CWI

The asymptotic variance of departures in critically loaded queues

A. Al Hanbali, M.R.H. Mandjes, Y. Nazarathy, W. Whitt

PNA-1003

Centrum Wiskunde & Informatica

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Science Park 123, 1098 XG Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

THE ASYMPTOTIC VARIANCE OF DEPARTURES IN CRITICALLY LOADED QUEUES

A. AL HANBALI, M. MANDJES, Y. NAZARATHY, AND W. WHITT

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 that the system load ϱ equals 1, and prove that the asymptotic variance rate satisfies

$$\lim_{t\to\infty} \frac{\mathbb{V}\mathrm{ar}D(t)}{t} = \lambda\left(1-\frac{2}{\pi}\right)\left(c_a^2+c_s^2\right),$$

where λ is the arrival rate and c_a^2, c_s^2 are squared coefficients of variation of the inter-arrival and service times respectively. As a consequence, the departures variability has a remarkable singularity in case ϱ equals 1, in line with the BRAVO effect (Balancing Reduces Asymptotic Variance of Outputs) which was previously encountered in the finite-capacity birth-death queues.
Under certain technical conditions, our result generalizes to multi-server 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 queues ⋆ critically loaded systems ⋆ uniform integrability ⋆ departure processes ⋆ renewal theory ⋆ Brownian bridge ⋆ multi-server queues

ACKNOWLEDGMENTS. AA was supported by NWO through the QNOISE project; part of his work was carried out while he was post-doc at EURANDOM. The work of YN was partly done while visiting CWI, Amsterdam, the Netherlands. WW was supported by NSF Grant CMMI 0948190.
The authors thank E. Aidekon (EURANDOM) for useful discussions on Theorem 4.3. We further thank and attribute Theorem 4.7 to A. Löpker (EURANDOM). We also thank R. Núñez Queija (University of Amsterdam) and B. Zwart (CWI) for useful discussions and advice.

Date: February 24, 2010.

1. INTRODUCTION

In the study of queueing systems, the analysis of departure processes has played an important role. Following Burke's theorem [5], stating that departures of a stationary M/M/1 queue form a Poisson process, many papers have dealt with properties of inter-departure times, departure counting processes, and approximations. A classic survey is by Daley [8], while other useful references in this area are [9] and [10, Ch. 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 $\mathbb{V}\mathrm{ar}D(t)$ is of crucial importance in the performance analysis of supply chain and manufacturing networks; several recent studies [11, 13, 14, 20, 25] have investigated approximations for departure processes in complex queueing systems. Related research deals with decoupling queueing networks into sub-systems where the output of one or several queues is fed as an input to other queues; see [18, 29, 30, 31] and references therein. In such cases, it is of crucial importance to understand the structure of $\mathbb{V}\mathrm{ar}D(t)$.

Contribution & main result. In this paper we contribute to the analysis of $\mathbb{V}\mathrm{ar}D(t)$ by considering the *critically loaded* GI/G/1 queue. This critically loaded regime, in which the mean inter-arrival 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 counter-intuitive results in line with the BRAVO (Balancing Reduces Asymptotic Variance of Outputs) effect observed previously in finite-capacity birth-death queues [22], see also [21].

We now describe the contribution of our work in more detail. In our GI/G/1 queue we denote by $Q(t)$ the number of customers present at time t. We let ζ_A represent a generic inter-arrival time and ζ_S a generic service time. We denote the system load by $\varrho := \lambda/\mu$, with $\lambda := 1/\mathbb{E}\zeta_A$ and $\mu := 1/\mathbb{E}\zeta_S$, 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

$$\sigma := \lim_{t\to\infty} \frac{\mathbb{V}\mathrm{ar}D(t)}{t},$$

when the queue is critically loaded, that is, $\varrho = 1$. Under suitable regularity conditions, it is not hard to prove that $m := \lim_{t\to\infty} \mathbb{E}D(t)/t = \min\{\lambda, \mu\}$, whereas $\sigma = \lambda c_a^2$ for $\varrho < 1$ and $\sigma = \mu c_s^2$ for $\varrho > 1$. However, there evidently is no explicit expression for σ in case $\varrho = 1$, in the literature. We show that

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

It thus follows that the variability function $v(\varrho) := \sigma/m = \lim_{t\to\infty} \mathbb{V}\mathrm{ar}D(t)/\mathbb{E}D(t)$ 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 the service process. 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) 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 multi-channel, multi-server queues with more general (non-renewal) 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); see Figure 1. This numerical experiment illustrates that the BRAVO phenomenon may also be observed in practice, in that it is not limited to the 'singular' case of $\rho = 1$.

Outline of technical results. Our starting point for obtaining (1) is a diffusion limit presented in [15, Sec. 4], where it is shown that for critically loaded queues the sequence of processes

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

converges weakly to

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

where $B_1(\cdot)$ and $B_2(\cdot)$ are independent standard Brownian motions. It then turns out that

$$\sigma = \lambda \mathbb{V}\mathrm{ar}\hat{D}(1),$$

given suitable uniform integrability (UI) conditions. The details are in the proof of Theorem 2.1. We then identify the distribution of $\hat{D}(1)$ (which for brevity we denote by simply $\hat{D}$). We show that $\mathbb{V}\mathrm{ar}\hat{D} = (1 - 2/\pi)(c_a^2 + c_s^2)$. This is done by relying on explicit formulae for the distribution of the maximum value attained by a Brownian bridge.

In attacking the UI conditions, our problem narrows down to proving that the sequence $\{Q(t)^2/t\}$ is UI. We subsequently prove UI for the M/M/1 queue, the GI/M/1 queue, and the GI/NWU/1 queue (where 'NWU' stands for *new worse than used*). The analysis of these three cases is of an incremental nature, in the sense that the argumentation becomes increasingly involved; we rely on properties of the reflection map for the queue length, some stochastic ordering results, and a number of new renewal-theoretic results (which are of independent interest). Finally we find that a sufficient condition for the UI requirement is that

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

where B denotes 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. The above condition has been shown to hold for the critically loaded M/G/1 in [34], and we conjecture that it holds for the critically loaded GI/G/1 queue as well (under appropriate moment conditions).

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) 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 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 $\mathbb{V}\mathrm{ar}D(t)$ at all time points in terms of Bessel functions. This yields an alternative derivation of (1) for this case as well as other more refined properties.

Organization. This paper is organized as follows. In Section 2 we present the main result. As mentioned above, we believe this result to hold under weaker assumptions, which we state in a conjecture. In Section 3 we derive the distribution of $\hat{D}$, and compute the explicit expression for $\mathbb{V}\mathrm{ar}\hat{D}$. 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 multi-server GI/G/s queue, as well as to queues with general more general arrival and service patterns.

Preliminaries & notation. This section is concluded by a review of some general definitions and notation, and preliminary results.
Recall that a collection of random variables $\{Z_t\}$ is uniformly integrable (UI) if

$$\lim_{M\to\infty}\left(\sup_t \mathbb{E}|Z_t|1_{\{|Z_t|\geq M\}}\right)=0.$$

A well known sufficient condition is to have

$$\sup_t \mathbb{E}\left(|Z_t|^{1+\epsilon}\right)<\infty, \quad \text{for some } \epsilon>0.$$

We denote by $Z_t \Rightarrow Z$ the fact that Z_t converges in distribution to Z. In case Z_t is UI, this also implies that $\lim_{t\to\infty}\mathbb{E}Z_t=\mathbb{E}Z$, see [4].
With X and Y non-negative random variables, $X\leq_{\rm st} Y$ means that

$$\mathbb{P}(X>x)\leq \mathbb{P}(Y>x),\ \forall x\geq 0.$$

Observe that this immediately implies that $\mathbb{E}X^n\leq \mathbb{E}Y^n$ for $n\geq 0$. Recall that a distribution of a random variable X is *new worse than used* (NWU) if

$$\mathbb{P}(X>x)\leq \frac{\mathbb{P}(X>t+x)}{\mathbb{P}(X>t)}, \quad \forall x,t\geq 0.$$

We denote by GI/NWU/1, the single server queue with the service times having a NWU service distribution; see [24] for more background.
Recall that Doob's L_p maximum inequality for both continuous time and discrete time states that for any $p>1$ and martingale $\{M_t\}$,

$$\mathbb{E}\left(\left(\sup_{0\leq s\leq t}|M_s|\right)^p\right)\leq\left(\frac{p}{p-1}\right)^p\mathbb{E}(|M_t|^p).$$

We shall make frequent use of the following inequality for real x and y and $r\geq 1$:

$$|x+y|^r\leq 2^{r-1}\left(|x|^r+|y|^r\right), \tag{3}$$

the validity of this statement follows from the fact that $(1+z)^r/(1+z^r)$ reaches a maximum at $z=1$ for $r\geq 1$ and $z\geq 0$.

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 inter-arrival time is identically distributed to the generic inter-arrival 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 $\mathscr{Q}$, defined as

$$\mathscr{Q}=\left\{\frac{Q(t)^2}{t},\ t\geq t_0\right\}, \tag{4}$$

for some $t_0 > 0$.

Theorem 2.1. *Consider the critically loaded GI/G/1 queue with $\mathbb{E}\zeta_A^2 < \infty$ and $\mathbb{E}\zeta_S^2 < \infty$. Assume $\mathscr{Q}$ is UI, then*

$$\sigma = \lambda \left(1 - \frac{2}{\pi}\right) \left(c_a^2 + c_s^2\right). \tag{5}$$

Proof: From the heavy-traffic functional central limit theorem in [15, Thm. 4.1], upon applying the projection map (at the time $t = 1$) and the continuous mapping theorem, we have

$$\frac{D(t) - \lambda t}{\sqrt{\lambda t}} \Rightarrow \hat{D} \quad \text{as} \quad t \to \infty. \tag{6}$$

Further, using the continuous mapping theorem we obtain

$$\frac{(D(t) - \lambda t)^2}{\lambda t} \Rightarrow \hat{D}^2 \quad \text{as} \quad t \to \infty. \tag{7}$$

Under UI conditions established below, we have from (6) and (7) that

$$\lim_{t\to\infty} \mathbb{E}\left(\left(\frac{D(t) - \lambda t}{\sqrt{\lambda t}}\right)^k\right) = \mathbb{E}(\hat{D}^k), \quad k = 1, 2. \tag{8}$$

Observe that $\mathbb{V}\mathrm{ar}D(t) = \mathbb{E}(D(t) - \lambda t)^2 - (\mathbb{E}D(t) - \lambda t)^2$, and combine this with (8) to obtain

$$\begin{aligned} \frac{\sigma}{\lambda} &= \lim_{t\to\infty} \frac{\mathbb{V}\mathrm{ar}D(t)}{\lambda t} \\ &= \lim_{t\to\infty} \frac{\mathbb{E}(D(t) - \lambda t)^2}{\lambda t} - \left(\lim_{t\to\infty} \frac{\mathbb{E}D(t) - \lambda t}{\sqrt{\lambda t}}\right)^2 = \mathbb{V}\mathrm{ar}\hat{D}, \end{aligned} \tag{9}$$

which yields the desired result using Proposition 3.2.
It now remains to establish the convergence of the moments in (8). To do so, we establish that the sequences $\{[(D(t) - \lambda t)/\sqrt{\lambda t}]^k, t \geq t_0\}$, $k = 1, 2$, are UI. First note $D(t) = A(t) - Q(t)$. Combining this with (3) yields

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

It thus suffices to show that the sequences $\{(A(t) - \lambda t)^2/\lambda t, t \geq t_0\}$ and $\mathscr{Q}$ are UI. UI of the first sequence is a standard result from renewal theory, cf. [12, 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 multi-channel, multi-server queues with more general arrival and service processes. We are able to establish the UI of $\mathscr{Q}$ needed by Theorem 2.1 for different cases:

Theorem 2.2. *If $\mathbb{E}\zeta_A^4 < \infty$ and $\mathbb{E}\zeta_S^4 < \infty$ then $\mathscr{Q}$ is UI in the following cases:*

(i) *Any critically loaded GI/G/1 queue with $\mathbb{P}(B > x) \sim L(x)x^{-1/2}$ 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 $\mathbb{P}(\zeta_S > b) = 1$ 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 in Section 6.

We conjecture that our result also holds under milder conditions. To this end, we first remark that the condition of (i) in Theorem 2.2 has been shown to be true in [34] for the critically loaded M/G/1 queue with $\mathbb{E}\zeta_S^2 < \infty$. We conjecture that this also holds for the critically loaded GI/G/1.

Conjecture 2.3. *For the critically loaded GI/G/1 queue with $\mathbb{E}\zeta_A^2 < \infty$ and $\mathbb{E}\zeta_S^2 < \infty$,*

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

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

Conjecture 2.3, 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 $2+\epsilon$ moments, for some strictly positive ϵ. Combining this with the multi-server result of Section 6, we conjecture the following:

Conjecture 2.4. *Consider the critically loaded GI/G/s multi-server queue. Assume that $\mathbb{E}\zeta_A^{2+\epsilon} < \infty$ and $\mathbb{E}\zeta_S^{2+\epsilon} < \infty$ for any $\epsilon > 0$. Then (5) holds.*

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

In this section we derive the distribution of the random variable $\hat{D}$, defined as $\inf_{0\le t\le 1}\{c_1B_1(t) + c_2B_2(1-t)\}$, with B_1 and B_2 be two independent standard Brownian motions. This answers an open question posed in [15]. As usual, $\Phi(x)$ is the distribution function of a standard normal random variable.

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

$$\begin{aligned}\mathbb{P}(\hat{D} \le x) &= \Phi\left(x/c_1\right) + \Phi\left(x/c_2\right) - \Phi\left(x/c_1\right)\Phi\left(x/c_2\right)\\ &\quad + 1/\sqrt{2\pi}\int_0^\infty e^{-L(u,x)}\Phi\left(-M(u,x)\right)\mathrm{d}u,\end{aligned} \tag{10}$$

$$\begin{aligned}L(u,x) &:= 1/2\left(u(c_1^2-c_2^2)/\check{c}^2 - x/c_1\right)^2,\ M(u,x) := (2uc_1c_2)/\check{c}^2 + x/c_2,\\ \check{c} &:= \sqrt{c_1^2+c_2^2}.\end{aligned}$$

For the case $c_1 = c_2 = c$ the last term in the right-hand side of (10) simplifies to

$$e^{-x^2/(2c^2)}/\sqrt{2\pi}\left(e^{-x^2/(2c^2)}/\sqrt{2\pi} - x\Phi\left(-x/c\right)/c\right).$$

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. Further, 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,

$$\mathbb{P}(\hat{D} \le x \,|\, \mathscr{E}(b_1,b_2)) = \begin{cases} \mathbb{P}(\inf_{0\le t\le 1}\{b_2c_2 + \check{c}B^{(d)}(t)\} \le x), & x \le \min(b_1c_1, b_2c_2),\\ 1, & x > \min(b_1c_1, b_2c_2),\end{cases}$$

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

Manipulating the above probability of the Brownian bridge, we obtain

$$\begin{aligned}\mathbb{P}\left(\inf_{0\le t\le 1}\{b_2c_2+\check{c}B^{(d)}(t)\}\le x\right) &= \mathbb{P}\left(\sup_{0\le t\le 1}\{-B^{(d)}(t)\}\ge (b_2c_2-x)/\check{c}\right)\\ &= \mathbb{P}\left(\sup_{0\le t\le 1}\{B^{(-d)}(t)\}\ge (b_2c_2-x)/\check{c}\right).\end{aligned}$$

The first equality is trivial and the second step follows from the symmetry of the Brownian bridge. Now use that [19, Ch. V]

$$\mathbb{P}\left(\sup_{0\le t\le 1}\{B^{(b)}(t)\}>y\right)=e^{-2y(y-b)},$$

to arrive at

$$\mathbb{P}(\hat{D}\le x\,|\,\mathscr{E}(b_1,b_2))=\begin{cases}\exp\{-\frac{2}{\check{c}^2}(x-b_1c_1)(x-b_2c_2)\}, & x\le\min(b_1c_1,b_2c_2),\\ 1, & x>\min(b_1c_1,b_2c_2).\end{cases}$$

By unconditioning, we obtain that

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

The first integral of the last expression can be represented as $\Phi(x/c_1)+\Phi(x/c_2)-\Phi(x/c_1)\Phi(x/c_2)$. For the integral on the right hand side, we first change the region of integration to the positive quadrant then, move the terms involving only b_1 out of the inner integral and then complete the square:

$$\begin{aligned}&\frac{1}{2\pi}\int_0^\infty\int_0^\infty e^{-(\frac{2c_1c_2}{\check{c}^2}b_1b_2+\frac{1}{2}(b_1+\frac{x}{c_1})^2+\frac{1}{2}(b_2+\frac{x}{c_2})^2)}\,\mathrm{d}b_1\,\mathrm{d}b_2\\ &\quad=\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-L(u,x)}\Phi\left(-M(u,x)\right)\mathrm{d}u.\end{aligned}\tag{11}$$

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 $\mathbb{V}\mathrm{ar}\hat{D}$.

Proposition 3.2.

$$\mathbb{E}\hat{D}=-\sqrt{2(c_1^2+c_2^2)/\pi},\quad \mathbb{E}\hat{D}^2=c_1^2+c_2^2,\quad \mathbb{V}\mathrm{ar}\hat{D}=(c_1^2+c_2^2)(1-2/\pi).$$

Proof: We first determine the density of $\hat{D}$ by differentiating the distribution function, and 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 (11) 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 $\mathscr{Q}$ is UI. In this section we find sufficient conditions for this assumption to hold, thus establishing (i)-(iv) in 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 M/M/1 and then for the GI/M/1. 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 n-th arrival, and then apply a UI result from [27].

Note that Corollary 4.6 is more general than Corollary 4.4, which is in turn more general than Corollary 4.2. As we feel that these results are of independent interest, and as they add insight, we chose to present all three results.

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 $\mathscr{Q}'$, defined as

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

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

$$Q'(t) = X(t) - \inf_{0\leq s\leq t} X(s) \tag{12}$$

denote the associated reflected process. Notice that for the GI/M/1 it holds that $Q'(t)$ equals $Q(t)$; see e.g. [23, p. 68]; this does not hold for the GI/G/1. For the M/M/1 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*

$$\mathbb{E}\left(\left(\sup_{0\leq s\leq t}\{|A(s)-\lambda s|\}\right)^4\right) \text{ and } \mathbb{E}\left(\left(\sup_{0\leq s\leq t}\{|S(s)-\lambda s|\}\right)^4\right) \tag{13}$$

are $O(t^2)$. Then it holds that

(i) $\mathbb{E}(Q'(t)^4) = O(t^2)$.
(ii) $\sup_{t\geq t_0} \mathbb{E}(Q'(t)^2/t^2) < \infty$.
(iii) *$\mathscr{Q}'$ is UI.*

Proof Use inequality (3), with $r = 4$, to obtain that

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

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

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

and therefore it follows from (13) that

$$\mathbb{E}\big(X(t)^4\big) = O(t^2). \tag{14}$$

We then consider the second term:

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

Again invoking (13) yields

$$\mathbb{E}\big(\sup_{0\le s\le t} -X(s)\big)^4 = O(t^2). \tag{15}$$

Upon combining (14) and (15), we obtain (i). The result (ii) follows directly from (i), and (iii) follows from the sufficient condition of UI in (ii). □

It now follows (almost) immediately that we have uniform integrability of $\mathscr{Q}$ in the M/M/1 case.

Corollary 4.2. *For the critically loaded M/M/1 queue, $\mathscr{Q}$ is UI.*

Proof All we need to show is that the arrival and service Poisson processes satisfy (13). To this end, observe that the process $\{A(t)-\lambda t\}$ is a martingale. Applying Doob's maximum inequality, we obtain

$$\mathbb{E}\left(\big(\sup_{0\le s\le t}(A(s)-\lambda s)\big)^4\right) \le \left(\frac{4}{3}\right)^4 (3\lambda^2 t^2+\lambda t) = O(t^2).$$

An identical argument is used for $\{S(t)-\lambda t\}$. □

Having established the uniform integrability of $\mathscr{Q}$ in the critically loaded M/M/1 case, we now attempt to generalize the above martingale argument for the GI/M/1 case. We do so in the theorem below, which we believe to be of independent interest as well; to the best of our knowledge, it has not appeared elsewhere in the literature.

Theorem 4.3. *Let $\{\zeta_i, i\ge 0\}$ be a sequence of nonnegative i.i.d. random variables, and $S_n := \sum_{i=1}^n \zeta_i$ their partial sums. Denote the corresponding renewal counting process by $N(t) := \sup\{n : S_n \le t\}$. Define $\mathbb{E}\zeta_1 := \gamma^{-1}$, and assume $\mathbb{E}\zeta_1^4 < \infty$. Then,*

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

Proof Denote $V(t) = \inf_n\{n : S_n \ge t\}$, 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 that

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

and on the other hand

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

Combining these two inequalities, we obtain

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

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

$$\sup_{0\le s\le t} |N(s) - \gamma s| \le \sup_{0\le n\le V(t)} |\gamma S_n - n| + 1 = \sup_{0\le n\le V(t)} |M_n| + 1. \tag{16}$$

We are interested in the 4-th moment of the quantity in the left-hand side of (16). Due to (3), we have

$$\mathbb{E}\left(\left(\sup_{0\le s\le t} |N(s) - \gamma s|\right)\right)^4 \le 8\mathbb{E}\left(\left(\sup_{0\le n\le V(t)} |M_n|\right)^4\right) + 8. \tag{17}$$

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, due to Doob's maximum inequality, for $k = 0, 1, \ldots$,

$$\mathbb{E}\left(\left(\sup_{0\le n\le k} |M_{n\wedge V(t)}|\right)^4\right) \le \left(\frac{4}{3}\right)^4 \mathbb{E}\left((M_{k\wedge V(t)})^4\right). \tag{18}$$

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

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

Applying the monotone convergence theorem, we obtain

$$\lim_{k\to\infty} \mathbb{E}\left(\left(\sup_{0\le n\le k} |M_{n\wedge V(t)}|\right)^4\right) = \mathbb{E}\left(\left(\sup_{0\le n\le V(t)} |M_n|\right)^4\right). \tag{19}$$

Further observe that, almost surely

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

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

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

and the fact that for fixed t the right-hand side has finite mean, see e.g. [12]. Now applying the dominated convergence theorem, we obtain

$$\lim_{k\to\infty} \mathbb{E}\left(\big(M_{k\wedge V(t)}\big)^4\right) = \mathbb{E}\left(\big(M_{V(t)}\big)^4\right). \tag{20}$$

Combining (18), (19) and (20) we obtain,

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

We now complete the proof by showing that the right-hand side of the previous display is $O(t^2)$. To this end, denote $\mathbb{E}(\xi_i^\ell) = m_\ell$, and recall that it was assumed that $m_\ell < \infty$, $\ell = 1, 2, 3, 4$. Further let $\gamma(r)$ denote the cumulant generating function of ξ_i, i.e., $\gamma(r) = \log(\mathbb{E}(e^{r\xi_i}))$, $\mathrm{Re}(r) \leq 0$. Let $\gamma^{(n)}(r)$ denote the n-th derivative of $\gamma(r)$. Observe that $\gamma(0) = 0$, $\gamma^{(1)}(0) = m_1 = 0$, $\gamma^{(2)}(0) = \mathbb{V}\mathrm{ar}(\xi_i) = m_2$ and that $\gamma^{(3)}(0)$ and $\gamma^{(4)}(0)$ can be expressed in terms of $m_\ell, \ell = 2, 3, 4$. Since $V(t)$ is a stopping time, Wald's identity [26] yields

$$\mathbb{E}\,\exp\bigl(rM_{V(t)} - V(t)\gamma(r)\bigr) = 1.$$

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

$$\mathbb{E}(M_{V(t)})^2 = \mathbb{E}V(t)m_2, \tag{21}$$

$$\begin{aligned}\mathbb{E}(M_{V(t)})^4 &= \gamma^{(4)}(0)\mathbb{E}V(t) + 4\gamma^{(3)}(0)\mathbb{E}V(t)M_{V(t)} - 3\mathbb{E}V(t)^2m_2^2 \\ &\quad + 6m_2\mathbb{E}V(t)(M_{V(t)})^2.\end{aligned} \tag{22}$$

Then, note that the Cauchy-Schwarz inequality gives

$$\mathbb{E}V(t)M_{V(t)} \leq \sqrt{\mathbb{E}V(t)^2\mathbb{E}(M_{V(t)})^2}, \tag{23}$$

$$\mathbb{E}V(t)(M_{V(t)})^2 \leq \sqrt{\mathbb{E}V(t)^2\mathbb{E}\bigl((M_{V(t)})^2\bigr)^2} = \mathbb{E}(M_{V(t)})^2\sqrt{\mathbb{E}V(t)^2}. \tag{24}$$

Also, $\mathbb{E}V(t) = O(t)$ and $\mathbb{E}V(t)^2 = O(t^2)$, see e.g. [3, Ch. V]. From (21), we deduce that $\mathbb{E}(M_{V(t)})^2 = O(t)$. Using (23) and (24), the latter equation gives that $\mathbb{E}V(t)M_{V(t)} = O(t^{3/2})$ and $\mathbb{E}V(t)(M_{V(t)})^2 = O(t^2)$. Plugging these results into (22) yields

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

as desired. □

Corollary 4.4. *For the critically loaded GI/M/1 queue, with $\mathbb{E}\zeta_A^4 < \infty$, $\mathscr{Q}$ is UI.*

Proof Theorem 4.3 gives (13) which completes the proof. □

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

Theorem 4.5. *Denote by $C(t)$ the number of busy cycles of the process $\{Q(t)\}$ during the time interval $[0, t]$. Let $r \geq 1$. Then,*

(i) *For GI/NWU/1: $Q(t) \leq_{\mathrm{st}} Q'(t)$, $t \geq 0$.*

(ii) *For GI/G/1: $\mathbb{E}Q(t)^r \leq 2^{r-1}\Bigl(\mathbb{E}Q'(t)^r + \mathbb{E}C(t)^r\Bigr)$, $t \geq 0$,*

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

$$\tilde{Q}(t) \leq \tilde{Q}'(t),\ t \geq 0,\ \text{w.p. } 1, \tag{25}$$

where $L(Q) = L(\tilde{Q})$ and $L(Q') = L(\tilde{Q}')$. Establishing such a construction is equivalent to stochastic order on the function space of sample paths, see [17], from

which (i) is an elementary consequence. We let $\tilde{Q} = Q$, so that it remains to produce (25) with $L(Q') = L(\tilde{Q}')$. 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 comes to an empty system. Otherwise, arrivals are assigned identical service times in both systems, which are taken from the given i.i.d 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, that 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 w.p. 1, and yet has its given probability law. Performing this simple construction maintains $L(Q') = L(\tilde{Q}')$. 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 w.p. 1, while their distributions remain unchanged. It is known and not difficult to show that the queue length sample paths will be ordered w.p. 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 [28]. Hence we achieve the sample-path order in (25) while keeping the relations $L(Q) = L(\tilde{Q})$ and $L(Q') = L(\tilde{Q}')$. This sample path order holds over the successive finite time segments $[0, \tau_n)$, where τ_n is the time that the n^{th} busy cycle begins. By mathematical induction, it thus holds over the entire positive halfline. We thus have (i).

We now turn to (ii). We shall 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

$$Q(t) \le \tilde{Q}'(t) + C(t), \ t \ge 0, \ \text{w.p. } 1. \tag{26}$$

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

$$Q(t) \le \hat{Q}(t) \le \tilde{Q}'(t) + C(t), \ t \ge 0, \ \text{w.p. } 1, \tag{27}$$

where still $L(Q') = L(\tilde{Q}')$.
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 i.i.d. random variables for Q.
The right hand side of (27) 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 (27).
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 (27). Combining now with (3) 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 $Q(t) \leq Q'(t) + C(t)$, w.p. 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 for queues with arbitrary arrival processes.
We now have UI of $\mathscr{Q}$ for the GI/NWU/1 queue.

Corollary 4.6. *For the critically loaded GI/NWU/1 queue with $\mathbb{E}\zeta_A^4 < \infty$ and $\mathbb{E}\zeta_S^4 < \infty$, $\mathscr{Q}$ is UI.*

Proof From Theorem 4.5 (i) we deduce that $\mathbb{E}Q(t)^4 \leq \mathbb{E}Q'(t)^4$. By Proposition 4.1 (ii) we have that $\mathbb{E}Q'(t)^4 = O(t^2)$. Thus, $\mathbb{E}Q(t)^4 = O(t^2)$, which completes the proof. □

In order to use the stochastic order in Proposition 4.1 (ii) for the UI of $\mathscr{Q}$ in the GI/G/1 queue, one needs 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 our knowledge this general result about renewal processes has not appeared elsewhere.

Theorem 4.7. *Let $N(t)$ and ζ_i be defined as in Theorem 4.3. Suppose that $\mathbb{P}(\zeta_i \geq x) = 1 - F(x) \sim L(x)x^{-\alpha}$ with $\alpha \in [0,1)$ and $L(\cdot)$ slowly varying. Then,*

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

Proof

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

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

$$P(B > x) \sim L(x)x^{-1/2}, \tag{28}$$

with $L(x)$ slowly varying and bounded then $\mathscr{Q}$ is UI.

Proof We apply Theorem 4.7 with $m = 4$ to $C(t)$ of Theorem 4.5, to obtain that $\mathbb{E}C(t)^4 = O(t^2)$. Further, observe that Theorem 4.3 applied to $A(t)$ and $S(t)$ implies condition (13), and thus by Proposition 4.1 (i), we have that $\mathbb{E}Q'(t)^4 = O(t^2)$. Since Theorem 4.5 (i) implies that $\mathbb{E}Q(t)^4 \leq 8\big(\mathbb{E}Q'(t)^4 + \mathbb{E}C(t)^4\big)$, we have

$$\mathbb{E}Q(t)^4 = O(t^2),$$

and as a result,

$$\sup_{t\geq t_0} \mathbb{E}\Big(\frac{Q'(t)^2}{t}\Big)^2 < \infty.$$

Conclude that $\mathscr{Q}$ is UI. □

Corollary 4.9. *For the critically loaded M/G/1 queue, with $\mathbb{E}\zeta_S^4 < \infty$, $\mathscr{Q}$ is UI.*

Proof The tail asymptotics for the busy period in (28) have been established for the critically loaded M/G/1 queue in [34, Theorem 4.1]. Consequently, the result follows from Theorem 4.8. □

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 [27].

Proposition 4.10. *For the critically loaded D/G/1 queue with $\mathbb{E}\zeta_S^4 < \infty$ and $\mathbb{P}(\zeta_S > b) = 1$ for some $b > 0$, $\mathscr{Q}$ is UI.*

Proof In the following we relate $Q(t)$ and W_n, the workload seen by the n-th arrival. Note that in [27, Thm. 4.1] it is shown that if $\mathbb{E}\zeta_S^{2m} < \infty$, then $(W_n/\sqrt{n})^k$, $k \leq 2m$, is UI. 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 UI, then so is Z_n.
We have that $Q(t) \leq W(t)/b + 1$ 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 $\lfloor \lambda t \rfloor > 0$,

$$\begin{aligned} \frac{Q(t)}{\sqrt{t}} &\leq b^{-1}\frac{W(t)+b}{\sqrt{t}} \leq b^{-1}\frac{W_{A(t)}+b}{\sqrt{t}} \leq b^{-1}\sqrt{\lambda}\frac{W_{A(t)}+b}{\sqrt{A(t)}} \\ &= b^{-1}\sqrt{\lambda}\Big\{\frac{W_{\lfloor \lambda t\rfloor}}{\sqrt{\lfloor \lambda t\rfloor}} + \frac{b}{\sqrt{\lfloor \lambda t\rfloor}}\Big\}, \end{aligned} \tag{29}$$

where the third inequality and the last follow from $A(t) = \lfloor \lambda t \rfloor \leq \lambda t$ (at time 0 the queue is empty and an inter-arrival time is deterministic and equal to $1/\lambda$). Using (3) with $r = 4$, Eqn. (29) then gives

$$\Big(\frac{Q(t)}{\sqrt{t}}\Big)^4 \leq 8b^{-4}\lambda^2\Big\{\Big(\frac{W_{\lfloor \lambda t\rfloor}}{\sqrt{\lfloor \lambda t\rfloor}}\Big)^4 + \Big(\frac{b}{\sqrt{\lfloor \lambda t\rfloor}}\Big)^4\Big\}. \tag{30}$$

Note that $(b/\sqrt{\lfloor \lambda t\rfloor})^4$ is bounded from above by b^4, $t \geq t_0 > 0$, which implies that it is UI. Moreover, under the assumption $\mathbb{E}\zeta_S^4 < \infty$, we have that $(W_{\lfloor \lambda t\rfloor}/\sqrt{\lfloor \lambda t\rfloor})^4$ is UI, see [27, Thm. 4.1]. Hence, we have that both terms in the right-hand side of (30) are UI, and hence so is $\mathscr{Q}$. □

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 $t \geq 0$. We first consider arbitrary $\lambda, \mu > 0$ and obtain cumbersome yet computationally tractable expressions for $\mathbb{E}D(t)$ and $\mathbb{V}\mathrm{ar}D(t)$ 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 for $\varrho \approx 1$ (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 5.2).

We are further able to perform an asymptotic expansion for $\mathbb{E}D(t)$ and $\mathbb{V}\mathrm{ar}D(t)$ for t large (Theorem 5.3). This expansion shows that in the critically loaded case, the variance and expectation curves have a lower order square root term that does not exist when $\lambda \neq \mu$. It also serves 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:

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

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

$$\begin{aligned}
\mathbb{E}D(t) &= \sqrt{\lambda\mu} \int_0^t (t-u) \frac{I_1(2u\sqrt{\lambda\mu}) e^{-(\lambda+\mu)u}}{u} \mathrm{d}u. \\
\mathbb{V}\mathrm{ar}D(t) &= \mu t(\mu t + 2) - \sqrt{\lambda\mu} \int_0^t (t-u)(\mu(t-u)+2) \frac{I_1(2u\sqrt{\lambda\mu})}{u} e^{-(\lambda+\mu)u} \mathrm{d}u \\
&\quad + 2\lambda\mu \int_0^t (t-u)^2 \frac{I_2(2u\sqrt{\lambda\mu})}{u} e^{-(\lambda+\mu)u} \mathrm{d}u \\
&\quad + \mu \int_0^t \big(\mu(\lambda-\mu)(t-u)^2 - 4\mu(t-u) - 2\big) I_0(2u\sqrt{\lambda\mu}) e^{-(\lambda+\mu)u} \mathrm{d}u \\
&\quad + \mu\sqrt{\lambda\mu} \int_0^t (t-u)\big((\mu-\lambda)(t-u)+2\big) I_1(2u\sqrt{\lambda\mu}) e^{-(\lambda+\mu)u} \mathrm{d}u \\
&\quad + \sqrt{\lambda\mu} \int_0^t (t-u) \frac{I_1(2u\sqrt{\lambda\mu}) e^{-(\lambda+\mu)u}}{u} \mathrm{d}u \\
&\quad - \lambda\mu \Big(\int_0^t (t-u) \frac{I_1(2u\sqrt{\lambda\mu}) e^{-(\lambda+\mu)u}}{u} \mathrm{d}u \Big)^2.
\end{aligned}$$

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 that

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

Note that, $\phi_\alpha(z)/\alpha$ can be interpreted as the Laplace transform of $\mathbb{E}z^{D(t)}$. Denote $\phi_\alpha^1 := \phi_\alpha'(1)/\alpha$ 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,

$$\mathbb{E}D(t) = \mathcal{L}^{-1}(\phi_\alpha^1), \quad \mathbb{E}D(t)^2 = \mathcal{L}^{-1}(\phi_\alpha^2 + \phi_\alpha^1). \tag{31}$$

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

$$\frac{\phi_\alpha(z)}{\alpha} = \frac{z}{\mu(1-z)+\alpha}\frac{1-r(z,\alpha)}{z-r(z,\alpha)}, \tag{32}$$

where

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

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 (32) according to z at the point $z=1$ yields

$$\phi_\alpha^1 = \frac{\lambda}{\alpha^2}r(1,\alpha), \tag{33}$$

$$\begin{aligned}\phi_\alpha^2 = {} & 2\mu\frac{\mu+\alpha}{\alpha^3} - 2\lambda\frac{\mu+\alpha}{\alpha^3}r(1,\alpha) + \frac{2\lambda^2}{\alpha^3}r(1,\alpha)^2 \\ & + \frac{2\mu}{\alpha^3}\Big(\mu - \frac{(\mu+\alpha)^2}{\lambda}\Big)\frac{1}{s(1,\alpha)-r(1,\alpha)} + \frac{2\mu}{\alpha^3}(\mu+\alpha-\lambda)\frac{r(1,\alpha)}{s(1,\alpha)-r(1,\alpha)}.\end{aligned} \tag{34}$$

Now using an explicit inversion, as in e.g. [6, p. 81], for (33), we obtain

$$\begin{aligned}\mathcal{L}^{-1}(\phi_\alpha^1) &= \lambda t * \sqrt{\mu/\lambda}\frac{I_1(2t\sqrt{\lambda\mu})e^{-(\lambda+\mu)t}}{t}, \\ \mathcal{L}^{-1}(\phi_\alpha^2) &= \mu t(\mu t+2) - \sqrt{\lambda\mu}t(\mu t+2) * \Big(\frac{I_1(2t\sqrt{\lambda\mu})}{t}e^{-(\lambda+\mu)t}\Big) \\ &\quad +2\lambda\mu t^2 * \Big(\frac{I_2(2t\sqrt{\lambda\mu})}{t}e^{-(\lambda+\mu)t}\Big) \\ &\quad +\mu\big(\mu(\lambda-\mu)t^2 - 4\mu t - 2\big) * \big(I_0(2t\sqrt{\lambda\mu})e^{-(\lambda+\mu)t}\big) \\ &\quad +\mu\sqrt{\lambda\mu}t\big((\mu-\lambda)t+2\big) * \big(I_1(2t\sqrt{\lambda\mu})e^{-(\lambda+\mu)t}\big).\end{aligned}$$

Using (31) 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 5.2. *For the critically loaded M/M/1 queue with $Q(0)=0$:*

$$\begin{aligned}\mathbb{E}D(t) &= \lambda t - \frac{1}{2}e^{-2\lambda t}\big((1+4\lambda t)I_0(2\lambda t) + 4\lambda t I_1(2\lambda t)\big) + \frac{1}{2}, \\ \mathbb{V}\mathrm{ar}D(t) &= \frac{1}{4}e^{-4\lambda t}\big(e^{4\lambda t}(8\lambda t+1) - (4\lambda t+1)^2 I_0(2\lambda t)^2 - 4e^{2\lambda t}\lambda t I_1(2\lambda t) \\ &\quad -16\lambda^2 t^2 I_1(2\lambda t)^2 - 4\lambda t I_0(2\lambda t)\big(e^{2\lambda t} + (2+8\lambda t)I_1(2\lambda t)\big)\big).\end{aligned}$$

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

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

Theorem 5.3. *For the M/M/1 queue with $Q(0)=0$:*

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

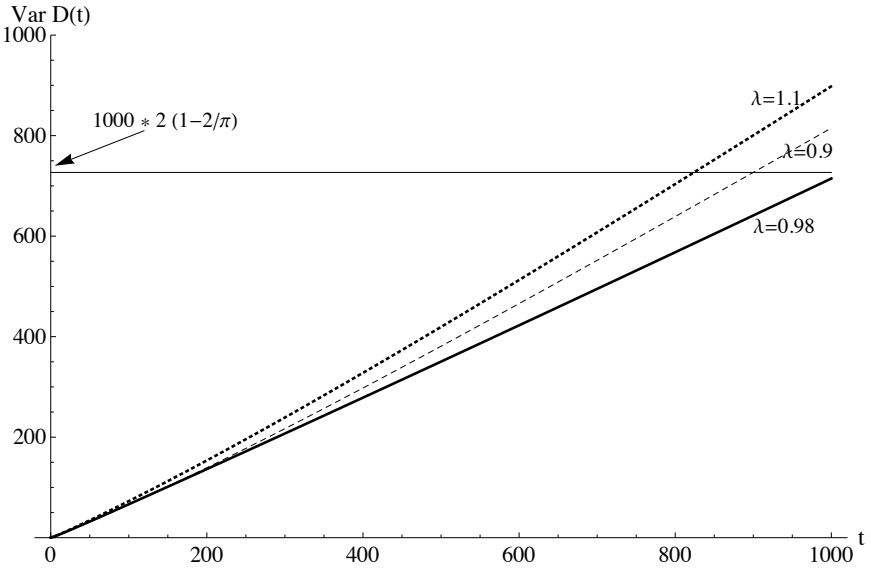

FIGURE 1. Demonstration of the BRAVO effect for $\lambda \approx \mu$ and finite t: $\mathbb{V}\mathrm{ar}D(t)$ is plotted for M/M/1 systems with $\mu = 1$. The dashed curved is for $\lambda = 0.9$, the solid curve is for $\lambda = 0.98$ and the dotted curve is for $\lambda = 1.1$. The thin horizontal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

and

$$\mathbb{V}\mathrm{ar}D(t) = \begin{cases} \lambda t - \frac{\varrho}{(1-\varrho)^2} + o(1) & \text{if } \lambda < \mu, \\ \lambda 2(1 - \frac{2}{\pi})t - \sqrt{\frac{\lambda}{\pi}}\, t^{1/2} + \frac{\pi-2}{4\pi} + o(1) & \text{if } \lambda = \mu, \\ \mu t - \frac{\varrho}{(1-\varrho)^2} + o(1) & \text{if } \lambda > \mu. \end{cases}$$

Proof The cases $\lambda = \mu$ and $\lambda \neq \mu$ are treated separately. The $\lambda = \mu$ case follows directly from Corollary 5.2: To obtain the linear term divide the expressions of Corollary 5.2 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 $\mathbb{E}D(t)$. Theorem 5.1 readily gives:

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

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 $\gamma \in \mathbb{Z}$:

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

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

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

Our result for $\mathbb{E}D(t)$ now follows. The result for $\mathbb{V}\mathrm{ar}D(t)$ follows along the same lines. □

We end this section with a numerical example. We use Theorem 5.1 to evaluate $\mathbb{V}\mathrm{ar}D(t)$ 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 non-asymptotic 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) 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 [15]. 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 multi-server case.

The general model we consider is a multi-channel, multi-server queue as described in [15], see also [16]: r arrival channels of customers arrive to a queue with s servers. When a customer arrives to find one or more free servers, he 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 queues up to wait for the next available server in a 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 $A_i(t)$, $i = 1, \ldots, r$ and the service processes $S_i(t)$, $i = 1, \ldots, s$. Assume the existence of $\lambda_i > 0$, $i = 1, \ldots, r$ and $\mu_i > 0$, $i = 1, \ldots, s$, such that,

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

Consider the queue in the critical regime with λ:

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

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

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

where the weak convergence is as in [15] as $n \to \infty$, and $B(\cdot)$ is a standard Brownian motion, cf. also [32]. In case of renewal processes, κ_i^a/λ_i and κ_i^s/μ_i are the squared coefficient of variation of the inter-renewal times. For ease of reference, we refer to this model as the critically loaded $\mathrm{G}_r/\mathrm{G}/s$ 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 UI:*

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

Then:

$$\sigma = \left(1 - \frac{2}{\pi} \right) \left(\sum_{i=1}^{r} \kappa_i^a + \sum_{i=1}^{s} \kappa_i^s \right). \tag{35}$$

Proof Follows the exact same lines as the proof of Theorem 2.1. See also [16] for a discussion of generalizing renewal processes. □

The critically loaded $\mathrm{GI}/\mathrm{G}/s$ 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 denote $\kappa_1^a = \lambda c_a^2$ and $\kappa_i^s = \lambda c_s^2/s$, $i = 1, \ldots, s$. The asymptotic variance (35) reduces once again to:

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

For the $\mathrm{GI}/\mathrm{G}/s$ we are able to establish the required UI conditions for a variety of cases. Observe first that the first UI condition hold for renewal arrivals as in Theorem 2.1. Further conditions for the second sequence are given in the following:

Theorem 6.2. *Consider the critically loaded GI/G/s queue operating under the first come first served discipline. Assume $\mathbb{E}\zeta_A^4 < \infty$ and $\mathbb{E}\zeta_S^4 < \infty$. Then, $\big\{ Q(t)/t^2, t \geq t_0 \big\}$ is UI in the following cases:*

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

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

$$P(\zeta_A \leq x) = \int_0^x \frac{\lambda^{1/s}}{\Gamma(1/s)} t^{1/s - 1} e^{-\lambda t} dt.$$

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

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

Proof We apply the results in [33] for the special case of $\mathrm{GI}/\mathrm{G}/s$ and the cyclic service. In the cyclic service discipline, arrival $sj + i, j = 0, 1, \ldots$ is assigned to the i^{th} server, $i = 1, 2, \ldots, s$. The partition of the arrivals in this manner generates a

collection of s GI/G/1 queues, each with service time ζ_S and inter-arrival 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 $i^{\rm th}$ single-server queue at time t with $Q_i(0) = 0$. Then it follows from [33], Equation (8), that,

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

We now have that for case (i)-(iv):

$$\mathbb{E}Q(t)^4 \leq \mathbb{E}\Big(\sum_{i=1}^{s} Q_i(t)\Big)^4 \leq 8^{s-1}\mathbb{E}\sum_{i=1}^{s}\big(Q_i(t)\big)^4 = s8^{s-1}\mathbb{E}\big(Q_1(t)\big)^4 = O(t^2). \tag{36}$$

The second inequality follows from $s-1$ applications of (3). The $O(t^2)$ term is obtained for cases (i)-(iv) by 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($1/s$,λ) random variables is an exponential. Also observe that since $\mathbb{E}\zeta_A^4 < \infty$, the s-fold convolution retains this property as is needed for (i) and (iii). □

REFERENCES

[1] A. Al Hanbali, R. de Haan, R.J. Boucherie, and J.C.W. van Ommeren. A tandem queueing model for delay analysis in disconnected ad hoc networks. In *Proceedings of the 15th international conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 189–205. Springer, 2008.

[2] G. Alsmeyer. On generalized renewal measures and certain first passage times. *The Annals of Probability*, 20(3):1229–1247, 1992.

[3] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, 2003.

[4] P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics, 1999.

[5] P.J. Burke. The output of a queuing system. *Operations Research*, 4(6):699–704, 1956.

[6] J.W. Cohen. *The Single Server Queue*. North-Holland, 1982.

[7] D.J. Daley. Further second-order properties of certain single-server queueing systems. *Stochastic Processes and their Applications*, 3:185–191, 1975.

[8] D.J. Daley. Queueing output processes. *Advances in Applied Probability*, 8:395–415, 1976.

[9] R.L. Disney and P.C. Kiessler. *Traffic Processes in Queueing Networks – A Markov Renewal Approach*. The Johns Hopkins University Press, 1987.

[10] R.L. Disney and D. Konig. Queueing networks: A survey of their random processes. *SIAM Review*, 27(3):335–403, 1985.

[11] S.B. Gershwin. Variance of output of a tandem production system. *in: Queueing Networks with Finite Capacity*, eds R. Onvural and I. Akyildiz, Proceedings of the Second International Conference on Queueing Networks with Finite Capacity (Elsevier, Amsterdam)., 1993.

[12] A. Gut. *Stopped random walks*. Springer-Verlag, 1988.

[13] K.B. Hendricks. The output processes of serial production lines of exponential machines with finite buffers. *Operations Research*, 40(6):1139–1147, 1992.

[14] K.B. Hendricks and J.O. McClain. The output processes of serial production lines of general machines with finite buffers. *Management Science*, 39(10):1194–1201, 1993.

[15] D.L. Iglehart and W. Whitt. Multiple Channel Queues in Heavy Traffic. I. *Advances in Applied Probability*, 2(1):150–177, 1970.

[16] D.L. Iglehart and W. Whitt. Multiple Channel Queues in Heavy Traffic. II: Sequences, Networks and Batches. *Advances in Applied Probability*, 2(2):355–369, 1970.

[17] T. Kamae, U. Krengel, and G.L. O'Brien. Stochastic inequalities on partially ordered spaces. *The Annals of Probability*, 5(6):899–912, 1977.

[18] P. Kuehn. Approximate analysis of general queueing networks by decomposition. *IEEE Trans. Commun.*, 27:113–126, 1979.

[19] M. Mandjes. *Large deviations for Gaussian queues: modelling communication networks*. Wiley, 2007.

[20] G.J. Miltenburg. Variance of the number of units produced on a transfer line with buffer inventories during a period of length T. *Naval Research Logistics*, 34:811–822, 1987.

[21] Y. Nazarathy. The Variance of Departure Processes: Puzzling Behavior and Open Problems. *EURANDOM Technical Report Series*, 2009-045, 2009.
[22] Y. Nazarathy and G. Weiss. The asymptotic variance rate of the output process of finite capacity birth-death queues. *Queueing Systems*, 59(2):135–156, 2008.
[23] N.U. Prabhu. *Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Communication*. Springer, 1998.
[24] D. Stoyan and D.J. Daley. *Comparison methods for queues and other stochastic models*. Wiley, 1983.
[25] B. Tan. Asymptotic variance rate of the output in production lines with finite buffers. *Annals of Operations Research*, 93:385–403, 2000.
[26] A. Wald. On cumulative sums of random variables. *The Annals of Mathematical Statistics*, 15(3):283–296, 1944.
[27] W. Whitt. Complements to heavy traffic limit theorems for the GI/G/1 queue. *Journal of Applied Probability*, 9(1):185–191, 1972.
[28] W. Whitt. Comparing counting processes and queues. *Advances in Applied Probability*, 13(1):207–220, 1981.
[29] W. Whitt. The queueing network analyzer. *The Bell Systems Technical Journal*, 62(9):2779–2815, 1983.
[30] W. Whitt. Departures from a queue with many busy servers. *Mathematics of Operations Research*, 9:534–544, 1984.
[31] W. Whitt. Variability Functions for Parametric-decomposition Approximations of Queueing Networks. *Management Science*, 41:1704–1715, 1995.
[32] W. Whitt. *Stochastic Process Limits*. Springer New York, 2002.
[33] R.W. Wolff. An upper bound for multi-channel queues. *Journal of Applied Probability*, pages 884–888, 1977.
[34] A.P. Zwart. Tail asymptotics for the busy period in the GI/G/1 queue. *Mathematics of Operations Research*, 26(3):485–493, 2001.

SCHOOL OF MANAGEMENT AND GOVERNANCE, UNIVERSITY OF TWENTE, ENSCHEDE, THE NETHERLANDS
E-mail address: alhanbali@eurandom.tue.nl

KORTEWEG-DE VRIES INSTITUTE FOR MATHEMATICS, UNIVERSITY OF AMSTERDAM, THE NETHERLANDS; EURANDOM, EINDHOVEN, THE NETHERLANDS; CWI, AMSTERDAM, THE NETHERLANDS
E-mail address: m.r.h.mandjes@uva.nl

EURANDOM, EINDHOVEN, THE NETHERLANDS; EINDHOVEN UNIVERSITY OF TECHNOLOGY, EINDHOVEN, THE NETHERLANDS
E-mail address: y.nazarathy@tue.nl

COLUMBIA UNIVERSITY, NEW YORK NY, UNITED STATES OF AMERICA
E-mail address: ww2040@columbia.edu

Centrum Wiskunde & Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands. The institute's strategy is to concentrate research on four broad, societally relevant themes: earth and life sciences, the data explosion, societal logistics and software as service.

Centrum Wiskunde & Informatica (CWI) is het nationale onderzoeksinstituut op het gebied van wiskunde en informatica. De strategie van het instituut concentreert zich op vier maatschappelijk relevante onderzoeksthema's: aard- en levenswetenschappen, de data-explosie, maatschappelijke logistiek en software als service.

Bezoekadres:
Science Park 123
Amsterdam

Postadres:
Postbus 94079, 1090 GB Amsterdam
Telefoon 020 592 93 33
Fax 020 592 41 99
info@cwi.nl
www.cwi.nl

CWI

Centrum Wiskunde & Informatica