Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues

Ward Whitt, Wei You

To cite this article:
Ward Whitt, Wei You (2018) Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues. Operations Research 66(1):184-199. https://doi.org/10.1287/opre.2017.1649

http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 66, No. 1, January–February 2018, pp. 184–199
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues

Ward Whitt,[a] **Wei You**[a]

[a] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027

Contact: ww2040@columbia.edu, http://orcid.org/0000-0003-4298-9964 (WW); wy2225@columbia.edu, http://orcid.org/0000-0003-0844-4194 (WY)

Received: March 4, 2016
Revised: October 10, 2016; March 11, 2017
Accepted: May 3, 2017
Published Online in Articles in Advance: July 31, 2017

Subject Classifications: queues: approximations, networks, algorithms
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2017.1649

Abstract. Queueing applications are often complicated by dependence among interarrival times and service times. Such dependence is common in networks of queues, where arrivals are departures from other queues or superpositions of such complicated processes, especially when there are multiple customer classes with class-dependent service-time distributions. We show that the robust queueing approach for single-server queues proposed in the literature can be extended to yield improved steady-state performance approximations in the standard stochastic setting that includes dependence among interarrival times and service times. We propose a new functional robust queueing formulation for the steady-state workload that is exact for the steady-state mean in the $M/GI/1$ model and is asymptotically correct in both heavy traffic and light traffic. Simulation experiments show that it is effective more generally.

Funding: Support was received from the National Science Foundation [Grants CMMI 1265070 and 1634133].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2017.1649.

Keywords: robust queueing • queueing approximations • dependence among interarrival times and service times • indices of dispersion • heavy traffic • queueing network analyzer

1. Introduction

Robust optimization is proving to be a useful approach to complex optimization problems involving significant uncertainty; e.g., see Bandi and Bertsimas (2012), Bertsimas et al. (2011), and references therein. In that context, the primary goal is to create an efficient algorithm to produce useful, practical solutions that appropriately capture the essential features of the uncertainty. Bandi et al. (2015) have applied this approach to create a robust queueing (RQ) theory, which can be used to generate performance predictions in complex queueing systems, including networks of queues as well as single queues. Indeed, they construct a full robust queueing analyzer (RQNA) to develop relatively simple performance descriptions such as those in the queueing network analyzer (QNA) in Whitt (1983).

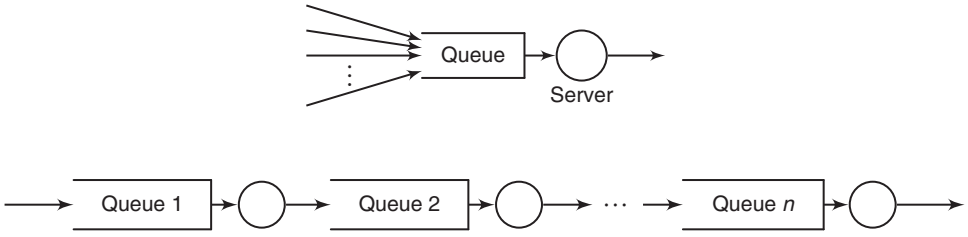
Our goal in this paper is to make further progress in the same direction. We do so by introducing new RQ formulations and evaluating their performance. We too want to obtain useful performance descriptions for complex queueing networks, but here we only consider a single queue. We judge our RQ formulations by their ability to efficiently generate useful performance approximations for the given stochastic model, which so far has been mostly intractable.

As emphasized in Bandi and Bertsimas (2012), the intractability is usually due to high dimension, but high dimensionality can occur in many different ways. The RQ in Bandi et al. (2015) emphasizes the high dimension arising when we consider a network of queues instead of a single queue. Instead, in this paper we focus on the high dimension that occurs in a single queue when there is complex stochastic dependence over time in the arrival and service processes. In a sequel, Whitt and You (2016), we focus on the high dimension that occurs in a single queue when the deterministic arrival-rate function is time varying. For both problems, we find that the robust optimization approach is remarkably effective. Here, we show that, with an appropriate choice of parameters, all our new RQ solutions are asymptotically correct in the heavy-traffic limit. Our most promising new RQ solutions in (18) and (28) are asymptotically correct in both light traffic and heavy traffic. Our simulation experiments show that the new RQ solutions provide useful approximations more generally.

1.1. Dependence Among Interarrival Times and Service Times

Even though we only focus on one single-server queue, ultimately we also want to develop methods that apply to complex networks of queues. We view the present paper as an important step in that direction, because experience from applications of QNA has shown that a major shortcoming is its inability to adequately capture the dependence among interarrival times and service

Figure 1. Common Queueing Network Structure That Can Induce Dependence Among Interarrival Times: Superpositions of Arrival Processes (Top) and Flow Through a Series of Queues (Bottom)



times at the individual queues in the network. That was dramatically illustrated by comparisons of QNA to model simulations in Sriram and Whitt (1986), Fendick et al. (1989), and Suresh and Whitt (1990).

Dependence among successive interarrival times at a queue is a common phenomenon, usually because that queue is actually part of a network of queues. For example, arrival processes in queueing networks are often superpositions of other arrival processes or departure processes from other queues, as depicted in Figure 1.

In most manufacturing production lines, an external (or initial) arrival process is often far less variable than a Poisson process by design, while complicated processing operations, such as those involving batching, often produce complicated dependence among the interarrival times at subsequent queues; e.g., see the example in section 3 of Segal and Whitt (1989). In both manufacturing and communication systems, dependence among successive interarrival times and among successive interdeparture times at a queue often occurs because there are multiple classes of customers with different characteristics (e.g., Bitran and Tirupati 1988). Multiple classes can even cause significant dependence (i) among interarrival times, (ii) among service times, and (iii) between interarrival times and service times, which all can contribute to a major impact on performance, as shown by Fendick et al. (1989) and reviewed in section 9.6 of Whitt (2002).

In service systems, an external customer arrival process often is well modeled by a Poisson process, because it is generated by many separate people making decisions independently, at least approximately, but dependence may be induced by overdispersion; e.g., see Kim and Whitt (2014) and references there. By contrast, internal arrivals within a network of queues are less likely to be well approximated by a Poisson process, because the flow through queues disrupts the statistical regularity of a Poisson process. In particular, service-time distributions are often not nearly exponential, while the interdeparture times in steady state from an $M/GI/1$ queue, with GI meaning that the service times are independent and identically distributed (i.i.d.), are themselves i.i.d. only if the service-time distribution is exponential, in which case the departure process is again Poisson. In other words, there are no nondeterministic non-Poisson renewal departure processes from an $M/GI/1$ queue; e.g., see Disney and Konig (1985).

The dependence among interarrival times and service times has long been recognized as a major difficulty in developing effective approximations for open queueing networks, such as in QNA in Whitt (1983); e.g., see Whitt (1995) and references therein. Refined performance approximations have been proposed using second-order partial characterizations of dependence, using indices of dispersion (variance-time functions), which involve correlations among interarrival times as well as means and variances; e.g., see Cox and Lewis (1966), Heffes (1980), Heffes and Luantoni (1986), Sriram and Whitt (1986), Fendick et al. (1989, 1991), and Fendick and Whitt (1989). Our new RQ formulations will exploit these same partial characterizations of the dependence among interarrival times and service times; see Sections 3.3 and 4. Even though we only consider a single queue here, in Section 6 we introduce a new framework in which we hope to develop a full RQNA based on the results in this paper.

1.2. Main Contributions

1. In this paper, we introduce several new RQ formulations for the steady-state waiting time and workload in a single-server queue, and we make useful connections to the general stationary $G/G/1$ stochastic model and the $GI/GI/1$ special case. In particular, we show how to choose the RQ parameters so that these RQ solutions all are asymptotically exact for the steady-state mean in the heavy-traffic limit.

2. In addition to new parametric versions of RQ as in Bandi et al. (2015), we introduce new functional formulations that capture the impact of dependence among the interarrival times and service times over time upon the steady-state performance of the queue as a function of the traffic intensity ρ. (See the uncertainty sets in (9) and (15).)

3. We evidently introduce the first RQ formulations for the continuous-time workload process and show that it is advantageous to do so. We show how to choose the RQ parameters so that the solution of the functional RQ for the workload coincides with the

steady-state mean in the $M/GI/1$ model for all traffic intensities and is simultaneously asymptotically correct in both heavy traffic and light traffic for the general $G/G/1$ model, including the dependence.

4. We conduct simulation experiments showing that the new functional RQ for the workload is effective in exposing the impact of the dependence among the interarrival times and service times over time upon the mean steady-state workload as a function of the traffic intensity.

5. We provide a road map for the application to networks of queues by introducing a new framework for an RQNA based on indices of dispersion. We show that such an RQNA is feasible and provide support with a simulation comparison for a series queue network.

1.3. More Related Literature

Mamani et al. (2016) also incorporated dependence within a robust optimization formulation for a problem in inventory management (which we might call RI), but otherwise, there is relatively little overlap with this paper; we discuss the connection in Remark 4. Mamani et al. (2016) point to early RI work by Scarf (1958) and then Moon and Gallego (1994). The new RQ work is also related to Whitt (1984a), which used optimization to study performance approximations used in QNA. In particular, Whitt (1984a) studied the range of possible values for the mean steady-state number in a $GI/M/1$ queue subject to specified first and second moments of the interarrival-time distribution. Klincewicz and Whitt (1984) and Whitt (1984b) construct tighter bounds based on additional constraints to enforce a realistic shape on the underlying interarrival-time distribution. This work showed that we can hope to obtain useful accuracy such as 20% relative error, but that we cannot hope to obtain extraordinarily high accuracy, such as an only 5% error, given the usual partial information based on the first two moments. And that is not yet considering the dependence. Ignoring the dependence can lead to much bigger errors, as in Fendick et al. (1989) and section 9.6 of Whitt (2002).

1.4. Organization of the Paper

In Section 2, after reviewing RQ for the steady-state waiting time in the single-server queue from sections 2 and 3.1 of Bandi et al. (2015), we develop an alternative formulation whose solution coincides with the Kingman (1962) bound and is asymptotically correct in heavy traffic. In Section 3 we introduce new parametric and functional RQ formulations for the continuous-time workload process and characterize their solutions. In Section 4 we introduce the index of dispersion for work (IDW) and incorporate it in the RQ. We develop closed-form RQ solutions and show that the functional RQ is asymptotically correct in both heavy and light traffic. In Section 5 we conduct simulation experiments for the two network structures in Figure 1. These experiments demonstrate (i) the strong impact of dependence upon performance and (ii) the value of the new RQ in capturing the impact of that dependence. Finally, in Section 6 we introduce a new framework for applying the results in this paper to develop a new RQNA that better captures the dependence. Additional supporting material appears in the e-companion (EC)—in particular, (i) a short summary of the main paper; (ii) additional discussion; (iii) additional theoretical support, including central limit theorems and heavy-traffic limit theorems; (iv) more results for the discrete-time waiting time and indices of dispersion; and (v) more simulation examples.

2. Robust Queueing for the Steady-State Waiting Time

We start by reviewing the RQ developed in sections 2 and 3.1 of Bandi et al. (2015), which involves separate uncertainty sets for the arrival times and service times. We then construct an alternative formulation with a single uncertainty set and show, for the $GI/GI/1$ queue, that a natural version of the RQ solution coincides with the Kingman (1962) bound and so is asymptotically correct in the heavy-traffic limit. We show that both formulations provide insight into the relaxation time for the $GI/GI/1$ queue, the approximate time required to reach steady state.

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

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

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.,

$$W_n = M_n \equiv \max_{0 \leqslant k \leqslant n} \{S_k\}, \quad n \geqslant 1, \tag{2}$$

using reverse-time indexing with $S_k \equiv X_1 + \cdots + X_k$ and $X_k \equiv V_{n-k} - U_{n-k}$, $1 \leqslant k \leqslant n$, 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 $W_n \uparrow W \equiv \sup_{k \geqslant 0} \{S_k\}$ 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 section 6.2 of Sigman (1995).

Bandi et al. (2015) propose an RQ approximation for the steady-state waiting time W by performing a deterministic optimization in (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):

$$\begin{aligned} \mathcal{U}^a &\equiv \{\tilde{U} \in \mathbb{R}^\infty : S_k^a \geqslant km_a - b_a\sqrt{k},\ k \geqslant 0\} \quad \text{and} \\ \mathcal{U}^s &= \{\tilde{V} \in \mathbb{R}^\infty : S_k^s \leqslant km_s + b_s\sqrt{k},\ k \geqslant 0\}, \end{aligned} \tag{3}$$

where $\tilde{U} \equiv \{U_k : k \geqslant 1\}$ and $\tilde{V} \equiv \{V_k : k \geqslant 1\}$ are arbitrary sequences of real numbers in $\mathbb{R}^\infty$; $S_k^a \equiv U_1 + \cdots + U_k$ and $S_k^s \equiv V_1 + \cdots + V_k$, $k \geqslant 1$; $S_0 \equiv 0$; and m_a, m_s, b_a, and b_s are parameters to be specified. The constraints in (3) are one-sided because that is what is required to bound the waiting times above, as we can see from (1) and (2). Thus, the RQ optimization can be expressed as

$$W^* \equiv \sup_{\tilde{U} \in \mathcal{U}^a} \sup_{\tilde{V} \in \mathcal{U}^s} \sup_{k \geqslant 0} \{S_k^s - S_k^a\}, \tag{4}$$

where S_k^a (S_k^s) is a function of $\tilde{U}$ ($\tilde{V}$) specified above. Versions of this formulation in (4) and others in this paper also apply to the transient waiting time W_n, but we will focus on the steady-state waiting time.

Thinking of the general stationary $G/G/1$ stochastic model, where the distributions of U_k and V_k are independent of k (but stochastic independence is not assumed), Bandi et al. (2015) assume that $m_a \equiv E[U_k]$ and $m_s \equiv E[V_k]$ and assume that $m_a > m_s$, so that $\rho \equiv m_s/m_a < 1$. The square-root terms in the constraints in (3) are motivated by the central limit theorem (CLT). Thinking of the $GI/GI/1$ model in which the interarrival times U_k and service times V_k come from independent sequences of i.i.d. random variables with finite variances σ_a^2 and σ_s^2, the CLT suggests that $b_a = \beta_a\sigma_a$ and $b_s = \beta_s\sigma_s$ for some positive constants β_a and β_s, perhaps with $\beta = \beta_a = \beta_s$. With this choice, these new parameters measure the number of standard deviations away from the mean in a Gaussian approximation. Bandi et al. (2015) also provide an extension to cover the heavy-tailed case, where finite variances might not exist; then $\sqrt{k}$ in (3) is replaced by $k^{1/\alpha}$ for $0 < \alpha \leqslant 2$, as we would expect from sections 4.5, 8.5, and 9.7 of Whitt (2002), but we will not discuss that extension here.

From (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), we propose the single uncertainty set (for each n)

$$\mathcal{U}^x \equiv \{\tilde{X} \in \mathbb{R}^\infty : S_k^x \leqslant -mk + b_x\sqrt{k},\ k \geqslant 0\}, \tag{5}$$

where $\tilde{X} \equiv \{X_k : k \geqslant 1\} \in \mathbb{R}^\infty$, $S_k^x \equiv X_1 + \cdots + X_k$, $k \geqslant 1$, 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 (5) by $\max\{k, k_L\}$, but that lower bound k_L has no impact if chosen appropriately. Combining (2) and (5), we obtain the alternative RQ optimization

$$W^* \equiv \sup_{\tilde{X} \in \mathcal{U}^x} \sup_{k \geqslant 0} \{S_k^x\}, \tag{6}$$

where S_k^x is the function of $\tilde{X}$ specified above. The RQ formulations in (4) and (6) are attractive because the optimization problems have simple solutions in which all constraints are satisfied as equalities. That follows easily from the fact that W_n is a nondecreasing (nonincreasing) function of V_k (of U_k) for all k and n. The simple closed-form solution follows from the triangular structure of the equations; see section 3.1 of Bandi et al. (2015). The following is a direct extension of theorem 2 of Bandi et al. (2015) to include the new RQ formulation in (6). The final statement involves an interchange of suprema, which is justified by Lemma EC.1.

Theorem 1 (RQ Solutions for the Steady-State Waiting Time). *The RQ optimizations* (4) *with* $m_a > m_s > 0$ *and* (6) *with* $m > 0$ *have the solution*

$$\begin{aligned} W^* &= \sup_{k \geqslant 0} \{-mk + b\sqrt{k}\} \leqslant \sup_{x \geqslant 0} \{-mx + b\sqrt{x}\} \\ &= -mx^* + b\sqrt{x^*} = \frac{b^2}{4m} \quad \text{for } x^* = \frac{b^2}{4m^2}, \end{aligned} \tag{7}$$

where $m = m_a - m_s > 0$. *For* (4), $b \equiv b_s + b_a$; *for* (6), $b \equiv b_x$. *In* (7), W^* *is maximized at one of the integers immediately above or below* x^*.

We now establish implications for the $GI/GI/1$ and general stationary $G/G/1$ models. To discuss heavy-traffic limits, it is convenient to introduce the traffic intensity ρ as a scaling factor applied to the interarrival times. Hence, we start with a sequence $\{(U_k, V_k)\}$ where $E[U_k] = E[V_k] = 1$ for all k. Then in model ρ we let the interarrival times be $\rho^{-1}U_k$, where $0 < \rho < 1$. Thus, $m_s = 1$ and $m_a = \rho^{-1}$, so that $m \equiv (1-\rho)/\rho$ and $W^* = b^2\rho/4(1-\rho)$ in (7).

Since the CLT underlies the heavy-traffic limit theory as well as the RQ formulation, it should not be surprising that we can make strong connections to heavy-traffic approximations. The new formulation in (6) is attractive because, with a natural choice of the constant b_x there, it matches the Kingman (1962) bound for the mean steady-state wait $E[W]$ in the $GI/GI/1$ stochastic model and so is asymptotically correct in heavy traffic, whereas that is not the case for (4) with a natural choice of b. To quantify the variability independent of the scale, let $c_s^2 \equiv \mathrm{Var}(V_1)/(E[V_1])^2 = \mathrm{Var}(V_1)$ and

$c_a^2 \equiv \mathrm{Var}(U_1)/(E[U_1])^2 = \rho^2 \mathrm{Var}(U_1)$ be the *squared coefficients of variation* (scvs). Let $\approx$ denote approximately equal, without any precise asymptotic meaning.

Corollary 1 (RQ Yields the Kingman Bound for $GI/GI/1$). *In the setting of* (6), *if we let* $b_x \equiv \beta\sqrt{\mathrm{Var}(X_1)}$ *and* $\beta \equiv \sqrt{2}$, *then* $b_x = \sqrt{2(c_s^2 + \rho^{-2}c_a^2)}$ *for the* $GI/GI/1$ *model with traffic intensity* ρ, *so that*

$$W^* \equiv W^*(\rho) = \frac{\mathrm{Var}(X_1)}{2|E[X_1]|} = \frac{\rho(c_s^2 + \rho^{-2}c_a^2)}{2(1-\rho)}, \tag{8}$$

which is the upper bound for $E[W]$ *in Theorem* 2 *of Kingman* (1962), *so that* $(1-\rho)W^*(\rho) \to (c_a^2 + c_s^2)/2$ *as* $\rho \uparrow 1$, *which supports the heavy-traffic approximation* $W^*(\rho) \approx \rho(c_a^2 + c_s^2)/2(1-\rho)$, *just as for* $E[W]$ *in the stochastic model. On the other hand, in the setting of* (4), *if we let* $b_s \equiv \beta\sqrt{\mathrm{Var}(V_1)}$ *and* $b_a \equiv \beta\sqrt{\mathrm{Var}(U_1)}$, *then we obtain* $b = b_s + b_a = \beta(c_s + \rho^{-1}c_a)$ *instead of* $b = \sqrt{b_s^2 + b_a^2} = \beta\sqrt{c_s^2 + \rho^{-2}c_a^2}$, *as needed.*

Remark 1 (The Significance for Approximations). The difference between the RQ solutions for (4) and (6) mentioned at the end of Corollary 1 can have serious implications for approximations; e.g., if $c_a^2 = c_s^2 = x$, then $(c_a^2 + c_s^2)/2 = x$, while $(c_a + c_s)^2/2 = 2x$, a factor of 2 larger. Hence, if we apply (4) with $b_a = b_s$ to the simple $M/M/1$ queue, one is forced to have a 100% error in heavy traffic. These two coincide only when at least one of b_a and b_s is 0 (i.e., in $D/GI/1$ or $GI/D/1$ models), and the percentage error is the largest when service times and arrival times have the same variability. Fortunately, robust optimization has flexibility that makes it possible to circumvent the difficulties in the form of the optimization in (4). For example, Bandi et al. (2015) use statistical regression in their section 7 to refine their solution to (4). Of course, such refinements complicate algorithms.

These RQ formulations provide insight into the rate of approach to steady state for the $GI/GI/1$ model, as captured by the relaxation time; see section III.7.3 of Cohen (1982) and section XIII.2 of Asmussen (2003). For RQ, steady state is achieved at a fixed time, whereas in the stochastic model, steady state is approached gradually, with the error $|E[W_n] - E[W]|$ typically being of order $O(n^{-3/2}e^{-n/r})$ as $n \to \infty$, where $r \equiv r(\rho)$ is called the relaxation time. As usual, we say $f(t)$ is $O(g(t))$ as $t \to \infty$, where f and g are positive real-valued functions, if $f(t)/g(t) \to c$ as $t \to \infty$, where $0 < c < \infty$.

Corollary 2 (Relaxation Time for the $GI/GI/1$ Queue). *With both* (4) *and* (6), *the place where the RQ supremum is attained is* $x^*(\rho) = O((1-\rho)^{-2})$ *as* $\rho \uparrow 1$, *which is the same order as the relaxation time in the* $GI/GI/1$ *model.*

Remark 2 (A Functional RQ to Expose the Impact of Dependence in the $G/G/1$ Model). The RQ problems in (4) and (6) can be considered instances of a *parametric RQ*, because they depend on the stochastic model only through a few parameters—in particular, (m_a, m_s, b_a, b_s) in (4) and (m, b_x) in (6). 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 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 (6) by

$$\mathcal{U}_f^x \equiv \{\tilde{X}\colon S_k^x \leqslant E[S_k^x] + b_x'\sqrt{\mathrm{Var}(S_k^x)},\ k \geqslant 0\}. \tag{9}$$

and similarly for the two constraints in (4). For the $GI/GI/1$ model, the new uncertainty set (9) is essentially equivalent to the previous one in (5), but they can be very different with dependence. It is significant that there are CLTs to motivate the form of the constraints in (9), just as there are in the i.i.d. case underlying (5). These supporting CLTs are reviewed here in Section EC.5. The CLT supports the spatial scaling by $\sqrt{\mathrm{Var}(S_k)}$ instead of $\sqrt{k}$, as we show in Section EC.5.3. 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. For brevity, we discuss this functional RQ for the waiting time in the EC because we will next develop such a functional RQ formulation for the continuous-time workload. As discovered in Fendick and Whitt (1989), it is convenient to focus on the steady-state workload when we want to expose the performance impact of the dependence among interarrival times and service times.

Remark 3 (Asymptotically Correct in Heavy Traffic for the $G/G/1$ Model). In Section EC.6.2 we observe that Corollary 1 can be extended, with the aid of Sections EC.5 and EC.6, to show that both the new parametric RQ in (6) and the new functional RQ with uncertainty set in (9) are asymptotically correct in heavy traffic for the more general stationary $G/G/1$ model, where we regard $\{(U_k, V_k)\}$ as a stationary sequence with the same mean values, including $E[V_k] = 1$ and $E[U_k] = \rho^{-1} > 1$ for all k. Now we must choose the parameter b_x appropriately to account for the dependence among the interarrival times and service times. Just as before, that can be motivated by the CLT, but now we need a CLT that accounts for the dependence, as in theorem 4.4.1 and section 9.6 of Whitt (2002); see Section EC.5.

Remark 4 (Connection to Mamani et al. 2016). At first glance, the connection to Mamani et al. (2016) may not be obvious, because we have introduced no explicit

covariances, like what appears in uncertainty set (6) in their section 2.5. The Lindley recursion in (1) here leads directly to the expression for the steady-state waiting time in terms of the partial sums S_k in (2), so it is natural for us to focus on the variances $\mathrm{Var}(S_k)$. However, the variances $\mathrm{Var}(S_k)$ in our uncertainty set (9) are variances of sums of random variables, which includes covariances of the summands X_j when these summands are not required to be independent. As indicated above, our uncertainty sets are motivated by CLTs, but CLTs without the usual independence assumption. The second paragraph of section 2.5 in Mamani et al. (2016) also mentions CLTs for dependent random variables but seems to be suggesting that the conditions are too restrictive to be useful. Unlike Mamani et al. (2016), the CLT and the heavy-traffic theory play a big role here to expose what properties of the model have the greatest impact upon the queue performance; see Section EC.5.

3. Robust Queueing for the Continuous-Time Workload

We now develop RQ formulations for the continuous-time workload in the single-server queue. We develop both a parametric RQ paralleling (6) and a functional RQ with an uncertainty set paralleling (9) in Remark 2.

The workload at time t is the amount of unfinished work in the system at time t; it is also called the virtual waiting time because it represents the waiting time a hypothetical arrival would experience at time t. The workload is more general than the virtual waiting time because it applies to any work-conserving service discipline. We consider the workload primarily because it can serve as a convenient, more tractable alternative to the waiting time.

We start by developing a reverse-time representation of the workload process paralleling (2). Then we develop both parametric and functional RQ formulations and give their solutions, which closely parallels Theorem 1. We then show that natural versions of both RQ formulations for the workload are exact for the $M/GI/1$ model and are asymptotically correct in both light traffic and heavy traffic for the general stationary $G/G/1$ model.

3.1. The Workload Process and Its Reverse-Time Representation

As before, we start with a sequence $\{(U_k, V_k)\}$ of interarrival times and service times. The arrival counting process can be defined by

$$A(t) \equiv \max\{k \geqslant 1: U_1 + \cdots + U_k \leqslant t\} \quad \text{for } t \geqslant U_1 \tag{10}$$

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

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k \quad \text{and} \quad N(t) \equiv Y(t) - t, \quad t \geqslant 0, \tag{11}$$

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

$$Z(t) = \Psi(N)(t) \equiv N(t) - \inf_{0 \leqslant s \leqslant t}\{N(s)\}, \quad t \geqslant 0. \tag{12}$$

As in section 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

$$Z(t) \equiv \sup_{0 \leqslant s \leqslant t}\{N(s)\}, \quad t \geqslant 0, \tag{13}$$

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 kth 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 Section 2, $Z(t)$ 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 section 6.3 of Sigman (1995).

3.2. Parametric and Functional RQ for the Steady-State Workload

Just as in Section 2, to create appropriate RQ formulations for the steady-state workload, it is helpful to have a reference stochastic model, which can be the stable stationary $G/G/1$ model, where such a steady-state workload is well defined. In discrete time, our formulation can be developed by scaling the interarrival times, assuming that $E[V_k] = E[U_k] = 1$ for all k for a base stationary sequence $\{(U_k, V_k)\}$ and introducing ρ by letting the interarrival times be $\rho^{-1}U_k$ when the traffic intensity is ρ, $0 < \rho < 1$. (That was done in Section 2, right after Theorem 1.) Now, in continuous time, we do essentially the same, but now we need to 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 $\{(A(t), Y(t)): t \geqslant 0\}$ with $E[A(t)] = E[Y(t)] = t$ for all $t \geqslant 0$ and introduce ρ by simple scaling via

$$A_\rho(t) \equiv A(\rho t) \quad \text{and} \quad Y_\rho(t) \equiv Y(\rho t), \quad t \geqslant 0 \text{ and } 0 < \rho < 1, \tag{14}$$

which implies that $E[A_\rho(t)] = E[Y_\rho(t)] = \rho t$ for all $t \geqslant 0$. Then $N_\rho(t) \equiv Y_\rho(t) - t$ and $Z_\rho(t) = \Psi(Y_\rho)(t)$, $t \geqslant 0$, just as in (11) and (12). With the reverse-time construction,

$Z_\rho(t)$ can be expressed as a supremum over the interval $[0,t]$, just as in (13).

Within that scaling framework, the natural parametric and functional (see Remark 2) uncertainty sets for the steady-state workload are, respectively,

$$\begin{aligned}\mathcal{U}_\rho^p &\equiv \{\tilde{N}_\rho\colon \mathbb{R}^+\to\mathbb{R}\colon \tilde{N}_\rho(s)\leqslant -(1-\rho)s + b_p\sqrt{s}, s\geqslant 0\} \quad \text{and}\\ \mathcal{U}_\rho &\equiv \mathcal{U}_\rho^f \equiv \big\{\tilde{N}_\rho\colon \mathbb{R}^+\to\mathbb{R}\colon \tilde{N}_\rho(s)\leqslant E[N_\rho(s)] \\ &\qquad + b_f\sqrt{\mathrm{Var}(N_\rho(s))}, s\geqslant 0\big\},\\ &= \big\{\tilde{N}_\rho\colon \mathbb{R}^+\to\mathbb{R}\colon \tilde{N}_\rho(s)\leqslant -(1-\rho)s \\ &\qquad + b_f\sqrt{\mathrm{Var}(N_\rho(s))}, s\geqslant 0\big\},\end{aligned} \tag{15}$$

where we regard $\tilde{N}_\rho \equiv \{\tilde{N}_\rho(s)\colon 0\leqslant s\leqslant t\}$ as an arbitrary real-valued function on $\mathbb{R}^+ \equiv [0,\infty)$, while we regard $\{N_\rho(s)\colon s\geqslant 0\}$ as the underlying stochastic process and $\{\mathrm{Var}(N_\rho(s))\colon s\geqslant 0\} = \{\mathrm{Var}(Y_\rho(s))\colon s\geqslant 0\}$ as its variance-time function, which can be either calculated for a stochastic model or estimated from simulation or system data; see Section 4.3. In (15), b_p and b_f are parameters to be specified.

Remark 5 (Choosing the Parameters b_p and b_f). The parameters b_p and b_f in (15) add a degree of freedom in the algorithm, but some choices lead to asymptotically correct values of the steady-state mean workload, while others do not. On the basis of Corollary 3 below, we will let $b=\sqrt{2}$ after this section.

Paralleling Section 2, the associated parametric and functional RQ formulations are ,

$$\begin{aligned}Z_{p,\rho}^* &\equiv \sup_{\tilde{N}_\rho\in\mathcal{U}_\rho^p}\sup_{s\geqslant 0}\{\tilde{N}_\rho(t)\},\\ Z_\rho^* \equiv Z_{f,\rho}^* &\equiv \sup_{\tilde{N}_\rho\in\mathcal{U}_\rho^f}\sup_{s\geqslant 0}\{\tilde{N}_\rho(t)\}.\end{aligned} \tag{16}$$

As in Section 2, our RQ formulations in (16) are motivated by a CLT but here for $Y_\rho(t)$ (which implies an associated CLT for $N_\rho(t)$), which we review in Section EC.5; in particular, see (EC.14) and (EC.16). The same reasoning as before yields the following analog of Theorem 1. The proof can be found in Section EC.7.

Theorem 2 (RQ Solutions for the Workload). *The solutions of the RQ optimization problems in* (16) *are*

$$Z_{p,\rho}^* = -(1-\rho)x^* + b_p\sqrt{x^*} = \frac{b_p^2}{4|1-\rho|}$$
$$\text{for } x^* \equiv x^*(\rho) = \frac{b_p^2}{4(1-\rho)^2} \tag{17}$$

and

$$Z_\rho^* \equiv Z_{f,\rho}^* = \sup_{s\geqslant 0}\big\{-(1-\rho)s + b_f\sqrt{\mathrm{Var}(Y_\rho(s))}\big\}. \tag{18}$$

We immediately obtain the following corollary, which states that the RQ formulation in (16) yields the exact mean steady-state workload for the $M/GI/1$ model.

Corollary 3 (Exact for $M/GI/1$). *For the $M/GI/1$ model, the total input process $\{Y_\rho(t)\colon t\geqslant 0\}$ in* (14) *is a compound Poisson process with $E[Y_\rho(t)] = \rho t$ and $\mathrm{Var}(Y_\rho(t)) = \rho t(c_s^2+1)$, so that $Z_{f,\rho}^* = Z_{p,\rho}^*$ if $b_p^2 = b_f^2\rho(c_s^2+c_a^2)$. If, in addition, $b_f\equiv\sqrt{2}$, then*

$$Z_{p,\rho}^* = Z_{f,\rho}^* = \frac{\rho(c_s^2+c_a^2)}{2(1-\rho)} = E[Z_\rho], \tag{19}$$

where $E[Z_\rho]$ is the mean steady-state workload in the $M/GI/1$ model with traffic intensity ρ.

This corollary suggests a natural choice of b_f in (15). From now on, we assume that $b_f=\sqrt{2}$ unless otherwise stated.

3.3. The Variance-Time Function for the Total Input Process

For further progress, we focus on the variance-time function $\mathrm{Var}(Y_\rho(t))$ in (18). As regularity conditions for $Y(t)$, we assume that $V(t)\equiv V_\rho(t)\equiv\mathrm{Var}(Y_\rho(t))$ is differentiable with derivative $\dot{V}(t)$ having finite positive limits as $t\to\infty$ and $t\to 0$; i.e.,

$$\begin{aligned}&\dot{V}(t)\to\sigma_Y^2 \quad \text{as } t\to\infty \quad \text{and}\\ &\dot{V}(t)\to\dot{V}(0)>0 \quad \text{as } t\to 0\end{aligned} \tag{20}$$

for an appropriate constant σ_Y^2. These assumptions are known to be reasonable; see section 4.5 of Cox and Lewis (1966), Fendick and Whitt (1989), and Section 4.3.

A common case in models for applications is to have positive dependence in the input process Y, which holds if

$$\begin{aligned}&\mathrm{Cov}(Y(t_2)-Y(t_1), Y(t_4)-Y(t_3))\geqslant 0\\ &\qquad\qquad \text{for all } 0\leqslant t_1<t_2\leqslant t_3<t_4.\end{aligned} \tag{21}$$

Negative dependence holds if the inequality is reversed. These are strict if the inequality is a strict inequality. From (17) and (18) of section 4.5 in Cox and Lewis (1966), which is restated in (48) and (49) of Fendick and Whitt (1989), with positive (negative) dependence, under appropriate regularity conditions, $\dot{V}(t)\geqslant 0$ and $\ddot{V}(t)\geqslant(\leqslant)0$.

Remark 6 (Example of Negative Dependence). Negative dependence in Y occurs if greater input in one interval tends to imply less input in another interval. Such negative dependence occurs when there is a specified number of arrivals in a long time interval, as in the $\Delta_{(i)}/GI/1$ model, where the arrival times (*not* interarrival times) are i.i.d. over an interval; see Honnappa et al. (2015). This phenomenon can also occur in

queues with arrivals by appointment, where there are i.i.d. deviations about deterministic appointment times; e.g., see Kim et al. (2017).

Theorem 3 (RQ Exposing the Impact of the Dependence). *Consider the functional RQ optimization for the steady-state workload in the general stationary $G/G/1$ queue with $\rho < 1$ formulated in (16) and solved in (18). Assume that (20) holds for the variance-time function $V(t) \equiv V_\rho(t) \equiv \mathrm{Var}(Y_\rho(t))$.*

(a) *For each ρ, $0 < \rho < 1$, there exists (possibly not unique) $x^* \equiv x^*(\rho)$, such that a finite maximum is attained at x^* for all $t \geqslant x^*$. In addition, $0 < x^* < \infty$ and x^* satisfies the equation*

$$(1-\rho) = \dot{h}(x), \quad \text{where } h(x) \equiv b'_z\sqrt{V(x)}. \tag{22}$$

The time x^* is unique if $h(x)$ is strictly concave or strictly convex—i.e., if $\dot{h}(x)$ is strictly increasing or strictly decreasing.

(b) *If there is positive (negative) dependence, as in (21) (with sign reversed), the variance function $V(x)$ is convex (concave), so that the function $h(x) \equiv \sqrt{V(x)}$ is concave. Moreover, a strict inequality is inherited. Thus, there exists a unique solution to the RQ if there is strict positive dependence or strict negative dependence. Moreover, the optimal time $x^*(\rho)$ is strictly increasing in ρ, approaching 1 as $\rho \uparrow 1$, so that $Z^*_\rho \to \dot{V}(\infty) = I_w(\infty) = \sigma^2_Y$ as $\rho \uparrow 1$.*

Proof. The inequalities can be satisfied as equalities just as before. There are finite values s_0 such that $\sqrt{V(s)} \leqslant \sqrt{2\sigma^2_Y s}$ for all $s \geqslant s_0$ by virtue of the limit in (20). (Also see (EC.1) and (EC.12).) That shows that the optimization can be regarded as being over closed bounded intervals. The assumed differentiability of V implies that it is continuous, which implies that the supremum is attained over the compact interval. Because $\dot{V}(x) \to \dot{V}(0) > 0$, we see that there exists a small s' such that

$$-(1-\rho)s + b'_z\sqrt{V(s)} \geqslant -(1-\rho)s + b'_z\sqrt{s\dot{V}(0)/2} > 0 \quad \text{for all } s \leqslant s'.$$

As a consequence, the maximum in (18) must be strictly positive and must be attained at a strictly positive time.

The results for $\sqrt{V(x)}$ with positive dependence follow from convexity properties of compositions. First, with positive dependence, $-\sqrt{V(x)}$ is a convex function of an increasing convex function and thus convex so that $\sqrt{V(x)}$ is concave. Second, with negative dependence, we have $V \geqslant 0$, $\dot{V}(t) \geqslant 0$ and $\ddot{V}(t) \leqslant (\leqslant) 0$. Thus, by direct differentiation,

$$\ddot{h}(x) = \frac{1}{\sqrt{V(x)}}\left(\frac{\ddot{V}(x)}{2} - \frac{\dot{V}(x)}{4V(x)}\right) \leqslant 0,$$

with strictness implying a strict inequality. □

4. The Indices of Dispersion for Counts and Work

The workload process is convenient not only because it leads to the continuous RQ optimization problem in (16) with a solution in (18) but also because the workload process scales with ρ in a more elementary way than the waiting times, as indicated in (14). By contrast, the scaling of the waiting times (specified in the first paragraph after Theorem 1) is more complicated because the interarrival times are scaled with ρ but the service times are not.

It is also convenient to relate the variances of the arrival counting process $A(s)$ and the cumulative work input process $Y(s)$ to associated continuous-time indices of dispersion, studied in Fendick and Whitt (1989) and Fendick et al. (1991). We define the *index of dispersion for counts* (IDC) associated with the rate-1 arrival process A as in section 4.5 of Cox and Lewis (1966) by

$$I_a(t) \equiv \frac{\mathrm{Var}(A(t))}{E[A(t)]} = \frac{\mathrm{Var}(A(t))}{t}, \quad t \geqslant 0 \tag{23}$$

and the *index of dispersion for work* (IDW) associated with the rate 1 cumulative input process Y by

$$I_w(t) \equiv \frac{\mathrm{Var}(Y(t))}{E[V_1]E[Y(t)]} = \frac{V(t)}{t}, \quad t \geqslant 0. \tag{24}$$

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. The reason for using these indices of dispersion is just like the reason for using the scvs (introduced before Corollary 1) instead of the variances. More generally, this is consistent with the well-established practice of carefully focusing on units in physics and engineering.

Fendick and Whitt (1989) show that the IDW I_w is intimately related to a scaled mean workload $c^2_Z(\rho)$, which can be defined by comparing to what it would be in the associated $M/D/1$ model; i.e.,

$$c^2_Z(\rho) \equiv \frac{E[Z_\rho]}{E[Z_\rho; M/D/1]} = \frac{2(1-\rho)E[Z_\rho]}{E[V_1]\rho} = \frac{2(1-\rho)E[Z_\rho]}{\rho}. \tag{25}$$

The normalization in (25) exposes the impact of variability separately from the traffic intensity. Hence, it should not be surprising that $c^2_Z(\rho)$ should be related to the IDW. Indeed, under regularity conditions (see Section EC.5.5), the following finite positive limits exist and are equal:

$$\lim_{t\to\infty}\{I_w(t)\} \equiv I_w(\infty) = \sigma^2_Y = c^2_Z(1) \equiv \lim_{\rho\to 1}\{c^2_Z(\rho)\}, \quad \text{and}$$
$$\lim_{t\to 0}\{I_w(t)\} \equiv I_w(0) = 1 + c^2_s = c^2_Z(0) \equiv \lim_{\rho\to 0}\{c^2_Z(\rho)\} \tag{26}$$

for $c_s^2 \equiv \mathrm{Var}(V_1)/E[V_1]^2$ and c_Y^2 in (20) and (EC.7). The limits for I_w above and the differentiability of I_w follow from the assumed differentiability for $V(t)$ and limits in (20). For $t \to 0$ and $\rho \to 0$, see section IV.A of Fendick and Whitt (1989).

The challenge is to relate $c_Z^2(\rho)$ to the IDW $I_w(t)$ for $0 < \rho < 1$ and $t \geqslant 0$. As observed by Fendick and Whitt (1989), a simple connection would be $c_Z^2(\rho) \approx I_w(t_\rho)$ for some increasing function t_ρ, reflecting that the impact of the dependence among the interarrival times and service times has impact on the performance of a queue over some time interval $[0, t_\rho]$, where t_ρ should increase as ρ increases. The extreme cases are supported by (26), but we want more information about the cases in between.

4.1. Robust Queueing with the IDW

To obtain more information, RQ can help. As a first step, we express the solution in (18) as

$$\begin{aligned} Z_\rho^* &= \sup_{s \geqslant 0} \left\{ -(1-\rho)s + b_f \sqrt{\mathrm{Var}(Y_\rho(s))} \right\} \\ &= \sup_{s \geqslant 0} \left\{ -(1-\rho)s + b_f \sqrt{\rho s I_w(\rho s)} \right\}, \end{aligned} \tag{27}$$

using (24). Making the change of variables $x \equiv \rho s$, we can write

$$Z_\rho^* = \sup_{x \geqslant 0} \left\{ -(1-\rho)x/\rho + b_f \sqrt{x I_w(x)} \right\}. \tag{28}$$

Clearly, from an algorithmic perspective, (28) is essentially the same as (18) and (27), but (28) is helpful for developing approximations and insights, including supporting theory. Our algorithm will exploit the one-dimensional optimization problem in (28), which is easy to solve given the IDW $I_w(x)$. We will discuss methods of estimating and calculating IDW in Sections 4.3 and 6.

To further relate the RQ solution in (28) to the steady-state workload in the $G/G/1$ queue, we define an RQ analog of the normalized mean workload in (25)—in particular,

$$c_{Z^*}^2(\rho) \equiv \frac{2(1-\rho)Z_\rho^*}{\rho}. \tag{29}$$

The RQ approach allows us to establish versions of the variability fixed-point equation suggested in (9), (15), and (127) of Fendick and Whitt (1989).

Theorem 4 (Restatement of Theorem 2 in Terms of the IDW). *Any optimal solution of the RQ in* (28) *is attained at* $s^*(\rho) \equiv x^*/\rho$, *where* $x^* \equiv x^*(\rho)$ *satisfies the equation*

$$x^* = \frac{b_f^2 \rho^2 I_w(x^*)}{4(1-\rho)^2} \left(1 + \frac{x^* \dot{I}_w(x^*)}{I_w(x^*)} \right)^2 \tag{30}$$

for b_f *in* (18). *The associated RQ optimal workload in* (28) *can be expressed as*

$$Z_\rho^* = \frac{b_f^2 \rho I_w(x^*)}{4(1-\rho)} \left(1 - \left(\frac{x^* \dot{I}_w(x^*)}{I_w(x^*)} \right)^2 \right), \tag{31}$$

which is a valid nonnegative solution provided that $x^* \dot{I}_w(x^*) \leqslant I_w(x^*)$. *If* $b_f = \sqrt{2}$, *then the associated scaled RQ workload satisfies*

$$c_{Z^*}^2(\rho) = I_w(x^*) \left(1 - \left(\frac{x^* \dot{I}_w(x^*)}{I_w(x^*)} \right)^2 \right). \tag{32}$$

Proof. Note that $x I_w(x) = V(x)$. Because we have assumed that $V(x)$ is differentiable, so too is I_w. We obtain (30) by differentiating with respect to x in (28) and setting the derivative equal to 0. After substituting (30) into (28), algebra yields (31). The limits in (20) imply that $x^* \dot{I}_w(x^*) \to 0$ and $I_w(x^*) \to I_w(\infty)$ as $\rho \to 1$. □

Given that $x \dot{I}_w(x) \to 0$ as $x \to \infty$, if $b_f = \sqrt{2}$, then it is natural to consider the approximation

$$\begin{aligned} x^*(\rho) &\approx \frac{\rho^2}{2(1-\rho)^2} I_w(x^*(\rho)) \quad \text{so that} \\ Z_\rho^* &\approx \frac{\rho I_w(x^*(\rho))}{2(1-\rho)} \quad \text{and} \quad c_{Z^*}^2(\rho) = I_w(x^*(\rho)). \end{aligned} \tag{33}$$

The first equation in (33) is a variability fixed-point equation of the form in suggested in (15) of Fendick and Whitt (1989).

4.2. Heavy-Traffic and Light-Traffic Limits

The following result shows the great advantage of doing RQ with (i) the continuous-time workload and (ii) the functional version of the RQ in (28). A proof is given in Section EC.7.

Theorem 5 (Heavy-Traffic and Light-Traffic Limits). *Under the regularity conditions assumed for the IDW* $I_w(t)$, *if* $b_f \equiv \sqrt{2}$, *then the functional RQ solution in* (28) *is an asymptotically correct characterization of steady-state mean workload both in heavy traffic* (*as* $\rho \uparrow 1$) *and light traffic* (*as* $\rho \downarrow 0$). *Specifically, we have the following supplement to* (26):

$$\begin{aligned} &\lim_{\rho \uparrow 1} c_{Z^*}^2(\rho) = I_w(\infty) = \lim_{\rho \uparrow 1} c_Z^2(\rho) \quad \text{and} \\ &\lim_{\rho \downarrow 0} c_{Z^*}^2(\rho) = I_w(0) = \lim_{\rho \downarrow 0} c_Z^2(\rho). \end{aligned} \tag{34}$$

Remark 7. Theorem 5 greatly generalizes results in Theorem 3(b) with both light and heavy traffic addressed in the general case beyond positive or negative correlations. We also note that the parametric RQ solution can be made correct in heavy traffic or in light traffic, as above, by choosing the parameter b_p appropriately, but both cannot be achieved simultaneously unless $I_w(\infty) = I_w(0)$.

4.3. Estimating and Calculating the IDW

For applications, it is significant that the IDW $I_w(t)$ used in Section 4 can readily be estimated from data from system measurements or simulation and calculated in a wide class of stochastic models. The time-dependent variance functions can be estimated from the time-dependent first and second moment functions, as discussed in section III.B of Fendick et al. (1991). Calculation depends on the specific model structure.

4.3.1. The $G/GI/1$ Model. If the service times are i.i.d. with a general distribution having mean τ and scv c_s^2 and are independent of a general stationary arrival process, then as indicated in (58) and (59) in section III.E of Fendick and Whitt (1989),

$$I_w(t) = c_s^2 + I_a(t), \quad t \geqslant 0, \tag{35}$$

where c_s^2 is the scv of a service time and I_a is the IDC of the general arrival process.

4.3.2. The Multiclass $\sum_i(G_i/G_i)/1$ Model. As indicated in (56) and (57) in section III.E of Fendick and Whitt (1989), if the input comes from independent sources, each with their own arrival process and service times, then the overall IDC and IDW are revealing functions of the component ones. Let λ_i be the arrival rate, let τ_i be the mean service time of class i, and let $\rho_i \equiv \lambda_i\tau_i$ be the traffic intensity for class i with $\lambda \equiv \sum_i \lambda_i$, $\tau \equiv \sum_i(\lambda_i/\lambda)\tau_i = 1$ so that $\rho = \lambda$. With our scaling conventions,

$$I_a(\lambda t) \equiv \frac{\mathrm{Var}(A(t))}{E[A(t)]} = \frac{\sum_i \mathrm{Var}(A_i(t))}{\lambda t} = \sum_i \frac{\lambda_i}{\lambda} I_{a,i}(\lambda_i t) \tag{36}$$

and

$$I_w(\lambda t) \equiv \frac{\mathrm{Var}(X(t))}{\tau E[X(t)]} = \frac{\sum_i V_i(t)}{\rho t} = \sum_i \frac{\rho_i\tau_i}{\rho\tau} I_{w,i}(\lambda_i t) \quad \text{for all } t \geqslant 0. \tag{37}$$

From (36) and (37), we see that I_a and I_w are convex combinations of the component $I_{a,i}$ and $I_{w,i}$ modified by additional time scaling.

4.3.3. The Multiclass $\sum_i G_i/GI/1$ Model. An important special case of Section 4.3.2 arising in open queueing networks is the $\sum_i G_i/GI/1$ model in which there are multiple general arrival streams coming to a queue where all arrivals experience common i.i.d. service times. We can combine (35) and (36) to get the expression for the IDW,

$$I_w(\lambda t) \equiv I_a(\lambda t) + c_s^2, \quad t \geqslant 0, \tag{38}$$

where $I_a(\lambda t)$ is given by (36). Of course, if all the component arrival streams are Poisson processes, then $I_a(\lambda t) = 1$ for all $t \geqslant 0$, but otherwise, the IDC $I_a(\lambda t)$ can be quite complicated.

4.3.4. The Balanced $\sum_i G_i/GI/1$ Model. An important special case of Section 4.3.3 is the balanced $\sum_i G_i/GI/1$ model in which the arrival process is the superposition of n i.i.d. non-Poisson processes each with rate ρ/n, so that the overall arrival rate is ρ, and asymptotic variability parameter is c_a^2. From the results above, we obtain

$$I_{a,n}(\rho t) = I_{a,1}(\rho t/n) \quad \text{and} \quad I_{w,n}(\rho t) = I_{w,1}(\rho t/n), \quad t \geqslant 0, \tag{39}$$

so that the superposition IDI and IDW differ from those of a single-component process only by the time scaling, but that time scaling involves both n and ρ.

As discussed in section 9.8 of Whitt (2002), this model is known to have complex behavior as a function of n and ρ, so that RQ may be helpful. In particular, under regularity conditions, (i) the superposition arrival process is known to be non-Poisson and nonrenewal, unless the component arrival streams are Poisson. (ii) If we let $n \to \infty$ but keep the total rate fixed, then the superposition process approaches a Poisson process. (iii) However, for any n, no matter how large, if we let $t \to \infty$, then the superposition process obeys the same CLT as a single component arrival process and so has asymptotic variability parameter c_a^2. Thus, we have $I_a(0) = 1$ and $I_a(\infty) = c_a^2$, but $I_a(t)$ depends on n and ρ in a complicated way for $0 < t < \infty$.

As shown in Whitt (1985), important insight can be gained by considering the *joint limit* as $n \uparrow \infty$ and $\rho \uparrow 1$. It turns out the asymptotic behavior depends on the limit of $n(1-\rho)^2$. The separate limits occur if that limit is either infinite or zero. A complex interaction occurs at finite limits. We will show that RQ provides important insight when we conduct simulation experiments for this model in Section 5.1.

4.3.5. The IDCs for Common Arrival Processes. The two previous subsections show that for a large class of models the main complicating feature is the IDC of the arrival process from a single source. The only really simple case is a Poisson arrival process with rate λ. Then $I_a(t) = 1$ for all $t \geqslant 0$. A compound (batch) Poisson process is also elementary because the process Y has independent increments; then the arrival process itself is equivalent to M/GI source. However, for a large class of models, the variance $\mathrm{Var}(A(t))$ and thus the IDC $I_a(t)$ can either be calculated directly or be characterized via their Laplace transforms and thus calculated by inverting those transforms and approximated by performing asymptotic analysis. For all models, we assume that the processes A and Y have stationary increments.

An important case for A is the renewal process; to have stationary increments, we assume that it is the equilibrium renewal process, as in section 3.5 of Ross (1996). Then $\mathrm{Var}(A(t))$ can be expressed in terms of the renewal function, which in turn can be related

to the interarrival-time distribution and its transform. The explicit formulas for renewal processes appear in (14), (16), and (18) in section 4.5 of Cox (1962). The required numerical transform inversion for the renewal function is discussed in section 13 of Abate and Whitt (1992). The hyperexponential (H_2) and Erlang (E_2) special cases are described in section III.G of Fendick and Whitt (1989).

It is also possible to carry out similar analyses for much more complicated arrival processes. Neuts (1989) applies matrix-analytic methods to give explicit representations of the variance $\mathrm{Var}(A(t))$ for the versatile Markovian point process or Neuts process; see section 5.4, especially theorem 5.4.1. Explicit formulas for the Markov-modulated Poisson process are given on pages 287–289.

All of these explicit formulas above have the asymptotic form

$$\mathrm{Var}(A(t)) = \sigma_A^2 t + \zeta + O(e^{-\gamma t}) \quad \text{as } t \to \infty.$$

5. Simulation Comparisons

We illustrate how the new RQ approach can be used with system data from queueing networks by applying simulation to analyze two common but challenging network structures in Figure 1: (i) a queue with a superposition arrival process and (ii) several queues in series. The specific examples are chosen to capture a known source of difficulty: there is complex dependence in the arrival process to the queue, so that the relevant variability parameter of the arrival process at the queue can depend strongly on the traffic intensity of that queue, as discussed in Whitt (1995). Our RQ approximations are obtained by applying (28) after estimating the IDC and applying (35).

5.1. A Queue with a Superposition Arrival Process

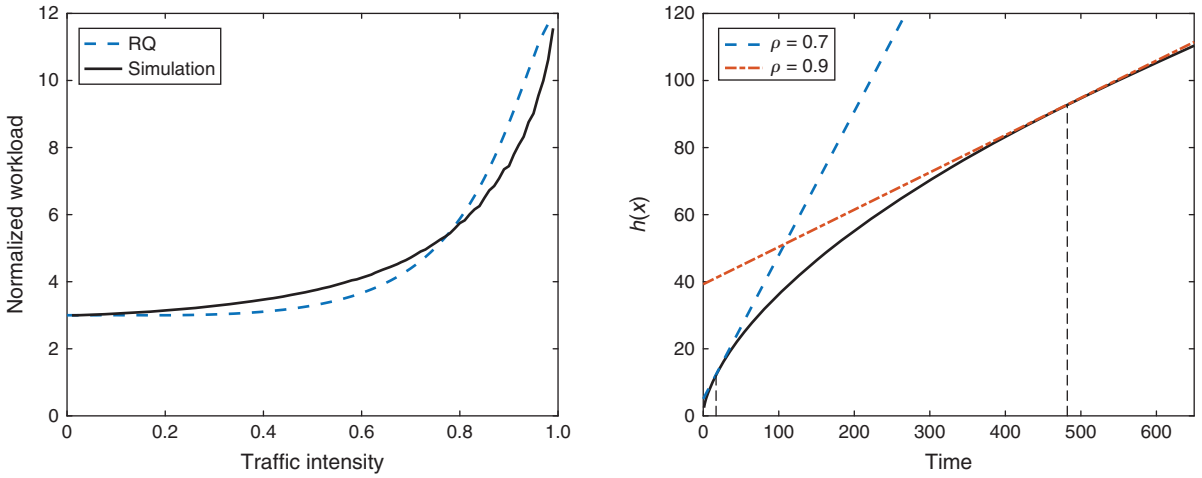
We start by looking at an example of a balanced $\sum_i G_i/GI/1$ model from Section 4.3.4, where (39) can be applied. Let the rate 1 arrival process A be the superposition of $n = 10$ i.i.d. renewal processes, each with rate $1/n$, where the times between renewals have a lognormal distribution with mean n and scv $c_a^2 = 10$. Let the service-time distribution be hyperexponential (H_2), a mixture of two exponential distributions with mean 1, $c_s^2 = 2$, and balanced means as on page 137 of Whitt (1982). Then (39) and (26) imply that the IDW has limits $I_w(0) = 1 + c_s^2 = 3$ and $I_w(\infty) = c_a^2 + c_s^2 = 12$, so that the IDW is not nearly constant.

The left panel of Figure 2 shows a comparison between the simulation estimate of the normalized workload $c_Z^2(\rho)$ in (25) and the approximation $c_{Z^*}^2(\rho)$ in (29) for this example. We make two important observations: (i) the normalized mean workload $c_Z^2(\rho)$ in (25) as a function of ρ is not nearly constant, and (ii) there is a close agreement between the RQ approximation $c_{Z^*}^2(\rho)$ in (29) and the direct simulation estimate; the close agreement for all traffic intensities is striking. It is important to note that the parametric RQ approximations produce constant approximations and so cannot be simultaneously good for all traffic intensities.

For this example, we see that $c_Z^2(\rho) \approx 3$ for $\rho \leqslant 0.5$, which is consistent with the Poisson approximation for the arrival process and the associated $M/G/1$ queue, where $c_Z^2(\rho) = 3$ for all ρ, but the normalized workload increases steadily to 12 after $\rho = 0.5$, as explained in section 9.8 of Whitt (2002).

The estimates for Figure 2 were obtained for ρ over a grid of 99 values, evenly spaced between 0.01 and 0.99. Similarly, the RQ optimization was performed using (28) with a discrete-time estimate of the IDW. By doing

Figure 2. (Color online) Left: A Comparison Between Simulation Estimates of the Normalized Mean Workload $c_Z^2(\rho)$ in (25) and Its Approximation $c_{Z^*}^2(\rho)$ in (29) as a Function of ρ for the $\sum_i^n GI_i/H_2/1$ Model with $c_s^2 = 2$ and a Superposition of n i.i.d. Lognormal Renewal Arrival Processes for $n = 10$ and $c_a^2 = 10$; Right: Graphical RQ Solution Showing $h(x) \equiv \sqrt{2xI_w(x)}$ and the Tangent Line with Slope $(1-\rho)/\rho$ at $x^* \approx 482$ for $\rho = 0.9$ and at $x^* \approx 17$ for 0.7, as Dictated by (22)



multiple runs, we ensured that the statistical variation was not an issue. For the main simulation of the arrival process and the queue we used 5×10^6 replications, discarding a large initial portion of the workload process to ensure that the system is approximately in steady state. (The component renewal arrival processes thus can be regarded as equilibrium renewal processes, as in section 3.5 of Ross 1996.) We let the run length and amount discarded be increasing in ρ, as dictated by Whitt (1989). We provide additional details about our simulation methodology in the appendix.

5.2. A Series of Ten Queues

This second example is a variant of examples in Suresh and Whitt (1990), exposing the complex impact of variability on performance in a series of queues if the external arrival process and service times at a previous queue have very different levels of variability. This example has 10 single-server queues in series. The external arrival process is a rate 1 renewal process with H_2 interarrival times having $c_a^2=5$. The first nine queues all have Erlang service times with $c_a^2=0.5$ denoted by E_2, i.e., the sum of two i.i.d. exponential random variables. The first eight queues have mean service time and thus traffic intensity 0.6, while the ninth queue has mean service time and thus traffic intensity 0.95. The last (10th) queue has an exponential service-time distribution with mean and traffic intensity ρ; we explore the impact of ρ on the performance of that last queue.

The Erlang services act to smooth the arrival process at the last queue. Thus, for sufficiently low traffic intensities ρ at the last queue, the last queue should behave essentially the same as a $E_2/M/1$ queue, which has $c_a^2=0.5$, but as ρ increases, the arrival process at the last queue should inherit the variability of the external arrival process and behave like an $H_2/M/1$ queue with scv $c_a^2=5$. This behavior is substantiated by Figure 3, which compares simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (25) at the last queue of 10 queues in series as a function of the mean service time and traffic intensity ρ there with the corresponding values in the $E_2/M/1$ queue (left panel) and with the RQ approximation $c_{Z^*}^2(\rho)$ in (29) (right panel). The left panel of Figure 3 shows that the last queue behaves like a $E_2/M/1$ queue for all traffic intensities $\leqslant 0.8$ but then starts behaving more like an $H_2/M/1$ queue as the traffic intensity approaches the value 0.95 at the ninth queue. The right panel of Figure 3 shows that RQ successfully captures this phenomenon and provides an accurate approximation for all ρ.

To elaborate on this series-queue example, we show the IDW for the last queue in Figure 4. The plot shows the IDW assuming continuous-time stationarity (which we use) together with the plot using the discrete-time Palm stationarity (see Sigman 1995) over the long interval $[10^{-2},10^5]$ in log scale. The good performance in Figure 3 for small values of ρ depends on using the proper (continuous-time) version.

We conclude this example by illustrating the discrete-time approach for approximating the expected steady-state waiting time $E[W]$ using the RQ optimization in (6) with uncertainty set in (9). Figure 5 is the discrete analog of Figure 3. Figure 5 compares simulation estimates of the normalized mean waiting time $c_W^2(\rho)$, defined just as in (25), at the last queue of 10 queues in series as a function of the mean service time and traffic intensity ρ there with the corresponding values in the $E_2/M/1$ queue (left) and with the RQ approximation $c_{W^*}^2(\rho)$, defined just as in (29). Figures 5 and 3 look

Figure 3. (Color online) A Comparison Between Simulation Estimates of the Normalized Mean Workload $c_Z^2(\rho)$ in (25) at the Last Queue of the 10 Queues in Series with Highly Variable External Arrival Process, but Low-Variability Service Times, as a Function of the Mean Service Time and Traffic Intensity ρ There with the Corresponding Value in the $E_2/M/1$ Queue (Left) and with the RQ Approximation $c_{Z^*}^2(\rho)$ in (29) (Right)

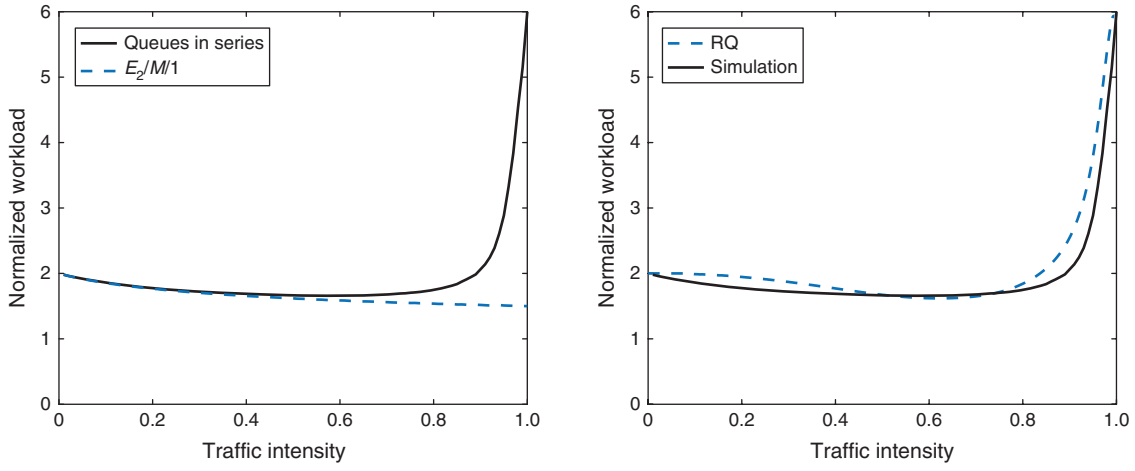
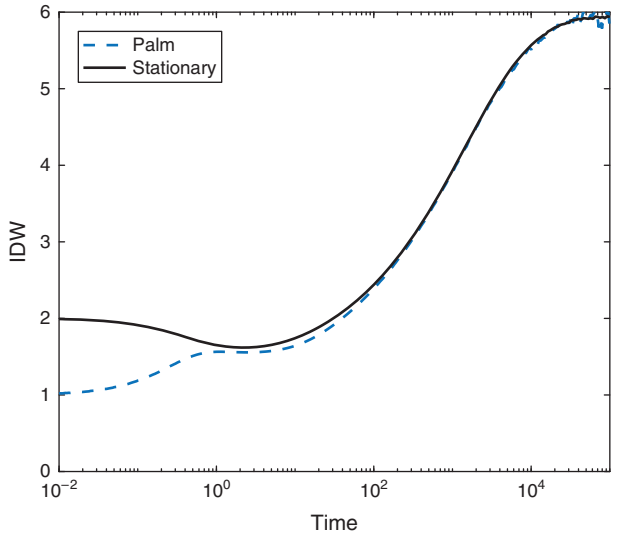


Figure 4. (Color online) The IDW at the Last Queue Over the Interval $[10^{-2}, 10^{5}]$ in Log Scale



Note. The continuous-time stationary version used for RQ with the workload is contrasted with the discrete-time Palm version.

similar, except that there is a significant difference for small values of ρ. In general, we do not expect RQ to be effective for extremely low ρ because (i) the CLT is not appropriate for only a few summands, and (ii) the mean waiting time is known to depend on other factors when ρ is small. The mean waiting time and mean workload actually are quite different in light traffic; see section IV.A of Fendick and Whitt (1989). As explained there, the mean workload tends to be more robust to model detail.

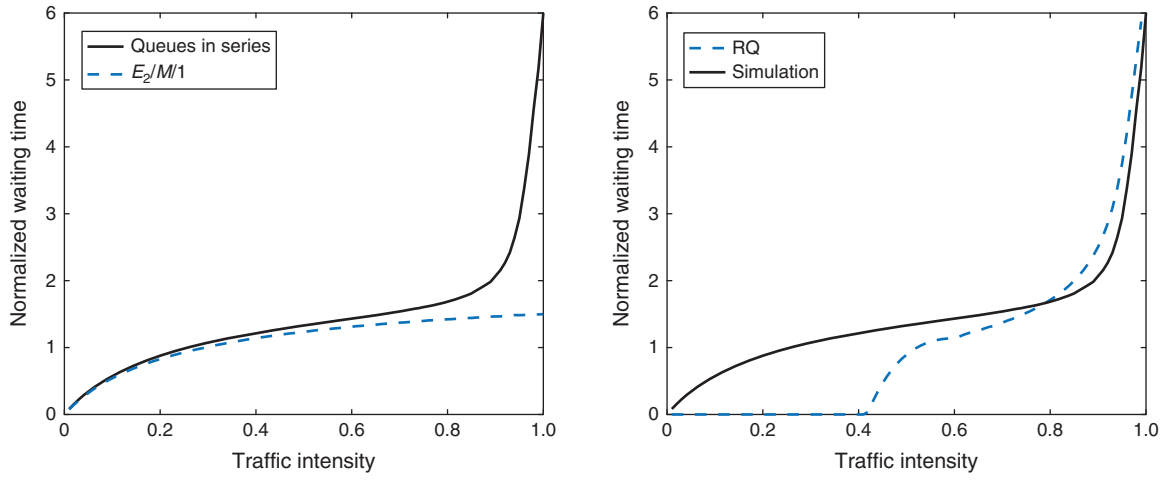
6. An IDC Framework for a New RQNA

A main contribution of Bandi et al. (2015) was to develop a full RQNA. While we have established good RQ results for one single-server queue, it still remains to develop a full RQNA exploiting the indices of dispersion and the results in the previous sections. To conclude this paper, we propose a candidate framework in which we hope to develop an initial IDC-based RQNA.

To start, we make several simplifying assumptions: (i) all queues are single-server queues with unlimited waiting space and the FCFS discipline; (ii) with m queues, the service times at these queues come from m independent sequences of i.i.d. random variables, independent of all the external arrival processes, where these service times have finite means and variances; (iii) each queue has its own external arrival process (which may be null), assuming that each is a general stationary point process; (iv) these m external arrival processes are mutually independent and exogenous, each having a finite arrival rate, with the arrival process satisfying a functional central limit theorem with a Brownian motion limit; (v) as in the basic form of QNA in Whitt (1983), we let departures be routed to other queues or out of the network by Markovian routing, independent of the rest of the model history; and (vi) given that the traffic-rate equations are used to find the net arrival rate at each queue, as in section 4.1 of Whitt (1983), the resulting traffic intensities satisfy $\rho_i < 1$ for all i, so that the final open network produces a stable general stationary $(G/GI/1)^m$ stochastic network model, which has a proper steady-state distribution.

As discussed in section 2.3 of Whitt (1983) and Segal and Whitt (1989), practical applications require much

Figure 5. (Color online) Contrasting the Discrete-Time and Continuous-Time Views: The Analog of Figure 3 for the Waiting Time



Note. Simulation estimates of the normalized mean waiting time $c_W^2(\rho)$, defined as in (25), at the last queue of the 10 queues in series with highly variable external arrival process, but low-variability service times, as a function of the mean service time and traffic intensity ρ there with the corresponding value in the $E_2/M/1$ queue (left) and with the RQ approximation $c_{W^*}^2(\rho)$, defined as in (29) (right).

more complicated models—e.g., including multiserver queues, non-FCFS disciplines and, as in section 2.3 of Whitt (1983), input by classes with basic routes that must be converted into the framework above—but here we suggest the $(G/GI/1)^m$ model above as a candidate reference stochastic model in which we want to develop an initial RQNA.

We propose going beyond QNA by letting the variability of each arrival process, external or internal, be partially characterized by its rate and IDC. Let the net arrival process at queue i have rate λ_i and IDC $I_{a,i}(t)$. We let the service-time cumulative distribution function (cdf) G_i at queue i be partially characterized by its mean τ_i and scv c_s^2, but we might use the full cdf G_i. By (35), the associated IDW is then $I_{w,i}(t) = I_{a,i}(t) + c_{s,i}^2$, $t \geq 0$. Thus, we can approximate the mean steady-state workload at queue i, $E[Z_i(\rho_i)]$ for each i, by solving the one-dimensional RQ optimization problem in (28). We consider that as the initial objective, even though we want to extend the algorithm to develop a full performance description. As a first cut to describe network performance, we would follow section VI of Whitt (1983).

For the $(G/GI/1)^m$ model introduced above, we specify the service time at queue i by its mean τ_i and scv $c_{s,i}^2$, as in QNA, but now we specify the external arrival process at queue i by its rate $\lambda_{o,i}$ and IDC $\{I_{a,o,i}(t): t \geq 0\}$, with o designated from outside. Paralleling QNA, the IDC-based RQNA would apply the familiar traffic-rate equations to determine the net arrival rate λ_i at queue i for each i, just as in section 4.1 of Whitt (1983), and associated traffic variability equations, based on a network calculus for the three operations—(i) superposition or merging, (ii) splitting, and (iii) flow through a queue or departure—to determine the final net IDC $I_{a,i}(t)$ at queue i for each i.

With the framework above, it suffices to specify and apply a network calculus to determine the IDC of the net arrival process to each queue. The difficult superposition operation (for component streams assumed to be mutually independent) is already covered by Section 4.3.3 here and has shown to be effective for approximating the mean workload in Section 5.1.

For splitting, as in QNA we assume independent splitting, with each customer routed in a given direction according to independent Bernoulli random variables. For independent splitting, we can express the split counting process $B(t)$ given the original counting process $A(t)$ by the random sum

$$B(t) = \sum_{i=1}^{A(t)} X_i, \tag{40}$$

where $\{X_i\}$ is i.i.d. and independent of $A(t)$ with $P(X_i = 1) = p = 1 - P(X_i = 0)$. Under those regularity conditions, we can apply the conditional variance formula to show that the IDC of the split stream can be represented exactly by

$$I_B(t) = pI_A(t) + 1 - p, \quad t \geq 0, \tag{41}$$

which is analogous to (36) in Whitt (1983).

Finally, it remains to treat the flow through a $G/GI/1$ queue. Of course, the rate out is just the rate in, so it suffices to calculate the IDC $I_d(t)$ for the departure process. We propose a candidate approximation that can serve as a basis for a full RQNA, but it remains to be more thoroughly tested and refined. In particular, a candidate approximation for the IDC $I_d(t)$ of the departure process from a $G/GI/1$ queue is

$$I_{d,\rho}(t) \approx w_\rho(t) I_a(t) + (1 - w_\rho(t)) I_s(t), \tag{42}$$

where $I_s(t)$ is the IDC of the equilibrium renewal process with specified service-time distribution, $w_\rho(t)$, $0 \leqslant w_\rho(t) \leqslant 1$, is a weight function, which depends on the traffic intensity ρ. Preliminary study indicates that the weight function might be

$$w_\rho(t) \equiv w(c(1-\rho)^2 t), \quad t \geqslant 0, \text{ where } w(t) \equiv 1 - e^{-\sqrt{t}} \tag{43}$$

and c is a properly chosen scale parameter; here, c is chosen to be 0.25. The component IDCs $I_a(t)$ and $I_s(t)$ in (42) can readily be estimated from simulations or calculated, as indicated in Section 4.3.5. The IDC of the equilibrium renewal process $I_s(t)$ can be obtained from the associated variance function via $I_s(t) = V(t)/t$, assuming that it has rate 1. In turn, the variance function of the rate 1 equilibrium renewal process is

$$V(t) = \int_0^t (1 + 2m(u) - 2u)\, du, \tag{44}$$

where $m(t)$ is the renewal function (mean function of the standard renewal process), which can be calculated by numerical transform inversion, given the Laplace transform of the service-time distribution, as discussed in section 13 of Abate and Whitt (1992).

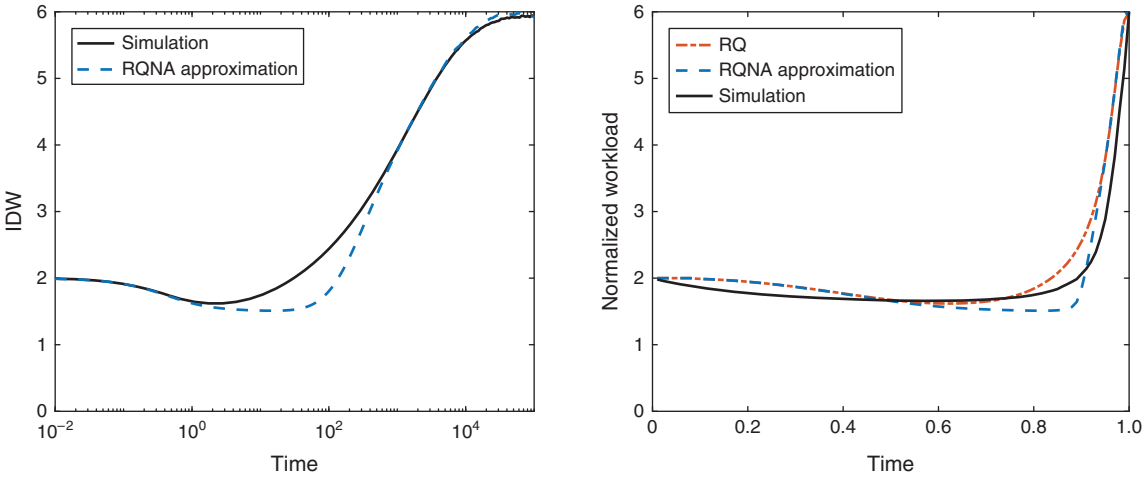
To show that this approach for approximating the IDC $I_d(t)$ has promise, we apply it to the series queue example in Section 5.2. Recall that the arrival process is an H_2 renewal process, while the service distribution at the first eight nodes is Erlang E_2 with traffic intensity 0.6 and the ninth node has a traffic intensity of 0.95. The IDCs for H_2 and E_2 are given in examples 3.1 and 3.2 of Fendick and Whitt (1989).

From (42), we iteratively obtain the approximation for the IDC of the departures from the ninth queue

$$I_{9,d,\rho}(t) \approx w_{\rho_1}^8(t) w_{\rho_9}(t) I_a(t) + (1 - w_{\rho_1}^8(t) w_{\rho_9}(t)) I_s(t). \tag{45}$$

This framework decomposes the IDC of the departure from the ninth queue into combinations of the IDC

Figure 6. (Color online) Left: Contrasting the IDW of Departure Process from the Ninth Queue from Simulation and the IDW Approximation Obtained from the Candidate RQNA Framework for the Example in Section 5.2; Right: Simulation Estimation of the Steady-State Mean Workload, the RQ Approximation in Section 5.2, and the RQNA Approximation



of the external arrival process and the IDC of the service renewal process. Figure 6 reports the approximation obtained from the RQNA framework for the IDW at the last queue in contrast with the one obtained from simulation, as well as the RQNA approximation of the steady-state mean workload at the last queue as a function of traffic intensity. Work is under way to develop and test the approximation for the IDC of the stationary departure process from a $G/GI/1$ queue and a full IDC-based RQNA for the $(G/GI/1)^m$ model.

References

Abate J, Whitt W (1992) The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10(1–2):5–88.

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

Bandi C, Bertsimas D (2012) Tractable stochastic analysis in high dimensions via robust optimization. *Math. Programming* 134(1): 23–70.

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

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Rev.* 53(3):464–501.

Bitran GR, Tirupati D (1988) Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Management Sci.* 34(1):75–100.

Cohen JW (1982) *The Single Server Queue*, 2nd ed. (North-Holland, Amsterdam).

Cox DR (1962) *Renewal Theory* (Methuen, London).

Cox DR, Lewis PAW (1966) *The Statistical Analysis of Series of Events* (Methuen, London).

Disney RL, Konig D (1985) Queueing networks: A survey of their random processes. *SIAM Rev.* 27(3):335–403.

Fendick KW, Whitt W (1989) Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proc. IEEE* 77(1):171–194.

Fendick KW, Saksena V, Whitt W (1989) Dependence in packet queues. *IEEE Trans. Comm.* 37(11):1173–1183.

Fendick KW, Saksena V, Whitt W (1991) Investigating dependence in packet queues with the index of dispersion for work. *IEEE Trans Comm.* 39(8):1231–1244.

Heffes H (1980) A class of data traffic processes-covariance function characterization and related queueing results. *Bell System Tech. J.* 59(6):897–929.

Heffes H, Luantoni D (1986) A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. Selected Areas Comm.* 4(6):856–868.

Honnappa H, Jain R, Ward A (2015) A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems* 80(1–2):71–103.

Kim S, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing Service Oper. Management* 16(3):464–480.

Kim S, Whitt W, Cha WC (2017) A data-driven model of an appointment-generated arrival processes at an outpatient clinic. *INFORMS J. Comput.* Forthcoming.

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

Klincewicz J, Whitt W (1984) On approximations for queues, II: Shape constraints. *AT&T Bell Laboratories Tech. J.* 63(1):115–138.

Lindley DV (1952) The theory of queues with a single server. *Math. Proc. Cambridge Philos. Soc.* 48(2):277–289.

Loynes RM (1962) The stability of a queue with non-independent inter-arrival and service times. *Math. Proc. Cambridge Philos. Soc.* 58(3):497–520.

Mamani H, Nassiri S, Wagner MR (2016) Closed-form solutions for robust inventory management. *Management Sci.* 62(3):1–20.

Moon I, Gallego G (1994) Distribution free procedures for some inventory models. *J. Oper. Res. Soc.* 45(6):651–658.

Neuts MF (1989) *Structured Stochastic Matrices of M/G/1 Type and Their Application* (Marcel Dekker, New York).

Ross SM (1996) *Stochastic Processes*, 2nd ed. (John Wiley & Sons, New York).

Scarf H (1958) A min-max solution of an inventory problem. Karlin S, Arrow K, Scarf H, eds. *Studies in the Mathematical Theory of Inventory and Production* (Stanford University Press, Stanford, CA), 201–209.

Segal M, Whitt W (1989) A queueing network analyzer for manufacturing. Bonatti M, ed. *Proc. 12th Internat. Teletraffic Congress* (Elsevier, Amsterdam), 1146–1152.

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

Sriram K, Whitt W (1986) Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Selected Areas Comm.* 4(6):833–846.
Suresh S, Whitt W (1990) The heavy-traffic bottleneck phenomenon in open queueing networks. *Oper. Res. Lett.* 9(6):355–362.
Whitt W (1982) Approximating a point process by a renewal process: Two basic methods. *Oper. Res.* 30(1):125–147.
Whitt W (1983) The queueing network analyzer. *Bell Laboratories Tech. J.* 62(9):2779–2815.
Whitt W (1984a) On approximations for queues, I. *AT&T Bell Laboratories Tech. J.* 63(1):115–137.
Whitt W (1984b) On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Laboratories Tech. J.* 63(1): 163–175.
Whitt W (1985) Queues with superposition arrival processes in heavy traffic. *Stochastic Processes Their Appl.* 21(1):81–91.
Whitt W (1989) Planning queueing simulations. *Management Sci.* 35(11):1341–1366.
Whitt W (1995) Variability functions for parametric-decomposition approximations of queueing networks. *Management Sci.* 41(10): 1704–1715.
Whitt W (2002) *Stochastic-Process Limits* (Springer, New York).
Whitt W, You W (2016) Time-varying robust queueing. Working paper, Columbia University, New York.

Ward Whitt is a professor in the Industrial Engineering and Operations Research Department at Columbia University. A major focus of his early work was the Queueing Network Analyzer performance analysis software tool, which is described in a 1983 paper in the *Bell Labs Technical Journal*. His new research explores ways to develop more effective approximations, drawing on new robust optimization methods as well as previous heavy-traffic limits and indices of dispersion.

Wei You is a doctoral student in the Industrial Engineering and Operations Research Department at Columbia University. His primary research focus is on queueing theory, applied probability, and their applications to service systems using stochastic modeling, optimization, and simulation.