On Approximations for the $GI/GI/1$ Queue and Generalized Jackson Open Queueing Networks Using Indices of Dispersion

Ward Whitt and Wei You

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

Abstract

Recent robust queueing approximations for open queueing networks exploit partial characterizations of each arrival process by its rate and index of dispersion for counts (IDC), which is a scaled version of the variance-time curve. Even though only means and variances are involved, for the $GI/GI/1$ queue and generalized Jackson networks, where the arrival processes are renewal processes, the arrival processes are fully characterized by the rate and IDC. This provides a basis for more accurate approximations than traditional partial characterizations.

Keywords: queues, open queueing networks, index of dispersion for counts, robust queueing, generalized Jackson networks, queueing approximations

1. Introduction

We briefly describe major advantages of new robust queueing (RQ) approximations for single-server queues and open networks of queues in [1] and subsequent papers. The approximations are intended for general $G/GI/1$ queues with general stationary arrival processes that are independent of independent and identically distributed service times, and open networks of such queues.

For the $GI/GI/1$ queue and generalized Jackson open queueing networks, which have mutually independent renewal external arrival processes and sequences of independent and identically distributed (i.i.d.) service times, at each queue and Markovian routing, traditional approximations for the steady-state performance measures depend on each interarrival-time and

service-time distribution through two parameters: its first two moments or, equivalently, its mean and squared coefficient of variation (scv, variance divided by the square of the mean). The simple two-parameter approach is applied in parametric-decomposition approximations such as the Queueing Network Analyzer (QNA) in [2] and the first Robust Queueing Network Analyzer (RQNA) in [3] as well as in approximations in [4], [5] and [6] based on multidimensional reflected Brownian motion stemming from the heavy-traffic limit established in [7].

Some of the key properties can be seen from a single queue. Thus, we will first focus on the $GI/GI/1$ queue with interarrival times U_n and service times V_n distributed as U and V, partially characterized by the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$, where $\lambda^{-1} \equiv E[U]$, $c_a^2 \equiv c_U^2 \equiv Var(U)/E[U]^2$ and $\tau \equiv \mu^{-1} \equiv E[V]$, $c_s^2 \equiv c_V^2 \equiv Var(V)/E[V]^2$, where $\rho \equiv \lambda/\mu < 1$ to ensure stability.

We will focus on the expected steady-state waiting time (for each arrival until starting service) $E[W]$ and workload (remaining work in the system at each time) $E[Z]$ at each queue. These are related by the conservation law $H = \lambda G$ or Brumelle's formula, [8] or (6.20) of [9], which for the $G/GI/1$ model is

$$E[Z] = \lambda \left(E[WV] + \frac{E[V^2]}{2} \right) = \rho E[W] + \rho\tau \frac{(c_s^2 + 1)}{2}. \tag{1}$$

These in turn are related to the mean number in queue and in system by Little's law.

Even though the heavy-traffic limits for the $GI/GI/1$ queue only depend on the model data through the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$, and similarly for generalized Jackson networks [7], the steady-state performance at typical traffic intensities can be quite complicated, as we explain in §2.

To do better for renewal arrival processes and to capture the dependence in more general arrival processes, we use the index of dispersion for counts (IDC) of the stationary arrival process, which is a scaled version of the variance-time function that is independent of the rate. In particular if $A(t)$ is the arrival counting process, assumed to be stationary with rate λ, then as in §4.5 of [10], the IDC is

$$I_a(t) \equiv \frac{Var(A(t))}{E[A(t)]} = \frac{Var(A(t))}{\lambda t}, \quad t \geq 0, \tag{2}$$

where $\equiv$ denotes equality by definition; see §4 of [1].

Our main RQ approximation in [1] is for the expected steady-state workload at each queue. It uses the *index of dispersion for work* (IDW) associated with the cumulative input process Y, defined by

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0, \tag{3}$$

and is defined, as in [11], by

$$I_w(t) \equiv \frac{Var(Y(t))}{E[V_1]E[Y(t)]}, \quad t \geq 0. \tag{4}$$

For the $G/GI/1$ model, where the arrival process is general but independent of an i.i.d. sequence of service times, the IDW is related to the IDC by

$$I_w(t) = I_c(t) + c_s^2, \quad t \geq 0; \tag{5}$$

see §4.3.1 of [1].

Given the IDW, the RQ approximation for the mean workload as a function of the traffic intensity ρ when the mean service time is fixed at $\tau = 1$ appears in (28) in §4.1 of [1], being simply

$$E[Z] \equiv E[Z_\rho] \approx Z_\rho^* \equiv \sup_{x \geq 0} \left\{ -(1-\rho)x/\rho + b_f \sqrt{x I_w(x)} \right\}, \tag{6}$$

where b_f is a parameter to be specified, which we take to be $\sqrt{2}$, which we explain below. (See [12] for additional background on the RQ approximations.)

Strong positive results for the RQ approximation in (6) with $b_f \equiv \sqrt{2}$ for the $G/GI/1$ queue appear in Theorems 2-5 of [1]. Theorem 2 states it is exact for the $M/GI/1$ queue, while Theorem 5 states that it is asymptotically correct in both light and heavy traffic. To state it, we define the normalized or scaled (steady-state) workload by comparing to what it would be in the associated $M/D/1$ model; i.e.,

$$c_Z^2(\rho) \equiv \frac{E[Z_\rho]}{E[Z_\rho; M/D/1]} = \frac{2(1-\rho)E[Z_\rho]}{E[V_1]\rho} = \frac{2(1-\rho)E[Z_\rho]}{\tau\rho}. \tag{7}$$

The normalization in (7) exposes the impact of variability separately from the traffic intensity.

Theorem 1.1. (*heavy-traffic and light-traffic limits from [1]*) *Under the regularity conditions assumed for the IDW $I_w(t)$, if $b_f \equiv \sqrt{2}$, then the RQ approximation in* (6) *is asymptotically correct for the $G/GI/1$ model with $\tau = 1$ both in heavy traffic (as $\rho \uparrow 1$) and light traffic (as $\rho \downarrow 0$). Specifically, we have the following limits:*

$$\lim_{\rho\uparrow 1} c^2_{Z^*}(\rho) = I_w(\infty) = \lim_{\rho\uparrow 1} c^2_Z(\rho) \quad \textit{and} \quad \lim_{\rho\downarrow 0} c^2_{Z^*}(\rho) = I_w(0) = \lim_{\rho\downarrow 0} c^2_Z(\rho). \tag{8}$$

We have developed this approximation method to treat general stationary arrival processes, which have complex dependence over time. However, this approach is important even for the basic $GI/GI/1$ queue as we will explain here. First in §2 we explain why we want to go beyond the traditional two-parameter characterizations of arrival processes, even for the $GI/GI/1$ model. Then in §3 we show that the rate and IDC fully characterize a renewal process, so that the rate and IDC of the arrival and service processes fully characterize the $GI/GI/1$ model. We discuss the special case of the $GI/M/1$ model in §4. Finally, we elaborate in §5 with simulation examples for queues in series, focusing on the heavy-traffic bottleneck phenomenon in [13], which has been discussed in [6].

2. Why We Do Want More Information about the Arrival Process?

For the $GI/GI/1$ queue partially specified by the vector $(\lambda, c_a^2, \tau, c_s^2)$, a commonly used approximation for the steady-state waiting time is

$$E[W] \approx \frac{\tau\rho(c_a^2 + c_s^2)}{2(1-\rho)}, \tag{9}$$

because it is exact (being the classical Pollaczek-Khintchine formula) for the $M/GI/1$ special case, when the interarrival time has an exponential distribution, in which case $c_a^2 = 1$.

We call (9) the heavy-traffic approximation because, under regularity conditions, it is asymptotically correct in that limit:

$$\frac{2(1-\rho)E[W(\rho)]}{\tau\rho} \to c_a^2 + c_s^2 \quad \text{as} \quad \rho \to 1. \tag{10}$$

In fact, the heavy-traffic limit does much more, showing that the scaled waiting-time distribution is asymptotically exponential and thus is asymptotically fully characterized by its mean. Hence, if we approximate the mean we also can approximate the full distribution.

A conservative approach to the mean waiting time $E[W]$ is to look for upper bounds. The best known upper bound (UB) on $E[W]$, given the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$, is the Kingman[14] UB,

$$E[W] \le \frac{\tau\rho([c_a^2/\rho^2] + c_s^2)}{2(1-\rho)}, \tag{11}$$

which is also asymptotically correct in heavy traffic, just like (9).

A better UB depending on these same parameters was obtained by Daley [15]. in particular, the Daley UB replaces the term c_a^2/ρ^2 by $(2-\rho)c_a^2/\rho$, i.e.,

$$E[W] \le \frac{\tau\rho([(2-\rho)c_a^2/\rho] + c_s^2)}{2(1-\rho)}. \tag{12}$$

Note that $(2-\rho)/\rho < 1/\rho^2$ because $\rho(2-\rho) < 1$ for all ρ, $0 < \rho < 1$, as can be seen by noting that $1 - 2\rho - \rho^2 = (1-\rho)^2 > 0$ for all ρ, $0 \le \rho < 1$.

But even the Daley bound in (12) is not tight; see [16] for a recent study of these bounds for $E[W]$, including a numerical algorithm to compute the tight UB and an approximation formula. A major concern is the range of possible values given the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$. For that purpose, these new results can be combined with the lower bound (LB), which has long been known. The explicit formula for the LB is

$$E[W(LB)] = \frac{\tau\rho((1+c_s^2)\rho - 1)^+}{2(1-\rho)}, \tag{13}$$

where $x^+ \equiv \max\{x, 0\}$.

Tables 1 and 2 plus Tables EC.1 and EC.2 in [16] give a numerical overview of the upper and lower bounds for $E[W]$, given the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$, in the four cases for which c_a^2 and c_s^2 assume all combinations of the two values 0.5 (less variable than exponential) and 4.0 (more variable than exponential). We illustrate by reproducing a portion of Table 1 of [16] here in Table 1. Paralleling (7), to focus on the impact of the variability independent of the traffic intensity ρ, so in Table 1 we display values for the normalized or *scaled mean waiting time*

$$c_W^2(\rho) \equiv \frac{2(1-\rho)E[W(\rho)]}{\rho\tau}, \tag{14}$$

which shows the total variability in the arrival and service processes, and assumes the constant value $c_a^2 + c_s^2$ for the heavy-traffic approximation in (9).

Table 1: A comparison of the scaled bounds and approximations for $E[W]$, i.e., for $c_W^2(\rho) \equiv 2(1-\rho)E[W(\rho)]/\rho\tau$ in (14), given the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$ as a function of ρ for the case $c_a^2 = c_s^2 = 4.0$. (Below HTA is the scaled version of (9), while the other entries are scaled versions of the tight bounds, the Daley UB in (12) and the Kingman UB in (11).)

ρ	Tight LB	HTA	Tight UB	Daley	Kingman
0.10	0.0	8.0	76.0	80.0	404.0
0.30	0.0	8.0	23.4	26.6	48.4
0.50	1.0	8.0	13.8	16.0	20.0
0.70	3.0	8.0	10.4	11.4	12.2
0.90	3.8	8.0	8.6	8.8	9.0
0.99	4.0	8.0	8.0	8.0	8.0

The first conclusion from Table 1 is that the basic approximation in (9) is consistent with the possible values for all ρ.

A second conclusion is that the quality of approximations depends on the traffic intensity. Consistent with [17], it is not possible to obtain reliable approximations for $E[W]$ in light traffic based only on $(\lambda, c_a^2, \tau, c_s^2)$. In contrast, the heavy-traffic approximation in (9) and all the upper bounds are asymptotically equivalent in the heavy-traffic limit for the scaled mean waiting time $2(1-\rho)E[W(\rho)]/\rho\tau$ in heavy-traffic. Even in heavy traffic, this is an iterated limit; i.e., we first fix the interarrival-time distribution with given c_a^2 and then we let $\rho \uparrow 1$.

We illustrate by giving a typical example.

Example 2.1. (*an $H_2/M/1$ queue example*) Suppose that we consider the $H_2/M/1$ queue, where the interarrival-time distribution is H_2 (hyperexponential) with $c_a^2 = 2$, where the third parameter is specified by assuming balanced means as in (37) on p. 137 of [18]. Then, the heavy-traffic for the scaled waiting time $c_W^2 \equiv 2(1-\rho)E[W(\rho)]/\rho\tau$ is $c_a^2 + c_s^2 = 2 + 1 = 3.0$. For the actual model, for traffic intensities $\rho = 0.3$, $\rho = 0.6$ and $\rho = 0.9$, we obtain the values $c_W^2 = 2.82, 2.96$ and 2.99, respectively, from Table I on p. 170 of [19], exploiting Little's law. These cases are well approximated by the HTA 3.0, so that it performs quite well.

However, the third conclusion from Table 1 is that the lower bound is surprisingly low, even in heavy traffic, so that the range of possible values consistent with the parameters is surprisingly wide. That occurs because the LB is attained asymptotically at the associated $D/GI/1$ queue with $c_a^2 = 0$. That D distribution is approached by a distribution that has a very small

mass at a very large value, and the rest of the mass just less than the mean. That allows very small values of $E[W]$.

The heavy-traffic limit of the lower bound is the heavy-traffic limit for the associated $D/GI/1$ model, which in Table 1 would be 4.0. The values in Table 1 are obtained by, first, fixing ρ and then letting the interarrival-time distribution approach D. Table 1 implies that, for any given ρ, the least possible value is attained in the corresponding $D/GI/1$ model. This is why the distance between the LB and the UB remains large for all ρ.

While we might regard the LB as something of an anomaly, these numerical results clearly indicate that the mean waiting time is not adequately approximated by the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$. Moreover, the difficulty is primarily caused by the arrival process. For example, for $M/GI/1$, the mean $E[W]$ is fully determined by (9), but for $GI/M/1$ there is a wide range.

The next question is: What do we gain by replacing the variability parameter c_a^2 of the renewal arrival process by its IDC? It may seem that we should gain little, because even though it is a function instead of a single number, it is still just a variance, but as we show next, that is not the case.

3. Full Characterizations of a Renewal Processes

In this section we observe that the rate and the IDC provide a full characterization of a renewal process and the $GI/GI/1$ queue.

For understanding and appreciating this conclusion, it is important to distinguish between the ordinary renewal process and the equilibrium renewal process, as discussed in [20] and §3.4 and §3.5 of [21]. These two alternative versions of a renewal process constitute a special case of the discrete-time stationary point process and the associated continuous-time stationary point process, linked by the Palm transformation, as discussed in [22]. This is important for the IDC because the IDC is defined in terms of the continuous-time stationary version, and is independent of the rate.

We start with a rate-λ renewal process $N \equiv \{N(t) : t \geq 0\}$. Let F be the *cumulative distribution function* (cdf) of the interval U between points (the interarrival time in a GI arrival process), having mean $E[U] = \lambda^{-1}$ and finite second moment. As a regularity condition, we also assume that F has a *probability density function* (pdf) f, where $F(t) = \int_0^t f(u)\, du$, $t \geq 0$. The pdf assumption ensures that the equilibrium renewal process arises as the time limit of the ordinary renewal process; e.g., see §3.4 and §3.5 of [21].

The stationary or equilibrium renewal process differs from the ordinary renewal process only by the distribution of the first interarrival time. Let F_e be the cdf of the equilibrium distribution, which has pdf $f_e(t) = \lambda(1 - F(t))$. Note that we can construct F_e given F, but we cannot construct F given F_e, but the pair (λ, F_e) fully characterizes F; we can construct F via

$$F(t) = 1 - \lambda^{-1} f_e(t), \quad t \geq 0. \tag{15}$$

Let $E^e[\cdot]$ denote the expectation under the stationary distribution (with first interval distributed according to F_e) and let $E^0[\cdot]$ denote the expectation under the Palm distribution (with first interval distributed as F).

Conditioning on the first arrival, distributed as F under the Palm distribution or as F_e under stationary distribution, the renewal equations for the mean and second moment of $N(t)$, the number of points in an interval $[0, t]$, are:

$$\begin{aligned}
m(t) &\equiv E^0[N(t)] = F(t) + \int_0^t m(t-s)dF(s), \\
m_e(t) &\equiv E^e[N(t)] = F_e(t) + \int_0^t m(t-s)dF_e(s), \\
\sigma(t) &\equiv E^0[N^2(t)] = F(t) + 2\int_0^t m(t-s)dF(s) + \int_0^t \sigma(t-s)dF(x), \\
\sigma_e(t) &\equiv E^e[N^2(t)] = F_e(t) + 2\int_0^t m(t-s)dF_e(s) + \int_0^t \sigma(t-s)dF_e(x).
\end{aligned}$$

The function $m(t)$ is the familiar renewal function. To express the relations among these quantities, we use the Laplace Transform (LT) instead of the Laplace-Stieltjes Transform (LST). Let the LT of a pdf $f(t)$ and the LST of F be denoted by $\mathcal{L}(f)(s) \equiv \hat{f}(s)$ and defined by

$$\hat{f}(s) \equiv \mathcal{L}(f)(s) \equiv \int_0^\infty e^{-st} f(t)dt = \int_0^\infty e^{-st} dF(t), \tag{16}$$

so that $f(t) = \mathcal{L}^{-1}(\hat{f})(t)$.

Let a subscript e denote a quantity associated with the equilibrium renewal process. Then the LT of f_e is

$$\hat{f}_e(s) = \frac{\lambda(1 - \hat{f}(s))}{s} \text{ and } \hat{F}_e(s) = \frac{\hat{f}_e(s)}{s},$$

where $\lambda^{-1} \equiv \int_0^\infty t f(t)\, dt$ is the mean.

Applying the LT to the renewal equations, we obtain

$$\hat{m}(s) = \frac{\hat{f}(s)}{s(1-\hat{f}(s))}, \tag{17}$$

$$\hat{m}_e(s) = \frac{\hat{f}_e(s)}{s(1-\hat{f}(s))} = \frac{\lambda}{s^2}, \tag{18}$$

$$\hat{\sigma}(s) = \frac{\hat{f}(s) + 2s\hat{m}(s)\hat{f}(s)}{s(1-\hat{f}(s))} = \frac{\hat{f}(s)(1+\hat{f}(s))}{s(1-\hat{f}(s))^2}, \tag{19}$$

$$\hat{\sigma}_e(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) = \frac{\lambda(1+\hat{f}(s))}{s^2(1-\hat{f}(s))}. \tag{20}$$

From (18), we see that $E^e[N(t)] = \lambda$, $t \geq 0$, as must be true for any stationary point process. We are now ready to state the basic characterization theorem, which is not new, e.g., [20] and [10], but deserves to be better known.

Theorem 3.1. (*renewal process characterization theorem*) *A renewal process with an inter-renewal distribution having pdf f and cdf F having finite first two moments with positive mean λ^{-1} is fully characterized by any one of the following:*

1. *the pdf $f(t)$ of the time between renewals;*
2. *the cdf $F(t)$ of the time between renewals;*
3. *the LT $\hat{f}(s)$;*
4. *the renewal function $m(t)$;*
5. *the LT $\hat{m}(s)$;*
6. *the rate λ and the variance function of the equilibrium renewal process $\sigma_e(t)$;*
7. *the rate λ and the LT $\hat{\sigma}_e(s)$;*
8. *the rate λ and the IDC $I_e(t) \equiv \sigma_e(t)/\lambda(t)$ of the equilibrium renewal process.*

Proof. The equivalence of the time functions and their transforms follows from the basic theory of Laplace transforms. Hence, we obtain the equivalence by explicit expressions in terms of the Laplace transforms. From (17)

and (20), we obtain

$$\begin{aligned} \hat{f}(s) &= \frac{s\hat{m}(s)}{1+s\hat{m}(s)} \quad \text{and} \\ \hat{f}(s) &= \frac{s^2\hat{\sigma}_e(s)-\lambda}{s^2\hat{\sigma}_e(s)+\lambda}. \end{aligned} \tag{21}$$

Then, from the definition of the IDC, we obtain

$$\sigma_e(t) = \lambda(t) I_e(t), \quad t \geq 0. \quad \blacksquare$$

Corollary 3.1. (*full characterization of a $GI/GI/1$ queue*) *The $GI/GI/1$ queue with interarrival-time cdf F and service-time cdf G having finite second moments is fully characterized by the four-tuple $(\lambda, I_a(t), \tau, I_s(t))$, where $I_a(t)$ ($I_s(t)$) is the IDC of the equilibrium renewal process associated with the interarrival (service) times.*

Corollary 3.1 is exploited strongly in [23] in the development of approximations for queues in series. Of course, the rate of a departure process is just the rate of the arrival process. The key step in the approximation is developing an approximation for the IDC of a stationary departure process via a convex combination of the IDC's $I_a(t)$ and $I_s(t)$. By that approach, we obtain a good characterization of each queue in the series model. The final formula is an approximation, exploiting heavy-traffic limits, but Corollary 3.1 implies that the true formula must be a function of the four-tuple $(\lambda, I_a(t), \tau, I_s(t))$.

4. The Case of an $H_2/M/1$ Queue

For the $GI/M/1$ queue with interarrival-time pdf f, the steady-state performance depends on a single root of a transform equation. In particular,

$$E[W] = \frac{\tau\sigma}{1-\sigma} \quad \text{and} \quad E[Z] = \rho\tau\left(\frac{\sigma}{1-\sigma} + \frac{c_s^2+1}{2}\right), \tag{22}$$

where σ is the unique root in $(0,1)$ of the equation

$$\hat{f}(\mu(1-\sigma)) = \sigma. \tag{23}$$

Consider an $H_2/M/1$ queue, which is a $GI/GI/1$ queue with an exponential service distribution and a hyperexponential (H_2) interarrrival-time distribution, i.e., a mixture of two exponential distributions, which has pdf

$$f(t) \equiv p\lambda_1 e^{-\lambda_1 t} + (1-p)\lambda_2 e^{-\lambda_2 t}, \quad t \geq 0, \tag{24}$$

and thus the parameter triple $(p, \lambda_1, \lambda_2)$. Equivalently, it has as parameters its first three moments or the mean λ^{-1}, scv c_a^2 and the ratio between the two components of the mean $r \equiv p_1/\lambda_1/(p_1/\lambda_1 + p_2/\lambda_2)$ where $\lambda_1 > \lambda_2$. The third parameter is often specified by stipulating balanced means,i.e., $r = 0.5$, as in (37) on p. 137 on [18]. The behavior as a function of the third parameter has been studied in [19]. As far as the congestion in the queue is concerned, the H_2 arrival process can be as smooth as a Poisson process with the same rate or as bursty as a batch Poisson process with the same scv; see (9) in §IV of [19]. The consequence is illustrated for $c_a^2 = 2$ and $c_a^2 = 12$ in Tables I and II on p. 170 of [19].

From p. 50 of [20], the IDC is

$$I_a(t) = c_a^2 - \frac{2\beta}{\gamma t}(1 - e^{-\gamma t}), \quad t > 0, \tag{25}$$

where

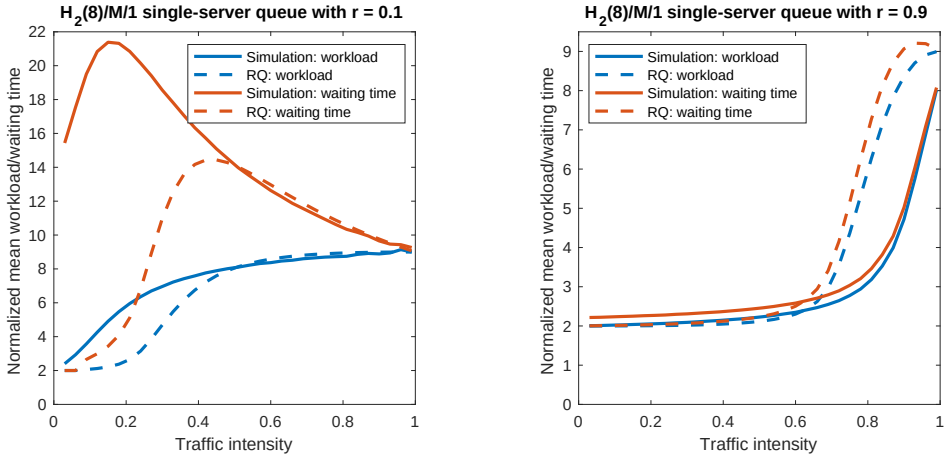
$$\beta \equiv \frac{c_a^2 - 1}{2} \quad \text{and} \quad \gamma \equiv 2\beta((m_3/3) - (c_s^2 + 1)^2/2) \tag{26}$$

with m_3 being the third moment, which is increasing in r. Note that $I_a(t)$ in (25) in strictly increasing in t from 1 at $t = 0$ to c_s^2 at $t = \infty$.

By virtue of Theorem 3.1, the set of H_2 arrival processes for given rate and scv c_a^2 are in fact fully characterized by their rate and IDC. The range of possible behavior of these IDC's can be seen from the IDC's of the two extremal H_2 arrivval processes. The IDC of the lower bound is identically 1, while the IDC of the upper bound is identically c_s^2. All other IDC's increase from 1 at 0 to c_s^2 at infinity.

Figure 1 compares the RQ approximation for the workload via (6) and then the waiting time via (1) with simulation estimates for the $H_2/M/1$ model with $c_a^2 = 8$ and $r = 0.1$ and $r = 0.9$. The IDW and normalized mean workload increase from 2 to 9. We see that the performance and the approximations remain closer to the $M/M/1$ lower bound with $r = 1$ for $r = 0.9$ than $r = 0.1$. Note that Figure 1 is consistent with Theorem 1.1.

Figure 1: A comparison of the robust queueing approximation for the mean workload and waiting time in (6) and (1) for the $H_2/M/1$ model with $c_a^2 = 8$ and two values of r to simulation estimates.



5. Simulation Comparisons for a Series of Queues

We now consider simulation experiments for a series of single-server queues. We consider the heavy-traffic bottleneck examples from [13], which are for a non-Poisson arrival process feeding a series of 9 single-server queues with $\rho_i = 0.6$ for $1 \le i \le 8$, and $\rho_9 = 0.9$, so that the last queue is a bottleneck queue. In particular, we compare the new RQNA algorithm in [23] and RQ, the algorithm in (6) from [1] with the exact estimated IDC, to the performance of QNA from [2], QNET from [4] and SBD from [6]. The QNET method uses the multi-dimensional reflected Brownian motion resulting from the heavy-traffic limit in [7]. The RQNA algorithm uses (74) and (75) of [23]. Each departure IDC is a convex combination of the arrival and service IDC's, i.e.,

$$I_{d,\rho}(t) \approx w_\rho(t) I_a(t) + (1 - w_\rho(t)) I_s(t), \quad t \ge 0, \tag{27}$$

where the ρ-dependent weight is

$$w_\rho(t) \equiv w^*((1-\rho)^2 \lambda t)/\rho c_s^2) \quad \text{for} \quad w^*(t) \equiv 1 - (1 - c^*(t))/2t, \tag{28}$$

where $c^*(t)$ is the correlation function of the stationary version of canonical (drift - 1, variance 1) RBM; see (24)-(27) of [23].

Table 2 compares five approximation methods to simulation for 9 exponential (M) queues in series fed by a highly-variable rate-1 H_2 renewal arrival

process with $c_a^2 = 8$ and the usual balanced means. Table 2 compares the various approximations of the mean steady-state waiting time at each station, as well as the total waiting time in the system, to simulation estimates. The simulation estimates for $E[W_i]$ and the IDC are based on a C^{++} program and run of length 5×10^7, discarding the initial 10^5 customers to approach steady state. The half width of the confidence interval at the final bottleneck queue was about 0.2% in every case. The QNA approximation appears in [13]; the heavy-traffic approximations QNET from [4] and SBD appear in [6].

Table 2: A comparison of five approximation methods to simulation for the expected waiting time $E[W]$ at each of 9 exponential (M) queues in series with $\rho_i = 0.6$, $1 \leq i \leq 8$, and $\rho_9 = 0.9$ fed by a highly-variable rate-1 H_2 renewal arrival process with $c_a^2 = 8$ and $r = 0.5$, i.e., the usual balanced means.

node	Sim	QNA	QNET	SBD	RQNA	RQ
1	3.36	4.05	4.05	4.05	3.95	3.95
2	2.32	2.92	1.81	1.82	1.58	2.61
3	1.96	2.19	1.47	1.49	0.98	2.04
4	1.77	1.73	1.16	1.19	0.92	1.72
5	1.64	1.43	1.07	1.10	0.90	1.53
6	1.56	1.24	1.03	1.06	0.90	1.41
7	1.49	1.12	1.00	1.03	0.90	1.32
8	1.44	1.04	0.98	1.01	0.90	1.27
9	29.2	8.9	6.0	36.4	29.1	37.1
sum	45.3	24.6	18.6	49.8	40.1	52.9

Table 2 shows poor performance of QNA [2] at the last bottleneck queue originally exposed in [13]. It also shows the significant improvement provided by the sequential bottleneck approximation (SBD) reported in [6], which is largely matched by RQNA and RQ.

To illustrate the impact of additional information about the arrival process, Table 3 is the analog of Table 2 in which we use three alternative rate-1 H_2 arrival processes, all with $c_a^2 = 8.0$ but different r, in particular for $r = 1.0$, 0.9 and 0.1. The case $r = 1$ is the lower-bound H_2 renewal arrival process with the same mean and $c_s^2 = 8$, which is the Poisson process, for which both RQNA and RQ are exact. For $r = 0.1$, the arrival process is close to a batch Poisson process. For these cases, the QNA, QNET and SBD approximations are the same as in Table 2.

Table 3: A comparison of RQNA and RQ to simulation for the expected waiting time at each queue for the same model in Table 2 except except the third parameter of the H_2 renewal arrival process with $c_a^2 = 8$ is changed from $r = 0.5$ to $r = 1.0$, 0.9 and 0.1.

r	1.0	0.9			0.1		
node	exact	Sim	RQNA	RQ	Sim	RQNA	RQ
1	0.90	1.16	1.13	1.13	5.69	5.84	5.83
2	0.90	1.16	0.95	1.12	2.46	2.71	2.40
3	0.90	1.15	0.91	1.11	1.98	1.28	1.83
4	0.90	1.14	0.90	1.10	1.76	0.97	1.56
5	0.90	1.14	0.90	1.10	1.63	0.91	1.41
6	0.90	1.13	0.90	1.09	1.54	0.90	1.31
7	0.90	1.13	0.90	1.08	1.48	0.90	1.24
8	0.90	1.12	0.90	1.08	1.42	0.90	1.20
9	8.10	19.6	27.2	36.5	29.6	29.3	36.3
sum	15.3	28.8	33.8	45.3	47.5	43.7	53.1

From Tables 2 and 3, we see that the mean waiting time increases as r decreases. We also see that both RQNA and RQ are very accurate at the first $H_2/M/1$ queue, where the arrival process is a renewal process, but are far less reliable at later queues, which have non-renewal arrival processes. For queues 3-8, RQNA seriously underestimates $E[W_i]$; since RQ does not, we include the difficulty lies in the IDC approximation in (27) under lighter loads. Consistent with the heavy-traffic limit in [23], RQNA performs well at the final bottleneck queue, although RQ does not, which is partly explained by the relevant times are those where the IDC experineces most of its increase. RQNA also does reasonably well predicting the sum of the waiting times.

The RQNA from [3] seems to perform far worse, as shown in Tables 1 and 2 of [12], but it provides tuning parameters that can yield significant improvement given additional information. In [12] we show that the specific version of RQNA from §7.2 of [3] corresponds to the asymptotic method from [18] for all the arrival processes and the Kingman upper bound from [14] at each queue.

Finally, we observe that the cases $r = 0.9$ and $r = 0.1$ in Table 3 provide a rough estimate of the range of reasonable approximation values for unspecified third parameter.

Acknowledgement. The authors thank NSF for research support (grant CMMI 1634133).

References

[1] W. Whitt, W. You, Using robust queueing to expose the impact of dependence in single-server queues, Operations Research 66 (1) (2018) 184–199.

[2] W. Whitt, The queueing network analyzer, Bell Laboratories Technical Journal 62 (9) (1983) 2779–2815.

[3] C. Bandi, D. Bertsimas, N. Youssef, Robust queueing theory, Operations Research 63 (3) (2015) 676–700.

[4] J. M. Harrison, V. Nguyen, The QNET method for two-moment analysis of open queueing networks, Queueing Systems 6 (1) (1990) 1–32.

[5] M. I. Reiman, Asymptotically exact decomposition approximations for open queueing networks, Operations research letters 9 (6) (1990) 363–370.

[6] J. Dai, V. Nguyen, M. I. Reiman, Sequential bottleneck decomposition: an approximation method for generalized Jackson networks, Operations Research 42 (1) (1994) 119–136.

[7] M. I. Reiman, Open queueing networks in heavy traffic, Math. Oper. Res. 9 (3) (1984) 441–458.

[8] S. Brumelle, On the relation between customer averages and time averages in queues, J. Appl. Prob. 8 (3) (1971) 508–520.

[9] W. Whitt, A review of $L = \lambda W$, Queueing Systems 9 (1991) 235–268.

[10] D. R. Cox, P. A. W. Lewis, The Statistical Analysis of Series of Events, Methuen, London, 1966.

[11] K. W. Fendick, W. Whitt, Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue, Proceedings of the IEEE 71 (1) (1989) 171–194.

[12] W. Whitt, W. You, Supplement to "On approximations for the $GI/GI/1$ queue and generalized Jackson open queueing networks using indices of dispersion", Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html (2018).

[13] S. Suresh, W. Whitt, The heavy-traffic bottleneck phenomenon in open queueing networks, Operations Research Letters 9 (6) (1990) 355–362.

[14] J. F. C. Kingman, Inequalities for the queue $GI/G/1$, Biometrika 49 (3/4) (1962) 315–324.

[15] D. J. Daley, Inequalities for moments of tails of random variables, with queueing applications, Zeitschrift fur Wahrscheinlichkeitsetheorie Verw. Gebiete 41 (1977) 139–143.

[16] Y. Chen, W. Whitt, Extremal $GI/GI/1$ queues given two moments, working paper, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html (2018).

[17] D. Daley, T. Rolski, A light-traffic approximation for a single-server queue, Mathematics of Operations Research 9 (4) (1984) 624–628.

[18] W. Whitt, Approximating a point process by a renewal process: two basic methods, Oper. Res. 30 (1982) 125–147.

[19] W. Whitt, On approximations for queues, III: Mixtures of exponential distribtuoons, AT&T Bell Laboratories Technical Journal 63 (1) (1984) 163–175.

[20] D. R. Cox, Renewal Theory, Methuen, London, 1962.

[21] S. M. Ross, Stochastic Processes, 2nd Edition, Wiley, New York, 1996.

[22] K. Sigman, Stationary Marked Point Processes: An Intuitive Approach, Chapman and Hall/CRC, New York, 1995.

[23] W. Whitt, W. You, Heavy-traffic limit of the $GI/GI/1$ stationary departure process and its variance function, Operations Research Letters 8 (2) (2018) 143–165.