Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

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

Ward Whitt
Industrial Engineering and Operations Research, Columbia University, ww2040@columbia.edu
Wei You
Industrial Engineering and Operations Research, Columbia University, wy2225@columbia.edu

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 by Bandi, Bertsimas and Youssef (2015) 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.

Key words: robust queueing, queueing approximations, dependence among interarrival times and service times, indices of dispersion, heavy traffic, queueing network analyzer

History: October 10, 2016

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 com-

plex 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 like 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) emphasize 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, we too ultimately 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 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.

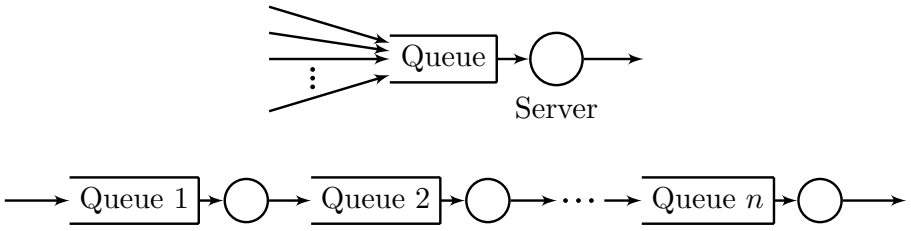

Figure 1 Common queueing network structure that can induce dependence among interarrival times: superpositions of arrival processes (top) and flow through a series of queues (bottom)

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 §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 §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 over-dispersion, e.g., see Kim and Whitt (2014) and references there. In 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 non-deterministic 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 treated by 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 §3.3 and §4. Even though we only consider a single queue here, in §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 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 captures 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 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 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.

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. They point to early inventory work by Scarf (1958) and then Moon and Gallego (1994). The new RQ work is also related to Whitt (1984a), which used optimization subject to constraints on the first two moments to understand the range of possible values in the performance approximations in QNA. 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 like 20% relative error, but that we cannot hope to obtain extraordinarily high accuracy, such as 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 §9.6 of Whitt (2002).

1.4. Organization of the Paper

In §2, after reviewing RQ for the steady-state waiting time in the single-server queue from §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 §3 we introduce new parametric and functional RQ formulations for the continuous-time workload process and characterize their solutions. In §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 §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. In §6 we introduce a new framework for applyiing the results in this paper to develop a new RQNA that better captures the dependence. Finally, in §7 we draw conclusions. Additional supporting material appears in the e-companion, in particular, (i) additional discussion, (ii) additional theoretical support, (iii) more results for the waiting time and (iv) more simulation examples.

2. Robust Queueing for the Steady-State Waiting Time

We start by reviewing the RQ developed in §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

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

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.,

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

using reverse-time indexing with $S_k \equiv X_1 + \cdots + X_k$ and $X_k \equiv V_{n-k} - U_{n-k}$, $1 \le k \le n$ 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).

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)

$$\mathcal{U}^a \equiv \{\tilde{U} \in \mathbb{R}^\infty : S_k^a \geq km_a - b_a\sqrt{k},\, k \geq 0\} \quad \text{and} \quad \mathcal{U}^s = \{\tilde{V} \in \mathbb{R}^\infty : S_k^s \leq km_s + b_s\sqrt{k},\, k \geq 0\}, \quad (3)$$

where $\tilde{U} \equiv \{U_k : k \geq 1\}$ and $\tilde{V} \equiv \{V_k : k \geq 1\}$ are arbitrary sequences of real numbers in $\mathbb{R}^\infty$, $S_k^a \equiv U_1 + \cdots + U_k$ and $S_k^s \equiv V_1 + \cdots + V_k$, $k \geq 1$, $S_0 \equiv 0$, 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

$$W^* \equiv \sup_{\tilde{U} \in \mathcal{U}^a} \sup_{\tilde{V} \in \mathcal{U}^s} \max_{k \geq 0} \{S_k^s - S_k^a\}. \quad (4)$$

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 independent and identically distributed (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 provided 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 $0 < \alpha \leq 2$, as we would expect from §§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)

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

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 (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

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

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

The RQ formulations in (4) and (6) are attractive because the optimizations 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 §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*

$$W^* = \sup_{k \ge 0}\{-mk + b\sqrt{k}\} \le \sup_{x \ge 0}\{-mx + b\sqrt{x}\} = -mx^* + b\sqrt{x^*} = \frac{b^2}{4m} \quad \textit{for} \quad x^* = \frac{b^2}{4m^2}, \tag{7}$$

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^* *for all* $n \ge x^*$.

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 $W_n^* = b^2\rho/4(1-\rho)$ 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. Let $c_s^2 \equiv Var(V_1)/(E[V_1])^2 = Var(V_1)$ and $c_a^2 \equiv Var(U_1)/(E[U_1])^2 = \rho^2 Var(U_1)$ be the squared coefficients of variation (scv's).

COROLLARY 1. (*RQ yields the Kingman bound for* $GI/GI/1$) *In the setting of* (6)*, if we let* $b_x \equiv \beta\sqrt{Var(X_1)}$ *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*

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

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* $b_s \equiv \beta\sqrt{Var(V_1)}$ *and* $b_a \equiv \beta\sqrt{Var(U_1)}$*, 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 $M/M/1$ queues, one is forced to have a 100% error in heavy traffic. These two coincides 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.

Exploiting the flexibility of robust optimization, Bandi et al. (2015) use statistical regression in their §7 to refine the solution to (4). Clearly, the regression step makes the overall algorithm much more complicated. If regression is used, it is helpful to start with a good functional form. ■

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 §III.7.3 of Cohen (1982) and §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 $G/G/1$ model as a function of the traffic intensity ρ by introducing a new *functional RQ* formulation. We replace the uncertainty set in (6) by

$$\mathcal{U}_f^x \equiv \{\tilde{X} : S_k^x \le E[S_k^x] + b'_x\sqrt{Var(S_k^x)}, \quad k \ge 0\}. \tag{9}$$

and similarly for the two constraints in (4). The constraints in (9) also can be motivated by a CLT, but with spatial scaling by $\sqrt{Var(S_k)}$ instead of $\sqrt{k}$, as we show in §EC.4.3. 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 §EC.5.2 we observe that Corollary 1 can be extended, with the aid of §EC.4 and §EC.5, 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 §9.6 of Whitt (2002); see §EC.4.

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 function 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

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

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

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

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

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

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

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

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 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 §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 §6.3 of Sigman (1995).

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

Just as in §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 §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 $\{(A(t), Y(t)) : t \geq 0\}$ with $E[A(t)] = E[Y(t)] = 1$ for all $t \geq 0$ and introduce ρ by simple scaling via

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

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 $Z_\rho(t) = \Psi(Y_\rho)(t)$, $t \geq 0$, 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,

$$\begin{aligned} \mathcal{U}_\rho^p &\equiv \left\{ \tilde{N}_\rho : \mathbb{R}^+ \to \mathbb{R} : \tilde{N}_\rho(s) \leq -(1-\rho)s + b_p\sqrt{s},\, s \geq 0 \right\} \quad \text{and} \\ \mathcal{U}_\rho \equiv \mathcal{U}_\rho^f &\equiv \left\{ \tilde{N}_\rho : \mathbb{R}^+ \to \mathbb{R} : \tilde{N}_\rho(s) \leq E[N_\rho(s)] + b_f\sqrt{Var(N_\rho(s))},\, s \geq 0 \right\}, \\ &= \left\{ \tilde{N}_\rho : \mathbb{R}^+ \to \mathbb{R} : \tilde{N}_\rho(s) \leq -(1-\rho)s + b_f\sqrt{Var(N_\rho(s))},\, s \geq 0 \right\}, \end{aligned} \tag{15}$$

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

Paralleling §2, the associated parametric and functional RQ formulations are, respectively,

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

As in §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 §EC.4; in particular, see (EC.14) and (EC.16). The same reasoning as before yields the following analog of Theorem 1. The interchange of suprema is justified by Lemma EC.1.

THEOREM 2. (*RQ solutions for the workload*) *The solutions of the RQ optimizations in* (16) *are*

$$\begin{aligned} Z^*_{p,\rho} &\equiv \sup_{\tilde{N}_\rho \in \mathcal{U}^p_\rho} \sup_{s\ge 0} \left\{ \tilde{N}_\rho(t) \right\} = \sup_{s\ge 0} \sup_{\tilde{N}_\rho \in \mathcal{U}^p_\rho} \left\{ \tilde{N}_\rho(t) \right\} = \sup_{s\ge 0} \left\{ -(1-\rho)s + b_z\sqrt{s} \right\} \\ &= -(1-\rho)x^* + b_z\sqrt{x^*} = \frac{b_z^2}{4|1-\rho|} \quad \textit{for} \quad x^* \equiv x^*(\rho) = \frac{b_z^2}{4(1-\rho)^2} \quad \textit{and} \end{aligned} \tag{17}$$

$$\begin{aligned} Z^*_\rho \equiv Z^*_{f,\rho} &\equiv \sup_{\tilde{N}_\rho \in \mathcal{U}^f_\rho} \sup_{s\ge 0} \left\{ \tilde{N}_\rho(t) \right\} = \sup_{s\ge 0} \sup_{\tilde{N}_\rho \in \mathcal{U}^f_\rho} \left\{ \tilde{N}_\rho(t) \right\} \\ &= \sup_{s\ge 0} \left\{ -(1-\rho)s + b_f \sqrt{Var(N_\rho(s))} \right\} \\ &= \sup_{s\ge 0} \left\{ -(1-\rho)s + b_f \sqrt{Var(Y_\rho(s))} \right\}. \end{aligned} \tag{18}$$

COROLLARY 3. (*exact for* $M/GI/1$) *For the* $M/GI/1$ *model, the total input process* $\{Y_\rho(t) : t \ge 0\}$ *in* (14) *is a compound Poisson process with* $E[Y_\rho(t)] = \rho t$ *and* $Var(Y_\rho(t)) = \rho t(c_s^2 + 1)$*, 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*.*

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

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

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

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

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

$$Cov(Y(t_2) - Y(t_1), Y(t_4) - Y(t_3)) \ge 0 \quad \text{for all} \quad 0 \le t_1 < t_2 \le t_3 < t_4. \tag{21}$$

Negative dependence holds if the inequality is reversed. These are strict if the inequality is a strict inequality. From (17) and (18) of §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, $\dot{V}(t) \geq 0$ and $\ddot{V}(t) \geq (\leq) 0$.

REMARK 4. (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. (2015).

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 $V(t) \equiv V_\rho(t) \equiv Var(Y_\rho(t))$.*

(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 $t \geq x^*$. In addition, $0 < x^* < \infty$ and x^* satisfies the equation

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

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) The variance function $V(x)$ is convex (concave), so that the function $h(x) \equiv \sqrt{V(x)}$ is concave if there is positive (negative) dependence, as in (21) (*with sign reversed*). *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 $\sqrt{V(s)} \le \sqrt{2\sigma_Y^2 s}$ for all $s \ge s_0$ 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 supremeum 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

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

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 $V \ge 0$, $\dot{V}(t) \ge 0$ and $\ddot{V}(t) \le (\le) 0$. Thus, by direct differentiation

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

with strictness implying a strict inequality. ∎

4. The Indices of Dispersion for Counts and Work

The workload process is not only convenient because it leads to the continuous RQ optimization problem in (16) with solution in (18), but also because the workload process scales with ρ in a more elementary way than the waiting times, as indicated in §3.2. In contrast, the scaling of the waiting times in previous sections is more complicated because the interarrival times are scaled with ρ but the service times are not

It is useful 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). With that convention, we define the *index of dispersion for counts* (IDC) associated with the rate-1 arrival process A as in §4.5 of Cox and Lewis (1966) by

$$I_c(t) \equiv \frac{Var(A(t))}{E[A(t)]} = \frac{Var(A(t))}{t}, \quad t \ge 0. \tag{23}$$

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

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

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

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

Indeed, under regularity conditions (see §EC.4.5), the following finite positive limits exist and are equal:

$$\begin{aligned} \lim_{t\to\infty} \{I_w(t)\} &\equiv I_w(\infty) = \sigma_Y^2 = c_Z^2(1) \equiv \lim_{\rho\to 1} \{c_Z^2(\rho)\} \\ \lim_{t\to 0} \{I_w(t)\} &\equiv I_w(0) = 1 + c_s^2 = c_Z^2(0) \equiv \lim_{\rho\to 0} \{c_Z^2(\rho)\} \end{aligned} \tag{26}$$

for $c_s^2 \equiv Var(V_1)/E[V_1]^2$ 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 §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 $t \geq 0$. As observed by Fendick and Whitt (1989), a simple connection would be $c_Z^2(\rho) \approx I_w(t_\rho)$ for some increasing fucntion 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.

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

$$Z_\rho^* = \sup_{s\geq 0} \left\{ -(1-\rho)s + b_f \sqrt{Var\left(Y_\rho(s)\right)} \right\} = \sup_{s\geq 0} \left\{ -(1-\rho)s + b_f \sqrt{\rho s I_w(\rho s)} \right\}, \tag{27}$$

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

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

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,

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

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. (*closed-form RQ solutions*) *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 $x^* \dot{I}_w(x^*) \le I_w(x^*)$. *If* $b_f = \sqrt{2}$, *then the associated scaled RQ workload satisfies*

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

Proof. Note that $xI_w(x) = V(x)$. Because we have assumed that $V(x)$ is differentiable, so 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

$$x^*(\rho) \approx \frac{\rho^2}{2(1-\rho)^2} I_w(x^*(\rho)) \quad \text{so that} \quad Z^*_\rho \approx \frac{\rho I_w(x^*(\rho))}{2(1-\rho)} \quad \text{and} \quad c^2_{Z^*}(\rho) = I_w(x^*(\rho)). \tag{33}$$

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 §EC.6.

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 asymptotically correct both in heavy traffic* (*as* $\rho \uparrow 1$) *and light traffic* (*as* $\rho \downarrow 0$)*, i.e., so that we have the following supplement to* (26)*:*

$$c^2_{Z^*}(1) = I_w(\infty) = c^2_Z(1) \quad \textit{and} \quad c^2_{Z^*}(0) = I_w(0) = c^2_Z(0). \tag{34}$$

REMARK 5. 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 §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 §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 §III.E of Fendick and Whitt (1989),

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

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

4.3.2. The Multi-Class $\sum_i (G_i/G_i)/1$ Model. As indicated in (56) and (57) in §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, τ_i the mean service time of class i, and $\rho_i \equiv \lambda_i \tau_i$ 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,

$$I_c(\lambda t) \equiv \frac{Var(A(t))}{E[A(t)]} = \frac{\sum_i Var(A_i(t))}{\lambda t} = \sum_i \left(\frac{\lambda_i}{\lambda}\right) I_{c,i}(\lambda_i t) \tag{36}$$

and

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

From (36) and (37), we see that I_c and I_w are convex combinations of the component $I_{c,i}$ and $I_{w,i}$ modified by additional time scaling. The interaction with the time scaling in (36) and (37) with the time scaling by $n = (1-\rho_n)^{-2}$ in (EC.17) for the HT limits in Theorem EC.2 can have an important implications for performance, as we illustrate in §4.3.4.

4.3.3. The IDC's 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 $I_c(t) = 1$ for all $t \geq 0$. 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 $Var(A(t))$ and thus the IDC $I_c(t)$ can either be calculated directly or can 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 §3.5 of Ross (1996). Then $Var(A(t))$ 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 §4.5 of Cox (1962). The required Numerical transform inversion for the renewal function is discussed in §13 of Abate and Whitt (1992). The hyperexponential (H_2) and Erlang (E_2) special cases are described in §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 $Var(A(t))$ for the versatile

Markovian point process or Neuts process; see §5.4, especially Theorem 5.4.1. Explicit formulas for the Markov modulated Poisson process (MMPP) are given on pp. 287-289.

All of these explicit formulas above have the asymptotic form

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

4.3.4. The Superposition of Many Component Sources. To better understand the complex multi-class examples, consider the $\sum_i GI_i/GI/1$ model where the arrival process is the superposition of n i.i.d. renewal processes, each with rate ρ/n, so that the overall arrival rate is ρ. From (36) and (37),

$$I_{c,n}(\rho t) = I_{c,1}(\rho t/n) \quad \text{and} \quad I_{w,n}(\rho t) = I_{w,1}(\rho t/n), \quad t \geq 0, \tag{38}$$

so that the superposition IDI and IDW differ from those of a single component process only by the time scaling. In support of the IDC and IDW as useful partial characterizations, we see that the expressions in (36)-(35) are consistent with the known complex behavior of queues with superposition arrival processes, as discussed in §9.8 of Whitt (2002). As $n \to \infty$, we see evidence of the convergence to a Poisson process; As $t \to \infty$ we see the same limit as for a single component renewal process, i.e., $I_{c,n}(\infty) = I_{c,1}(\infty)$. We see that the RQ approach can capture the complex interaction between n and ρ.

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: These 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 approximmations 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 $\sum_i G_i/GI/1$ single-server queue with a superposition arrival process, where (38) 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-times distribution be hyperexponential (H_2), a mixture of two exponential distributions) with mean 1, $c_s^2 = 2$ and balanced means as on p. 137 of Whitt (1982). Then (38) 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.

Figure 2 (left) 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. Two important observations are: (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 $\rho \leq 0.5$, 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 §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 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 §3.5 of §Ross (1996).) We let the run length and amount discarded be increasing in ρ, as dictated by Whitt (1989b). We provide additional details about our simulation methodology in the appendix.

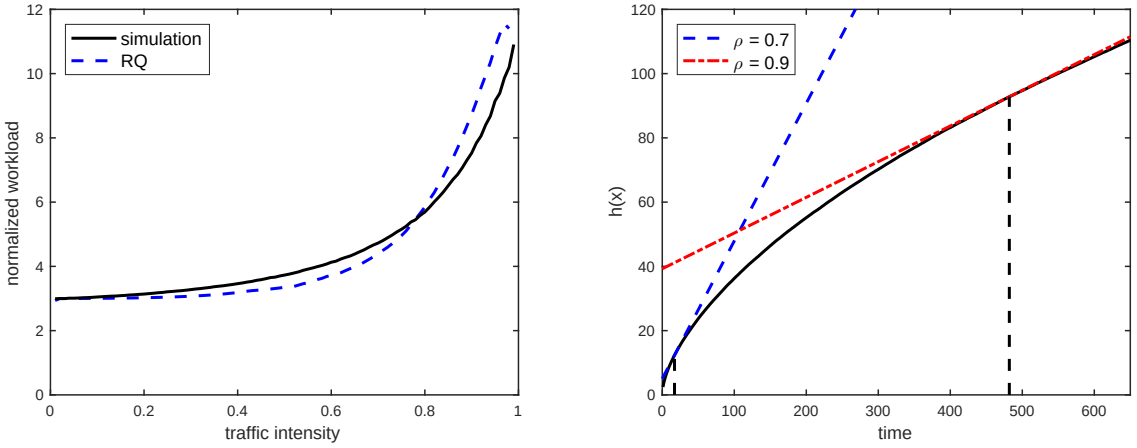

Figure 2 A comparison between simulation estimates of of the normalized mean workload $c_Z^2(\rho)$ in (25) and its approximation $c_{Z^*}^2(\rho)$ in (29) as a function of ρ for the $\sum_i^n GI_i/H_2/1$ model with $c_s^2 = 2$ and a superposition of n i.i.d. lognormal renewal arrival processes for $n = 10$ and $c_a^2 = 10$ (left). On the right is the 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).

5.2. Ten Queues in Series

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 $c_a^2 = 10$. (We use the same distribution as for the service time in §5.1.) The first 9 queues all have deterministic service times. The first 8 queues have mean service time and thus traffic intensity 0.6, while the 9^{th} queue has mean service time and thus traffic intensity 0.95. The last (10^{th}) 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 deterministic queues 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 $D/M/1$ queue, which has $c_a^2 = 0$, 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 $c_a^2 = 10$. 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 ten queues in series as a function of the mean service time and traffic intensity ρ there with the corresponding values in the $D/M/1$ queue (left) and with the RQ approximation $c^2_{Z^*}(\rho)$ in (29) (right). Figure 3 (left) shows that the last queue behaves like a $D/M/1$ queue for all traffic intensities ≤ 0.8, but

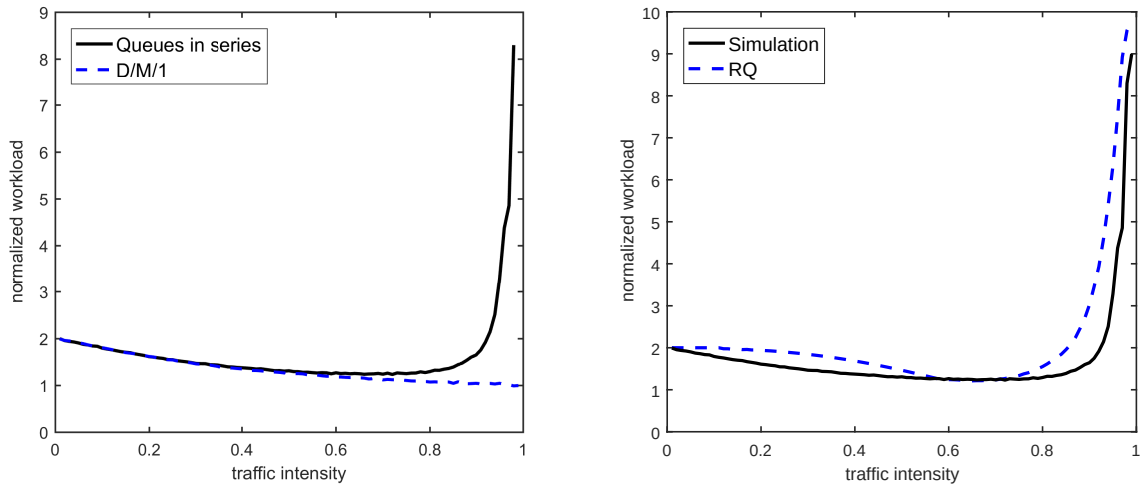

Figure 3 A comparison between simulation estimates of the normalized mean workload $c^2_Z(\rho)$ in (25) at the last queue of the ten queues in series with highly variable external arrival process, but low-variability service times, as a function of the mean service time and traffic intensity ρ there with the corresponding value in the $D/M/1$ queue (left) and with the RQ approximation $c^2_{Z^*}(\rho)$ in (29) (right).

then starts behaving more like an $H_2/M/1$ queue as the traffic intensity approaches the value 0.95 at the 9^{th} queue. Figure 3 (right) 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 on the left shows the IDW over the long interval $[0, 10^5]$, while the plots in the middle and right give a closer view of the IDW over the initial segments $[0, 20]$ and $[0, 400]$. On the right, we plot the IDW assuming continuous-time stationarity (which we use) together with the plot using the discrete-time Palm stationarity (see Sigman (1995)), which acts as if there is an arrival at time 0, so that the plot is 0 over the initial interval of length 0.95 (the deterministic service time at the previous queue). The good performance in Figure 3 for small values of ρ depends on using the proper (continuous-time) version.

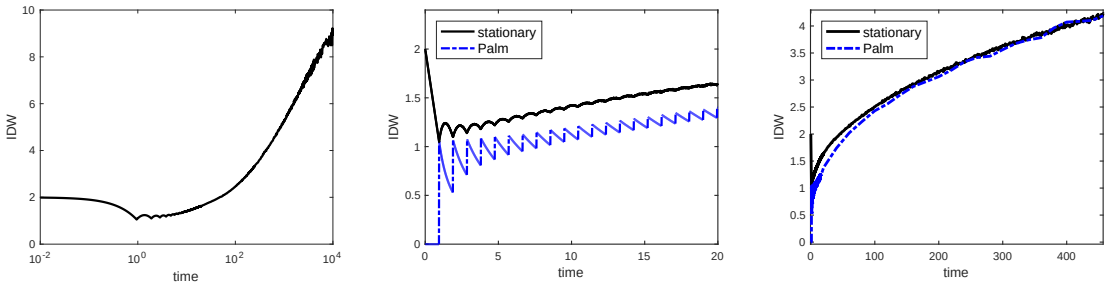

Figure 4 The IDW at the last queue over the interval $[0, 10000]$ in log scale (left), $[0, 20]$ (middle) and $[0, 400]$ (right). The continuous-time stationary version used for RQ with the workload is contrasted with the discrete-time Palm version over the initial segment on the middle and right.

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 ten queues in series as a function of the mean service time and traffic intensity ρ there with the corresponding values in the $D/M/1$ queue (left) and with the RQ approximation $c_{W*}^2(\rho)$, defined just as in (29). Figure 5 and 3 look similar, except that there is a significant difference for small velues 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 properties when ρ is small. The mean waiting time and mean workload actually are quite different in light traffic; see §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 robust queueing network analyzer (RQNA). While we have established good RQ results for one single-server queue, it still remains to develop a full RQNA exploiting the IDW and the results in the previous sections. In this section we propose a candidate framework in which we hope to develop such a new RQNA. This framework exploits the IDC in §3. Given an effective algorithm in this framework, we would then want to generalize the framework.

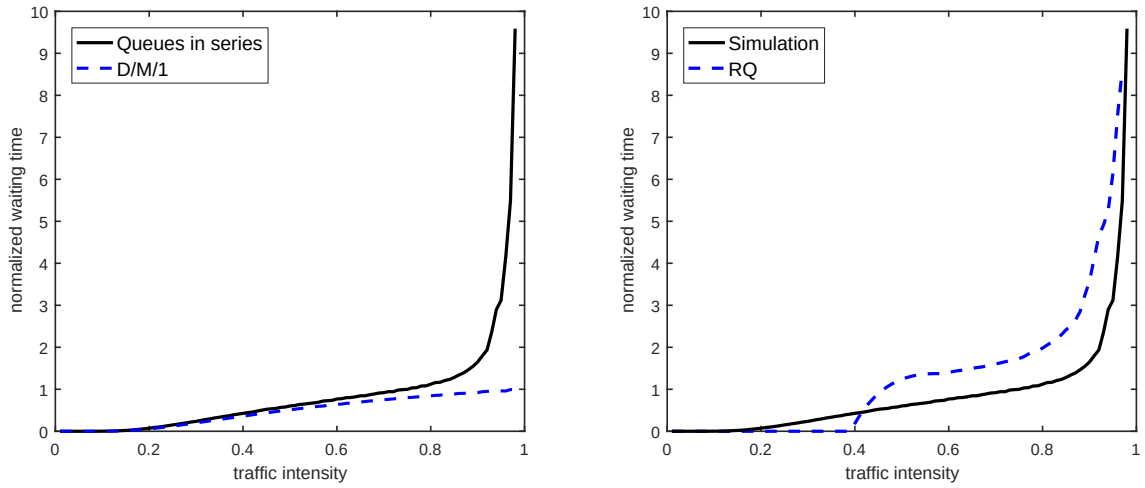

Figure 5 Contrasting the discrete-time and continuous-time views: the analog of Figure 3 for the waiting time. Simulation estimates of the normalized mean waiting time $c^2_W(\rho)$, defined as in (25), at the last queue of the ten queues in series with highly variable external arrival process, but low-variability service times, as a function of the mean service time and traffic intensity ρ there with the corresponding value in the $D/M/1$ queue (left) and with the RQ approximation $c^2_{W*}(\rho)$, defined as in (29) (right).

We start by specifying an initial reference stochastic queueing network model. To start, we make several simplifying assumptions, which we would want to generalize: (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 FCLT with a BM 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; (vi) given that the traffic rate equations are used to find the net arrival rate at each queue, as in §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 §2.3 of Whitt (1983) and Segal and Whitt (1989), practical applications require much more complicated models, e.g., perhaps having 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 a new RQNA exploiting the results in this paper.

We propose going beyond QNA by letting the variability of each arrival process, external or internal, be partially characterized by its IDC. Let the net arrival process at queue i have IDC $I_{c,i}(t)$. By (35), the associated IDW is then $I_{w,i}(t) = I_{c,i}(t) + c^2_{s,i}$, $t \geq 0$. Given $I_{w,i}(t)$ and the net arrival rate λ_i determined by the traffic equations, and thus the traffic intensity ρ_i, we can approximate the mean steady-state workload at queue i, $E[Z_i(\rho_i)]$ for each i. We consider that as the initial objective, even though we want to extend the algorithm to develop a full performance description.

For the $(G/GI/1)^m$ model introduced above, we specify the service time at queue i by its mean τ_i and scv $c^2_{s,i}$, as in QNA, but now we specify the external arrival process at queue i by its rate $\lambda_{o,i}$ and IDC $\{I_{c,o,i}(t) : t \geq 0\}$, with o designated from outside. A simplified alternative asymptotic IDC framework would replace the full IDC by the pair $(I_{c,o,i}(0), I_{c,o,i}(\infty))$. Paralleling QNA, the IDC-based RQNA would apply a network calculus to determine the final net IDC at each queue. The difficult superposition operation is already covered by §4.3.4 and has shown to be effective in §5.1. It remains to treat the flow through a queue and the splitting. And it remains to carefully examine the performance of alternative methods.

REMARK 6. (one uncertainty set versus two) Bandi et al. (2015) exploited their RQ for a single-server queue based on the two separate uncertainty sets for the interarrival times and service times in (3) in order to obtain their full RQNA algorithm. Our Remark 1 shows that choice came at a cost. Hence, it is important to note that the new framework we propose here does not require two separate uncertainty sets.

7. Conclusions

We have formulated and solved new forms of robust queueing (RQ) for a single-server queue and shown that the solutions relate nicely to the mean steady-state waiting time and workload in the general stationary $G/G/1$ single-server queue and its $GI/GI/1$ special case. Unlike Bandi et al. (2015), we only consider a

single queue, but in §6 we provide a framework that can be used to develop a new robust queueing network analyzer (RQNA).

In §2 we introduced a new RQ formulation for the waiting time with a single uncertainty set instead of two separate uncertainty sets. Corollary 1 shows that, if we choose a single parameter correctly, then the RQ solution coincides with the classic Kingman (1962) bound for the $GI/GI/1$ queue and so is asymptotically correct in heav traffic. Corollary 2 show that the deterministic time where the RQ solution attains its supremum is the same order as the relaxation time in the $GI/GI/1$ queue, exposing how steady state is approached in the stochastic model.

We introduced new parametric and functional versions of RQ for the continuous-time workload in §3. The functional versions include the variance of the total input of work as a function of time. In §4 we introduced the indices of dispersion for counts (IDC) and work (IDW). We expressed the solution of the functional RQ in terms of the IDW in (28), which is in a form convenient for applications, provided the IDW is available. In §4.3 we reviewed useful properties of these important indices and indicated how they can be calculated in stochastic models or estimated from data. Theorem 4 gives a closed-form expression for the solution, which also provides insight; e.g. it relates to the variability fixed-point equation in equation (15) of Fendick and Whitt (1989). Theorem 5 shows that the solution of the functional RQ for the mean steady-state workload is asymptotically correct in both heavy traffic and light traffic.

We evaluated the new functional RQ for the workload by making comparisons with simulations of queues with common network structure, as depicted in Figure 1. The simulations show that the RQ solutions serve as good approximations for the mean steady-state workload as a function of the traffic intensity. They also confirm that those common network structures can induce strong dependence, which has a significant impact upon performance.

Finally, in §6 we introduced a framework for developing a new robust queueing network analyzer (RQNA) based on the indices of dispersion. It remains to exploit that framework to develop such a new RQNA. The paper shows that the functional RQ is effective in exposing the impact of the dependence among the interarrival times and service times as a function of time upon the mean steady-state workload

as a function of the underlying traffic intensity at the queue. Overall, the paper supports the initative begun by Bandi et al. (2015). Clearly, many more opprotunities remain.

Acknowledgments

Support was received from NSF grants CMMI 1265070 and 1634133.

e-companion

EC.1. Overview

This is an online e-companion to the main paper. It has seven sections. First, in §EC.2 we provide additional motivation for and discussion about our RQ approach. Then §EC.3 elaborates on ways that the results can be applied. Next, §EC.4 establishes (mostly reviews) supporting functional central limit theorems (FCLT's), the CLT's that follow from them and their implications. Then in §EC.5 we develop the functional RQ for the discrete-time waiting time mentioned in Remark 2. In §EC.6 we present additional proofs for some of the results in the main paper. Finally, in §EC.7 we present additional simulation examples.

EC.2. Additional Motivation and Discussion

In this section we make several remarks to amplify the discussion in the main paper.

EC.2.1. Underlying Philosophy

In doing this RQ work, it is good to communicate our underlying philosophy: We view RQ, not as a way to replace an intractable stochastic model by an alternative deterministic model, without drawing on the axioms of probability, as suggested in Bandi and Bertsimas (2012), but instead as a way to develop improved approximations for the performance of a given stochastic model. We think that the stochastic model often does effectively capture essential features of the uncertainty; the main problem is its intractability. (Of course, there also may be uncertainty about model parameters and the model itself.) Thus, we judge our RQ formulations by their ability to efficiently generate useful performance approximations for the given stochastic model.

EC.2.2. Why Does RQ Perform So Well?

Given that robust optimization is a way to obtain bounds in an alternative deterministic framework, without reference to an underlying probability model, it is natural to wonder why the RQ provides such effective approximations if we just choose a single parameter appropriately. We have tried to explain right after Theorem 1 by explaining the close connection between RQ and heavy-traffic approximations. In particular,

they are both based on the central limit theorem (CLT), as we review here in §EC.4. The CLT in turn says that the probability distribution primarily depends on the mean and variance, which are precisely what provides the basis for all the RQ constraints.

It is natural to want a still better explanation. We might ask how the RQ for the workload can provide such a spectacularlly good approximation (exact) for the mean workload $E[Z]$ in the $M/GI/1$ queue, as shown in Corollary 3, and more generally. A partial explanation is that the net-input process in the $M/GI/1$ queue and for the RBM heavy-traffic limit is a Levy process (has stationary and independent increments) with negative drift ($E[N(t) = -mt$), finite variance ($Var(N(t)) = vt$) and no negative jumps. With such exceptionally nice structure,

$$E[Z] = v/2m;$$

e.g., see see Kella and Whitt (1992) or §IX.3 of Asmussen (2003). A nice simple proof for $M/GI/1$ appears in §5.13 of Wolfe (1989). That is the same form as the RQ solutions. It remains to say more.

EC.2.3. The Mythical Renewal Arrival Process

Experience with queueing applications has shown that most arrival processes can be classified as (i) approximately a Poisson process, (ii) approximately a deterministic evenly spaced arrival process, or (iii) a complex arrival process with dependence among successive interarrival times. In other words, non-Poisson non-deterministic renewal arrival processes are extremely rare in practice. The $GI/GI/1$ model with independent sequences of i.i.d. interarrival times and service times evidently has received so much attention largely because it is relatively tractable; i.e., it is possible to analyze exactly with sophisticated tools, as in Asmussen (2003). Explicit numerical results can then be obtained by numerical algorithms, such as numerical transform inversion, as in Abate et al. (1993). The $GI/GI/1$ model does give a good idea about the impact of departures from the tractable M Markovian assumption, but experience indicates that it can be misleading. We might think that it suffices to estimate the scv of a service time or an interarrival times in order to assess the level of variability, but that misses the dependence, and so might be a big mistake, as illustrated by Fendick et al. (1989), as reviewed in §9.6 of Whitt (2002).

EC.2.4. The Probability That The Constraints Are Satisfied

It is natural to ask what would be the probability in the stochastic model that the RQ constraints in (3) or (5) would be satisfied. In fact, it is not difficult to see that, even for the basic $M/M/1$ model, the probability is 0. That follows from the law of the iterated logarithm. Nevertheless, the deterministic RQ is useful. Of course, we could consider only finitely many constraints as in Bandi et al. (2015). With a proper choice the solution is unchanged.

EC.3. How Can the Functional RQ Results Be Applied?

This paper helps develop useful diagnostic tools to study complex queueing systems. This paper adds additional support to Fendick and Whitt (1989) by showing how to measure flows (arrival processes, possibly together with service times) in complex queueing systems and the value for doing so in understanding congestion at a queue, as characterized by the mean workload and the mean waiting time. In particular, we see how the variance time curves and indices of dispersion can provide useful descriptions of the flows, enabling us with the aid of RQ to predict congestion as a function of the traffic intensity quite accurately. These measurements can fruitfully be applied with either system measurements or simulations. As we indicated in §4.3, the indices of dispersion can also be calculated for quite complex models.

As in Bandi et al. (2015), the new RQ can help develop improved performance analysis tools for complex queueing networks. In particular, the methods here provide a basis for improving parametric-decomposition approximations such as QNA in Whitt (1983) by exploiting variability functions instead of variability parameters, as proposed in Whitt (1995). In §6 we provide a road map for the way to proceed by introducing a candidate IDC framework for creating a new RQNA that can capture the dependence in the flows.

One concrete way the RQ here can be applied is to analyze the consequence of changing the service mechanism and/or the arrival process associated with a single-server queue in a complex queueing network. For example, assuming that (i) the same arrival process would come to a new service mechanism and (ii) the new service mechanism produces i.i.d. service times with a distribution that can be predicted, then we could first measure the IDC of the arrival process and combine that with (35) to obtain an estimate of the full IDW. Then we could apply RQ to estimate the mean workload at the queue. If we are contemplating

several alternative service mechanisms, we can apply the same techniques to compare their performance impact.

As a second example, suppose that the arrival rate will increase. If that will occur in a way that corresponds approximately to deterministic scaling of the arrival counting process, then we can directly apply RQ to predict the performance consequence. On the other hand, if the arrival rate increases by superposing more streams, as in Sriram and Whitt (1986), then we can apply RQ with (36)-(38) to predict the performance consequence.

EC.4. Supporting Functional Central Limit Theorems (FCLT's)

In this section we establish establish (mostly review) the supporting FCLT's and the CLT's that follow from them. These are for the general stationary $G/G/1$ model, allowing stochastic dependence among the interarrival times and service times. §EC.4.1 starts with a basic FCLT for partial sums of random variables from weakly dependent stationary sequences, as in Theorems 19.1-19.3 of Billingsley (1999) and Theorem 4.4.1 of Whitt (2002).

To state the basic FCLT underlying the RQ approach to the waiting time and workload processes, we consider a sequence of models indexed by n with stationary sequence of interarrival times and service times. In §EC.4.1 we establish the underlying FCLT for the partial sums of the interarrival times and service times. Then in EC.4.2 we establish a FCLT for other basic processes. In §EC.4.3 we establish different ordinary CLT's that support the parametric RQ and functional RQ. Finally, in §EC.4.4 we establish heavy-traffic FCLLT's for the waiting time and workload processes.

EC.4.1. The Basic FCLT for the Partial Sums

As in §2, we assume that the models are generated by a fixed sequence of mean-1 random variables $\{(U_k, V_k)\}$, with the interarrival times in model n being $U_{n,k} \equiv \rho_n^{-1} U_k$. For each n, let the sequence of pairs of partial sums be $\{(S^a_{n,k}, S^s_{n,k} : k \geq 1\}$. Let $\lambda_n = \rho_n$ and $\mu_n = 1$ denote the arrival rate and service rate in model n. Let $\lfloor x \rfloor$ denote the greatest integer less than or equal to the real number x. Let D^2 be the two-fold product space of the function space D and let $\Rightarrow$ denote convergence in distribution. For this initial FCLT, we let $\rho_n \to \rho$ as $n \to \infty$ for arbitrary $\rho > 0$. Let random elements in the function apace D^2 be defined by

$$\left(\hat{\mathbf{S}}^a_n(t), \hat{\mathbf{S}}^s_n(t)\right) \equiv n^{-1/2} \left(\left[S^a_{n,\lfloor nt \rfloor} - \rho_n^{-1} nt\right], \left[S^s_{n,\lfloor nt \rfloor} - nt\right]\right), \quad t \geq 0.$$

THEOREM EC.1. (*FCLT for partial sums of interarrival times and service times*) *Let* $\{(U_k, V_k) : k \geq 1\}$ *be a weakly dependent stationary sequence with* $E[U_k] = E[V_k] = 1$. *Let* $U_{n,k} = \rho_n^{-1} U_k$ *and* $V_{n,k} = V_k$, $n \geq 1$, *and assume that the variances and covariances satisfy*

$$0 < \rho^{-2}\sigma_A^2 \equiv \lim_{n\to\infty} \{n^{-1} Var(S_n^a)\} < \infty, \quad 0 < \sigma_S^2 \equiv \lim_{n\to\infty} \{n^{-1} Var(S_n^s)\} < \infty$$
$$\text{and} \quad \rho^{-1}\sigma_{A,S}^2 \equiv \lim_{n\to\infty} \{n^{-1} Cov(S_n^a, S_n^s)\}. \tag{EC.1}$$

Then (*under additional regularity conditions assumed, but not stated here*)

$$\left(\hat{\mathbf{S}}_n^a, \hat{\mathbf{S}}_n^s\right) \Rightarrow \left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s\right) \quad \text{in} \quad D^2 \quad \text{as} \quad n \to \infty, \tag{EC.2}$$

where $\left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s\right)$ *is distributed as zero-drift two-dimensional Brownian motion* (*BM*) *with covariance matrix*

$$\Sigma = \begin{pmatrix} \rho^{-2}\sigma_A^2 & \rho^{-1}\sigma_{A,S}^2 \\ \rho^{-1}\sigma_{A,S}^2 & \sigma_S^2 \end{pmatrix}.$$

Proof. The one-dimensional FCLT's for weakly dependent stationary sequences in D can be used to prove the two-dimensional version in Theorem EC.1. First, the limits for the individual processes $\hat{\mathbf{S}}_n^a$ and $\hat{\mathbf{S}}_n^s$ imply tightness of these processes in D, which in turn implies joint tightness in D^2. Second, the Cramer-Wold device in Theorem 4.3.3 of Whitt (2002) implies that limits for the finite-dimensional distributions for all linear combinations (which should be implied by the unstated regularity condition) implies the joint limit for the finite-dimensional distributions (fidi's). Finally, tightness plus convergence of the fidi's implies the desired weak convergence by Corollary 11.6.2 of Whitt (2002). ∎

EC.4.2. The FCLT for Other Basic Processes

As a consequence of Theorem EC.1, we also have an associated FCLT for scaled random elements associated with $S_{n,k}^x \equiv S_{n,k}^s - S_{a,k}^a$, $k \geq 1$, $A_n(s)$ and $Y_n(s) \equiv \sum_{i=1}^{A_n(s)} V_{n,i} = \sum_{i=1}^{A(\rho_n s)} V_i = Y(\rho_n s)$, $s \geq 0$, for A and Y in (10) and (11). Let the associated scaled processes be defined by

$$\left(\hat{\mathbf{S}}_n^x(t), \hat{\mathbf{A}}_n(t), \hat{\mathbf{Y}}_n(t)\right) \equiv n^{-1/2}\left(\left[S_{n,\lfloor nt\rfloor}^x - (1-\rho_n^{-1})nt\right], [A_n(nt) - \rho_n nt], [Y_n(nt) - \rho_n nt]\right), \tag{EC.3}$$

for $t \geq 0$. Let $\mathbf{B}(t)$ be standard (zero drift and unit variance) one-dimensional BM and let $\mathbf{e}$ be the identity function in D, i.e., $\mathbf{e}(t) = t$. Let $\stackrel{\mathrm{d}}{=}$ mean equal in distribution, as processes if used for stochastic processes.

COROLLARY EC.1. (*joint FCLT for basic processes*) *Under the conditions of Theorem* EC.1*,*

$$\left(\hat{\mathbf{S}}_n^a, \hat{\mathbf{S}}_n^s, \hat{\mathbf{S}}_n^x, \hat{\mathbf{A}}_n, \hat{\mathbf{Y}}_n\right) \Rightarrow \left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s, \hat{\mathbf{S}}^x, \hat{\mathbf{A}}, \hat{\mathbf{Y}}\right) \quad \textit{in} \quad D^5 \quad \textit{as} \quad n \to \infty, \tag{EC.4}$$

where $\hat{\mathbf{S}}^x = \hat{\mathbf{S}}^s - \hat{\mathbf{S}}^a \stackrel{\mathrm{d}}{=} \sigma_X \mathbf{B}$*, with variance function*

$$\sigma_X^2 \equiv \sigma_X^2(\rho) = \rho^{-2}\sigma_A^2 + \sigma_S^2 - 2\rho^{-1}\sigma_{A,S}^2, \quad 0 < \sigma_X^2 < \infty, \tag{EC.5}$$

for $\rho^{-2}\sigma_A^2$*,* σ_S^2 *and* $\rho^{-1}\sigma_{A,S}^2$ *in* (EC.1)*, while*

$$\begin{aligned} \hat{\mathbf{A}} &= -\rho\hat{\mathbf{S}}^a \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} -\rho\sigma_A \mathbf{B}_a \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} \rho^{3/2}\sigma_Y \mathbf{B}_a, \\ \hat{\mathbf{Y}} &= \hat{\mathbf{S}}^s \circ \rho\mathbf{e} - \rho\hat{\mathbf{S}}^a \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} \sigma_Y \mathbf{B} \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} \sqrt{\rho}\sigma_Y \mathbf{B}, \end{aligned} \tag{EC.6}$$

where

$$\sigma_Y^2 \equiv \sigma_Y^2(\rho) = \sigma_A^2 + \sigma_S^2 - 2\sigma_{A,S}^2, \quad 0 < \sigma_Y^2 < \infty, \quad \textit{for all} \quad \rho. \tag{EC.7}$$

Hence, $\hat{\mathbf{Y}} \stackrel{\mathrm{d}}{=} \hat{\mathbf{S}}^{\mathrm{x}}$ *for* $\rho = 1$*, but not otherwise.*

Proof. We apply the continuous mapping theorem (CMT) using several theorems from Whitt (2002). The CMT itself is Theorem 3.4.4. We treat the process $S_{n,k}^x$ using addition. We treat the counting processes A_n by apply the inverse map with centering to go from the FCLT for $S_{n,k}^a$ to the FCLT for the associated scaled counting processes, applying Theorem 7.3.2, which is a consequence of Corollary 13.8.1 to Theorem 13.8.2, which follows from Theorem 13.7.1. Then the limit for Y_n follows from Corollary 13.3.1. However, it is also possible to give a more elementary direct argument. First, let $\bar{A}_n(t) \equiv n^{-1}A_n(t)$, $t \geq 0$, and note that $\bar{A}_n \Rightarrow \rho\mathbf{e}$ as a consequence of the limit for $\mathbf{A}_n$. The initial limits all hold jointly by Theorems 11.4.4 and 11.4.5. Then observe that we can apply the continuous mapping theorem with composition and addition to treat $\mathbf{Y}_n$, because we can write

$$\mathbf{Y}_n = \mathbf{S}_n^s \circ \bar{A}_n + \mathbf{A}_n \tag{EC.8}$$

i.e.,

$$\mathbf{Y}_n(t) \equiv n^{-1/2}\left(\sum_{k=1}^{A(nt)} - \rho nt\right), \quad t \geq 0, \tag{EC.9}$$

while

$$(\mathbf{S}_n^s \circ \bar{A}_n)(t) = n^{-1/2}\left(\sum_{k=1}^{A(nt)} - A_n(nt)\right), \quad t \geq 0, \tag{EC.10}$$

We then add to get (EC.9), observing that two terms cancel.

We now derive alternative expressions for the limit process $\mathbf{Y}$. First, directly from (EC.8) we obtain

$$\mathbf{Y} = \mathbf{S}^s \circ \rho\mathbf{e} + \mathbf{A} = \mathbf{S}^s \circ \rho\mathbf{e} - \rho\mathbf{S}^a\rho\mathbf{e} \stackrel{\mathrm{d}}{=} \sigma_Y\mathbf{B} \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} \sqrt{\rho}\sigma_Y\mathbf{B}. \tag{EC.11}$$

which justifies the expression for σ_Y^2 in (EC.7). ∎

REMARK EC.1. (uniform integrability) Condition (EC.1) implies that $k^{-1}Var(S_k^x) \to \sigma_X^2$ as $k \to \infty$ for σ_X^2 in (EC.5). In addition to the conclusions of Theorem EC.2 and Corollary EC.1, we assume that the appropriate uniform integrability holds, so that we also have the continuous-time analog

$$s^{-1}Var(Y(s)) \to \sigma_Y^2 \quad \text{as} \quad s \to \infty \tag{EC.12}$$

for σ_Y^2 in (EC.7).

EC.4.3. Alternative Scaling in the CLT

Theorem EC.1 and Corollary EC.1 imply ordinary CLT's for the processes S_n^x and $Y_n(s)$ by simply applying the applying the CMT with the projection map $\pi : D \to \mathbb{R}$ with $\pi(x) \equiv x(1)$.

COROLLARY EC.2. (*associated CLT's*) *Under the assumptions of Theorem EC.1, there are CLT's for the partial sums S_n^x and the total input processes Y_n, stating*

$$(S_n^x - nE[X_1])/\sqrt{n\sigma_X^2} \Rightarrow N(0,1) \quad \textit{as} \quad n \to \infty, \tag{EC.13}$$

and

$$(Y_n - \rho n])/\sqrt{n\sigma_Y^2} \Rightarrow N(0,1) \quad \textit{as} \quad n \to \infty, \tag{EC.14}$$

where $N(0,1)$ is a standard (mean-0, variance-1) normal random variable, σ_X^2 is the asymptotic variance constant in (EC.1) *and* (EC.5)*, and σ_Y^2 is the asymptotic variance constant in* (20) *and* (EC.7)*.*

Clearly, Corollary EC.2 supports the parametric RQ formulations and indicates how to choose the parameters b_x and b_p in order to produce versions that should be asymptotically correct in heavy-traffic (see the next section). We now show that there are alternative versions of these CLT's that support the functional RQ formulations. First, instead of (EC.13), we can also write

$$[S_n^x - E[S_n^x]]/\sqrt{Var(S_n^x)} \Rightarrow N(0,1) \quad \text{as} \quad n \to \infty. \tag{EC.15}$$

Second, instead of (EC.14), we can also write

$$[Y(t) - E[Y(t)]]/\sqrt{Var(Y(t))} \Rightarrow N(0,1) \quad \text{as} \quad t \to \infty. \tag{EC.16}$$

The numerators in (EC.13) and (EC.15) are identical because $E[S_n^x] = nE[X_1]$ and $E[Y(t)] = \rho t$. The full statements in (EC.13) and (EC.15) are asymptotically equivalent as $n \to \infty$ by the CMT, because

$$\frac{S_n^x - nE[X_1]}{\sqrt{Var(S_n^x)}} = \frac{S_n^x - nE[X_1]}{\sqrt{n\sigma_X}} \times \frac{\sqrt{n\sigma_X}}{\sqrt{Var(S_n^x)}} \Rightarrow N(0,1) \times 1 = N(0,1).$$

The same is true for the CLT's in (EC.14) and (EC.16).

EC.4.4. The Associated Heavy-Traffic FCLT

Theorem EC.1 and Corollary EC.1 also provide a basis for heavy-traffic (HT) FCLT's for the waiting-time and workload processes. To state the HT FCLT, we let $\rho_n \to 1$ as $n \to \infty$ at the usual rate; see (EC.18) below. Let $\hat{\mathbf{W}}^n$ and $\hat{\mathbf{Z}}^n$ be the random elements associated with the waiting time and workload processes, defined by

$$\left(\hat{\mathbf{W}}^n(t), \hat{\mathbf{Z}}^n(t)\right) = \left(n^{-1/2} W_{n,\lfloor nt \rfloor}, n^{-1/2} Z_n(nt)\right), \quad t \geq 0. \tag{EC.17}$$

Let $\psi : D \to D$ be the one-dimensional reflection map with impenetrable barrier at the origin, assuming $x(0) = 0$, i.e., $\psi(x)(t) \equiv x(t) - \inf_{0 \leqslant s \leqslant t} x(s)$; see §13.5 of Whitt (2002). Here is the HT FCLT; it is is a variant of Theorem 2 of Iglehart and Whitt (1970); see §5.7 and 9.6 in Whitt (2002). Given Corollary EC.1, it suffices to apply the CMT with the reflection map ψ.

THEOREM EC.2. (*heavy-traffic FCLT*) *Consider the sequence of $G/G/1$ models specified above. If, in addition to the conditions of Theorem* EC.1*,*

$$n^{1/2}(1-\rho_n) \to \eta, \quad 0 < \eta < \infty, \tag{EC.18}$$

then

$$\left(\hat{\mathbf{W}}_n, \hat{\mathbf{Z}}_n\right) \Rightarrow \left(\psi(\hat{\mathbf{S}}^x - \eta\mathbf{e}), \psi(\hat{\mathbf{S}}^x - \eta\mathbf{e})\right) \quad \textit{in} \quad D^2 \quad \textit{as} \quad n \to \infty, \tag{EC.19}$$

jointly with the limits in (EC.4)*, where ψ is the reflection map and $\hat{\mathbf{S}}^x - \eta\mathbf{e} \stackrel{\mathrm{d}}{=} \sigma_Y\mathbf{B} - \eta\mathbf{e}$ is BM with variance constant σ_Y^2 in* (EC.7) *and drift $-\eta < 0$, so that $\psi(\hat{\mathbf{S}}^x - \eta\mathbf{e})$ is reflected BM (RBM).*

The HT approximation for the mean steady-state wait and workload stemming from Theorem EC.2 is

$$E[W(\rho)] \approx E[Z_\rho] \approx \frac{\sqrt{n}\sigma_Y^2}{2\eta} \approx \frac{\sigma_Y^2}{2(1-\rho)} \tag{EC.20}$$

for σ_Y^2 in (EC.7), which is independent of ρ, using the mean of the exponential limiting distribution of the RBM $\psi(\sigma_x\mathbf{B} - \eta\mathbf{e})(t)$ as $t \to \infty$.

REMARK EC.2. (the two forms of stationarity) As discussed in the beginning of §3.2, there are two forms of stationarity, one for discrete time and the other for continuous time. When we focus on the waiting time, we use discrete-time stationarity; when we focus on the workload, we use continuous-time stationarity. So far in this section, we have built everything in the framework of discrete-time stationarity. However, in doing so, we automatically can get FCLT's in both settings. The theoretical basis is provided by Nieuwenhuis (1989).

REMARK EC.3. (the limit-interchange problem) the standard HT limits for the processes do not directly imply limits for the steady-state distributions. Strong results have been obtained with i.i.d. assumptions, e.g., see Budhiraja and Lee (2009), but the case with dependence is more difficult. Nevertheless, supporting results for the $G/G/1$ queue when dependence is allowed appear in Szczotka (1990, 1999). We assume that this interchange step is also justified.

REMARK EC.4. (the asymptotic method) The RQ approach in Theorem 2 corresponds to approximating the arrival and service processes in the $G/G/1$ queue by the asymptotic method in Whitt (1982), which develops approximations for the arrival and service processes using all the correlations. That is in contrast to the stationary-interval method discussed just before §EC.4, which uses none of the correlations. Our RQ approach develops an intermediate methods in between those two extremes.

EC.4.5. The Normalized Workload and the IDW: Justifying (26)

We are motivated to develop the functional RQ for the steady-state workload because of the close connection between the IDW $\{I_w(t) : t \geq 0\}$ and the normalized mean workload $\{c_Z^2(\rho) : 0 \leq \rho \leq 1\}$ established by Fendick and Whitt (1989). The key asymptotic components are the heavy-traffic (HT) and light-traffic (LT) limits stated here in (26). Now that we have just developed the supporting HT FCLT, we review the theoretical support for (26).

First, the HT limit is supported by the FCLT for $\hat{\mathbf{Z}}_n$ in Theorem EC.2. We use the continuous-time stationarity, justified by Remark EC.2. For the FCLT's, we require weak dependence, which is specified by relatively complex mixing conditions. Given the weak dependence and the FCLT, we need extra regularity conditions to get to what is actually stated in (26). First we need the limit-interchange property discussed in Remark EC.3 to get associated limits for the steady-state distributions. Second, we need appropriate uniform integrability to get from convergence of random variables to convergence of their moments; see Remark EC.1.

The LT limit is established in §IV.A of Fendick and Whitt (1989). An important observation made there is that the LT limiting behavior is much more robust for the steady-state workload than for the steady-state waiting time. In particular, the LT limit for the steady-state waiting time depends more on the fine structure of the model. The LT limits provide theoretical insight into why it is easier to describe the mean steady-state workload than the mean steady-state waiting time, even though they agree in the HT limit.

EC.5. Functional RQ for the Discrete-Time Waiting Times

We now provide extra details about the functional RQ for the steady-state waiting time, paralleling §3, as promised in Remark 2. We introduce the indices of dispersion for intervals in §EC.5.1. We briefly mention the heavy-traffic and light-traffic limits in §EC.5.2.

First, paralleling the functional RQ optimization for $Z^*_{f,\rho}$ in (16), we have the discrete-time analog based on (9):

$$W^* \equiv W^*_{f,\rho} \equiv \sup_{\tilde{X}\in\mathcal{U}^x_f} \sup_{k\geq 0} \{S^x_k\}\,. \tag{EC.21}$$

where $\mathcal{U}^x_f$ is defined in (9). For the $G/G/1$ model stationary in discrete time, the reasoning for Theorem 1 leads to the alternative representation as

$$W^* = \sup_{k\geq 0} \left\{-mk + b_{f,d}\sqrt{Var(S^x_k)}\right\} \tag{EC.22}$$

instead of (7), where $m \equiv (1-\rho)/\rho$ as before. We can alternative representations using indices of dispersion, but now for intervals instead of for counts, which we discuss next.

EC.5.1. Discrete Time: Indices of Dispersion for Intervals

We now recast the discrete-time RQ solution in (EC.22) in terms of indices of dispersion for intervals. For that purpose, we create scaled versions of the discrete-time variance-time functions (sequences) $Var(S^x_k)$, $Var(S^a_k)$ and $Var(S^s_k)$ as functions of k. That yields the *indices of dispersion for intervals* (IDI), as in Chapter 4 of Cox and Lewis (1966), defined by

$$I_a(k) \equiv \frac{kVar(S^a_k)}{(E[S^a_k])^2}, \quad I_s(k) \equiv \frac{kVar(S^s_k)}{(E[S^s_k])^2} \quad \text{and} \quad I_{a,s}(k) \equiv \frac{kCov(S^a_k, S^s_k)}{E[S^a_k]E[S^s_k]}. \tag{EC.23}$$

With (EC.23),

$$\sqrt{Var(S^x_k)} = E[U_1]\sqrt{kI_x(k)}, \quad k\geq 1, \quad \text{and} \quad \sigma^2_X \equiv \lim_{k\to\infty}\left\{k^{-1}Var(S^x_k)\right\} = E[U_1]^2 I_x(\infty) \tag{EC.24}$$

where

$$I_x(k) \equiv I_a(k) + \rho^2 I_s(k) - 2\rho I_{a,s}(k) \quad \text{for} \quad \rho \equiv E[V_1]/E[U_1] < 1. \tag{EC.25}$$

These three IDI's $I_a(k)$, $I_s(k)$ and $I_{a,s}(k)$ were used to develop queueing approximations in Fendick et al. (1989).

As a consequence, (EC.22) can be rewritten a

$$W^*_{f,\rho} = \sup_{k\geq 0}\left\{-(1-\rho)k/\rho + b_{f,d}\sqrt{kI_x(k)}\right\}. \tag{EC.26}$$

Similar to the continuous-time workload, we focus on the normalized mean waiting time and RQ approximation defined by

$$c^2_W(\rho) \equiv \frac{2(1-\rho)}{\rho}E[W_\rho], \quad \text{and} \quad c^2_{W^*}(\rho) \equiv \frac{2(1-\rho)}{\rho}W^*_{f,\rho}. \tag{EC.27}$$

EC.5.2. Heavy-Traffic and Light Traffic Limits

By essentially the same reasoning, we can show that both the parametric RQ and the functional RQ for the steady-state waiting time W are asymptotically exact in heavy-traffic, with the same HT limit as for the continuous-time workload, if we choose the constant $b_{f,d}$ in (EC.26) appropriately. The light-traffic behavior is much more complicated for the steady-state waiting time, as discussed in §IV.A of Fendick and Whitt (1989) and §1 of Whitt (1989a). That is a major reason for using the workload instead of the waiting time.

EC.6. Additional Technical Support for the Main Paper

In this section we provide additional technical support for the main paper. First, a key step in obtaining tractable solutions of the RQ optimizations is an interchange of suprema. The following lemma shows that this interchange is justified in all cases.

LEMMA EC.1. (*interchange of suprema*) *The interchange of suprema below holds for any real-valued function* $f(x,y)$

$$M := \sup_{\substack{x\in A\\ y\in B}} f(x,y) = \sup_{x\in A}\sup_{y\in B} f(x,y) = \sup_{y\in B}\sup_{x\in A} f(x,y),$$

where the joint supremum M *is allowed to be infinite.*

Proof By symmetry, we need only prove that

$$\sup_{\substack{x\in A\\ y\in B}} f(x,y) = \sup_{x\in A}\sup_{y\in B} f(x,y).$$

Suppose the joint supremum M is finite, then there exist a sequence $(x_n,y_n)\in A\times B$ such that $f(x_n,y_n) > M-1/n$, where M is the finite joint supremum. Then, we have

$$\sup_{x\in A}\sup_{y\in B} f(x,y) \ge \sup_{y\in B} f(x_n,y) \ge f(x_n,y_n) \ge M-\frac{1}{n}, \quad \text{for all } n>0.$$

This implies that

$$\sup_{x\in A}\sup_{y\in B} f(x,y) \ge M = \sup_{\substack{x\in A\\ y\in B}} f(x,y).$$

The other direction of inequality is trivial by noting that $M\ge f(x,y)$ and taking iterated supremum on both sides.

For the case where the joint supremum M is infinite, then there exist a sequence $(x_n, y_n) \in A \times B$ such that $f(x_n, y_n) > n$. Then

$$\sup_{x\in A}\sup_{y\in B} f(x,y) \geq \sup_{y\in B} f(x_n, y) \geq f(x_n, y_n) \geq n, \quad \text{for all } n > 0.$$

Hence the iterated supremum is also infinite, which completes the proof. ∎

We now prove Theorem 5. We state and prove two separate results here.

THEOREM EC.3. (*RQ in heavy traffic*) *Let $b'_z = \sqrt{2}$ and assume that $I_w(x)$ is non-negative, continuous and that $I_w(\infty) \equiv \lim_{x\to\infty} I_w(x)$ exist, then we have the following heavy-traffic limit for the normalized RQ optimal value*

$$c^2_{Z^*}(1) \equiv \lim_{\rho\to 1} \frac{2(1-\rho)}{\rho} Z^*(\rho) = I_w(\infty). \tag{EC.28}$$

To prove Theorem EC.3, we need two lemmas.

LEMMA EC.2. (*order-preservation of the RQ solution*) *Let f, g be two positive functions on non-negative real numbers, satisfying $f(x) \geq g(x)$ for all $x \geq 0$. Then we have*

$$Z^*_f \geqslant Z^*_g,$$

where Z^*_f is the solution to the RQ problem with f replacing I_w.

Proof Let x^*_f denote the optimal solution to the RQ problem specified by f. Then

$$\begin{aligned} Z^*_f = -\frac{1-\rho}{\rho} x^*_f + b\sqrt{x^*_f f(x^*_f)} &\geq -\frac{1-\rho}{\rho} x^*_g + b\sqrt{x^*_g f(x^*_g)} \\ &\geq -\frac{1-\rho}{\rho} x^*_g + b\sqrt{x^*_g g(x^*_g)} = Z^*_g. \end{aligned}$$ ∎

LEMMA EC.3. (*continuity property of the normalized RQ solution*) *Let $c^2_{Z^*}(\rho)(f)$ be the normalized solution to* (28) *with I_w replaced by f. Then $c^2_{Z^*}(\rho)$ is a continuous function from space $(C_b(\mathbb{R}^+, \mathbb{R}^+), \|\cdot\|_\infty)$ to $\mathbb{R}^+$, with the former one being the space of all continuous and bounded functions from $\mathbb{R}^+$ to $\mathbb{R}^+$ equipped with the supremum norm.*

Proof Let $f, g \in (C_b(\mathbb{R}^+, \mathbb{R}^+), \|\cdot\|_\infty)$, satisfying $\|f-g\|_\infty \le \epsilon$. Then we have

$$f(x) - \epsilon \le g(x) \le f(x) + \epsilon, \quad \text{for all } x \ge 0.$$

Since $f \in C_b(\mathbb{R}^+, \mathbb{R}^+)$, there exist $M > 0$ such that $f(x) < M$ for all $x \ge 0$. Then for all $x > M_\rho$, where $M_\rho \equiv (\rho b'_z/(1-\rho))^2 M$, we have

$$-\frac{1-\rho}{\rho}x + b'_z\sqrt{xf(x)} < -\frac{1-\rho}{\rho}x + b'_z\sqrt{xM} < 0$$

Hence,

$$\begin{aligned}
c_{Z^*}(\rho)(g) \le c_{Z^*}(\rho)(f+\epsilon) &= \frac{2(1-\rho)}{\rho} \sup_{0\le x\le \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{x(f(x)+\epsilon)} \right\} \\
&\le \frac{2(1-\rho)}{\rho} \sup_{0\le x\le \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{xf(x)} + b'_z\sqrt{x\epsilon} \right\} \\
&\le \frac{2(1-\rho)}{\rho} \sup_{0\le x\le \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{xf(x)} \right\} + b'_z\sqrt{\tilde{M}_\rho\epsilon} \qquad \text{(EC.29)} \\
&= c_{Z^*}(\rho)(f) + \frac{2(1-\rho)}{\rho} b'_z\sqrt{\tilde{M}_\rho\epsilon} \\
&= c_{Z^*}(\rho)(f) + 2(b'_z)^2\sqrt{(M+\epsilon)\epsilon}, \qquad \text{(EC.30)}
\end{aligned}$$

where $\tilde{M}_\rho \equiv (\rho b'_z/(1-\rho))^2(M+\epsilon)$ and the first inequality follows from Lemma EC.2. Similarly, we can prove that

$$c_{Z^*}(\rho)(g) \ge c_{Z^*}(\rho)(f-\epsilon) \ge c_{Z^*}(\rho)(f) - 2(b'_z)^2\sqrt{(M+\epsilon)\epsilon}. \qquad \text{(EC.31)}$$

Combining (EC.30) and (EC.31), we have

$$|c_{Z^*}(\rho)(g) - c_{Z^*}(\rho)(f)| \le 2(b'_z)^2\sqrt{(M+\epsilon)\epsilon}.$$

Hence the lemma holds. ∎

Proof of Theorem EC.3. Recall that Theorem 4 suggest that the optimal solution is of order $O(\rho^2/(2(1-\rho)^2))$, we perform a change of variable $t = 2(1-\rho)^2x/\rho^2$ in (28) and scale the space by a constant $\rho/(2(1-\rho))$. Hence, we have

$$c^2_{Z^*}(\rho) = \sup_{0\le t\le\infty} \left\{ -t + 2\sqrt{tI_w\left(\frac{\rho^2}{2(1-\rho)^2}t\right)} \right\}. \qquad \text{(EC.32)}$$

Since $I_w(\infty) \equiv \lim_{x\to\infty} I_w(x)$ exist, there exist a T sufficiently large such that $|I_w(t) - I_w(\infty)| < \epsilon$ for all $t > T$. Now, we define

$$\tilde{I}_w(t) = \begin{cases} I_w(t), & t \le T, \\ \text{linear}, & T - \epsilon < t \le T, \\ I_w(\infty), & t > T. \end{cases}$$

By virtue of Lemma EC.3, we need only prove that $c_{Z^*}(1)(\tilde{I}_w) = \tilde{I}_w(\infty) = I_w(\infty)$.

Note that continuity and finite limit at $x = \infty$ implies that $I_x(x)$ is bounded, say $I_w(x) < M - \epsilon$ for all $x \ge 0$. Hence we have

$$-t + 2\sqrt{t\tilde{I}_w\left(\frac{\rho^2}{2(1-\rho)^2}t\right)} \le -t + 2\sqrt{tM}. \tag{EC.33}$$

We assume first that the limit $I_w(\infty)$ is strictly positive. The case where $I_w(\infty) = 0$ can be deduced by considering a sequence of functions $f_n(x)$ such that $f_n(\infty) > 0$ and $|I_w - f_n|_\infty < 1/n$, and applying Lemma EC.3.

Now, for the case where $I_w(\infty) > 0$, we can choose ρ_0 such that

$$T_\rho \equiv \frac{2(1-\rho_0)^2}{\rho_0^2} T < \min\left\{ I_w(\infty), 2M - I_w(\infty) - 2\sqrt{M^2 - I_w(\infty)M} \right\},$$

since the right-hand-side of the inequality wil be strictly positive. Then for all $\rho > \rho_0$, we have

$$\begin{aligned} \sup_{0\le t\le T_\rho} \left\{ -t + 2\sqrt{t\tilde{I}_w\left(\frac{\rho^2}{2(1-\rho)^2}t\right)} \right\} &\le \sup_{0\le t\le T_\rho} \left\{ -t + 2\sqrt{tM} \right\} \\ &\le I_w(\infty). \end{aligned}$$

But plugging $I_w(\infty)$ into the objective function, we have the objective value $I_w(\infty)$ by the fact that $\frac{\rho^2}{2(1-\rho)^2} I_w(\infty) > T$ and that $\tilde{I}_w(t)$ is constant after $t > T$. This implies that

$$\begin{aligned} c^2_{Z^*}(\rho)(\tilde{I}_w) &= \sup_{T_\rho \le t \le \infty} \left\{ -t + 2\sqrt{t\tilde{I}_w\left(\frac{\rho^2}{2(1-\rho)^2}t\right)} \right\} \\ &= \sup_{T_\rho \le t \le \infty} \left\{ -t + 2\sqrt{tI_w(\infty)} \right\} \\ &= I_w(\infty), \quad \text{for all } \rho > \rho_0. \end{aligned}$$

Hence, we've proved that $c_{Z^*}(1)(\tilde{I}_w) = \tilde{I}_w(\infty) = I_w(\infty)$. ∎

Next, we state the corresponding result for RQ in light traffic.

THEOREM EC.4. (*RQ in light traffic*) *Let $b'_z = \sqrt{2}$ and assume that $I_w(x)$ is non-negative, continuous and that $I_w(0) \equiv \lim_{x\to 0} I_w(x)$ exist, then we have the following light-traffic limit for the normalized RQ optimal value*

$$c^2_{Z^*}(0) \equiv \lim_{\rho\to 0} \frac{2(1-\rho)}{\rho} Z^*(\rho) = I_w(0). \tag{EC.34}$$

Proof As in the proof for heavy-traffic limit, we perform the same time and space scaling to get (EC.32). For the same reason, we have (EC.33), which implies that

$$-t + 2\sqrt{t\tilde{I}_w\left(\frac{\rho^2}{2(1-\rho)^2}t\right)} \le -t + 2\sqrt{tM} < 0, \quad \text{for all } t > 4M.$$

Hence, we need only consider the supremum in (EC.32) over bounded interval $[0, 4M]$. Note also that, since $I_w(0) \equiv \lim_{x\to 0} I_w(x)$ exist, for any $\epsilon > 0$, there exist a $\delta > 0$ such that $|I_w(t) - I_w(0)| < \epsilon$ for all $x \in [0, \delta]$. We now choose ρ_0 such that $2\rho_0^2 M/(1-\rho_0)^2 < \delta$, and take a modification

$$\tilde{I}_w(t) = \begin{cases} I_w(0), \ t < \delta, \\ \text{linear}, \ \delta \le t < \delta + \epsilon, \\ I_w(t), \ t \ge \delta + \epsilon, \end{cases}$$

which satisfies $|I_w - \tilde{I}_w|_\infty < \epsilon$ and

$$c^2_{Z^*}(\rho)(\tilde{I_w}) = I_w(0), \quad \text{for all } \rho < \rho_0.$$

We then apply Lemma EC.3 to get the desired light-traffic limit. ∎

EC.7. Additional Examples

In this final section we present some additional examples illustrating more complex behavior that can be seen in the IDW $I_W(t)$ and in the normalized mean workload $c^2_Z(\rho)$. All examples are for single-server queues in series, as in §5.2. For background on this example, we refer to §4.5 of Whitt (1983), Suresh and Whitt (1990) and §§5 and 6 of Whitt (1995).

EC.7.1. The First Example of Queues in Series

Recall that Figure 3 illustrated the performance impact in an $H_2/D/1 \to \cdot/D/1 \ldots \to \cdot/D/1 \to \cdot/M/1$ model with a rate-1 H_2 renewal external arrival process, where the interarrival times has scv $c_a^2 = 10$, followed by nine single-server queues with deterministic D service times and then a final 10^{th} queue with an exponential service time distribution. The first 8 queues all have mean service times and thus traffic intensities of $\rho_k = 0.6$, while the 9^{th} queue has mean service time and thus traffic intensity $\rho_9 = 0.95$. We look at the performance at the last queue as a function of the traffic intensity $\rho \equiv \rho_{10}$ there. Figure 3 shows that the normalized workload at the last queue as a function of ρ. From (26), we know that the left and right limits of the normalized mean workload are $c_Z^2(0) = 1 + c_s^2 = 2.0$ and $c_Z^2(1) = c_a^2 + c_s^2 = 11.0$. Figure 3 shows that the performance is consistent with these limits, even though we cannot see the right hand limit, because the simulation considered traffic intensities bounded above by a quantity less than 1. Nevertheless, we see that the performance varies as a function of ρ approximately as predicted by these two limits.

Figure 3 also shows a dip in the middle consistent with the smoothing provided by the the low variability at the first 9 queues, but the performance does not oscillate too much. Now we illustrate more complex performance functions that can be obtained with more complex models.

In general, experience indicates that for 10 queues in series the normalized mean workload can be bounded above and below, approximately, by

$$\min\{1, c_a^2, c_{s,k}^2, 1 \le k \le 9\} + c_{s,10}^2 \le c_Z^2(\rho) \le \max\{c_a^2, c_{s,k}^2, 1 \le k \le 9\} + c_{s,10}^2. \tag{EC.35}$$

(The "1" appears in the minimum because the left limit at 0 is $1 + c_s^2$.) For example, this approximate bound is consistent with the approximatioon for the variability parmeter c_d^2 of the departure process froma $GI/GI/1$ queue in formula (38) in Whitt (1983), i.e.,

$$c_d^2 \approx (1 - \rho^2)c_a^2 + \rho^2 c_s^2. \tag{EC.36}$$

The bound can be obtained by iterating that approximation forward to get an approximation for $c_{d,9}^2$ and then allowing the previous traffic intensities to vary.

For this example, the bound in (EC.35) is not too informative, concluding that $1 \leq c_Z^2(\rho) \leq 11$, which corresponds to the left and right limits. Our goal is to say more about $c_Z^2(\rho)$ for $0 < \rho < 1$ by using the IDW and RQ.

However, so far, the examples do not show that too much is going on in the middle except for moving from one limit to the other. That motivates us to look at the next examples.

EC.7.2. The $EHEHE \to M$ Example with Four Internal Modes

We now consider an example of 5 single-server queues in series where the variability increases and then decreases 5 times, with the traffic intensities at successive queues decreasing. That makes the external arrival process and the earlier queues relevant only as the traffic intensity increases. Specifically, the example can be donoted by

$$E_{10}/H_2/1 \to \cdot/E_{10}/1 \to \cdot/H_2/1 \to \cdot/E_{10}/1 \to\to \cdot/M/1. \tag{EC.37}$$

In particular, the external arrival process is a rate-1 renewal process with E_{10} interarrival times, thus $c_a^2 = 0.1$. The $1^{\rm st}$ queue has H_2 service times with mean 0.99 and $c_s^2 = 10$ (and also balanced means, as before), thus the traffic intensity at this queue is 0.99. The $2^{\rm nd}$ queue has E_{10} service time with mean and thus traffic intensity 0.98. The $3^{\rm rd}$ queue has H_2 service times with mean 0.70 and $c_s^2 = 10$. The $4^{\rm th}$ queue has E_{10} service times with mean and thus traffic intensity 0.5. The last ($5^{\rm th}$) queue has an exponential service-time distribution. with mean and traffic intensity ρ. As before, we explore the impact of ρ on the performance of that last queue.

Looking backwards starting from the $4^{\rm th}$ queue, i.e., the queue just before the last queue, the Erlang service 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 $E_{10}/M/1$ queue, which has $c_a^2 = 0.1$, but as ρ increases, the arrival process at the last queue should inherit the variability of the previous service times and the external arrival process, and altering between $H_2/M/1$ and $E_{10}/M/1$ as the traffic intensity at the last queue increases. This implies that the normalized workload $c_Z^2(\rho)$ in (25) as a function of ρ should have four internal modes. (If we also count the left and right ends, there will be six modes.

This behavior is substantiated by Figure EC.1 (left), which compares simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (25) at the last queue with the RQ approximation $c_{Z^*}^2(\rho)$ in (29). It shows that the the normalized workload at the last queue fluctuates and each mode corresponds to a previous service process or the external arrival process. Figure EC.1 (left) also shows that RQ successfully captures all modes and provides a reasonably accurate approximation for all ρ. Note that a new scale in the horizontal x axis is used in Figure EC.1 (left), namely $-\ln(1-\rho)$. Since 4 out of 6 modes lies in $\rho > 0.8$, the new scale act to stretch out the crowded plot under heavy traffic.

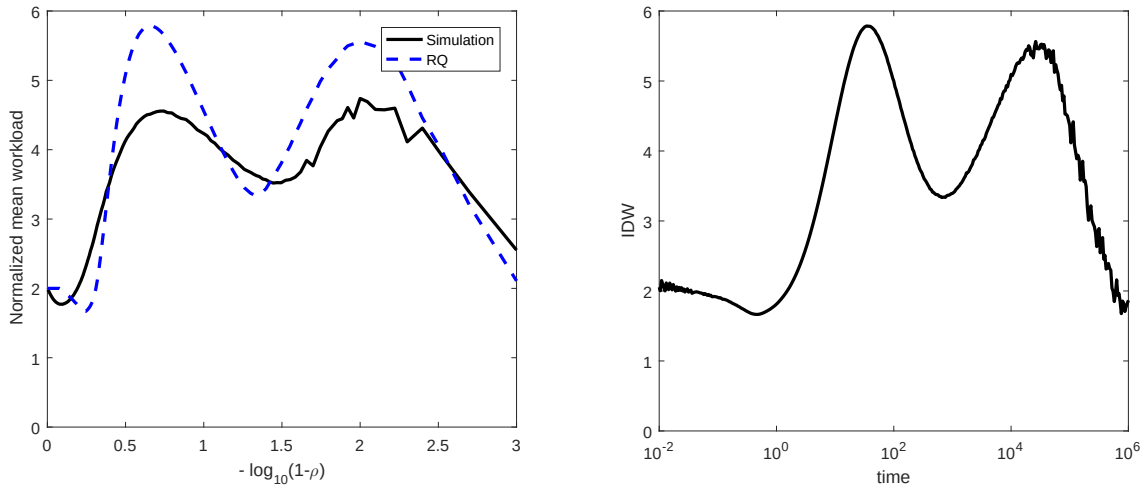

Figure EC.1 A comparison between simulation estimation of the normalized workload $c_Z^2(\rho)$ at the last queue as a function of traffic intensity ρ with the RQ approximation $c_{Z^*}^2(\rho)$ in (29) (left), and the IDW at the last queue over the interval $[0, 10000]$ in log scale (right).

To conclude on this series-queue example, we show the IDW for the last queue in Figure EC.1 (right). The x axis of the figure is in log scale for easier display. We see a more irregular plot at the right because it is harder to directly estimate the IDW $I_w(t)$ for very large t, but the limit as $t \to \infty$ can be calculated from (26). Clearly, the IDW has the same qualitative property as the normalized workload as well as the RQ approximation, as we expect from equation (33).

EC.7.3. A Similar Example with Highly Variable Input

In this section, we consider a similar example where the normalized workload as a function of ρ also has several modes, but the external arrival here has high variability.

In this example we use groups of queues in series with the same distribution and traffic intensity in order to better bring about an adjustment in the level of variability. This device is motivated by the convex-copmbination approximation in (EC.36). Specifically, this example has 13 single-server queues in series. The external arrival process is a rate-1 renewal process with H_2 interarrival times with $c_a^2 = 10$. A group of three queues having E_{10} service times with mean 0.99 is then added to smooth the highly variable external arrivals. The next group of three queues has H_2 service times with mean 0.92 and squared coefficient of variation 5. These queues will bring up the variability of the departure process. Then, another group of three queues with mean 0.9 has E_{10} service times to smooth the departure process again. The variability is then raised by yet another group of three queues having H_2 service times with mean 0.3 and $c_S^2 = 10$. Finally, the last (13^{th}) queue has exponential service times with mean and traffic intensity ρ. As before, we explore the impact of ρ on the performance of that last queue.

As explained in last example, for sufficiently low traffic intensities ρ at the last queue, the last queue should behave approximately the same as an $H_2/M/1$ queue, which has $c_a^2 = 10$, but as ρ increases, the arrival process at the last queue should inherit the variability of the previous service times and the external arrival process, and altering between $E_{10}/M/1$ and $H_2/M/1$ as the traffic intensity at the last queue increases. This implies that the normalized workload $c_Z^2(\rho)$ in (25) as a function of ρ should have several modes, corresponding to the variability of the external arrival process and the service processes at the first 4 groups of queues.

We then have the similar plots in Figure EC.2, which compares simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (25) at the last queue with the RQ approximation $c_{Z^*}^2(\rho)$ in (29) (left) and shows the IDW for this example (right). Again, we are using the same scale as in Figure EC.1 (left), i.e., $-\ln(1-\rho)$, to stretch out the plot under heavy traffic.

Figure EC.2 (left) shows that the the normalized workload at the last queue again has four internal modes and that RQ successfully captures all modes and provides a reasonably accurate approximation for all ρ. Figure EC.2 (right) shows that the IDW has the same qualitative property as the RQ approximation, which is explained in (33). However, the fluctuations in the simulation values for $0 < \rho < 1$ in Figure EC.2 are much less than in Figure EC.1.

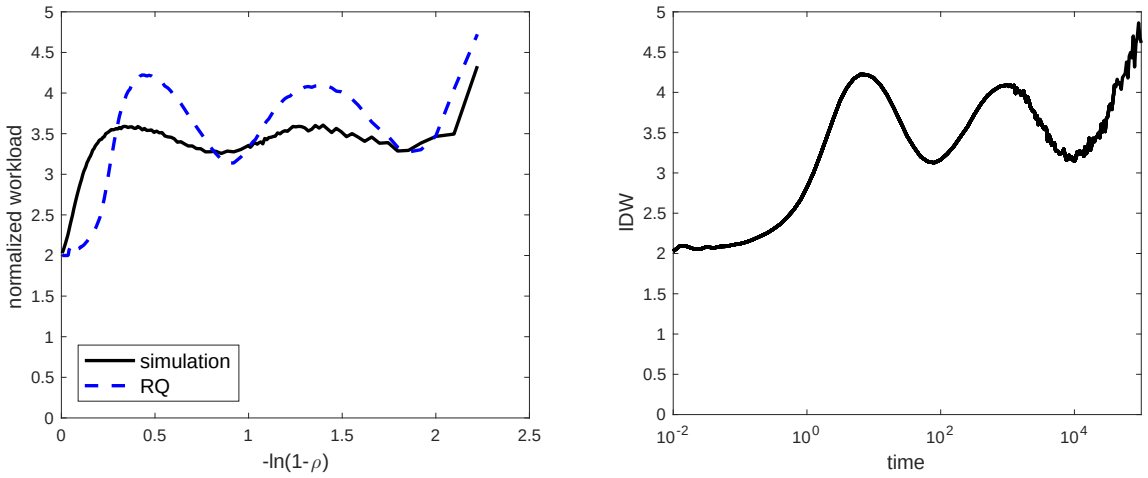

Figure EC.2 A comparison between simulation estimation of the normalized workload $c_Z^2(\rho)$ at the last queue as a function of traffic intensity ρ with the RQ approximation $c_{Z^*}^2(\rho)$ in (29) (left), and the IDW at the last queue over the interval $[0, 10000]$ in log scale (right).

We conclude that (i) the IDW and RQ do capture the qualititative behavior and (ii) the RQ approximation based on the IDW is reasonably accurate in these difficult examples.

References

Abate, J., G. L. Choudhury, W. Whitt. 1993. Calculation of the GI/G/1 steady-state waiting-time distribution and its cumulants from Pollaczek's formula. *Archiv fur Elektronik und bertragungstechnik* **47**(5/6) 311–321.

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

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

Bandi, C., D. Bertsimas. 2012. Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming* **134** 23–70.

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

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

Billingsley, P. 1999. *Convergence of Probability Measures*. Wiley, New York.

Bitran, G. R., D. Tirupati. 1988. Multiproduct queueing networks with deterministic routing: decompositioon approach and thenotion of interference. *Management Science* **34** 75–100.

Budhiraja, A., C. Lee. 2009. Stationary distribution convergence for generalized jackson networks in heavy traffic. *Mathematics of Operations Research* **34**(1) 45–56.

Cohen, J. W. 1982. *The Single Server Queue*. 2nd ed. North-Holland, Amsterdam.

Cox, D. R. 1962. *Renewal Theory*. Methuen, London.

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

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

Fendick, K. W., V. Saksena, W. Whitt. 1989. Dependence in packet queues. *IEEE Trans Commun.* **37** 1173–1183.

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

Fendick, K. W., W. Whitt. 1989. 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.

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

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

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

Iglehart, D. L., W. Whitt. 1970. Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Advances in Applied Probability* **2**(2) 355–369.

Kella, O., W. Whitt. 1992. Useful martingales for stochastic storage processes with Levy input. *Journal of Applied Probability* **29** 396–403.

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

Kim, S., W. Whitt, W. C. Cha. 2015. A data-driven model of an appointment-generated arrival processes at an outpatient clinic. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

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

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

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

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

Mamani, H., S. Nassiri, M. R. Wagner. 2016. Closed-form solutions for robust inventory management. *Management Science* **62**(3) 1–20. Articles in advance, Published April 29, 2016.

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

Neuts, M. F. 1989. *Structured Stochastic Matrices of $M/G/1$ Type and their Application*. Marcel Dekker, New York.

Nieuwenhuis, G. 1989. Equivalence of functional limit theorems for stationary point processes and their Palm distributions. *Probability Theory and Related Fields* **81** 593–608.

Ross, S. M. 1996. *Stochastic Processes*. 2nd ed. Wiley, New York.

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

Segal, M., W. Whitt. 1989. A queueing network analyzer for manufacturing. M. Bonatti, ed., *Teletraffic Science for New Cost-Effective Systems, Networks and ServicesProceedings: ITC 12, Proceedings of the 12th International Teletraffic Congress*. Elsevier, North-Holland, 1146–1152.

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

Sriram, K., W. Whitt. 1986. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications* **SAC-4**(6) 833–846.

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

Szczotka, W. 1990. Exponential approximation of waiting time and queue size for queues in heavy traffic. *Advances in Applied Probability* **22**(1) 230–240.

Szczotka, W. 1999. Tightness of the stationary waiting time in heavy traffic. *Advances in Applied Probability* **31**(3) 788–794.

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

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

Whitt, W. 1984a. On approximations for queues, I. *AT&T Bell Laboratories Technical Journal* **63**(1) 115–137.

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

Whitt, W. 1989a. An interpolation approximation for the mean workload in a $GI/G/1$ queue. *Operations Research* **37**(6) 936–952.

Whitt, W. 1989b. Planning queueing simulations. *Management Science* **35**(11) 1341–1366.

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

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

Whitt, W., W. You. 2016. Time-varying robust queueing. Columbia University, New York, NY http://www.columbia.edu/∼ww2040/allpapers.html.

Wolfe, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.