Supplement on Robust Queueing Approximations for the $GI/GI/1$ Queue and Series of These Queues

Ward Whitt and Wei You

Industrial Engineering and Operations Research
Columbia University, New York, NY, 10027; {ww2040,wy2225}@columbia.edu

October 13, 2018

Abstract

This is a supplement to our recent paper, "The Advantage of Indices of Dispersion in Queueing Approximations," in Whitt and You (2018c). We review robust queueing approximations from Bandi et al. (2015) and Whitt and You (2018b,a) for the mean steady-state waiting time at a $GI/GI/1$ queue and at each queue in a $GI/GI/1 \to GI/1 \to \ldots \to GI/1$ series (tandem) open queueing network. We review the asymptotic method and the stationary-interval method for approximately characterizing the variability of a point process, and their application to departure processes. Then we discuss alternative versions of the robust queueing approximations. We observe that the robust queueing network analyzer from Bandi et al. (2015) employs a variant of the asymptotic method to approximately characterize the variability of the total arrival process to each queue. Moreover, the version discussed in §7.2 of Bandi et al. (2015) coincides with the asymptotic method for each queue together with the Kingman upper bound from Kingman (1962) at that queue.

Keywords: queueing network analyzer, heavy-traffic bottleneck phenomenon, robust queueing, heavy traffic

1 Introduction

This is a supplement to our recent paper, "The Advantage of Indices of Dispersion in Queueing Approximations," in Whitt and You (2018c). In this supplement we review recent robust queueing (RQ) approximations from Bandi et al. (2015) and Whitt and You (2018b,a) for the mean steady-state waiting time $E[W]$ in both a single $GI/GI/1$ queue and at each queue in a series network of such queues, i.e., in a $GI/GI/1 \to GI/1 \to \ldots \to GI/1$ series (or tandem) open queueing network (OQN). Table 2 here provides a detailed comparison of alternative methods for the series-queue example in Tables 2 and 3 of Whitt and You (2018c). Thus we expose weaknesses of alternative two-parameter methods.

We first give background on one $GI/GI/1$ queue in §2. Then we give background on the basic robust queueing (RQ) approximations for the mean waiting time $E[W]$ and the mean workload $E[Z]$ at one queue. Afterwards, we discuss the RQ approximations for a series OQN.

2 Background on One $GI/GI/1$ Queue

In this section we provide background on the classical $GI/GI/1$ queue. The $GI/GI/1$ model has a sequence of independent and identically distributed (i.i.d.) interarrival times $\{U_n : n \geq 1\}$ each distributed as U, which is independent of a sequence of i.i.d. service times $\{V_n : n \geq 1\}$, each distributed as V. Let an interarrival time U have mean $E[U] \equiv \lambda^{-1}$ and squared coefficient of variation (scv, variance divided by the square of the mean) c_a^2; let a service time V have mean $E[V] \equiv \tau$ and scv c_s^2. Assume that the second moments, and thus the scv's c_a^2 and c_s^2, are finite, which implies that the means are finite as well. Assume that $\rho \equiv \lambda\tau < 1$, so that the model is stable.

We use the representation of the waiting time (before receiving service) in a general single-server queue with unlimited waiting space and the first-come first-served (FCFS) service discipline, without imposing any stochastic assumptions. The waiting time of arrival n satisfies the Lindley (1952) recursion

$$W_n = (W_{n-1} + V_{n-1} - U_{n-1})^+ \equiv \max\{W_{n-1} + V_{n-1} - U_{n-1}, 0\}, \tag{2.1}$$

where V_{n-1} is the service time of arrival $n-1$, U_{n-1} is the interarrival time between arrivals $n-1$ and n, and $\equiv$ denotes equality by definition. If we initialize the system by having an arrival 0 finding an empty system, then W_n can be represented as the maximum of a sequence of partial

sums, using the Loynes (1962) reverse-time construction; i.e.,

$$W_n = M_n \equiv \max_{0 \le k \le n} \{S_k\}, \quad n \ge 1, \tag{2.2}$$

using reverse-time indexing with $S_k \equiv X_1 + \cdots + X_k$ and $X_k \equiv V_{n-k} - U_{n-k}$, $1 \le k \le n$ and $S_0 \equiv 0$. (Bandi et al. (2015) actually look at the system time, which is the sum of an arrival's waiting time and service time. These representations are essentially equivalent.)

If we extend the reverse-time construction indefinitely into the past from a fixed present state, then $W_n \uparrow W \equiv \sup_{k \ge 0} \{S_k\}$ with probability 1 as $n \to \infty$, allowing for the possibility that W might be infinite. For the stable stationary $G/G/1$ stochastic model with $E[U_k] < \infty$, $E[V_k] < \infty$ and $\rho \equiv E[V_k]/E[U_k] < 1$, $P(W < \infty) = 1$; e.g., see Loynes (1962) or §6.2 of Sigman (1995).

Let W be the steady-state waiting time. It is known that W is a proper random variable with a finite mean. Specifically, the mean can be expressed as

$$E[W] = \sum_{k=1}^{\infty} \frac{E[S_k^+]}{k} < \infty, \tag{2.3}$$

where $x^+ \equiv \max\{x, 0\}$, $S_k \equiv X_1 + \cdots + X_k$ and $X_k \equiv V_k - U_k$, $k \ge 1$; e.g., see §§X.1-X.2 of Asmussen (2003) or (13) in §8.5 of Chung (2001). It should be evident that it is not straightforward to use (2.3) to calculate numerical values.

3 Robust Queueing Approximations for One Queue

Robust queueing (RQ) approximations are primarily intended for more complex queueing systems, but as a special case they apply to the general stationary $G/G/1$ queue (where W is assumed to be well defined) and the $GI/GI/1$ special case. In particular, RQ approximations for $E[W]$ appear in Bandi et al. (2015) and Whitt and You (2018b).

3.1 The Basic RQ Approximations for $E[W]$ in the $GI/GI/1$ Model

3.1.1 The Original Approach in Bandi et al. (2015)

Bandi et al. (2015) proposed an RQ approximation for the steady-state waiting time W by performing a deterministic optimization in (2.2) subject to deterministic constraints, where we can ignore the time reversal. Treating the partial sums S_k^a of the interarrival times U_k and the partial sums S_k^s of the service times V_k separately leads to the two uncertainty sets (for W)

$$\mathcal{U}^a \equiv \{\tilde{U} \in \mathbb{R}^\infty : S_k^a \ge k m_a - b_a \sqrt{k},\ k \ge 0\} \quad \text{and}$$

$$\mathcal{U}^s \equiv \{\tilde{V} \in \mathbb{R}^\infty : S_k^s \le km_s + b_s\sqrt{k},\ k \ge 0\}, \tag{3.1}$$

where $\tilde{U} \equiv \{U_k : k \ge 1\}$ and $\tilde{V} \equiv \{V_k : k \ge 1\}$ are arbitrary sequences of real numbers in $\mathbb{R}^\infty$, $S_k^a \equiv U_1 + \cdots + U_k$ and $S_k^s \equiv V_1 + \cdots + V_k$, $k \ge 1$, $S_0 \equiv 0$, and m_a, m_s, b_a and b_s are parameters to be specified. The constraints in (3.1) are one sided because that is what is required to bound the waiting times above, as we can see from (2.1) and (2.2). Thus, the RQ optimization can be expressed as

$$W^* \equiv \sup_{\tilde{U}\in\mathcal{U}^a} \sup_{\tilde{V}\in\mathcal{U}^s} \sup_{k\ge 0} \{S_k^s - S_k^a\}. \tag{3.2}$$

where S_k^a (S_k^s) is a function of $\tilde{U}$ ($\tilde{V}$) specified above.

Bandi et al. (2015) also provided an extension to cover the heavy-tailed case, where finite variances might not exist; then $\sqrt{k}$ in (3.1) is replaced by $k^{1/\alpha}$ for $0 < \alpha \le 2$, as we would expect from §§4.5, 8.5 and 9.7 of Whitt (2002), but we will not discuss that extension here.

3.1.2 The Single-Uncertainty-Set Version in Whitt and You (2018b)

From (2.1), it is evident that the waiting times depend on the service times and interarrival times only through their difference X_n. Thus, instead of the two uncertainty sets in (3.1), we can cnsider the single uncertainty set (for each n)

$$\mathcal{U}^x \equiv \{\tilde{X} \in \mathbb{R}^\infty : S_k^x \le -mk + b_x\sqrt{k},\ k \ge 0\}, \tag{3.3}$$

where $\tilde{X} \equiv \{X_k : k \ge 1\} \in \mathbb{R}^\infty$, $S_k^x \equiv X_1 + \cdots + X_k$, $k \ge 1$ and $S_0 \equiv 0$, while m and b_x are constant parameters to be specified. To avoid excessively strong constraints for small values of k, not justified by the CLT, we could replace k in the constraint bounds on the right in (3.3) by $\max\{k, k_L\}$, but that lower bound k_L has no impact if chosen appropriately. Combining (2.2) and (3.3), we obtain the alternative RQ optimization

$$W^* \equiv \sup_{\tilde{X}\in\mathcal{U}^x} \sup_{k\ge 0} \{S_k^x\}. \tag{3.4}$$

where S_k^x is the function of $\tilde{X}$ specified above.

We comment that the two uncertainty sets instead of only one evidently were introduced in Bandi et al. (2015) to facilitate RQ approximations for open networks of queues (which we discuss later).

The following is Theorem 1 of Whitt and You (2018b), which is a variant of Theorem 2 of Bandi et al. (2015) to include the new RQ formulation in (3.4). The final statement involves an interchange of suprema, e.g., see Lemma EC.1 of Whitt and You (2018b).

Theorem 3.1 (*RQ solutions for the steady-state waiting time*) *The RQ optimizations* (3.2) *with* $m_a > m_s > 0$ *and* (3.4) *with* $m > 0$ *have the solution*

$$\begin{aligned} W^* &= \sup_{k\geq 0}\{-mk + b\sqrt{k}\} \\ &\leq \sup_{x\geq 0}\{-mx + b\sqrt{x}\} = -mx^* + b\sqrt{x^*} = \frac{b^2}{4m} \\ &\quad for \quad x^* = \frac{b^2}{4m^2}, \end{aligned} \tag{3.5}$$

where $m = m_a - m_s > 0$. *For* (3.2), $b \equiv b_s + b_a$; *for* (3.4), $b \equiv b_x$. *In* (3.5), W^* *is maximized at one of the integers immediately above or below* x^*.

3.1.3 Functional and Parametric RQ

As noted in Remark 2 of Whitt and You (2018b), the RQ approaches above are examples of a *parametric RQ*, because the variability of the arrival and service processes or their difference is characterized by a single parameter. To capture the impact of dependence in the queues of a queueing network, it can be important to allow a more general functional characterization of the variability.

We can expose the impact of dependence among the interarrival times and service times on the steady-state waiting time in the general stationary $G/G/1$ model (now relaxing the i.i.d. condtions of the $GI/GI/1$ model) as a function of the traffic intensity ρ by introducing a new *functional RQ* formulation. (With the $G/G/1$ model, we assume stationarity, so that there is a well defined steady state, but we allow dependence among the interarrival times and service times.) To treat the $G/G/1$ model, we replace the uncertainty set in (3.4) by

$$\mathcal{U}_f^x \equiv \{\tilde{X} : S_k^x \leq E[S_k^x] + b'_x\sqrt{Var(S_k^x)}, \quad k \geq 0\}. \tag{3.6}$$

and similarly for the two constraints in (3.2).

For the $GI/GI/1$ model, the new uncertainty set (3.6) is essentially equivalent to the previous one in (3.3), but they can be very different with dependence. It is significant that there are CLT's to motivate the form of the constraints in (3.6), just as there are in the i.i.d. case underlying (3.3). These supporting CLT's are reviewed in §EC.5 of Whitt and You (2018b). The CLT supports the spatial scaling by $\sqrt{Var(S_k)}$ instead of $\sqrt{k}$, as shown in §EC.5.3 of Whitt and You (2018b). Of course, the functional RQ produces a more complicated optimization problem, but it is potentially more useful, in part because it too can be analyzed.

3.1.4 Relations Between Performance Measures

Approximations for the mean steady-state waiting time $E[W]$ in the $GI/GI/1$ model follow directly from Theorem 3.1. They also follow from the subsequent approximations for the mean steady-state workload, denoted by $E[Z]$, developed in §3 and §4 of Whitt and You (2018b) and Brumelle's theorem, which relates $E[Z]$ to $E[W]$ in great generality, which we here discuss in §3.3

The steady-state mean values $E[Z]$ to $E[W]$ and other steady-state performance measures are related as a consequence of Little's law $L = \lambda W$ and its generalization $H = \lambda G$; e.g., see Whitt (1991) and Chapter X of Asmussen (2003) for the $GI/GI/1$ special case. Let W, Q and N be the steady-state waiting time, queue length and the number in system (including the one in service, if any, at an arbitrary time). Assume that the mean service time is $\tau = 1$ and the arrival rate is $\lambda = \rho < 1$. By Little's law,

$$\begin{aligned} E[Q] &= \lambda E[W] = \rho E[W] \quad \text{and} \\ E[N] &= E[Q] + \rho = \rho(E[W] + 1). \end{aligned} \tag{3.7}$$

By Brumelle's formula Brumelle (1971) or $H = \lambda G$, (6.20) of Whitt (1991), in the $G/G/1$ model with service times distributed as V,

$$E[Z] = \lambda E[WV] + \lambda \frac{E[V^2]}{2}. \tag{3.8}$$

In the $G/GI/1$ model with i.i.d. service times distributed as V (having mean 1 and variance c_s^2) that are independent of the arrival process,

$$E[Z] = \rho E[W] + \rho \frac{(c_s^2 + 1)}{2}, \tag{3.9}$$

For the simple formula in (3.9), it is important that the service times be independent of the waiting times. That property breaks down when there is customer feedback in a queueing network.

Given an approximation Z^* for $E[Z]$, we can use the approximations

$$\begin{aligned} E[W] &\approx \max\{0, Z^*/\rho - (c_s^2 + 1)/2\mu\} \quad \text{and} \\ E[Q] &\approx \lambda E[W]. \end{aligned} \tag{3.10}$$

3.2 Direct RQ Approximations for the Mean Waiting Time

As can be seen from Theorem 3.1, the RQ approximations for the mean waiting time in the $GI/GI/1$ queue from Bandi et al. (2015) has two free parameters that need to be specified, while the RQ

approximation from Whitt and You (2018b) has one free parameter that needs to be specified. This can the basis of corrections or adjustments in applications, which leaves the overall approximation unspecified, but it is also natural to specify that free parameter so that the formula coincides with standard cases for which the exact formula is known. That has been done in Bandi et al. (2015) and Whitt and You (2018b). We will discuss in more detail here.

3.2.1 The New RQ Approximation from Whitt and You (2018b).

First, for the $GI/GI/1$ queue with parameter vector $(\lambda, c_a^2, \tau, c_s^2) = (1, c_a^2, \rho, c_s^2)$, Theorem 3.1 yields

$$E[W] \approx \left(\frac{\rho}{1-\rho}\right)\left(\frac{b^2}{4}\right) \tag{3.11}$$

for

$$b \equiv b_x, \tag{3.12}$$

where b_x is the single variability in the single uncertainty set in (3.3).

The next issue then is how to specify b_x. By the motivation for the uncertainty set in terms of the central limit theorem (in the classical case of i.i.d random variables with finite second moments), we should have

$$b_x \equiv \beta\sqrt{Var(X)}, \tag{3.13}$$

for some constant β.

As indicated in Corollary 1 of Whitt and You (2018b), if we let

$$b \equiv b_x \equiv \beta\sqrt{Var(X)} \quad \text{and} \quad \beta \equiv \sqrt{2}, \tag{3.14}$$

i.e., if we let $b \equiv \sqrt{2Var(X)} = \sqrt{2(c_a^2 + \rho^2 c_s^2)}$, then

$$E[W] \approx \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{c_s^2 + (c_a^2/\rho^2)}{2}\right), \tag{3.15}$$

which coincides with the Kingman upper bound in Kingman (1962) when we set $\tau = \rho$. Hence, as stated in Corollary 1 of Whitt and You (2018b), their new RQ approximation coincides with the Kingman bound and so is asymptotically correct in heavy traffic.

Unfortunately, the Kingman upper bound tends to be too large away from heavy traffic, as can be seen from the tables in Chen and Whitt (2018). Thus, the approximation in (3.15) is actually not so good away from heavy traffic. Fortunately, the alternative RQ approximation based on the continuous-time workload process developed in §3 of Whitt and You (2018b) tends to be more accurate; see §3.3 here. Thus, we apply (3.9) to obtain $E[W]$ from $E[Z]$ in the $G/GI/1$ model via (3.10).

3.2.2 The Original RQ Approximation from Bandi et al. (2015).

On the other hand, the RQ approximation in Bandi et al. (2015) also yields (3.11), but now with

$$b \equiv b_a + b_s, \tag{3.16}$$

where b_a and b_s are the variability parameters in the two uncertainty sets in (3) of Whitt and You (2018b). However, the details are somewhat different, so a careful comparison may be confusing. First, they focus on the mean system time, but that is just the mean waiting time plus the mean service time. That additional mean service time gives an extra term. Nevertheless, it is reasonable to arrive at the same formula as (3.11) for Bandi et al. (2015) except we have (3.16) instead of (3.14) and (3.13).

Given the two separate uncertainty sets for the arrival process and the service process, it is natural to apply the CLT reasoning to each process separately, but it is not necessary to do so. Indeed, in §7 of Bandi et al. (2015) they take a different approach, which we will discuss later.

For now, assume that we do indeed apply the CLT reasoning to each parameter separately. That leads to the definitions

$$b_a \equiv \beta_a \sqrt{Var(U)} \quad \text{and} \quad b_s \equiv \beta_s \sqrt{Var(V)}, \tag{3.17}$$

which still leaves two parameters to specify. We reduce these parameters to only one by assuming that we treat the service times and interarrival times the same way; i.e., we assume that

$$\beta_a = \beta_s = \beta, \tag{3.18}$$

which leaves only the single unspecified parameter β. (This seems natural, but is not actually done in Bandi et al. (2015).)

Combining these assumptions yields (3.11), but now with

$$b \equiv b_a + b_s = \beta(\sqrt{Var(U)} + \sqrt{Var(V)}) = \beta(c_a + \rho c_s), \tag{3.19}$$

so that

$$E[W] \approx \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{\beta^2}{4}\right)\left(c_s^2 + (c_a^2/\rho^2) + 2(c_s c_a/\rho)\right). \tag{3.20}$$

The problems with this approach are described in Corollary 1 of Whitt and You (2018b), but again we emphasize that this natural approach is not followed in Bandi et al. (2015).

If we again use approximation $\beta \equiv \sqrt{2}$, as in (3.14) above, then formula (3.20) coincides with formula (3.15) for both the $D/GI/1$ model (when $c_a^2 = 0$) and the $GI/D/1$ model (when $c_s^2 = 0$). More generally, if we let $\beta \equiv \sqrt{2}$, then we obtain

$$E[W] \approx \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{c_s^2 + (c_a^2/\rho^2) + 2(c_s c_a/\rho)}{2}\right). \tag{3.21}$$

Clearly, the only difference between formula (3.21) and formula (3.15) is the final term $2(c_s c_a/\rho)$, which disappears if and only if the model is $D/GI/1$ or $GI/D/1$.

We have observed that the approximation (3.15) proposed in Whitt and You (2018b) coincides with the Kingman upper bound (and so already is too large), but the approximation from (3.21) is even larger. For the special case of $c_a^2 = c_s^2$, the ratio of the two approximations is

$$\frac{E[W; RQ(BBY15)]}{E[W; RQ(WY18)]} = \frac{1+\rho^2+2\rho}{1+\rho^2}, \tag{3.22}$$

which is increasing in ρ, going from 1 at $\rho = 0$ up to 2 as ρ increases toward 1.

3.2.3 An $M/M/1$-tuned RQ approximation starting from Bandi et al. (2015).

Clearly our parameter choice for the original RQ approximation in Bandi et al. (2015) is not good, because it is consistently too large. A simple way to do better, at least in some cases, is to choose the parameters so that the approximation is exact for the $M/M/1$ model. Then we have $c_a^2 = c_s^2 = 1$. In that case, (3.20) becomes correct if, instead of letting $\beta \equiv \sqrt{2}$, we let

$$\left(\frac{\beta^2}{4}\right)\left(1 + (1/\rho^2) + (2/\rho)\right) \equiv 1 \tag{3.23}$$

in (3.20) or, equivalently, if

$$\beta \equiv \frac{2\rho}{1+\rho} \tag{3.24}$$

Combining (3.24) and (3.20), the overall $GI/GI/1$ approximation becomes

$$E[W] \approx \left(\frac{\rho^2}{1-\rho}\right)\left(\frac{\rho^2}{(1+\rho)^2}\right)\left(c_s^2 + (c_a^2/\rho^2) + 2(c_s c_a/\rho)\right). \tag{3.25}$$

To check (3.25), we see that it agrees with the $M/M/1$ exact formula when $c_a^2 = c_s^2 = 1$.

Note that the new $M/M/1$-tuned approximation in (3.25) is smaller than the approximation in (3.21) by a factor of

$$\psi(\rho) \equiv \frac{2\rho^2}{(1+\rho)^2}. \tag{3.26}$$

which approaches $1/2$ as ρ increases toward 1.

3.2.4 The RQ Approximation in §7 of Bandi et al. (2015)

No doubt, the difficulties above were noticed by the authors of Bandi et al. (2015). Thus, they chose their parameters in a different way that circumvents many of these difficulties. In particular, in §7.2 they choose

$$\Gamma_a \equiv \sigma_a, \tag{3.27}$$

which is quite natural, but in their function f introduced on p. 696 below Table 1 they circumvent this choice by making Γ_s a complex function of the arrival and service parameters, in particular, they let

$$\Gamma_s \equiv \sqrt{\theta_0 + \theta_1\sigma_s^2 + \theta_2\sigma_a^2\rho^2} - \sigma_a. \tag{3.28}$$

This step effectively reduces the number of parameters to one instead of two.

The subtraction in (3.28) acts to cancel out the definition in (3.27). They discuss applying regression with examples to fit the three parameters θ_i, but do not discuss in detail. However, it seems natural to choose $\theta_0 \equiv 0$ and $\theta_1 \equiv \theta_2 \equiv 1$, which yields

$$\Gamma_s \equiv \sqrt{\sigma_s^2 + \sigma_a^2\rho^2} - \sigma_a. \tag{3.29}$$

Then

$$\Gamma^2 = (\Gamma_a + \Gamma_s)^2 = \sigma_s^2 + \sigma_a^2\rho^2, \tag{3.30}$$

so that the difficulty observed in Corollary 1 of Whitt and You (2018b) is avoided. In particular, for $\beta \equiv \sqrt{2}$, then the bound in Bandi et al. (2015) coincides with the Kingman upper bound in Kingman (1962) and (5) in Chen and Whitt (2018); i.e., if we let the arrival rate be 1 and the mean service time be ρ, then

$$E[W] \approx \frac{\rho^2([c_a^2/\rho^2] + c_s^2)}{2(1-\rho)} = \frac{\sigma_a^2 + \sigma_s^2}{2(E[U] - E[V])}. \tag{3.31}$$

3.3 The RQ Approximation for the Mean Workload $E[Z]$

Whitt and You (2018b) found that it was advantageous to approach the RQ approximation via the continuous-time workload process. Hence, they primarily focus on the RQ approximation for the mean workload $E[Z]$.

The basic continuous-time processes are the arrival counting process, defined by

$$A(t) \equiv \max\{k \geq 1 : U_1 + \cdots + U_k \leq t\} \quad \text{for} \quad t \geq U_1 \tag{3.32}$$

with $A(t) \equiv 0$ for $0 \leq t < U_1$, the total input of work over $[0,t]$ and the net-input process, respectively,

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k \quad \text{and} \quad N(t) \equiv Y(t) - t, \quad t \geq 0, \tag{3.33}$$

while the workload (the remaining workload) at time t, starting empty at time 0, is the reflection map Ψ applied to N, i.e.,

$$Z(t) = \Psi(N)(t) \equiv N(t) - \inf_{0 \leq s \leq t} \{N(s)\}, \quad t \geq 0; \tag{3.34}$$

e.g., see §13.5 of Whitt (2002).

As in §6.3 of Sigman (1995), we again use a reverse-time construction to represent the workload in a single-server queue as a supremum, so that the RQ optimization problem becomes a maximization over constraints expressed in an uncertainty set, just as before, but now it is a continuous optimization problem. Using the same notation, but with a new meaning, let $Z(t)$ be the workload at time 0 of a system that started empty at time $-t$. Then $Z(t)$ can be represented as

$$Z(t) \equiv \sup_{0 \leq s \leq t} \{N(s)\}, \quad t \geq 0, \tag{3.35}$$

where N is defined in terms of Y as before, but Y is interpreted as the total work in service time to enter over the interval $[-s, 0]$. That is achieved by letting V_k be the k^{th} service time indexed going backwards from time 0 and $A(s)$ counting the number of arrivals in the interval $[-s, 0]$. Paralleling the waiting time in §3, $Z(t)$ increases monotonically to Z as $t \to \infty$. For the stable stationary $G/G/1$ stochastic queue, Z corresponds to the steady-state workload and satisfies $P(Z < \infty) = 1$; see §6.3 of Sigman (1995).

In continuous time, we work with continuous-time stationarity instead of discrete-time stationarity; e.g., see Sigman (1995). Hence, we assume that there is a base stationary process $\{(A(t), Y(t)) : t \geq 0\}$ with $E[A(t)] = E[Y(t)] = t$ for all $t \geq 0$ and introduce ρ by simple scaling via

$$A_\rho(t) \equiv A(\rho t) \quad \text{and} \quad Y_\rho(t) \equiv Y(\rho t), \quad t \geq 0 \quad \text{and} \quad 0 < \rho < 1, \tag{3.36}$$

which implies that $E[A_\rho(t)] = E[Y_\rho(t)] = \rho t$ for all $t \geq 0$. Then $N_\rho(t) \equiv Y_\rho(t) - t$ and $Z_\rho(t) = \Psi(Y_\rho)(t)$, $t \geq 0$. With the reverse-time construction $Z_\rho(t)$ can be expressed as a supremum over the interval $[0,t]$, just as in (3.35).

Within that scaling framework, the natural parametric and functional (see §3.1.3) uncertainty sets for the steady-state workload are, respectively,

$$\mathcal{U}_\rho^p \equiv \left\{ \tilde{N}_\rho : \mathbb{R}^+ \to \mathbb{R} : \tilde{N}_\rho(s) \leq -(1-\rho)s + b_p\sqrt{s},\ s \geq 0 \right\} \quad \text{and}$$

$$\begin{aligned}\mathcal{U}_\rho \equiv \mathcal{U}_\rho^f &\equiv \left\{\tilde{N}_\rho : \mathbb{R}^+ \to \mathbb{R} : \tilde{N}_\rho(s) \le E[N_\rho(s)] + b_f\sqrt{Var(N_\rho(s))},\ s \ge 0\right\}, \\ &= \left\{\tilde{N}_\rho : \mathbb{R}^+ \to \mathbb{R} : \tilde{N}_\rho(s) \le -(1-\rho)s + b_f\sqrt{Var(N_\rho(s))},\ s \ge 0\right\}, \end{aligned} \tag{3.37}$$

where we regard $\tilde{N}_\rho \equiv \{\tilde{N}_\rho(s) : 0 \le s \le t\}$ as an arbitrary real-valued function on $\mathbb{R}^+ \equiv [0, \infty)$, while we regard $\{N_\rho(s) : s \ge 0\}$ as the underlying stochastic process, and $\{Var(N_\rho(s)) : s \ge 0\} = \{Var(Y_\rho(s)) : s \ge 0\}$ as its variance-time function, which can either be calculated for a stochastic model or estimated from simulation or system data. In (3.37), b_p and b_f are parameters to be specified.

To apply this RQ formulation for the mean workload, Whitt and You (2018b) apply the IDC and the associated *index of dispersion for work* (IDW) associated with the rate-1 cumulative input process Y by

$$I_w(t) \equiv \frac{Var(Y(t))}{E[V_1]E[Y(t)]} = \frac{V(t)}{t}, \quad t \ge 0. \tag{3.38}$$

The IDW was introduced in Fendick and Whitt (1989). Clearly, these indices of dispersion are just scaled versions of the associated variance functions, but they are important for understanding, because they expose the variability over time, independent of the scale.

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

$$I_w(t) = I_c(t) + c_s^2, \quad t \ge 0; \tag{3.39}$$

see §4.3.1 of Whitt and You (2018b)

Given the IDW, the RQ approximation for the mean workload from (28) in §4.1 of Whitt and You (2018b) is simply

$$Z_\rho^* = \sup_{x \ge 0}\left\{-(1-\rho)x/\rho + b_f\sqrt{xI_w(x)}\right\}. \tag{3.40}$$

Strong positive results for the RQ approximation for the mean workload $E[Z]$ in the $G/GI/1$ queue in Theorems 2-5 of Whitt and You (2018b). 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.

Finally, we observe that this approximation for the mean workload also provides an approximation for the mean waiting time by applying the exact relation in (3.8).

4 The Series Queue Model

We are now ready to consider approximations for the mean steady-state waiting time at each queue in a $GI/GI/1 \to GI/1 \to \ldots \to GI/1$ tandem open queueing network. We start by reviewing

the basic asymptotic and stationary-interval methods for approximating a point process and its application to departure processes, which constitute the arrival processes at the following queues. Afterwards, we consider alternative methods, including ones stemming from the heavy-traffic and RQ literature.

4.1 The Asymptotic and Stationary-Interval Methods

The asymptotic and stationary-interval methods for approximating a point process are discussed in Whitt (1982), while their application to queues in series is discussed in Whitt (1984). The stationary-interval method is used in the QNA approximation, as indicated in §4.5 of Whitt (1983a). These methods partially characterize the variability of a stationary point process by a single variability parameter, chosen to be independent of the rate of the point process. The idea of the asymptotic method is to match the asymptotic variability parameter in the CLT, while the idea of the stationary-interval method is to match the scv of the stationary interval between points. Thus, these are two extreme perspectives on the variability. The asymptotic method yields the large-time perspective, while the stationary-interval method yields the short-time (or local) perspective. One might also take the perspective directly as seen by a queue, as discussed in Whitt (1981) and Whitt (1983b).

As indicated above, QNA exploits the stationary-interval method. In particular, the approximating scv for a renewal process approximation for the departure process from a $GI/GI/1$ queue is taken to be c_d^2, defined as the following convex combination of the scv's of the service time and interarrival time

$$c_d^2 \approx \rho^2 c_s^2 + (1 - \rho^2) c_a^2; \tag{4.1}$$

see (37) in §4.5 of Whitt (1983a). This is based on the analysis in §2 of Whitt (1984) and references cited there.

The asymptotic method would just approximate the departure process by the arrival process, as indicated in §1 of Whitt (1984). Instead of (4.1), that yields

$$c_d^2 \approx c_a^2. \tag{4.2}$$

It is known that the asymptotic method is asymptotically correct for the arrival process at a queue in heavy traffic. That follows from Theorem 1 of Iglehart and Whitt (1970). The heavy-traffic limit for a general queue depends on the CLT of the arrival process. Moreover, the parameters of

the limiting reflected Brownian motion (RBM) depend on the arrival process beyond its rate only through its CLT, i.e, through the asymptotic-method variability parameter of the arrival process.

For two queues in series, this result implies that the asymptotic method is asymptotically correct for the arrival process at the second queue (the departure process from the first queue) if the second queue is in heavy traffic, while the first queue is not. However, early experiments leading to QNA in Whitt (1983a) showed that the stationary-interval method usually is far better than the asymptotic method. We will illustrate the difficulties associated with the asymptotic method in our simulation examples.

This discussion is highly relevant when considering alternative approximations, because the sequential bottleneck decomposition (SBD) approximation in Dai et al. (1994) in the case of a single bottleneck node and the robust queueing approach in Bandi et al. (2015) both use the asymptotic method (although that is not discussed). The RQ algorithm in Bandi et al. (2015) uses the asymptotic method because their approach makes the uncertainty set for a departure process coincide with the uncertainty set of the arrival process.

Overall, the stationary-interval method and the asymptotic method are quite different perspectives that often lead to different conclusions. Neither of these simple methods works well in all cases.

4.2 The Heavy-Traffic Bottleneck Phenomenon

Challenges in approximating departure processes and queues in series were exposed in the paper Suresh and Whitt (1990), which discusses the heavy-traffic bottleneck phenomenon. The QNA approximation based on the stationary-interval method in (4.1) makes gradual changes in the scv of each departure process in passing through a series of identical queues. The HT bottleneck approximation shows that, even after passing through a large number of queues with low-to-moderate traffic intensity, the behavior at a later queue with a much higher traffic intensity can be determined by the asymptotic method; i.e., the behavior at the later queue in heavy traffic is as if all the immediate queues were not there. This phenomenon is a result of complicated long-range dependence embedded in the arrival processes, introduced by flowing through a queue (the departure processes). This example was introduced to show the limitation of traditional decomposition methods, e.g. the QNA algorithm, and is often used as a benchmark for different approximation methods, see §3.3 of Dai et al. (1994).

Ways to address those problems have been discussed by Fendick and Whitt (1989), Harrison

and Nguyen (1990), Reiman (1990), Dai et al. (1994), Whitt (1995) and Whitt and You (2018a,c). We shall use these examples to illustrate the RQ algorithms.

4.3 Comparisons with Simulation

In this section, we compare the performance of various approximations for the mean waiting time at each queue for the example with 9 queues in series considered by Suresh and Whitt (1990).

This section supplements §5 of Whitt and You (2018c), which compares the new RQ algorithms in Whitt and You (2018b,a) to previous methods. Here we focus on basic methods, especially the variants of RQ from Bandi et al. (2015).

We compare several methods to simulation estimates from Suresh and Whitt (1990). We start by comparing the $M/M/1$ model, which ignores the variability parameters, i.e., sets $c_a^2 = c_s^2 = 1$. Then we compare QNA from Whitt (1983a), which coincides with the stationary-interval method, to simulation. Then we consider the modification of QNA with the arrival process at each queue approximated by the asymptotic method, which approximate the scv of the arrival process by the scv of the external arrival process, but otherwise using the standard $GI/GI/1$ approximation in

$$E[W] \equiv E[W(\rho, c_a^2, c_s^2)] \approx \frac{\rho^2(c_a^2 + c_s^2)}{2(1-\rho)}. \tag{4.3}$$

for the case with arrival rate 1 and mean service time ρ.

As discussed in Suresh and Whitt (1990), we see that the stationary-interval method does far better than the asymptotic method at queues $2-8$, but not at the final queue 9 with high traffic intensity. We then display the approximations resulting from three variants of the RQ algorithm in Bandi et al. (2015). In particular, we use the asymptotic method with (i) (3.20) in §3.2.2, (ii) the $M/M/1$-tuned method in (3.25) in §3.2.3 and (iii) the Kingman upper bound from §7 of Bandi et al. (2015) discussed here in §3.2.4. We find that all these methods fail to produce consistently good approximations.

In particular, we consider an OQN with 9 stations in tandem, each with i.i.d. exponential service times. Station 1 has the only external arrival process, which is a rate-1 general renewal process. The traffic intensities at the first 8 queues are set to $\rho_i = 0.6$ for $1 \le i \le 8$, while the last queue has the significantly higher traffic intensity $\rho_9 = 0.9$. As in Suresh and Whitt (1990), two specific external renewal arrival processes are considered: (i) deterministic (D) interarrival times with $c_{a_0}^2 = 0$; and (ii) highly variable H_2 interarrival times with $c_{a_0}^2 = 8$ (and balanced means).

4.3.1 Basic Approximations

Table 1 (for low variability) and Table 2 (for high variability) compare 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.

Table 1: A comparison of six approximation methods to simulation for 9 exponential (M) queues in series with $\rho_i = 0.6$, $1 \leq i \leq 8$, and $\rho_9 = 0.9$ fed by a deterministic arrival process with $c_a^2 = 0$. The two columns for QNA are for the stationary interval method (SI, actually used) and the asymptotic method (AM). The last two columns contain the three versions of RQ from Bandi et al. (2015) in §3.2.2- §3.2.4, but two coincide in this case. All use the asymptotic method. The final column is the Kingman upper bound combined with the asymptotic method, as in §7 of Bandi et al. (2015).

	Method		QNA	Whitt (1983a)	RQ, Bandi et. al.	
node	Sim	M/M/1	SI	AM	(3.21) and (3.31)	(3.25)
1	0.290 (2.41%)	0.90	0.45 (55.2%)	0.45 (55.2%)	0.45 (55.17%)	0.13
2	0.491 (1.43%)	0.90	0.61 (24.2%)	0.45 (-8.4%)	0.45 (-8.4%)	0.13
3	0.607 (1.32%)	0.90	0.72 (18.6%)	0.45 (-25.9%)	0.45 (-25.9%)	0.13
4	0.666 (1.20%)	0.90	0.78 (17.1%)	0.45 (-32.4%)	0.45 (-32.4%)	0.13
5	0.706 (1.42%)	0.90	0.83 (17.6%)	0.45 (-36.3%)	0.45 (-36.3%)	0.13
6	0.731 (1.78%)	0.90	0.85 (16.3%)	0.45 (-38.4%)	0.45 (-38.4%)	0.13
7	0.748 (1.34%)	0.90	0.87 (16.3%)	0.45 (-39.8%)	0.45 (-39.8%)	0.13
8	0.775 (1.68%)	0.90	0.88 (13.6%)	0.45 (-41.9%)	0.45 (-41.9%)	0.13
9	5.031 (4.31%)	8.10	7.99 (58.8%)	4.05 (-19.5%)	4.05 (-19.5%)	1.82
Total	10.05	15.30	13.97 (39.00%)	7.65 (-23.8%)	7.65 (-23.8%)	2.86

First, we display the simple $M/M/1$ approximation, which ignores the variability parameters entirely (acts as if $c_a^2 = c_s^2 = 1$). Next we present the QNA algorithm from Whitt (1983a), which uses the stationary-interval method, which means that the arrival process variability parameters are computed recursively via (4.1) and then the mean waiting time $E[W]$ is computed from (4.3). Second, the asymptotic method differs by letting the arrival-process variability parameter at each queue be identical to the given external arrival process variability parameter. Hence, for the asymptotic method the approximation is identical for queues 1-8, but different for the final queue 9.

The last three columns show results for the two RQ algorithms from §3.2.2-§3.2.4, based on Bandi et al. (2015). The first version (3.21) is the same as approximation (9) in Whitt and You (2018c) for the $D/GI/1$ and $GI/D/1$ models, so we see that the third and fourth columns of Table 1 coincide. Since the departure uncertainty set is made equal to the arrival uncertainty set, this corresponds to the asymptotic method in column 3. Since RQ from Bandi et al. (2015) makes the departure uncertainty set agree with the arrival uncertainty set, the approximations for queues 1-8

are all identical, just as with QNA with the asymptotic method. However, the RQ approximations in (3.21) and (3.25) are remarkably inaccurate for $c_a^2 = 8$. Table 2 is to be contrasted with Tables 2 and 3 in Whitt and You (2018c).

Table 2: A comparison of six approximation methods to simulation for 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 H_2 rate-1 renewal arrival process with $c_a^2 = 8$ and the third parameter specified by balanced means. The two columns for QNA are for the stationary interval method (SI, actually used) and the asymptotic method (AM). The last three columns contain the three versions of RQ from Bandi et al. (2015) in §3.2.2- §3.2.4. The final column is the Kingman upper bound combined with the asymptotic method, as in §7 of Bandi et al. (2015).

	method		QNA		RQ	Bandi et al.	
node	Sim	$M/M/1$	(SI)	(AM)	(3.21)	(3.25)	(3.31)
1	3.284 (3.50%)	0.90	4.05 (23.3%)	4.05 (23.3%)	22.3	6.3	10.4
2	2.321 (4.18%)	0.90	2.92 (25.8%)	4.05 (74.5%)	22.3	6.3	10.4
3	1.914 (3.40%)	0.90	2.19 (14.4%)	4.05 (112%)	22.3	6.3	10.4
4	1.719 (4.07%)	0.90	1.73 (0.6%)	4.05 (135%)	22.3	6.3	10.4
5	1.598 (3.69%)	0.90	1.43 (-10.5%)	4.05 (153%)	22.3	6.3	10.4
6	1.478 (4.13%)	0.90	1.24 (-16.1%)	4.05 (174%)	22.3	6.3	10.4
7	1.423 (3.23%)	0.90	1.12 (-21.3%)	4.05 (185%)	22.3	6.3	10.4
8	1.413 (4.67%)	0.90	1.04 (-26.4%)	4.05 (189%)	22.3	6.3	10.4
9	30.12 (16.8%)	8.10	8.90 (-70.5%)	41.0 (36%)	115.7	51.8	44.05
Total	45.27	15.30	24.60 (-45.7%)	73.4 (62.1%)	294	102.2	127.2

References

d

S. Asmussen. *Applied Probability and Queues*. Springer, New York, second edition, 2003.

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

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

Y. Chen and W. Whitt. Extremal $GI/GI/1$ queues given two moments. submitted to Operations Research. Available at http://www.columbia.edu/∼ww2040/allpapers.html, 2018.

K. L. Chung. *A Course in Probability Theory*. Academic Press, New York, third edition, 2001.

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

K. W. Fendick and 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):171–194, 1989.

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

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.

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

D. V. Lindley. The theory of queues with a single server. *Math. Proceedings Cambridge Phil. Soc.*, 48: 277–289, 1952.

R.M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(3):497–520, 1962.

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

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

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

W. Whitt. Approximating a point process by a renewal process: The view through a queue, an indirect approach. *Management Science*, 27(6):619–636, 1981.

W. Whitt. Approximating a point process by a renewal process: two basic methods. *Oper. Res.*, 30:125–147, 1982.

W. Whitt. The queueing network analyzer. *Bell Laboratories Technical Journal*, 62(9):2779–2815, 1983a.

W. Whitt. Queue tests for renewal processes. *Operations Research Letters*, 2(1):7–12, 1983b.

W. Whitt. Approximations for departure processes and queues in series. *Naval Research Logistics (NRL)*, 31(4):499–521, 1984.

W. Whitt. A review of $L = \lambda W$. *Queueing Systems*, 9:235–268, 1991.

W. Whitt. Variability functions for parametric-decomposition approximations of queueing networks. *Management Science*, 41(10):1704–1715, 1995.

W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.

W. Whitt and W. You. Heavy-traffic limit of the $GI/GI/1$ stationary departure process and its variance function. *Stochastic Systems*, 8(2):143–165, 2018a.

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

W. Whitt and W. You. The advantage of indices of dispersion in queueing approximations. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html, 2018c.