e-companion

EC.1. Overview

This is an online e-companion to the main paper. It has eight sections. First, in §EC.2 we provide a summary of the main paper. In §EC.3 we provide additional motivation for and discussion about our RQ approach. Then §EC.4 elaborates on ways that the results can be applied. Next, §EC.5 establishes (mostly reviews) supporting functional central limit theorems (FCLT's), the CLT's that follow from them and their implications. Then in §EC.6 we develop the functional RQ for the discrete-time waiting time mentioned in Remark 2. In §EC.7 we present additional proofs for some of the results in the main paper. Finally, in §EC.8 we present additional simulation examples.

EC.2. Summary of the Main Paper

In the main paper we formulated and solved new forms of robust queueing (RQ) for a single-server queue and showed 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 shows 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. Remark 2 introduces a new functional version of RQ for the discrete-time waiting time that can be used to expose the impact of dependence among the interarrival times and service times. That is expanded upon in this e-companion in §EC.6. Finally, we discuss the connection to related work in Mamani et al. (2016) in Remark 4.

We introduced new parametric and functional versions of RQ for the continuous-time workload in §3. At the beginning of §3 we noted that it is convenient to work with the continuous-time workload instead of

the discrete-time waiting time, because the workload process scales with the traffic intensity via direct time scaling as in (14), whereas the waiting times are more complicated because the interarrival times are scaled but the service times are not; see the first paragraph after Theorem 1.

The functional version of RQ for the continuous-time workload 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). These indices of dispersion are just scaled versions of the variance functions, but they are helpful because the scaling makes them independent of the scale; that facilitates developing the key variability fixed point equation in (33). 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 initiative begun by Bandi et al. (2015). Clearly, many more opportunities remain.

EC.3. Additional Motivation and Discussion

In this section we make several remarks to amplify the discussion in §EC.2 and the main paper.

EC.3.1. 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.

In this paper we only considered the problem of describing the performance of a fixed queueing system. We should emphasize that the approximation methods here and in previous work such as QNA in Whitt (1983) have important applications in system design and control problems, e.g., as reviewed in §3 of the survey paper Bitran and Morabito (1996). We think that RQ offers new opportunities in this direction. indeed, we think that is an important direction for further research.

EC.3.2. 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.5. 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 spectacularly 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.3.3. 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.3.4. 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.4. 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)-(39) to predict the performance consequence.

EC.5. 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.5.1 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.5.1 we establish the underlying FCLT for the partial sums of the interarrival times and service times. Then in EC.5.2 we establish a FCLT for other basic processes. In §EC.5.3 we establish different ordinary CLT's that support the parametric RQ and functional RQ. Finally, in §EC.5.4 we establish heavy-traffic FCLLT's for the waiting time and workload processes.

EC.5.1. 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*

$$0 < \rho^{-2}\sigma_A^2 \equiv \lim_{n\to\infty} \{n^{-1} Var(S^a_n)\} < \infty, \quad 0 < \sigma_S^2 \equiv \lim_{n\to\infty} \{n^{-1} Var(S^s_n)\} < \infty$$
$$\textit{and} \quad \rho^{-1}\sigma^2_{A,S} \equiv \lim_{n\to\infty} \{n^{-1} Cov(S^a_n, S^s_n)\}. \tag{EC.1}$$

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

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

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.5.2. 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*

$$\hat{\mathbf{A}} = -\rho\hat{\mathbf{S}}^a \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} -\rho\sigma_A\mathbf{B}_a \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} \rho^{3/2}\sigma_Y\mathbf{B}_a,$$

$$\hat{\mathbf{Y}} = \hat{\mathbf{S}}^s \circ \rho\mathbf{e} - \rho\hat{\mathbf{S}}^a \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} \sigma_Y\mathbf{B} \circ \rho\mathbf{e} \stackrel{\mathrm{d}}{=} \sqrt{\rho}\sigma_Y\mathbf{B}, \tag{EC.6}$$

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, $\hat{\mathbf{Y}} \stackrel{\mathrm{d}}{=} \hat{\mathbf{S}}^{\mathsf{x}}$ 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.5.3. 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.5.4. 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 Continuous Mapping Theorem (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.5, which uses none of the correlations. Our RQ approach develops an intermediate methods in between those two extremes.

EC.5.5. 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.6. 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.6.1. We briefly mention the heavy-traffic and light-traffic limits in §EC.6.2.

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.6.1. 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) $Var(S^x_k)$,

$Var(S_k^a)$ and $Var(S_k^s)$ as functions of k. That yields the *indices of dispersion for intervals* (IDI), as in Chapter 4 of Cox and Lewis (1966), defined by

$$I_k^a \equiv \frac{kVar(S_k^a)}{(E[S_k^a])^2}, \quad I_k^s \equiv \frac{kVar(S_k^s)}{(E[S_k^s])^2} \quad \text{and} \quad I_k^{a,s} \equiv \frac{kCov(S_k^a, S_k^s)}{E[S_k^a]E[S_k^s]}. \tag{EC.23}$$

With (EC.23),

$$\sqrt{Var(S_k^x)} = E[U_1]\sqrt{kI_k^x}, \quad k \geq 1, \quad \text{and} \quad \sigma_X^2 \equiv \lim_{k\to\infty} \left\{k^{-1}Var(S_k^x)\right\} = E[U_1]^2 I_\infty^x \tag{EC.24}$$

where

$$I_k^x \equiv I_k^a + \rho^2 I_k^s - 2\rho I_k^{a,s} \quad \text{for} \quad \rho \equiv E[V_1]/E[U_1] < 1. \tag{EC.25}$$

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

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

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

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

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

EC.6.2. 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.7. 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

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

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

Proof of Theorem 2. The solutions of the RQ optimizations in *(16)* are

$$Z^*_{p,\rho} \equiv \sup_{\tilde{N}_\rho \in \mathcal{U}^p_\rho} \sup_{s\ge 0} \left\{ \tilde{N}_\rho(t) \right\} = \sup_{s\ge 0} \sup_{\tilde{N}_\rho \in \mathcal{U}^p_\rho} \left\{ \tilde{N}_\rho(t) \right\} = \sup_{s\ge 0} \left\{ -(1-\rho)s + b_p\sqrt{s} \right\}$$
$$= -(1-\rho)x^* + b_p\sqrt{x^*} = \frac{b_p^2}{4|1-\rho|} \quad \text{for} \quad x^* \equiv x^*(\rho) = \frac{b_p^2}{4(1-\rho)^2} \quad \text{and} \tag{EC.28}$$

$$Z^*_\rho \equiv Z^*_{f,\rho} \equiv \sup_{\tilde{N}_\rho \in \mathcal{U}^f_\rho} \sup_{s\ge 0} \left\{ \tilde{N}_\rho(t) \right\} = \sup_{s\ge 0} \sup_{\tilde{N}_\rho \in \mathcal{U}^f_\rho} \left\{ \tilde{N}_\rho(t) \right\}$$
$$= \sup_{s\ge 0} \left\{ -(1-\rho)s + b_f\sqrt{Var(N_\rho(s))} \right\}$$
$$= \sup_{s\ge 0} \left\{ -(1-\rho)s + b_f\sqrt{Var(Y_\rho(s))} \right\}. \tag{EC.29}$$

where the interchange of suprema is justified by Lemma EC.1. ∎

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*

$$c^2_{Z^*}(1) \equiv \lim_{\rho\to 1} \frac{2(1-\rho)}{\rho} Z^*(\rho) = I_w(\infty). \tag{EC.30}$$

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 $f(x) \ge g(x)$ for all $x \ge 0$. 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

$$Z^*_f = -\frac{1-\rho}{\rho} x^*_f + b\sqrt{x^*_f f(x^*_f)} \ge -\frac{1-\rho}{\rho} x^*_g + b\sqrt{x^*_g f(x^*_g)}$$
$$\ge -\frac{1-\rho}{\rho} x^*_g + b\sqrt{x^*_g g(x^*_g)} = Z^*_g. \quad ∎$$

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,

$$\begin{aligned}
c_{Z^*}(\rho)(g) \le c_{Z^*}(\rho)(f+\epsilon) &= \frac{2(1-\rho)}{\rho} \sup_{0\le x\le \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{x(f(x)+\epsilon)} \right\} \\
&\le \frac{2(1-\rho)}{\rho} \sup_{0\le x\le \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{xf(x)} + b'_z\sqrt{x\epsilon} \right\} \\
&\le \frac{2(1-\rho)}{\rho} \sup_{0\le x\le \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{xf(x)} \right\} + b'_z\sqrt{\tilde{M}_\rho\epsilon} \qquad \text{(EC.31)} \\
&= c_{Z^*}(\rho)(f) + \frac{2(1-\rho)}{\rho} b'_z\sqrt{\tilde{M}_\rho\epsilon} \\
&= c_{Z^*}(\rho)(f) + 2(b'_z)^2\sqrt{(M+\epsilon)\epsilon}, \qquad \text{(EC.32)}
\end{aligned}$$

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

$$c_{Z^*}(\rho)(g) \ge c_{Z^*}(\rho)(f-\epsilon) \ge c_{Z^*}(\rho)(f) - 2(b'_z)^2\sqrt{(M+\epsilon)\epsilon}. \qquad \text{(EC.33)}$$

Combining (EC.32) and (EC.33), 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

$$c^2_{Z^*}(\rho) = \sup_{0\le t\le\infty} \left\{ -t + 2\sqrt{tI_w\left(\frac{\rho^2}{2(1-\rho)^2}t\right)} \right\}. \qquad \text{(EC.34)}$$

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

$$-t + 2\sqrt{t\tilde{I}_w\left(\frac{\rho^2}{2(1-\rho)^2}t\right)} \le -t + 2\sqrt{tM}. \tag{EC.35}$$

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*

$$c^2_{Z^*}(0) \equiv \lim_{\rho\to 0} \frac{2(1-\rho)}{\rho} Z^*(\rho) = I_w(0). \tag{EC.36}$$

Proof As in the proof for heavy-traffic limit, we perform the same time and space scaling to get (EC.34). For the same reason, we have (EC.35), 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.34) 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.8. 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.8.1. 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

$$\min\{1, c_a^2, c_{s,k}^2, 1 \le k \le 9\} + c_{s,10}^2 \le c_Z^2(\rho) \le \max\{c_a^2, c_{s,k}^2, 1 \le k \le 9\} + c_{s,10}^2. \tag{EC.37}$$

(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.,

$$c_d^2 \approx (1 - \rho^2)c_a^2 + \rho^2 c_s^2. \tag{EC.38}$$

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.37) is not too informative, concluding that $1 \le c_Z^2(\rho) \le 11$, 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.8.2. 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

$$E_{10}/H_2/1 \to \cdot/E_{10}/1 \to \cdot/H_2/1 \to \cdot/E_{10}/1 \to \cdot/M/1. \tag{EC.39}$$

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.

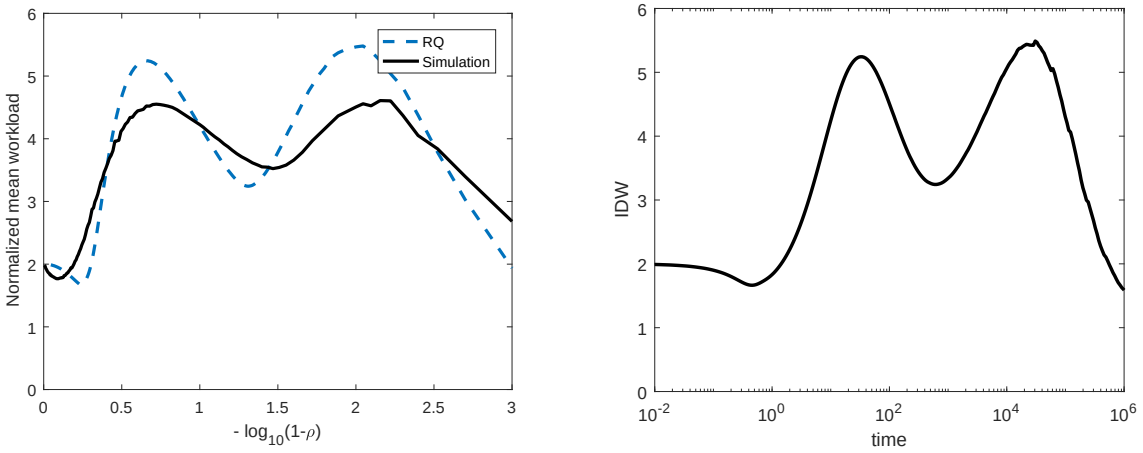

Figure EC.1 A comparison between simulation estimation of the normalized workload $c_Z^2(\rho)$ at the last queue as a function of traffic intensity ρ with the RQ approximation $c_{Z^*}^2(\rho)$ in (29) (left), and the IDW at the last queue over the interval $[10^{-2}, 10^5]$ in log scale (right).

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.8.3. 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.38). 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.

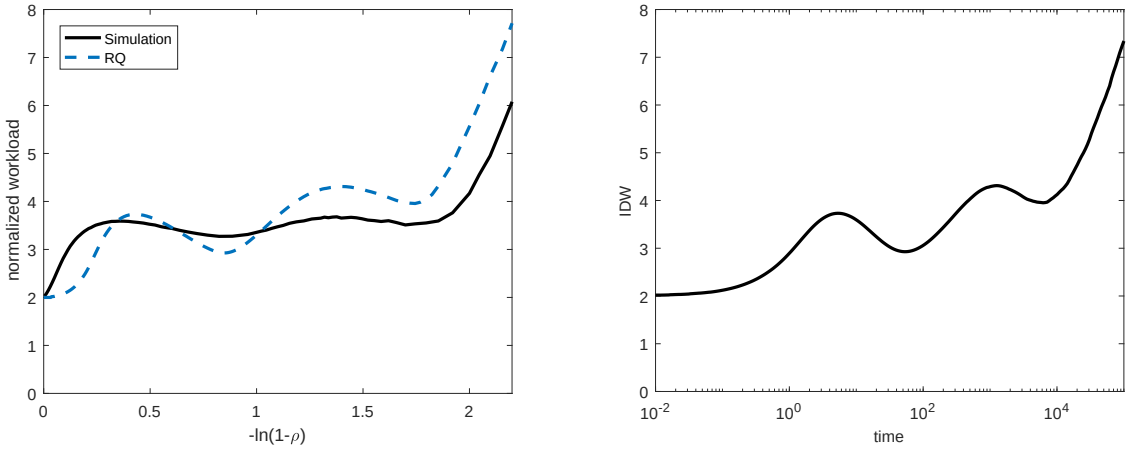

Figure EC.2 A comparison between simulation estimation of the normalized workload $c^2_Z(\rho)$ at the last queue as a function of traffic intensity ρ with the RQ approximation $c^2_{Z^*}(\rho)$ in (29) (left), and the IDW at the last queue over the interval $[10^{-2}, 10^5]$ in log scale (right).

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., R. Morabito. 1996. Open queueing networks: optimization and performance evaluation models for discrete manufacturing systems. *Production and Operations Management* **5**(2) 163–193.

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. 1985. Queues with superposition arrival processes in heavy traffic. *Stochastic Processes and Their Applications* **21** 81–91.

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.