, as in, e.g. [6, p. 81], for (5.3), we obtain

$$\begin{aligned}
\mathcal{L}^{-1}(\phi^1_\alpha) &= \lambda t * \sqrt{\frac{\mu}{\lambda}}\frac{I_1(2t\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)t}}{t}, \\
\mathcal{L}^{-1}(\phi^2_\alpha) &= \mu t(\mu t+2) - \sqrt{\lambda\mu}t(\mu t+2) * \left(\frac{I_1(2t\sqrt{\lambda\mu})}{t}\mathrm{e}^{-(\lambda+\mu)t}\right) \\
&\quad + 2\lambda\mu t^2 * \left(\frac{I_2(2t\sqrt{\lambda\mu})}{t}\mathrm{e}^{-(\lambda+\mu)t}\right) \\
&\quad + \mu(\mu(\lambda-\mu)t^2 - 4\mu t - 2) * (I_0(2t\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)t}) \\
&\quad + \mu\sqrt{\lambda\mu}t((\mu-\lambda)t+2) * (I_1(2t\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)t}).
\end{aligned}$$

Using (5.1) and reorganizing the above convolution term, we obtain the result.

In the case $\varrho = 1$, the integrals of Theorem 5.1 evaluate into somewhat simpler expressions given in terms of Bessel functions (rather than integrals of Bessel functions).

Corollary 5.1. *For the critically loaded M/M/1 queue with $Q(0) = 0$,*

$$\begin{aligned}
\mathrm{E}\, D(t) &= \lambda t - \tfrac{1}{2}\mathrm{e}^{-2\lambda t}((1+4\lambda t)I_0(2\lambda t) + 4\lambda t I_1(2\lambda t)) + \tfrac{1}{2}, \\
\operatorname{var} D(t) &= \tfrac{1}{4}\mathrm{e}^{-4\lambda t}(\mathrm{e}^{4\lambda t}(8\lambda t+1) - (4\lambda t+1)^2 I_0(2\lambda t)^2 - 4\mathrm{e}^{2\lambda t}\lambda t I_1(2\lambda t) \\
&\quad - 16\lambda^2 t^2 I_1(2\lambda t)^2 - 4\lambda t I_0(2\lambda t)(\mathrm{e}^{2\lambda t} + (2+8\lambda t)I_1(2\lambda t))).
\end{aligned}$$

Proof. Directly evaluate the integrals of Theorem 5.1 with $\lambda = \mu$.

Furthermore, the integrals of Theorem 5.1 yield the following asymptotic expansion.

Theorem 5.2. *For the M/M/1 queue with $Q(0) = 0$,*

$$\mathrm{E}\, D(t) = \begin{cases} \lambda t - \dfrac{\varrho}{1-\varrho} + o(1) & \textit{if } \lambda < \mu, \\ \lambda t - 2\sqrt{\dfrac{\lambda}{\pi}}t^{1/2} + \dfrac{1}{2} + o(1) & \textit{if } \lambda = \mu, \\ \mu t - \dfrac{1}{\varrho - 1} + o(1) & \textit{if } \lambda > \mu, \end{cases}$$

and

$$\operatorname{var} D(t) = \begin{cases} \lambda t - \dfrac{\varrho}{(1-\varrho)^2} + o(1) & \textit{if } \lambda < \mu, \\ \lambda 2\Big(1 - \dfrac{2}{\pi}\Big)t - \sqrt{\dfrac{\lambda}{\pi}}t^{1/2} + \dfrac{\pi - 2}{4\pi} + o(1) & \textit{if } \lambda = \mu, \\ \mu t - \dfrac{\varrho}{(1-\varrho)^2} + o(1) & \textit{if } \lambda > \mu. \end{cases}$$

Proof. The cases $\lambda = \mu$ and $\lambda \neq \mu$ are treated separately. The $\lambda = \mu$ case follows directly from Corollary 5.1. To obtain the linear term, divide the expressions of Corollary 5.1 by t and evaluate the limit as $t \to \infty$. To obtain the $\sqrt{t}$-term, subtract the linear term, divide by $\sqrt{t}$, and evaluate the limit. To obtain the constant term, subtract the linear and $\sqrt{t}$-terms, and evaluate the limit. The remaining error is $o(1)$.

The $\lambda \neq \mu$ case is more complicated. Consider first $\mathrm{E}\, D(t)$. Theorem 5.1 readily gives

$$\begin{aligned}\mathrm{E}\, D(t) &= \sqrt{\lambda\mu}\biggl(t\int_0^\infty \frac{I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}}{u}\,\mathrm{d}u - \int_0^\infty I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}\,\mathrm{d}u \\ &\qquad - t\int_t^\infty \frac{I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}}{u}\,\mathrm{d}u + \int_t^\infty I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}\,\mathrm{d}u\biggr) \\ &= \sqrt{\lambda\mu}\biggl(\frac{2\sqrt{\lambda\mu}}{\lambda+\mu+|\lambda-\mu|}t - \frac{2\sqrt{\lambda\mu}}{|\lambda-\mu|(\lambda+\mu+|\lambda-\mu|)} \\ &\qquad - t\int_t^\infty \frac{I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}}{u}\,\mathrm{d}u + \int_t^\infty I_1(2u\sqrt{\lambda\mu})\mathrm{e}^{-(\lambda+\mu)u}\,\mathrm{d}u\biggr),\end{aligned}$$

where the second equality follows by interchanging the integration and the summation resulting from the definition of the $I_1(2\sqrt{\lambda\mu t})$ functions in the first two terms. For the second two terms, we use the following result for $p, s > 0,\ p \neq s$, and $\gamma \in \mathbb{Z}$:

$$\int_t^\infty u^\gamma I_m(pu)\mathrm{e}^{-su}\,\mathrm{d}u = \frac{1}{\sqrt{2\pi p}(s-p)}\frac{t^{\gamma-1/2}}{\mathrm{e}^{(s-p)t}} + O\biggl(\frac{t^{\gamma-3/2}}{\mathrm{e}^{(s-p)t}}\biggr).$$

See, for example, [6, p. 83]. Combining we obtain

$$\mathrm{E}\, D(t) = \frac{2\lambda\mu}{\lambda+\mu+|\lambda-\mu|}t - \frac{2\lambda\mu}{|\lambda-\mu|(\lambda+\mu+|\lambda-\mu|)} + o(1).$$

Our result for $\mathrm{E}\, D(t)$ now follows. The result for $\operatorname{var} D(t)$ follows along the same lines.

We end this section with a numerical example. We use Theorem 5.1 to evaluate $\operatorname{var} D(t)$ for three M/M/1 queues with $\varrho < 1$, $\varrho \approx 1$, and $\varrho > 1$. The integrals of expressions involving Bessel functions are easily evaluated numerically. Variance curves of three example systems are plotted in Figure 1. The time horizon is $[0, 1000]$. It can be observed that as ϱ is varied from 0.9 to 1.1, the variance curve decreases when $\varrho \approx 1$.

The main point made is that the BRAVO effect appears for $\lambda \approx \mu$, for finite t and not only for the critical $\lambda = \mu$ case. It is further evident that the asymptotic slope of $2(1 - 2/\pi)$ which holds for $\varrho = 1$ also approximately holds as a nonasymptotic slope (for finite t) for $\varrho \approx 1$.

6. Extensions

In this section we address a number of extensions. The contribution is twofold. Our first aim is to indicate that the $(1 - 2/\pi)$ effect as in (1.1) holds in great generality. In this respect we simply require that the arrival and service processes satisfy a functional law of large numbers (FLLN) and a functional central limit theorem (FCLT), relying on the same diffusion limit result of [12]. In this general case, we assume that the UI conditions hold without attempting to prove so. Our second aim is to establish the UI conditions for the GI/G/s queue in the same manner as the GI/G/1 queue, thus generalizing our main result to the multiserver case.

The general model we consider is a multichannel, multiserver queue as described in [12]; see also [13]: r arrival channels of customers arrive to a queue with s servers. When a customer arrives to find one or more free servers, he/she is served by a free server under some arbitrary tie breaking rule. When a customer arrives to a system with all s servers busy, he/she queues up to wait for the next available server in an FCFS manner. The service times do not depend on the arrival channel but may depend on the server used. The $r + s$ arrival and service processes are mutually independent. Denote the arrival processes by $A_i(t)$, $i = 1, \dots, r$, and the service processes by $S_i(t)$, $i = 1, \dots, s$. Assume the existences of $\lambda_i > 0$, $i = 1, \dots, r$, and $\mu_i > 0$, $i = 1, \dots, s$, such that

$$\lim_{t\to\infty} \frac{\mathrm{E}\, A_i(t)}{t} = \lambda_i \quad \text{and} \quad \lim_{t\to\infty} \frac{\mathrm{E}\, S_i(t)}{t} = \mu_i \quad \text{(FLLN)}.$$

Consider the queue in the critical regime with λ:

$$\lambda = \sum_{i=1}^{r} \lambda_i = \sum_{i=1}^{s} \mu_i.$$

Furthermore, assume that there exist asymptotic variances $\kappa_i^a > 0$, $i = 1, \dots, r$, and $\kappa_i^s > 0$, $i = 1, \dots, s$, such that

$$\frac{A_i(nt) - \lambda_i nt}{\sqrt{\kappa_i^a n}} \Rightarrow B(t) \quad \text{and} \quad \frac{S_i(nt) - \mu_i nt}{\sqrt{\kappa_i^s n}} \Rightarrow B(t) \quad \text{(FCLT)},$$

where the weak convergence is as in [12] as $n \to \infty$, and $B(\cdot)$ is a standard Brownian motion; cf. also [25]. In case of renewal processes, κ_i^a/λ_i and κ_i^s/μ_i are the squared coefficient of variation of the interrenewal times. For ease of reference, we refer to this model as the critically loaded G$_r$/G/s queue. We now have the following result.

Theorem 6.1. *Consider the critically loaded G$_r$/G/s queue. Assume that the following two processes are uniformly integrable:*

$$\left\{ \frac{(\sum_{i=1}^{r} A_i(t) - \lambda t)^2}{\lambda t},\ t \ge t_0 \right\} \quad \textit{and} \quad \left\{ \frac{Q(t)}{t^2},\ t \ge t_0 \right\}.$$

Then

$$\sigma = \left(1 - \frac{2}{\pi}\right)\left(\sum_{i=1}^{r} \kappa_i^a + \sum_{i=1}^{s} \kappa_i^s\right). \tag{6.1}$$

Proof. The proof follows the exact same lines as the proof of Theorem 2.1. See also [13] for a discussion of generalizing renewal processes.

The critically loaded GI/G/s queue with arrival rate λ is a special case. Take $r = 1$, and set all $s + 1$ processes as renewal processes with the s service processes having the same distributions. In this case define $\kappa_1^a = \lambda c_a^2$ and $\kappa_i^s = \lambda c_s^2/s$, $i = 1, \dots, s$. The asymptotic variance (6.1) reduces once again to

$$\sigma = \lambda\left(1 - \frac{2}{\pi}\right)(c_a^2 + c_s^2).$$

For the GI/G/s queue, we are able to establish the required UI conditions for a variety of cases. Observe first that the first UI condition holds for renewal arrivals as in Theorem 2.1. Furthermore, conditions for the second sequence are given in the following theorem.

Theorem 6.2. *Consider the critically loaded GI/G/s queue operating under the FCFS discipline. Assume that* $\mathrm{E}\,\zeta_A^4 < \infty$ *and* $\mathrm{E}\,\zeta_S^4 < \infty$*. Then,* $\{Q(t)^2/t,\ t \ge t_0\}$ *is uniformly integrable* *in the following cases.*

(i) $\mathrm{P}(B > x) \sim L(x)x^{-1/2}$, *where* $L(\cdot)$ *is a bounded, slowly varying function and* B *is the busy period of a GI/G/1 queue with an interarrival time distribution which is an* s*-fold convolution of* ζ_A.

(ii) *The critically loaded Gamma(*$1/s,\lambda$*)/G/s queue. That is,*

$$\mathrm{P}(\zeta_A \le x) = \int_0^x \frac{\lambda^{1/s}}{\Gamma(1/s)} t^{1/s-1} \mathrm{e}^{-\lambda t}\, \mathrm{d}t.$$

(iii) *The critically loaded GI/NWU/s queue.*

(iv) *The critically loaded D/G/s queue with* $\mathrm{P}(\zeta_S > b) = 1$ *for some* $b > 0$.

Proof. We apply the results in [26] for the special case of GI/G/s and the cyclic service. In the cyclic service discipline, arrival $sj + i$, $j = 0, 1, \ldots,$ is assigned to the ith server, $i = 1, 2, \ldots, s$. The partition of the arrivals in this manner generates a collection of sGI/G/1 queues, each with service time ζ_S and interarrival time being an s-fold convolution of ζ_A. It is easily seen that, when the GI/G/s is critically loaded, all the s individual GI/G/1 queues are also critically loaded.

Let $Q_i(t)$, $i = 1, \ldots, s$, denote the queue length of the ith single-server queue at time t with $Q_i(0) = 0$. Then it follows from [26, Equation (8)] that

$$Q(t) \le_{\mathrm{st}} \sum_{i=1}^{s} Q_i(t).$$

We now have, for cases (i)–(iv),

$$\mathrm{E}\, Q(t)^4 \le \mathrm{E}\biggl(\sum_{i=1}^{s} Q_i(t)\biggr)^4 \le 8^{s-1}\, \mathrm{E} \sum_{i=1}^{s} (Q_i(t))^4 = s8^{s-1}\, \mathrm{E}(Q_1(t))^4 = O(t^2).$$

The second inequality follows from $s - 1$ applications of (2.1). The $O(t^2)$ term is obtained for cases (i)–(iv) using the results of Section 4. Note that case (ii) is based on the M/G/1 result of Corollary 6.1 since a convolution of s Gamma$(1/s, \lambda)$ random variables is an exponential. Also, observe that, since $\mathrm{E}\, \zeta_A^4 < \infty$, the s-fold convolution retains this property as is needed for (i) and (iii).

Acknowledgements

AA was supported by NWO through the QNOISE project; part of his work was carried out while he was a postdoc at EURANDOM. The work of YN was partly done while visiting CWI, Amsterdam, The Netherlands. WW was supported by the NSF, grant number CMMI 0948190.

We thank E. Aidekon (EURANDOM) for useful discussions on Theorem 4.1. We further thank A. Löpker (EURANDOM) and attribute Theorem 4.3 to him. We also thank R. Núñez Queija (University of Amsterdam) and B. Zwart (CWI) for useful discussions and advice.

References

[1] Al Hanbali, A., de Haan, R., Boucherie, R. J and van Ommeren, J.-K. (2008). A tandem queueing model for delay analysis in disconnected ad hoc networks. In *Analytical and Stochastic Modeling Techniques and Applications* (Lecture Notes Comput. Sci. **5055**), Springer, Berlin, pp. 189–205.

[2] Alsmeyer, G. (1992). On generalized renewal measures and certain first passage times. *Ann. Prob.* **20,** 1229–1247.

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

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

[5] Burke, P. J. (1956). The output of a queuing system. *Operat. Res.* **4,** 699–704.

[6] Cohen, J. W. (1982). *The Single Server Queue*, 2nd edn. North-Holland, Amsterdam.

[7] Daley, D. J. (1975). Further second-order properties of certain single-server queueing systems. *Stoch. Process. Appl.* **3,** 185–191.

[8] Daley, D. J. (1976). Queueing output processes. *Adv. Appl. Prob.* **8,** 395–415.

[9] Disney, R. L. and Kiessler, P. C. (1987). *Traffic Processes in Queueing Networks*. The Johns Hopkins University Press, Baltimore, MD.

[10] Disney, R. L. and König, D. (1985). Queueing networks: a survey of their random processes. *SIAM Rev.* **27,** 335–403.

[11] Gut, A. (1988). *Stopped Random Walks*. Springer, New York.

[12] Iglehart, D. L. and Whitt, W. (1970). Multiple channel queues in heavy traffic. I. *Adv. Appl. Prob.* **2,** 150–177.

[13] Iglehart, D. L. and Whitt, W. (1970). Multiple channel queues in heavy traffic. II. Sequences, networks, and batches. *Adv. Appl. Prob.* **2,** 355–369.

[14] Kamae, T., Krengel, U. and O'Brien, G. L. (1977). Stochastic inequalities on partially ordered spaces. *Ann. Prob.* **5,** 899–912.

[15] Mandjes, M. (2007). *Large Deviations for Gaussian Queues*. John Wiley, Chichester.

[16] Nazarathy, Y. (2009). The variance of departure processes: puzzling behavior and open problems. Tech. Rep. 2009-045, EURANDOM.

[17] Nazarathy, Y. and Weiss, G. (2008). The asymptotic variance rate of the output process of finite capacity birth–death queues. *Queueing Systems* **59,** 135–156.

[18] Prabhu, N. U. (1998). *Stochastic Storage Processes*, 2nd edn. Springer, New York.

[19] Stoyan, D. and Daley, D. J. (1983). *Comparison Methods for Queues and Other Stochastic Models*. John Wiley, Chichester.

[20] Tan, B. (2000). Asymptotic variance rate of the output in production lines with finite buffers. *Ann. Operat. Res.* **93,** 385–403.

[21] Wald, A. (1944). On cumulative sums of random variables. *Ann. Math. Statist.* **15,** 283–296.

[22] Whitt, W. (1972). Complements to heavy traffic limit theorems for the GI/G/1 queue. *J. Appl. Prob.* **9,** 185–191.

[23] Whitt, W. (1981). Comparing counting processes and queues. *Adv. Appl. Prob.* **13,** 207–220.

[24] Whitt, W. (1983). The queueing network analyzer. *Bell Systems Tech. J.* **62,** 2779–2815.

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

[26] Wolff, R. W. (1977). An upper bound for multi-channel queues. *J. Appl. Prob.* **14,** 884–888.

[27] Zwart, A. P. (2001). Tail asymptotics for the busy period in the GI/G/1 queue. *Math. Operat. Res.* **26,** 485–493.