Operations Research

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

Time-Varying Robust Queueing

Ward Whitt, Wei You

To cite this article:
Ward Whitt, Wei You (2019) Time-Varying Robust Queueing. Operations Research 67(6):1766-1782. https://doi.org/10.1287/opre.2019.1846

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

OPERATIONS RESEARCH
Vol. 67, No. 6, November–December 2019, pp. 1766–1782
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Time-Varying Robust Queueing

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

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

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

Received: August 28, 2016
Revised: June 30, 2017; December 8, 2018
Accepted: December 17, 2018
Published Online in Articles in Advance: September 4, 2019

Subject Classifications: queues: nonstationary, approximations, limit theorems
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2019.1846

Abstract. We develop a time-varying robust-queueing (TVRQ) algorithm for the continuous-time workload in a single-server queue with a time-varying arrival-rate function. We apply this TVRQ to develop approximations for the periodic steady-state expected workload in models with a periodic arrival-rate function. We apply simulation and asymptotic methods to examine the performance of periodic TVRQ (PRQ). We find that PRQ predicts the mean of the periodic distribution and even the full distribution (specified by the quantiles) remarkably well. We show that the PRQ converges to a proper limit in appropriate long-cycle and heavy-traffic regimes and coincides with long-cycle fluid limits and heavy-traffic diffusion limits for long cycles.

Funding: Financial support was received from the National Science Foundation, Division of Civil, Mechanical and Manufacturing Innovation [Grants 1265070 and 1634133].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2019.1846.

Keywords: robust queueing theory • time-varying arrival rates • nonstationary queues • periodic queues • heavy traffic

1. Introduction

Queueing has long played a prominent role in operations research (OR) applications. For example, early OR studies include traffic delays at toll booths by Edie (1954), letter delays at post offices by Oliver and Samuel (1962), airplane landing delays at airports by Koopman (1972), and dispatching delays for police patrol cars by Kolesar et al. (1975). As in many other OR applications, the arrival processes in these applications all have time-varying (TV) arrival rates. Thus, the natural queueing models require simulation or nonstandard analysis techniques beyond elementary stochastic textbooks.

Those four OR studies also illustrate two of the most important analytical techniques for analyzing TV queueing models. First, the papers by Edie (1954) and Oliver and Samuel (1962) illustrate that a relatively simple deterministic analysis can be used when the TV arrival rate tends to dominate the randomness. The other papers by Koopman (1972) and Kolesar et al. (1975) illustrate how numerical methods for systems of TV ordinary differential equations (ODEs) can be applied to calculate TV performance measures for the TV Markovian $M_t/M_t/s_t$ queueing model, which has a nonhomogeneous Poisson process (M_t) as its arrival process, and possibly a TV service rate and number of servers as well, because the number of customers in the system evolves as a TV birth-and-death process, so that its TV transition probability density function evolves according to a system of ODEs, often called the Kolmogorov forward equations.

The ODE approach to the TV $M_t/M_t/s_t$ queueing model has become the accepted analytical approach. The ODE approach is complicated by the fact that there are infinitely many ODEs in the system of equations, but that difficulty can be circumvented by truncating to a finite system, as was done by Koopman (1972) and Kolesar et al. (1975). Improved computer power has made this approach easier to apply.

Further progress with the ODE approach has also been made by introducing other approximations. Much more efficient ODE algorithms for the TV mean and variance were subsequently obtained by Rothkopf and Oren (1979) by using closure approximations to dramatically reduce the number of equations; also see Taaffe and Ong (1987), Ong and Taaffe (1989), and others.

Despite the successes of the ODE approach to TV queues, there are two deficiencies. First, the ODE approach only applies to TV Markov processes. Second, just like computer simulation and some other numerical approaches, such as the numerical-transform-inversion algorithm of Choudhury et al. (1997a), the ODE approach yields the numerical values of performance measures, but it does not otherwise provide any structural insight.

This second deficiency has recently been addressed by Massey and Pender (2013) and Pender and Massey (2017) by developing closure approximations for

the $M_t/M_t/s$ model and more general TV Markovian systems in the context of many-server heavy-traffic (MSHT) limits as in Mandelbaum et al. (1998), which yield deterministic fluid and stochastic diffusion approximations. They use the closure approximation to greatly improve the numerical accuracy of the MSHT diffusion approximations.

However, no such link has yet been provided between numerical algorithms and the very different conventional heavy-traffic (HT) limits for single-server models. In fact, the HT limits for TV single-server queues tend to be quite intractable themselves, as can be seen from Mandelbaum and Massey (1995) and Whitt (2014, 2016), so that we need new tractable approximation methods.

1.1. Main Contributions

1. In this paper, we introduce a time-varying robust queueing (TVRQ) approach to single-server queueing systems that addresses the two deficiencies mentioned above. In particular, we develop a TVRQ algorithm to approximate the TV workload in the non-Markov $G_t/G_t/1$ single-server queue. Like Rothkopf and Oren (1979), we focus on the special case of the dynamic steady-state behavior of a system with a periodic arrival rate. In doing so, we establish new periodic TVRQ (PRQ). This paper evidently is the first application of robust optimization to study the performance of a queueing model with time-varying arrival rates.

2. Even for the stationary model, we contribute by extending Whitt and You (2018b) to approximate all quantiles as well as the mean. The PRQ provides remarkably tractable approximations; for example, see (20), (22), and (28). Extensive simulation experiments confirm that the quantile connection is remarkably effective.

3. As in Whitt and You (2018b), we develop a nonparametric approximation by exploiting the index of dispersion for work (IDW) to represent the variability of the total input of work over time, independent of its mean. We use the IDW to develop TVRQ and PRQ for models with stochastic dependence as well as a time-varying arrival-rate function. The IDW is convenient for separately characterizing these two important causes of congestion. The nonparametric approach also provides a vehicle to connect the modeling to large data sets.

4. We establish new HT limits for PRQ in the $G_t/G/1$ model. These new HT limits exploit the HT scaling introduced in Whitt (2014, 2016) and so go beyond the earlier HT literature. In particular, time scaling is used within the deterministic arrival-rate function, so that the length of the periodic cycle grows with the traffic intensity ρ. We show that the HT limits for PRQ and the original model do not coincide in general, but they do in associated long-cycle and heavy-traffic double limits; see Section 6.

1.2. Related Literature

There is a substantial literature on TV single-server queues, which can be divided into three main categories: (i) structural results (e.g., definition and existence of processes), illustrated by Harrison and Lemoine (1977), Lemoine (1981), Heyman and Whitt (1984), Lemoine (1989), and Rolski (1989), (ii) numerical algorithms, as discussed above, and (iii) asymptotic methods and approximations by Newell (1968a, b, c), Keller (1982), Massey (1985), Mandelbaum and Massey (1995), and Whitt (2014, 2016). The present paper falls in the last two categories.

Robust optimization is a relatively new approach to difficult stochastic models. As in Beyer and Sendhoff (2007), Ben-Tal et al. (2009), and Bertsimas et al. (2011a), the main idea is to replace a difficult stochastic model by a tractable optimization problem. We replace an "average-case" expected value by a "worst-case" optimization, where stochastic process sample paths are constrained to belong to uncertainty sets. From a pure-optimization-centric view of the operations research landscape, robust optimization might be viewed as a way to replace stochastic modeling entirely. However, we think of robust optimization as a useful tool that supplements existing tools in our stochastic toolkit. Accordingly, much of this paper is devoted to establishing connections between PRQ and established queueing theory.

Our work on TVRQ builds on our previous paper, Whitt and You (2018b), which developed robust queueing (RQ) algorithms to approximate the expected steady-state waiting-time and workload in stationary single-server queues, aiming especially to capture the impact of dependence among interarrival times and service times. In turn that paper builds on the RQ formulation of Bandi et al. (2015), which has precedents in earlier work such as Bertsimas and Thiele (2006), Bertsimas et al. (2011b), and references cited there. The principal difference here is that we focus on the TV performance of a TV model instead of the steady-state performance of a stationary model.

Bandi et al. (2018) have also developed an RQ formulation for the transient behavior of stationary models, which tends to be a quite different (but still challenging) problem (and for which there is a large literature, which we do not review here). We remark that the performance of a queueing model with time-varying arrival-rate function can be approximated by the iterative transient analysis of the associated model with a piecewise-constant arrival-rate function, but that approach introduces another level of approximation and is not easy to implement. Indeed, the iterative transient approach to TV queues has evidently has been attempted only once, by Choudhury et al. (1997a).

1.3. Organization

In Section 2 we formulate TVRQ. In Section 3 we narrow our scope to focus on PRQ, introduce our framework to approximate the quantiles of the steady-state workload, and describe the simulation experiments. In Sections 4 and 5 we study PRQ for underloaded models and overloaded models, respectively. In Section 6 we establish heavy-traffic limits for PRQ. Supplementary material appears in the e-companion (EC, available online), including proofs and additional simulation examples.

2. TVRQ for the Steady-State Workload in the $G_t/G_t/1$ Queue

In Section 2.1 we introduce the general time-varying $G_t/G_t/1$ model and define the steady-state workload at each time in that model. In Section 2.2 we develop the TVRQ approximation, and in Section 2.3 we express it in terms of the index of dispersion for work.

2.1. The Steady-State Workload in the $G_t/G_t/1$ Queue

We consider a time-varying version of the standard single-server queue with unlimited waiting space and the first-come first-served service discipline, which we call the $G_t/G_t/1$ queue. As in Whitt and You (2018b), we will exploit a reverse-time construction of the workload process, but here we will directly construct the steady-state workload at time t. For that purpose, let $A_t(s)$ be the number of arrivals in interval $[t-s, t]$. As in Whitt (2015), let the service requirements be specified separately from the rate at which service is provided. Let service be provided at a time-varying rate $\mu(u)$ at time u, where μ is a right-continuous deterministic nonnegative function with left limits, so that the cumulative service rate available in the interval $[t-s, t]$ is

$$M_t(s) \equiv \int_{t-s}^{t} \mu(u)\, du, \quad s \geq 0. \tag{1}$$

Let the service requirement of customer k be V_k, indexed going backward from time t. Let the (potential) net-input of work in the interval $[t-s, t]$, $s \geq 0$, be

$$X_t(s) \equiv \sum_{k=1}^{A_t(s)} V_k - M_t(s), \quad t \geq 0. \tag{2}$$

Then the steady-state workload at time t is

$$W_t \equiv \sup_{s \geq 0}\{X_t(s)\}, \tag{3}$$

which we assume is almost surely finite.

For our supporting mathematical results and simulation examples, we will impose more structure. We impose one-dimensional partial characterizations of the variability of the arrival and service processes by assuming that the arrival process A takes the form of

$$A_t(s) = N(\Lambda_t(s)), \quad s \geq 0, \tag{4}$$

where the base process N is a unit-rate stationary point process satisfying the functional central limit theorem (FCLT)

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt] \Rightarrow c_a B_a \quad \text{in} \quad \mathcal{D}, \tag{5}$$

with $\Rightarrow$ denoting convergence in distribution, B_a being standard (drift 1, diffusion 1) Brownian motion (BM), and $\mathcal{D}$ the function space of (right-continuous with left limits) sample paths as in Whitt (2002), whereas the cumulative arrival rate function is

$$\Lambda_t(s) \equiv \int_{t-s}^{t} \lambda(u)\, du, \quad t \geq 0, \tag{6}$$

with the arrival-rate function λ being a deterministic nonnegative function in $\mathcal{D}$ (e.g., ensuring that the integral is well defined). If N is a Poisson process, then A is a nonhomogeneous Poisson process, but we allow other possibilities. Similarly, we assume that $\{V_k\}$ is a stationary sequence, independent of the process N, with $E[V_k] = 1$ satisfying the FCLT

$$\hat{S}_n(t) \equiv n^{-1/2}\left[\sum_{k=1}^{\lfloor nt \rfloor} V_k - nt\right] \Rightarrow c_s B_s \quad \text{in} \quad \mathcal{D}, \tag{7}$$

where B_s is a BM independent of B_a. The actual service times are relatively complicated; see section 3.1 of Whitt (2015). However, we will primarily focus on the standard special case $\mu(t) \equiv 1$, where the service times coincide with the service requirements. If $\mu(t) \equiv 1$, then W_t is the usual virtual waiting time. More generally, the virtual waiting time can be expressed in terms of the workload as a first passage time, as in lemma 4.1 of Ma and Whitt (2019).

From all past work (e.g., theorem 1 of Massey 1985), it is known that the performance at time t depends strongly on the loading, which depends on the history of the rates before time t, as characterized by the *time-varying traffic intensity*

$$\rho^*(t) \equiv \sup_{s \geq 0}\{\Lambda_t(s)/M_t(s)\} \tag{8}$$

for Λ_t in (6) and M_t in (1), which is to be distinguished from the *instantaneous traffic intensity*

$$\rho(t) \equiv \lambda(t)/\mu(t). \tag{9}$$

The model is called overloaded (OL), underloaded (UL), and critically loaded (CL) at time t if $\rho^*(t) > 1$, < 1, and $= 1$, respectively.

Remark 1 (An Alternative Representation). Combining (2), (3), and (6), we have the following equivalent representation of the steady-state workload

$$W_t = \sup_{s\ge 0}\left\{\sum_{k=1}^{N(\Lambda_t(M_t^{-1}(s)))} V_k - s\right\},$$

which can be viewed as an equivalent system with alternative arrival-rate function $\Lambda_t(M_t^{-1}(s))$.

2.2. Time-Varying Robust Queueing

From (3), we see that the steady-state workload at time t can be formulated directly a supremum. For our TVRQ, we apply robust optimization in the setting of Section 2.1 by replacing the stochastic model of the reverse-time net input process $X_t(s)$ in (2) and (3) by an appropriate deterministic uncertainty set $\mathcal{U}_t$ and then analyzing the worst case scenario. In particular, we let the TVRQ approximation of the steady-state workload at time t be

$$W_t^* \equiv \sup_{X_t\in\mathcal{U}_t}\sup_{s\ge 0}\{X_t(s)\}, \tag{10}$$

where $\mathcal{U}_t$ is the deterministic uncertainty set

$$\mathcal{U}_t \equiv \{X_t(s)\in\mathbb{R} : X_t(s) \le E[X_t(s)] + bSD(X_t(s)),\ \ s\ge 0\}, \tag{11}$$

with SD being the standard deviation and b being a parameter to be specified.

The uncertainty set in (11) is a natural time-varying generalization of the uncertainty sets in Whitt and You (2018b), which are similar to the ones used in Bandi et al. (2015). The main idea is that (11) can be based on a Gaussian approximation for $X_t(s)$, assuming that the supremum is attained for s not too small, which in turn is supported by an FCLT for $X_t(s)$ in (2), which follows from the assumed FCLTs in (5) and (7); see the electronic companion of Whitt and You (2018b).

For applications, the practical meaning of the Gaussian approximation for the net input process $X_t(s)$ supporting (11) is that our TVRQ approximation is intended for high-volume systems. High-volume means high arrival rates and service rates, which we achieve by scaling time. We are also primarily aiming to treat large-scale systems. Large scale is achieved by having the system operate under heavy-traffic conditions (i.e., by having high instantaneous traffic intensities over extended periods). For large-scale high-volume systems, the supporting FCLTs are appropriate, being intimately related to the heavy-traffic limits for the queueing model. We will establish new heavy-traffic limits that will further justify the connection.

As in lemma EC.1 of Whitt and You (2018b), we can interchange the order of the suprema in (10) and write

$$W_t^* \equiv \sup_{s\ge 0}\{E[X_t(s)] + bSD(X_t(s))\}, \tag{12}$$

where again $X_t(s)$ is defined in (2).

2.3. TVRQ Formulation Using the Index of Dispersion for Work

As in Whitt and You (2018b), let the index of dispersion for work in the underlying (time-homogenous) process be

$$I_w(t) \equiv \frac{\mathrm{Var}\left(\sum_{k=1}^{N(t)} V_k\right)}{E[\sum_{k=1}^{N(t)} V_k]} = t^{-1}Var\left(\sum_{k=1}^{N(t)} V_k\right), \quad t\ge 0, \tag{13}$$

with the last relation holding because $E[N(t)] = t$ and $E[V_k] = 1$. Clearly, the IDW is just a scaled version of the variance function of the total input process, but it is conveniently scaled to be independent of the rate. When the service requirements are independent and identically distributed (i.i.d.) with squared coefficient of variation (scv, variance divided by the square of the mean) c_s^2,

$$I_w(t) = I_a(t) + c_s^2, \quad t\ge 0, \tag{14}$$

where $I_a(t)$ is the *index of dispersion for counts* (IDC) of the base arrival process N, defined by

$$I_a(t) \equiv \frac{\mathrm{Var}(N(t))}{E[N(t)]} = t^{-1}\mathrm{Var}(N(t)), \quad t\ge 0, \tag{15}$$

as in section 4.5 of Cox and Lewis (1966). When N is Poisson, $I_a(t) = 1$, $t\ge 0$.

For the net input process $X_t(s)$ in (2),

$$E[X_t(s)] = \Lambda_t(s) - M_t(s) \quad \text{and}$$
$$\mathrm{Var}(X_t(s)) = \mathrm{Var}\left(\sum_{k=1}^{N(\Lambda_t(s))} V_k\right) = \Lambda_t(s)I_w(\Lambda_t(s)), \tag{16}$$

so that we can express the TVRQ representation for the steady-state workload at time t in terms of the IDW as

$$W_t^* \equiv \sup_{s\ge 0}\left\{\Lambda_t(s) - M_t(s) + b\sqrt{\Lambda_t(s)I_w(\Lambda_t(s))}\right\}, \tag{17}$$

where Λ_t and M_t are defined in (6) and (1), whereas I_w is the IDW defined in (13).

Example 1 (A Markov Model). An important special case is the associated Markov model, where N is a rate-1 Poisson process whereas $\{V_k\}$ is an i.i.d. sequence of mean-1 random variables with scv c_s^2, so that the total input of work over $[0,t]$ is a nonhomogeneous compound Poisson process. In this case, by (14), $I_w(t) = 1 + c_s^2$

for all t, so that the IDW plays a relatively trivial role. In this case,

$$W_t^* = \sup_{s\ge 0}\left\{\Lambda_t(s) - s + b\sqrt{(1+c_s^2)\Lambda_t(s)}\right\} \tag{18}$$

for Λ_t in (6). □

3. Periodic Robust Queueing

Henceforth in this paper we will narrow the scope and focus on the special case of periodic TVRQ, but much of what follows should be applicable more generally. In particular, we will assume that $\mu(s) \equiv 1, s \ge 0$, and λ is a periodic nonnegative function with period c and average rate

$$\rho \equiv c^{-1}\int_0^c \lambda(s)\,ds < 1, \tag{19}$$

which makes the steady-state workload W_t in (3) and the TVRQ W_t^* in (17) periodic with period c as well. We then let

$$W_y^* \equiv \sup_{s\ge 0}\left\{\Lambda_{yc}(s) - s + b\sqrt{\Lambda_{yc}(s)I_w(\Lambda_{yc}(s))}\right\}, \quad 0 \le y \le 1 \tag{20}$$

be the TVRQ at time yc, which we refer to as "position y in the cycle." As before, Λ_t comes from (6), and I_w comes from (13)–(15). We understand that W_y^* is an approximation for W_{yc}.

In Section 3.1 we introduce a new framework for exploiting the PRQ parameter b to approximate the full distribution of W_y. In Section 3.2 we describe our simulation experiments that we use to study PRQ.

3.1. Approximating the Full Distribution of W_y

In this section, we show how PRQ W_y^* in (20) with the PRQ parameter b can be used to approximate the full distribution of the stochastic steady-state workload W_{yc} in (3) as a function of y, $0 \le y \le 1$, which we do via quantiles. Hence, we refer to this as the PRQ(b) algorithm.

In Whitt and You (2018b), we established the connection between RQ and stochastic queues in the case of a stationary model. In particular, we found that the steady-state mean is often well approximated by letting $b = \sqrt{2}$; that choice makes RQ correct for the Kingman bound for $GI/GI/1$ [corollary 1 in Whitt and You (2018b)], the Pollaczek-Khintchine formula for $M/GI/1$ [corollary 3 in Whitt and You (2018b)], heavy-traffic and light-traffic limits for $G/G/1$ [theorem 5 in Whitt and You (2018b)], and can be explained by an exact analysis of Levy processes [section EC.3.2 in Whitt and You (2018b)].

From the form of PRQ(b), it is evident that as b increases, the approximation should apply more to the tail of the distribution. We find that a useful connection can be made between the parameter b and the quantiles of the distribution of the steady-state workload W_{yc} at position y within a cycle. For a nonnegative random variable Z and $0 < p < 1$, let the p^{th} *quantile* of (the distribution of) Z be

$$Z(p) \equiv \inf\{z \ge 0 : P(Z \le z) = p\}, \quad 0 < p < 1, \tag{21}$$

(i.e., the inverse of the cumulative distribution function [cdf]). We propose the approximation

$$W_{yc}(\Pi(b)) \approx W_y^*(b), \tag{22}$$

where $W_y^*(b)$ denotes PRQ in (20), whereas $\Pi : (-\infty, \infty) \to (0,1)$ is a one-to-one continuous function chosen to map the PRQ parameter b into the quantile level p of W_{yc}.

As indicated in Section 2.1, we find that the form of the mapping $\Pi(b)$ should depend on the loading. To proceed, we focus on the maximum TV traffic intensity, defined by

$$\rho^{\uparrow} \equiv \sup\{\rho^*(t) : 0 \le t \le c\}, \tag{23}$$

for ρ^* in (8). The periodic model is called overloaded, underloaded, and critically loaded if $\rho^{\uparrow} > 1$, $\rho^{\uparrow} < 1$, and $\rho^{\uparrow} = 1$, respectively. In Sections 4 and 5 we examine PRQ in the UL and OL cases. We discuss PRQ in the CL case in Section EC.6.

3.2. Simulation Experiments

For simulation comparisons, we will focus on the sinusoidal special case

$$\lambda(t) \equiv \rho + \beta\sin(2\pi\gamma t), \quad t \ge 0, \quad \text{and} \quad c \equiv c(\gamma) \equiv 1/\gamma, \tag{24}$$

with parameter vector (ρ, β, γ). We assume that $\beta \le \rho < 1$ to ensure that the arrival rate is always nonnegative and periodic steady state is well defined. In Section 6 when we consider heavy-traffic limits, we will let the parameter pair (β, γ) depend on ρ.

For these simulations, we consider the $GI_t/GI/1$ model with arrival rate function in (24) and i.i.d. service times $\{V_k\}$ with $E[V_k] = 1$ and scv c_s^2 that are independent of a base rate-1 stationary renewal process N used to generate the arrival process via (4). Let c_a^2 be the scv of an interarrival time in the ordinary renewal process associated with N. Our examples use exponential (M), Erlang (E_k), hyperexponential (H_2, mixture of two exponentials with balanced means; p. 137 of Whitt 1982), and lognormal distributions, with the scv specified in parentheses for each experiment. By varying the level of variability in the arrival and processes, we can expose and separate the impact of the stochastic variability from the impact of the deterministic time-variability provided by the

time-varying arrival rate in (24). We describe the simulation methodology in Section EC.2.

4. Underloaded Models

In this section we investigate PRQ for UL models, which of course includes the stationary model as a special case. At first glance, the proposed scheme in (22) deviates from our previous approximation that focused on the steady-state mean in the stationary model in Whitt and You (2018b), but in Section 4.1 we show that RQ can be generalized to an RQ(b) algorithm that approximates the quantiles in addition to the mean. In Section 4.2, we show that PRQ(b) is quite effective in approximating the quantiles of the steady-state workload for UL models.

4.1. RQ(b) for Stationary Queueing Models

For stationary queues, the standard heavy-traffic approximation implies that the steady-state workload W should be approximately exponentially distributed; see sections 5.7 and 9.3 in Whitt (2002). In particular, for mean-1 service and traffic intensity ρ,

$$P(W > x) \approx e^{-x/m}, \quad x \geq 0, \quad \text{for} \quad m \equiv \frac{\rho c_x^2}{2(1-\rho)}. \tag{25}$$

Thus, for quantile p of W, denoted by $W(p)$, we have $P(W \leq W(p)) \approx 1 - e^{-W(p)/m} = p$, so that

$$W(p) \approx -\ln(1-p)m \tag{26}$$

for m in (25).

On the other hand, if we apply theorem 2 of Whitt and You (2018b) to the $M/GI/1$ queue or the reflected Brownian motion approximation, then we get

$$W^*(b) = \frac{b^2 m}{2}. \tag{27}$$

To match the actual mean in $M/GI/1$ for all ρ and to match the mean in heavy-traffic and light-traffic limits, corollary 3 and theorem 5 of Whitt and You (2018b) imply that we should chose $b^2 = \sqrt{2}$ in Whitt and You (2018b). Hence, further connection can be made by equating (26) and (27) to obtain an approximation for the desired function Π in (22), getting

$$p \approx \Pi(b) \equiv 1 - e^{-b^2/2}. \tag{28}$$

For the stationary model, we propose the RQ(b) algorithm as in (20) with (22) and (28), where we restrict (20) to stationary arrival rate functions.

By (26), for an exponential random variable, the mean coincides with the $p = 1 - e^{-1} \approx 0.632$ quantile. By (28), this quantile corresponds to $b = \sqrt{2}$. Hence, the RQ(b) algrithm reduces to the RQ algorithm for the steady-state mean workload in Whitt and You (2018b).

4.2. PRQ(b) for Underloaded Models

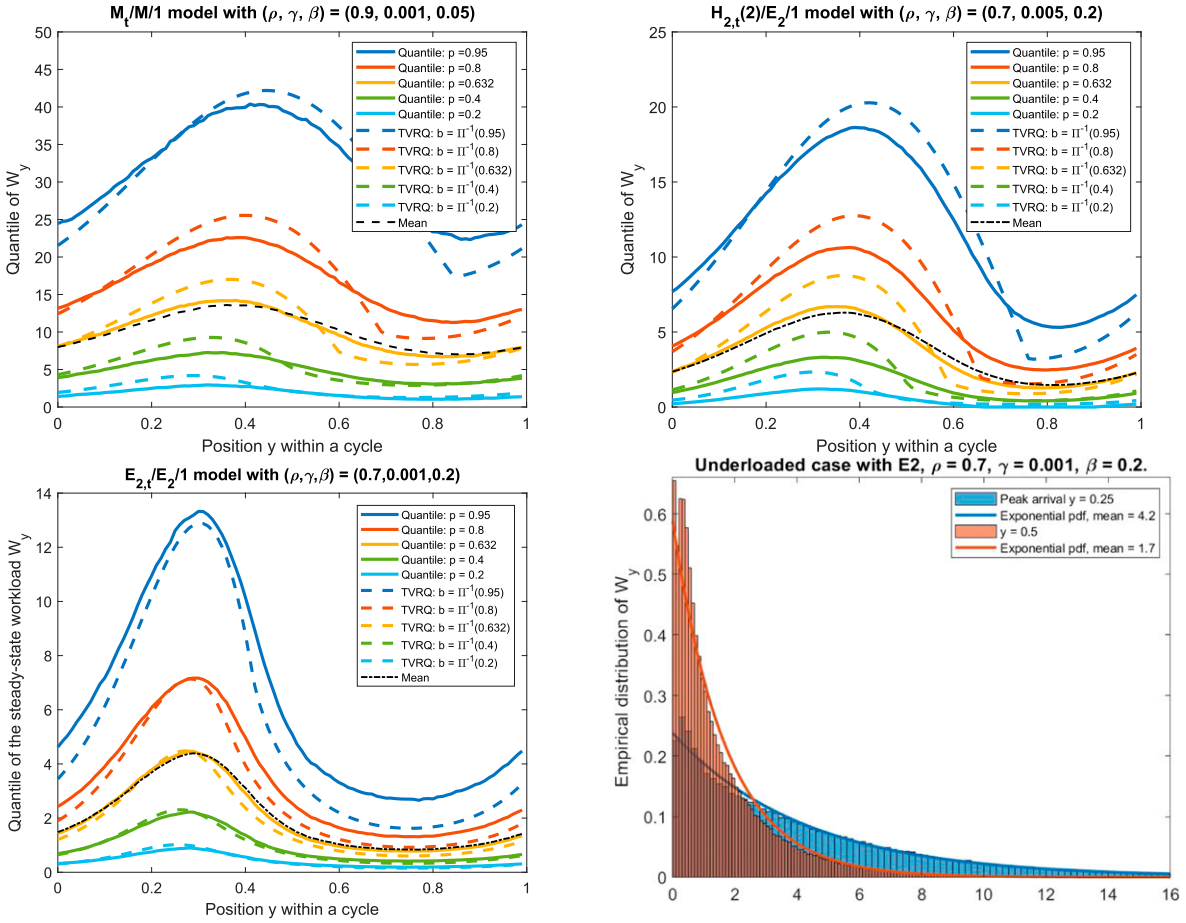
We now return to the periodic model. To start, we note that an alternative approximation for UL models is the *pointwise stationary approximation* (PSA) as in Whitt (1991), Green and Kolesar (1991), and Massey and Whitt (1998). The idea in PSA is to approximate the time-varying performance at time t in the UL $G_t/G_t/1$ model by using the steady-state performance of the stationary $G/G/1$ model having the parameters that prevail at time t. In our setting, the PSA is appropriate if the cycle length is sufficiently long that the arrival rate does not change too quickly (relative to the service times). The periodic queue then performs at each time approximately the same as the PSA stationary queue, which is discussed in Section 4.1. As a result, we propose the same mapping $\Pi(b)$ in (28) for UL periodic queues.

Figure 1 demonstrates the performance of PRQ(b) in the UL case. The first three plots in Figure 1 show the simulation estimates of the quantiles at the level of $p = 0.95, 0.8, 0.632, 0.4$, or 0.2 for three models. In each plot, we overlay the PRQ approximations of the quantiles in broken curves, calculated from (22) and (28). Figure 1 shows that (i) our PRQ framework for approximating the full distribution of W_y is very effective; (ii) the estimated mean is close to the 0.6321 quantile, and PRQ(b) with $b = \sqrt{2}$ serves as a good approximation for the mean in the UL case, as discussed in Section 4.1; and (iii) even though the exponential approximation draws on the HT limits, we see that our approximation works well under moderate traffic intensity, as demonstrated by the upper right and lower left plots. For the lower right plot, we show the empirical distribution of W_y at two locations of the cycle, $y = 0.25$ and $y = 0.5$. Both of them are well fitted by exponential distributions, showing that the exponential approximation is appropriate in our settings here.

In Figure 1 the cycle lengths are $c = 1{,}000$ with $\gamma = 0.001$, which is quite long, representing high-volume systems. In contrast, Figure 2 shows the performance of the $M_t/M/1$ model for shorter cycles. Figure 2 shows plots for all combinations of $\rho = 0.7$ and 0.9 and $\gamma = 0.01$ and 0.1. Figure 2 again shows that PRQ can be effective to approximate both the quantiles and the mean. Figure 2 also shows that PRQ accurately captures the asymptotically stationary performance that prevails in heavy traffic without the extra scaling of the arrival-rate function introduced in Whitt (2014). It also motivates our use of the scaling from Whitt (2014) in our heavy-traffic limits in Section 6.

To conclude this section, we return to consider PSA, which motivated our use of (28) for periodic UL models as well as stationary models. Unlike the right-hand plots in Figure 2, PSA predicts relatively rapid oscillations for short cycles, much like the PSA plot in figure 1 of Jennings et al. (1996) for many-server

Figure 1. (Color online) Comparison of the PRQ Quantile Approximation in (22) and (28) with Simulation Estimates of the Quantiles in the $M_t/M/1$ Model (Upper Left), $H_{2,t}(2)/E_2/1$ Model (Upper Right), and $E_{2,t}(2)/E_2/1$ Model (Lower Left)



Notes. The arrival-rate function is (24), with parameters specified in the title of the plot. For the quantile level, we consider $p = 0.95, 0.8, 0.632, 0.4$ and 0.2. The lower right shows the empirical distribution of W_y for the $E_{2,t}(2)/E_2/1$ model at two locations of the cycle: $y = 0.25$ and $y = 0.50$.

models. Figure 3 shows that PSA makes sense for long cycles but that PRQ provides an improvement. In the present context, we can combine RQ with PSA to obtain PSA-RQ. It suffices to change (17) (with $M(t) \equiv t$) to

$$X^*_{PSA,t} \equiv \sup_{s\geq 0}\left\{\Lambda(t)s - s + b\sqrt{\Lambda(t)sI_w(\Lambda(t)s)}\right\}$$
$$= \sup_{s\geq 0}\left\{-(1-\rho(t))s + \sqrt{\rho(t)sI_w(\rho(t)s)}\right\}, \qquad (29)$$

which corresponds to the RQ formula (27) in Whitt and You (2018b) with $\rho \equiv \rho(t) = \lambda(t) < 1$.

Figure 3 compares PRQ and PSA-RQ with simulation estimates for three different models with (24) for $\rho = 0.7$, $\beta = 0.2$, and $\gamma = 0.001$ (left) and $\gamma = 0.01$ (right). As in (25) of Whitt and You (2018b), Figure 3 shows the normalized mean workload $2(1-\rho)E[W_y]/\rho$ (which would be 1 in the $M/D/1$ model) as a function of the position y within the cycle.

Figure 3 shows that PRQ provides only a slight improvement over PSA-RQ for $\gamma = 0.001$ (left), but a significant improvement for $\gamma = 0.01$ (right). As before, Figure 3 shows that the quality of the approximation is excellent for the exponential distribution (M) and lower levels of variability but degrades for higher variability, serving as an upper bound at the peak (but not uniformly in y). Unlike PSA-RQ, PRQ provides remarkably good estimates of the location of the peak congestion. See Section EC.8.2 for more simulation comparisons.

5. Overloaded Models

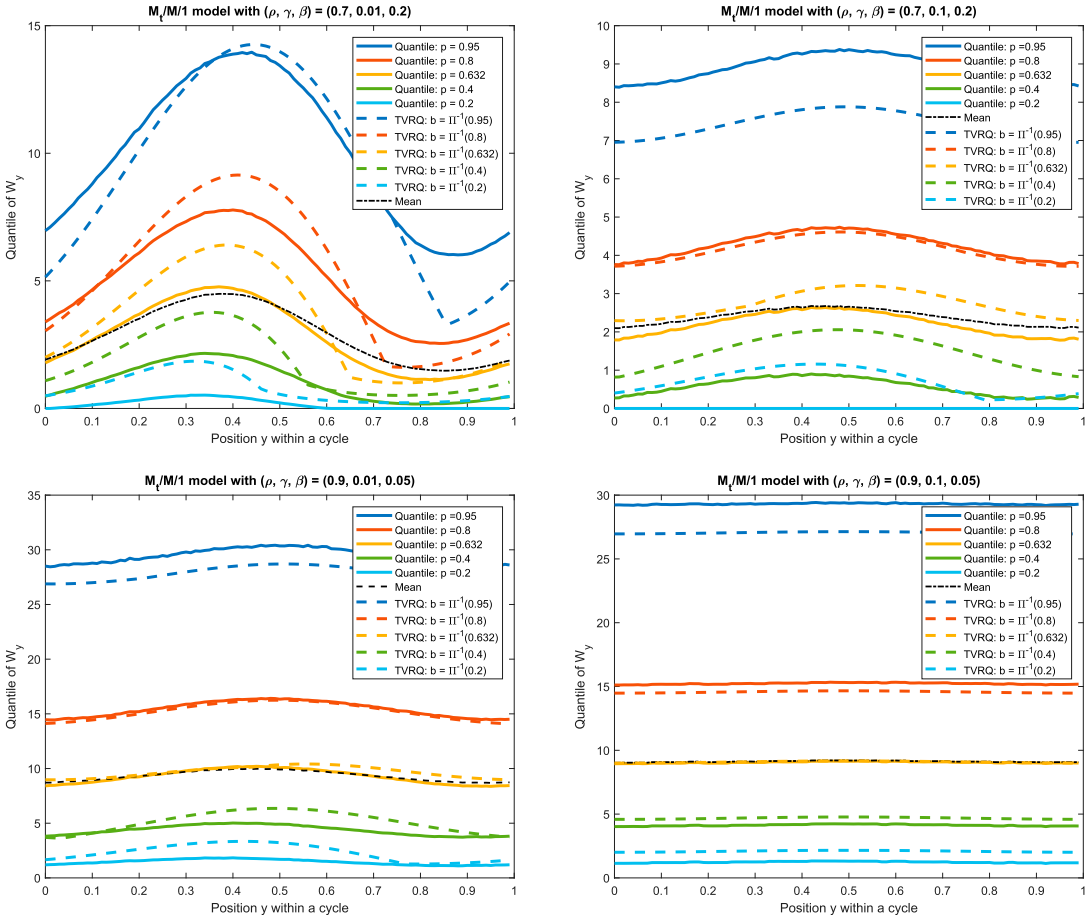
The behavior of OL models is quite different, especially at the peak. Because $\rho^{\uparrow} > 1$ for $\rho^{\uparrow}$ in (23), PSA does not apply at the peak.

5.1. Deterministic Approximations

For OL models, it makes sense to consider relatively simple deterministic approximations, which we obtain by assuming that there is no stochastic variability. One way to do so is to assume that $X(t) \equiv \Lambda(t) - M(t) = \Lambda(t) - t$ for all t. As a consequence,

$$W^*_{det,t} = W_{det,t} = \sup_{s\geq 0}\{X_t(s)\} = \sup_{s\geq 0}\{\Lambda_t(s) - s\}. \qquad (30)$$

Figure 2. (Color online) Comparing the PRQ Quantile Approximation in (22) and (28) with Simulation Estimates of the Quantiles in the $M_t/M/1$ Model for $(\rho, \gamma) = (0.7, 0.01)$ (Upper Left), $(0.7, 0.1)$ (Upper Right), $(0.9, 0.01)$ (Lower Left), and $(0.9, 0.01)$ (Lower Right)



Because the model is deterministic, TVRQ cannot provide an improved performance approximation, but we see that in this case TVRQ is giving the exact time-varying workload. We discuss this model further in Section EC.4, but we make two important observations. First, Proposition EC.1 shows that in the periodic case it suffices to do the supremum over one cycle. Second, the deterministic model is very helpful to identify the position $y^\uparrow$ where W_y attains its peak; for example, for the OL sinusoidal model in (24) with $\rho^\uparrow > 1$ in (23), measuring time in units of a cycle length, Corollary EC.2 implies that

$$W_{det,y^\uparrow} \equiv \sup_{0\le y\le 1}\{W_{det,y}\} \quad \text{for}$$
$$y^\uparrow = 0.5 - \arcsin(1-\rho)/\beta/2\pi. \tag{31}$$

Because the arrival rate has its peak at $y = 0.25$, the time lag in the peak of $W_{det,y}$ is $0.25 - \arcsin(1-\rho)/\beta/2\pi$, both measured in units of a cycle length.

5.2. Long-Cycle Fluid Limits

The deterministic model in (30) also arises by taking a long-cycle limit, for which we consider a family of periodic $G_t/GI/1$ stochastic models with growing cycle length indexed by the parameter γ. We assume that model γ has arrival-rate function

$$\lambda_\gamma(t) \equiv \lambda(\gamma t), \quad t \ge 0, \tag{32}$$

for a base periodic arrival-rate function λ. Thus, the arrival rate in model γ is periodic with cycle length $c_\gamma \equiv c/\gamma$. We will let $\gamma \downarrow 0$, so that $c_\gamma \to \infty$.

As regularity conditions for N, we assume that

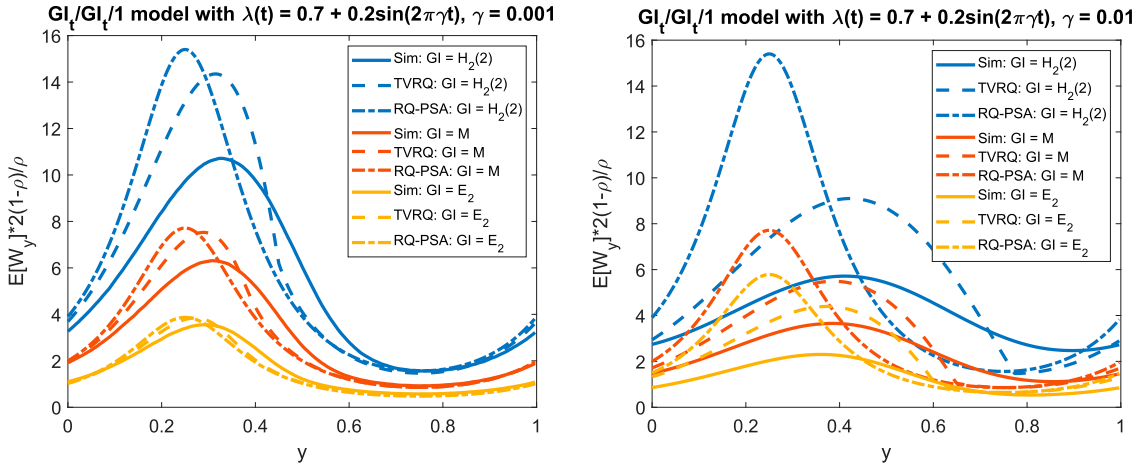
$$t^{-1}N(t) \to 1 \quad \text{as} \quad t \to \infty \quad \text{w.p.1} \tag{33}$$

and, for all $\epsilon > 0$, there exists $t_0 \equiv t_0(\epsilon)$ such that

$$|t^{-1}N(t) - 1| < \epsilon \quad \text{for all} \quad t \ge t_0 \quad \text{w.p.1.} \tag{34}$$

Both conditions hold when N is a Poisson process and can be anticipated more generally. We prove the

Figure 3. (Color online) A Comparison of PRQ in (20) and PSA-RQ in (29) with Simulation Estimates of the Normalized Steady-State Mean Workload $2(1-\rho)E[W_y]/\rho$ in the UL $GI_t/GI/1$ Model with Sinusoidal Arrival Rate in (24) Having $(\rho,\beta) = (0.7, 0.2)$ for $\gamma = 0.001$ (Left) and $\gamma = 0.01$ (Right), as a Function of the Position y in a Cycle



Note. Three cases for the underlying distributions are displayed $(H_2(2), M, E_2)$, being identical for arrival and service.

following result and provide additional discussion in Section EC.4.3.

Theorem 1 (Long-Cycle Fluid Limit). *For the periodic $G_t/GI/1$ model under conditions (33) and (34), including the scaling in (20) as a function of γ,*

$$(\gamma W_{\gamma,y}, \gamma W^*_{\gamma,y}(b)) \to (W_{det,y}, W_{det,y}) \quad as \quad \gamma \downarrow 0 \quad w.p.1 \tag{35}$$

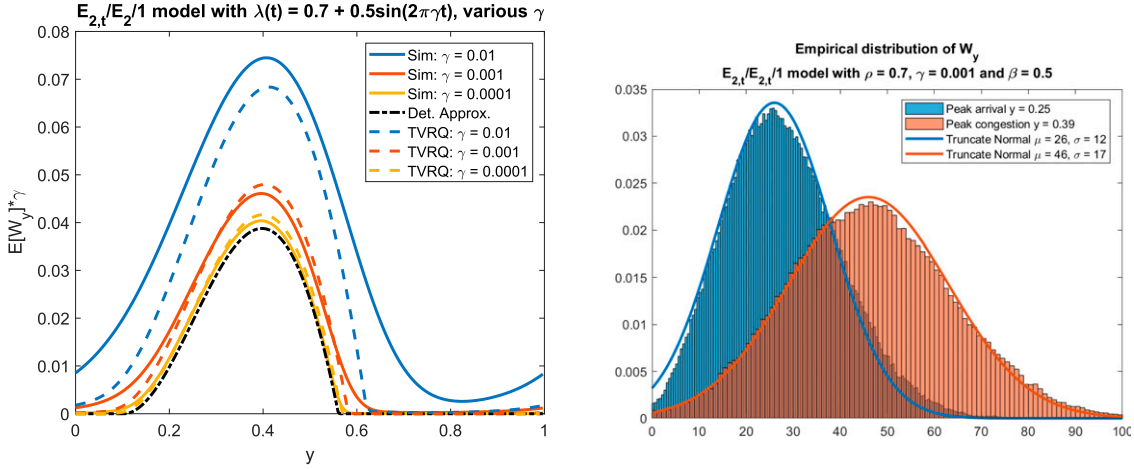
for any b, where $W_{det,y}$ is the deterministic workload in (30) at time yc within a cycle of length c, whereas $W^*_{\gamma,y}(b)$ is the PRQ(b) approximation in (20).

5.3. A Gaussian Approximation for the Quantiles

The connection to quantiles changes for OL models. Heavy-traffic theory indicates that W_{yc} in the OL period in the cycle should be approximately Gaussian, approximately equal in law to $X_{yc}(s)$ for an appropriate s (where the OL begins in the cycle); for example, see Newell (1968b), regions B and E in figure 4.1 of Mandelbaum and Massey (1995), and theorems 5.3.3(b) and 13.4.2 of Whitt (2002).

To illustrate, Figure 4 (left) compares PRQ in (20) using $b = 0.50$ developed below and the deterministic approximation in (30) to simulation estimates of the normalized steady-state mean workload $E[W_y]\gamma$, which is consistent with Theorem 1, in the $E_{2,t}/E_2/1$ model with sinusoidal arrival rate in (24), $\rho = 0.7$, $\beta = 0.5$, and three values of γ, as functions of the position y in a cycle. The deterministic approximation is not sensitive to changing cycle length as well as stochastic variability, but it is asymptotically exact as the

Figure 4. (Color online) PRQ in (20) and the Deterministic Approximation in (30) Compared with Simulation Estimates of the Normalized Steady-State Mean Workload $\gamma E[W_y]$ in the OL $E_{2,t}/E_2/1$ Model with Sinusoidal Arrival Rate in (24), $\rho = 0.7$, $\beta = 0.5$, and Three Values of γ, as Functions of the Position y in a Cycle (Left); and Estimates of the Distribution of W_y at the Location of the Peak of the Arrival Rate and of W_y (Right)



cycle length grows to infinity. Moreover, both PRQ and the deterministic approximation predict the location of the peak congestion very well, showing that it lags substantially after the peak of $\lambda(t)$, which is 0.25, again measuring time in cycle lengths. In particular, formula (31) predicts the peak congestion occurs at $y^{\uparrow} = 0.3975$, which is a significant time lag of 0.1475. Figure 4 (left) shows that both the deterministic approximation and PRQ predict this time lag very accurately. We have found that to be consistently true for both OL and UL models.

At this point, we proceeded experimentally. We looked at multiple $G_t/G_t/1$ models to estimate the function $\Pi(b)$ in (22) that relates the TVRQ parameter b to the sample quantiles. To illustrate, Figure 5 compares the quantiles for p ranging from 0.9 to 0.1 estimated by simulation to the PRQ(b) values associated with the parameter b to make PRQ(b) agree as closely as possible. In particular, we focus on $E_{4,t}/E_{4,t}/1$, $H_{2,t}(8)/H_{2,t}(8)/1$ and $E_{4,t}/H_{2,t}(8)/1$ models and an arrival rate function of $\lambda(t) = 0.9 + 0.8\sin(0.001 * 2\pi t)$.

First, Figure 5 shows that the match is remarkably good for all y. Second, Figure 5 (lower right) shows these numerical results fit to normal cdf's, for which there is remarkable consensus. As a simple overall approximation, we choose

$$\Pi(b) \approx \Phi(b; 0.5, 1.0), \tag{36}$$

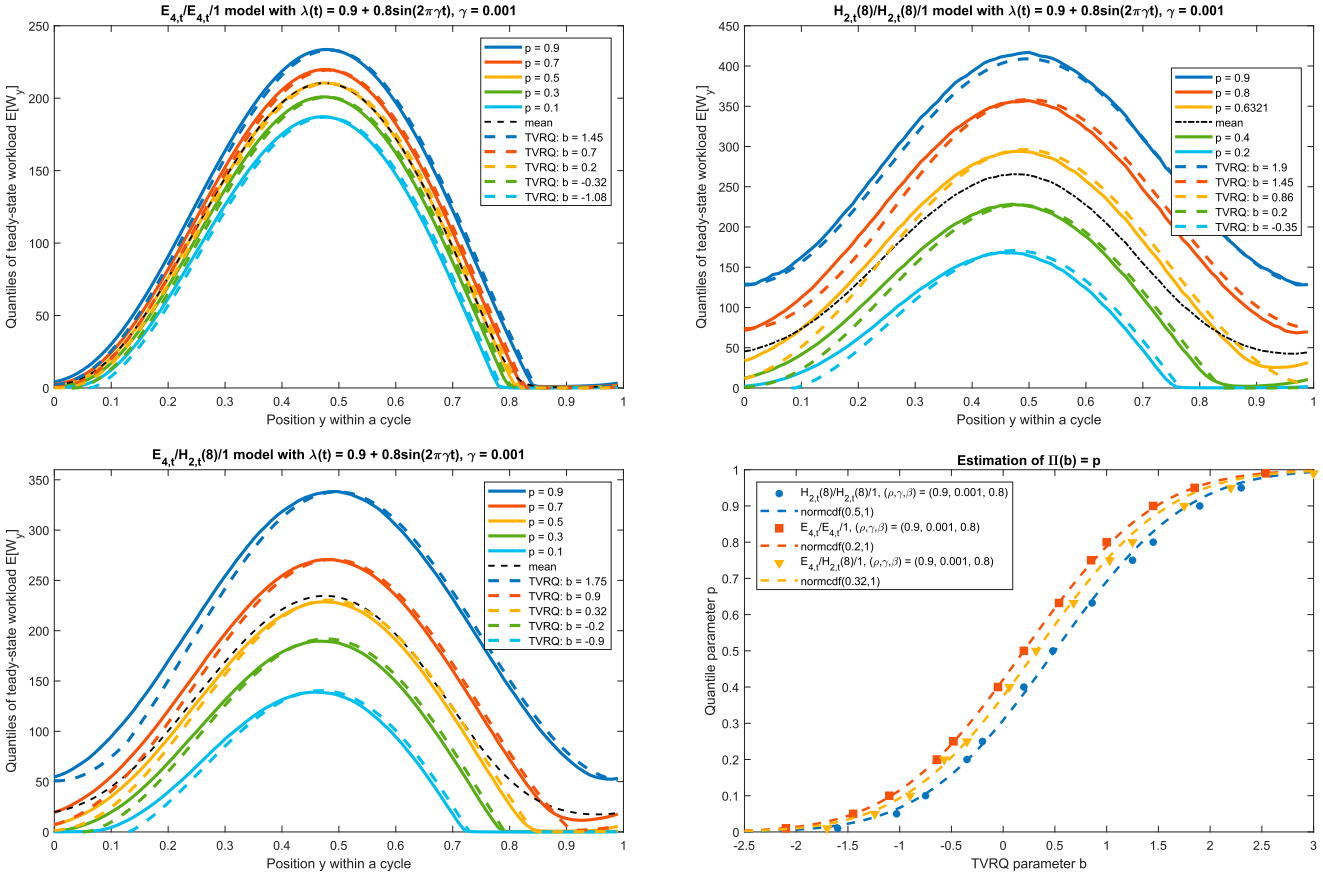
where $\Phi(x; m, \sigma^2) \equiv P(N(m, \sigma^2) \le x) = P(N(0,1) \le (x - m)/\sigma)$ for mean m and variance σ^2. If we want to approximate the mean, then we use $b = 0.5$ because $\Pi(0.5) = 0.5$, the median.

We then tested PRQ(b) with Π in (36) for a range of OL models. Figure 6 illustrates by showing the results for the $M_t/M/1$ model for the parameter vectors $(\rho, \beta, \gamma) = (0.9, 0.5, 0.001)$ and $(0.7, 0.5, 0.01)$. See Section EC.8.3 for more simulation comparisons.

6. Heavy-Traffic Limits for Periodic Queues

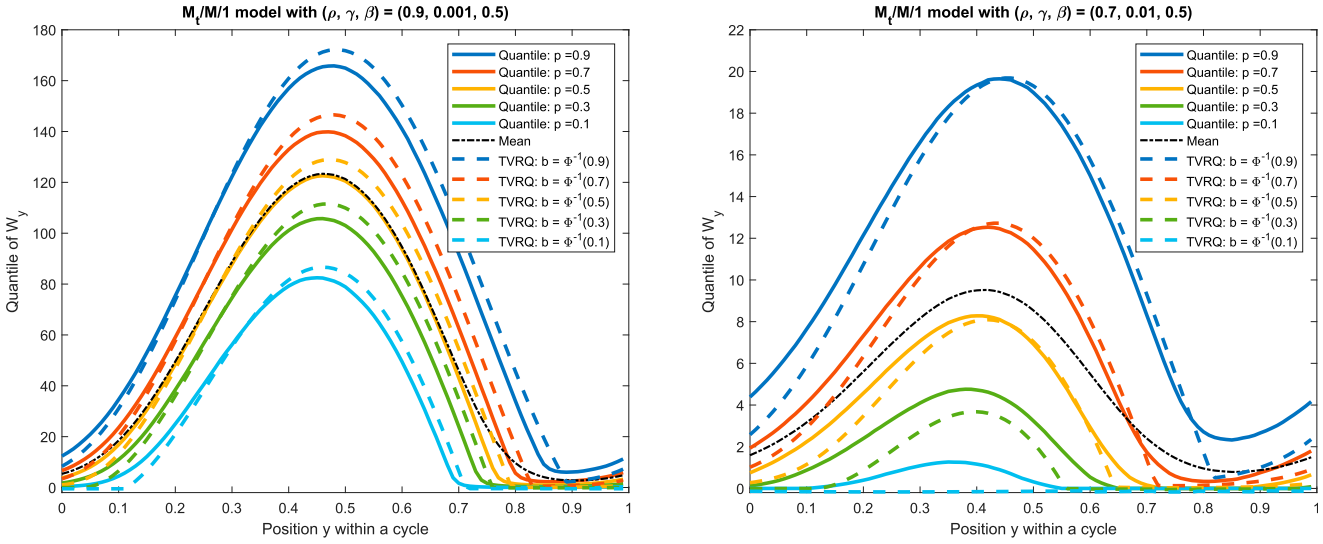
We now apply heavy-traffic limits to further study periodic robust queueing. In Section 6.1 we first review a heavy-traffic limit for periodic queues from Whitt (2014) and Ma and Whitt (2018a, b). In addition to the conventional heavy-traffic scaling of time in space, as in chapter 9 of Whitt (2002), these heavy-traffic limits

Figure 5. (Color online) A Comparison of Quantiles p Ranging from 0.9 to 0.1 Estimated by Simulation to the PRQ(b) Values Associated with the Parameter b to make PRQ(b) Agree as Closely as Possible



Notes. For the $E_{4,t}/E_{4,t}/1$, $H_{2,t}(8)/H_{2,t}(8)/1$ and $E_{4,t}/H_{2,t}(8)/1$ models and the sinusoidal arrival rate function in (24) with $(\rho, \beta, \gamma) = (0.9, 0.8, 0.001)$. These are fit to Gaussian cdfs in the lower right.

Figure 6. (Color online) A Comparison of Quantiles p Ranging from 0.9 to 0.1 Estimated by Simulation to the PRQ(b) Based on Π in (36) for the $M_t/M/1$ Model and the Sinusoidal Arrival Rate Function in (24) with $(\rho, \beta, \gamma) = (0.9, 0.5, 0.001)$ (Left) and $(0.7, 0.5, 0.01)$ (Right)



involve an additional scaling of the arrival rate function. In Section 6.2 we show how it can be used to generate a diffusion-based parametric PRQ. We then compare our proposed functional PRQ, the diffusion-based parametric PRQ, and the direct heavy-traffic diffusion approximation to simulation estimates of the time-varying mean workload. In Section 6.3 we develop new heavy-traffic limits for PRQ approximation. In Section 6.4 we establish new heavy-traffic limits combined with long-cycle limits. These involve the three cases: underloaded, overloaded, and critically loaded.

6.1. Heavy-Traffic Limit for the Workload Process in the Stochastic Model

We consider a family of models indexed by the long-run average traffic intensity ρ. To avoid notational confusion, we add a subscript d to the diffusion quantities. We let the cumulative arrival-rate function in model ρ be

$$\Lambda_{\gamma,\rho}(t) \equiv \rho t + (1-\rho)^{-1}\Lambda_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0, \tag{37}$$

so that the associated arrival-rate function is

$$\lambda_{\gamma,\rho}(t) \equiv \rho + (1-\rho)\lambda_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0, \tag{38}$$

where

$$\Lambda_{d,\gamma}(t) \equiv \int_0^t \lambda_{d,\gamma}(s)\,ds, \quad \lambda_{d,\gamma}(t) \equiv h(\gamma t), \quad \text{and} \quad \int_0^1 h(t)\,dt = 0 \tag{39}$$

with $h(t)$ being a periodic function with period 1. As a consequence, $\lambda_{d,\gamma}(t)$ is a periodic function with period $c_\gamma = 1/\gamma$, and $\lambda_{\gamma,\rho}(t)$ is a periodic function with period $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$. To ensure that $\lambda_{\gamma,\rho}$ is nonnegative, we assume that

$$h(t) \geq -\rho/(1-\rho), \quad 0 \leq t < 1, \tag{40}$$

which will be satisfied for all ρ sufficiently close to the critical value 1 provided that h is bounded below. In fact, we directly assume that

$$-\infty < h^{\downarrow} \equiv \inf_{0\leq t\leq 1}\{h(t)\} < \sup_{0\leq t\leq 1}\{h(t)\} \equiv h^{\uparrow} < \infty. \tag{41}$$

There are two primary cases of interest, $h^{\uparrow} < 1$ and $h^{\uparrow} > 1$. When $h^{\uparrow} < 1$, the instantaneous traffic intensity, which is $\lambda_{\gamma,\rho}(t)$, satisfies $\lambda_{\gamma,\rho}(t) < 1$ for all t and ρ. On the other hand, when $h^{\uparrow} > 1$, $\lambda_{\gamma,\rho}(t) > 1$ for some t. When $\lambda_{\gamma,\rho}(t) > 1$ for some t, the workload can reach very high values when time is scaled, because the cycles are very long. That takes us into the setting of Choudhury et al. (1997b).

Theorem 3.2 of Whitt (2014) and theorem 2 of Ma and Whitt (2018a) provide a heavy-traffic limit as $\rho \uparrow 1$ when $h^{\uparrow} < 1$ for the workload at time t starting empty at time 0, which we denote by $W_{\gamma,\rho}(t)$, in the periodic $G_t/GI/1$ model. This heavy-traffic limit is for the time-varying behavior starting empty, but it also applies to the periodic steady-state distribution except for the usual problem of interchanging the order of the limits as $\rho \uparrow 1$ and as $t \uparrow \infty$. We use the periodic steady-state of the limit to approximate the periodic steady-state of the periodic $G_t/GI/1$ queue.

To express the heavy-traffic limits, we use (37) and let

$$A_{\gamma,\rho}(t) \equiv N(\Lambda_{\gamma,\rho}(t)), \quad Y_{\gamma,\rho}(t) \equiv \sum_{k=1}^{A_{\gamma,\rho}(t)} V_k, \quad \text{and} \quad X_{\gamma,\rho}(t) \equiv Y_{\gamma,\rho}(t) - t, \quad t \geq 0. \tag{42}$$

Then $X_{\gamma,\rho}(t)$ is the net-input process and $W_{\gamma,\rho}(t)$ is the workload process, which is the image of $X_{\gamma,\rho}$ under the reflection map Ψ; that is,

$$W_{\gamma,\rho}(t) = \Psi(X_{\gamma,\rho})(t) = \sup_{0\le s\le t}\{X_{\gamma,\rho}(t) - X_{\gamma,\rho}(t-s)\}. \quad (43)$$

For the heavy-traffic functional central limit theorem, we introduce the scaled processes

$$\begin{aligned}
&\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt],\\
&\hat{A}_{\gamma,\rho}(t) \equiv (1-\rho)[A_{\gamma,\rho}\big((1-\rho)^{-2}t\big) - (1-\rho)^2 t],\\
&\hat{X}_{\gamma,\rho}(t) \equiv (1-\rho)X_{\gamma,\rho}\big((1-\rho)^{-2}t\big) \quad \text{and}\\
&\hat{W}_{\gamma,\rho}(t) \equiv (1-\rho)W_{\gamma,\rho}\big((1-\rho)^{-2}t\big), \quad t \ge 0. \qquad (44)
\end{aligned}$$

Let $\mathcal{D}^k$ be the k-fold product space of the function space $\mathcal{D}$. Again let e be the identity map in $\mathcal{D}$ (i.e., $e(t) \equiv t$, $t \ge 0$). Recall that $g(x) = o(x)$ as $x \to 0$ if $g(x)/x \to 0$ as $x \to 0$.

Theorem 2 (Heavy-Traffic FCLT, Theorem 3.2 of Whitt 2014 and Theorem 2 of Ma and Whitt 2018a). *For the family of $G_t/GI/1$ models indexed by (γ,ρ) with cumulative arrival-rate functions in (37), if $\hat{N}_n \Rightarrow c_a B_a$ as $n \to \infty$, where B_a is a standard Brownian motion, then*

$$(\hat{A}_{\gamma,\rho}, \hat{X}_{\gamma,\rho}, \hat{W}_{\gamma,\rho}) \Rightarrow (\hat{A}_\gamma, \hat{X}_\gamma, \hat{W}_\gamma) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \uparrow 1, \quad (45)$$

where

$$(\hat{A}_\gamma, \hat{X}_\gamma, \hat{W}_\gamma) \equiv (c_a B_a + \Lambda_{d,\gamma} - e, \hat{A}_\gamma + c_s B_s, \Psi(\hat{X}_\gamma)), \quad (46)$$

Ψ is the reflection map in (43), $\Lambda_{d,\gamma}$ is defined in (39), and B_a and B_s are two independent standard (mean 0 variance 1) *Brownian motions; that is, $\hat{W}_\gamma$ is reflected periodic Brownian motion (RPBM) with*

$$\hat{W}_\gamma = \Psi(c_a B_a + c_s B_s + \Lambda_{d,\gamma} - e) \stackrel{\mathrm{d}}{=} \Psi(c_x B + \Lambda_{d,\gamma} - e), \quad (47)$$

where $c_x^2 = c_a^2 + c_s^2$. The result remains valid if a term of order $o(1-\rho)$ is added to $\Lambda_{\gamma,\rho}$ in (37).

6.2. Three Periodic Approximations from Theorem 2

We directly can obtain three approximations for the mean workload in the periodic $G_t/G_t/1$ model from Theorem 2. In particular, the workload at fixed place y within a cycle for a system that started empty and has run for t time units is

$$W_{\gamma,\rho,y}(t) \stackrel{\mathrm{d}}{=} \sup_{0\le s\le t}\left\{\sum_{k=1}^{A_{\gamma,\rho,y}(s)} V_k - s\right\}, \quad (48)$$

where $A_{\gamma,\rho,y}(s) \equiv A_{\gamma,\rho}(y) - A_{\gamma,\rho}(y-s)$, $A_{\gamma,\rho}(t)$ is defined in (42), and V_k is a generic service time.

As a consequence, first there is the direct diffusion approximation based on (47),

$$\tilde{W}_{\gamma,\rho,y} \equiv \sup_{s\ge 0}\{\Lambda_{\gamma,\rho,y}(s) - s + c_x B(s)\}. \quad (49)$$

Second, there is the parametric PRQ (for the diffusion approximation) obtained from (49) using the mean and variance of BM in (49), namely,

$$\tilde{W}^{**}_{\gamma,\rho,y}(b) \equiv \sup_{s\ge 0}\{\Lambda_{\gamma,\rho,y}(s) - s + bc_x\sqrt{s}\}, \quad (50)$$

where we use $b = \sqrt{2}$ if we are interested in the mean, because this model is UL.

Finally, there is our proposed functional PRQ,

$$\tilde{W}^{*}_{\gamma,\rho,y}(b) \equiv \sup_{s\ge 0}\left\{\Lambda_{\gamma,\rho,y}(s) - s + b\sqrt{\Lambda_{\gamma,\rho,y}(s) I_w(\Lambda_{\gamma,\rho,y}(s))}\right\}, \quad (51)$$

where we again use $b = \sqrt{2}$ if we are interested in the mean. Note that (51) does not exploit the diffusion approximation and so should have advantages away from heavy traffic.

For all simulation examples in the section, we use the base sinusoidal arrival function in (24) with the scaling in (37)–(39), so that

$$\lambda_{\gamma,\rho} = \rho + (1-\rho)h^{\uparrow}\sin(2\pi(1-\rho)^2\gamma t). \quad (52)$$

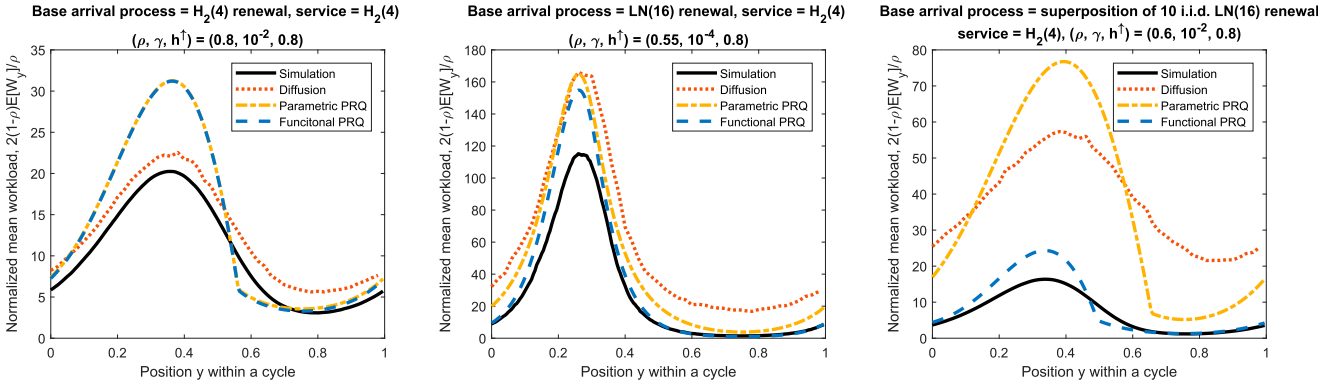
Figure 7 compares these three approximations for the mean in three cases. First, we consider a case for which the heavy-traffic approximation should perform well. In particular, we first consider the $H_{2,t}(4)/H_{2,t}(4)/1$ model with $(\rho,\gamma) = (0.8, 0.01)$ (left). Figure 7 (left) shows that the diffusion performs best, as expected.

Then we consider two cases that should favor PRQ more. For the $LN_t(16)/H_{2,t}(4)/1$ model with $(\rho,\gamma) = (0.55, 0.0001)$ (middle), which has lighter traffic and longer cycles, we see that all three approximations perform about the same, although functional PRQ does better away from the peak. Finally, for the for the $\sum_{i=1}^{10} LN_{i,t}(16)/H_{2,t}(4)/1$ model with the more complex arrival process from the superposition of 10 i.i.d. stationary $LN_t(16)$ renewal processes having $(\rho,\gamma) = (0.6, 0.01)$, we see that functional PRQ performs far better than the others, evidently because the IDC is able to capture the complex dependence in the superposition arrival process.

6.3. The Heavy-Traffic Limit for PRQ

We now establish a heavy-traffic limit for PRQ as given in (51) above. The proofs for the following results appear in Section EC.5.

Figure 7. (Color online) A Comparison of the Diffusion Approximation in (49), the Parametric PRQ in (50), and the Functional PRQ in (51) for the Normalized Mean Workload $2(1-\rho)E[W_y]/\rho$ as a Function of the Position y Within a Cycle to Simulation Estimates in Three Cases: Standard Model (Left), Lighter Traffic and Longer Cycles (Middle), and Complex Superposition Arrival Process (Right)



Lemma 1. *For a fixed place y within a cycle in the periodic $G_t/G_t/1$ model indexed by (ρ,γ),*

$$\Lambda_{\gamma,\rho,y}(s) = \rho s + \frac{1}{\gamma(1-\rho)} H_{\gamma,\rho,y}(s), \tag{53}$$

where

$$H_{\gamma,\rho,y}(s) \equiv \int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt \tag{54}$$

and $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$ is the cycle length.

To express the heavy-traffic limit, we define two functions. The first function,

$$f(t) \equiv -t + 2\sqrt{t}, \tag{55}$$

is a variant of the function to be optimized with the stationary $M/GI/1$ model, as can be seen from theorem 1 of Whitt and You (2018b). The second function,

$$g_{\gamma,\rho,y}(t) \equiv \frac{4}{b^2c_x^2\gamma\rho^2} H_{\gamma,\rho,y}\left(\frac{b^2c_x^2\rho}{4(1-\rho)^2}t\right) = \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} h(s)ds, \tag{56}$$

is a periodic function that captures the time-varying part of the arrival rate function. The period of $g_{\gamma,\rho,y}(t)$ is $4/b^2c_x^2\gamma\rho$. When the arrival-rate function is constant, $g_{\gamma,\rho,y}(t) = 0$ because $h(t) = 0$.

We remark that the constant $\rho c_x^2/2(1-\rho)$ is the exact steady-state mean waiting time in a $M/GI/1$ model, $f(t)$ attains maximum value of 1 at $t = 1$, and $g_{\gamma,\rho,y}$ is a periodic function fluctuating around 0 with limits in Lemma EC.3 in Section EC.5. Now, we present the heavy traffic limit for PRQ.

Theorem 3 (Heavy Traffic Limit for PRQ). *For the $G_t/G/1$ model with $W^*_{\gamma,\rho,y}(b)$ in (51), f in (55), and g in (56),*

$$\lim_{\rho\uparrow 1}\frac{2}{b^2}\cdot\frac{2(1-\rho)}{\rho c_x^2}\cdot W^*_{\gamma,\rho,y}(b) = \sup_{t\ge 0}\{f(t) + g_{\gamma,1,y}(t)\}. \tag{57}$$

We immediately obtain an upper bound for the PRQ in the special case of a sinusoidal arrival rate, which reveals the essential shape of the solution, as we shall see in later examples.

Corollary 1. *Suppose $h(x) = \beta\sin(2\pi x)$; then*

$$\lim_{\rho\uparrow 1}\frac{2}{b^2}\cdot\frac{2(1-\rho)}{\rho c_x^2} W^*_{\gamma,\rho,y} \le \lim_{\rho\uparrow 1} f(t) + \lim_{\rho\uparrow 1} g_{\gamma,\rho,y}(t) \le 1 + \frac{2\beta}{\pi b^2 c_x^2\gamma}(1-\cos(2\pi y)), \quad 0\le y<1. \tag{58}$$

Remark 2 (The Heavy Traffic Limits Do Not Coincide in This Case). Our numerical experiments show that PRQ in Theorem 3 does not coincide with the mean in Theorem 2 in general, but we will get agreement in double limits in the next section.

6.4. Long-Cycle Limits for PRQ in Heavy Traffic

For useful approximations of periodic queues, it is helpful to combine the heavy-traffic perspective with the long-cycle perspective considered in Sections 5.2 and EC.4.3. When we let the cycles get long in heavy-traffic, we see that there are three very different cases, depending on h in (38) or, equivalently upon the loading $\rho^\uparrow$ defined in (23). In the heavy-traffic setting of Sections 6.1–6.3, the three cases are the underloaded case in which $h^\uparrow < 1$, the overloaded case in which $h^\uparrow > 1$, and the critically loaded case in which $h^\uparrow = 1$. We consider the critically loaded case in Section EC.6.

6.4.1. Underloaded Queues.

In the underloaded case, there will be no times at which the net input rate is positive. We will show that if we let the cycles get long for PRQ in an underloaded model, PRQ is asymptotically consistent with the heavy-traffic limit and PSA.

Theorem 4 (Long-Cycle Heavy-Traffic Limit for PRQ in an Underloaded Queue). *Assume that h in (38) is continuously*

differentiable with $h^{\uparrow}<1$; then the PRQ workload in (51) for the $G_t/G/1$ model admits the double limit

$$\lim_{\substack{\gamma\downarrow 0\\ \rho\uparrow 1}} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y}(b) = \frac{b^2}{2}\frac{1}{1-h(y)}, \tag{59}$$

so that PRQ is asymptotically consistent with PSA; that is, the instantaneous traffic intensity is $\rho(y) = \rho + (1-\rho)h(y)$ and

$$W^*_{\gamma,\rho,y}(b) = \frac{b^2}{2} \cdot \frac{\rho(y)c_x^2}{2(1-\rho(y))} + o(1-\rho) + o(\gamma). \tag{60}$$

By (28), we have

$$\lim_{\substack{\gamma\downarrow 0\\ \rho\uparrow 1}} \frac{2(1-\rho)}{\rho c_x^2} \cdot W_{\gamma,\rho,y}(p) = -\ln(1-p)\frac{1}{1-h(y)} = \lim_{\substack{\gamma\downarrow 0\\ \rho\uparrow 1}} \frac{2(1-\rho)}{\rho c_x^2} W^*_{\gamma,\rho,y}\left(\Pi^{-1}(b)\right), \tag{61}$$

where $W_{\gamma,\rho,y}(p)$ is the p^{th} quantile of $W_{\gamma,\rho,y}$ and $\Pi(b)$ is defined in (28), so that PRQ captures the exact steady-state distribution of the workload $W_{\gamma,\rho,y}$ in the long-cycle heavy-traffic limit.

Remark 3 (The Iterated Limit). We remark that the double limit in Theorem 4 is stronger than a natural iterated limit, which has been established for the $M_t/M/1$ queue and should hold more generally. In particular, PSA has been proved to be asymptotically correct as $\gamma \downarrow 0$ for the $M_t/M/1$ model in Whitt (1991). Then RQ has been shown to be asymptotically correct for the stationary model as $\rho \uparrow 1$ in Whitt and You (2018b).

Figure 8 (left) compares the PRQ approximation in (20) and the PSA approximation with the simulated steady-state mean workload. Under moderate traffic intensity $\rho = 0.5$ and moderate cycle length $\gamma = 0.01$, the PRQ provides substantial improvement over PSA. Figure 8 (right) demonstrate the performance of the PRQ approximation for a higher traffic intensity of $\rho = 0.7$ and a longer cycle length with $\gamma = 0.005$, validating Theorem 4.

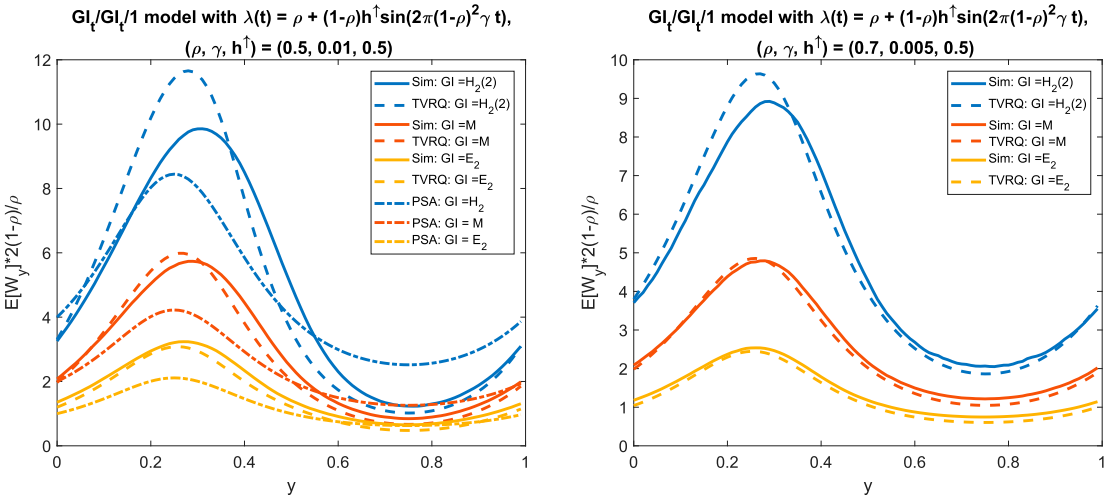
6.4.2. Overloaded Queues. The overloaded case is very different. With long cycles, there will be long stretches of time over which the workload will build up. This will lead to limits with new scaling, as in Choudhury et al. (1997b).

Theorem 5 (Long-Cycle Limit for PRQ in an Overloaded Queue). *For the $G_t/G/1$ model with the heavy-traffic scaling in (37) and $h^{\uparrow}>1$, PRQ in (51) admits the long-cycle limit*

$$(1-\rho)\lim_{\gamma\downarrow 0}\gamma \cdot W^*_{\gamma,\rho,y}(b) = \sup_{t\geq 0}\left\{-t + \int_{y-t}^{y} h(s)ds\right\}, \quad 0 \leq \rho < 1. \tag{62}$$

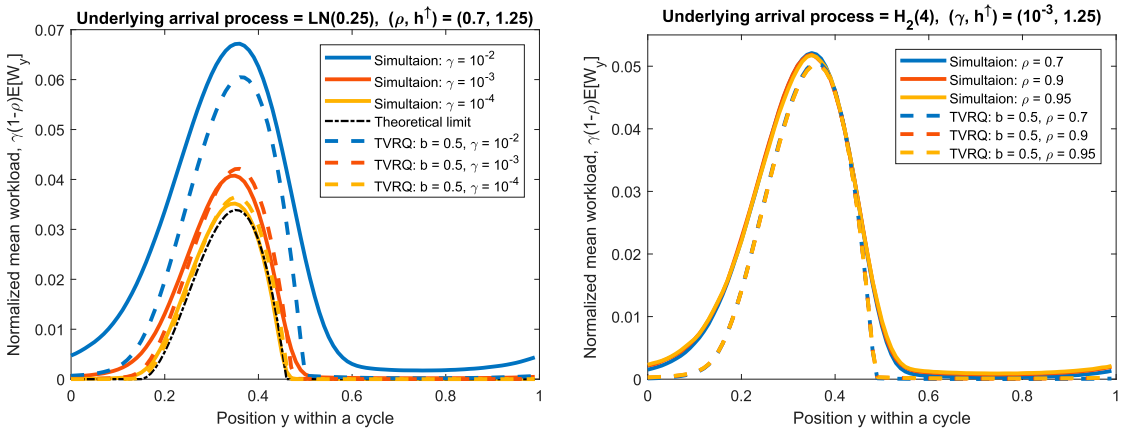
Note that the long-cycle limit is independent of the parameter b, suggesting a deterministic workload. This is consistent with the long-cycle fluid limit in Theorem 1. Theorem 5 here goes beyond the long-cycle fluid limit by revealing the linear dependence on $(1-\rho)$. This is confirmed in Figure 9, where we

Figure 8. (Color online) A Comparison of PRQ in (20) as a Function of the Position y Within a Cycle with Simulation Estimations of the Normalized Mean Workload $2(1-\rho)E[W_{\gamma,\rho,y}]/\rho$ for $W_{\gamma,\rho,y}$ in (48) and the Limit in Theorem 4 in the Underloaded $GI_t/GI_t/1$ Model with Arrival-Rate Function in (24) and (37) for the Arrival Rate Function in (52) with $(\gamma,\rho,h^{\uparrow}) \in \{(0.5,0.01),(0.7,0.005)\}$



Notes. Several interarrival time and service time distributions are considered to demonstrate the robustness of the PRQ algorithm. Left plot also displays the corresponding PSA approximation.

Figure 9. (Color online) A Comparison of PRQ(b) in (20), (22), and (36) as a Function of b and the Position y Within a Cycle to Simulation Estimates of the Normalized Mean Workload $\gamma(1-\rho)E[W_{\gamma,\rho,y}]$ for $W_{\gamma,\rho,y}$ in (48) and the Limit in Theorem 5 in the Overloaded $G_t/LN(1)/1$ Model with Arrival-Rate Function in (38) and (24) for Three Values of γ (Left) and Three Values of ρ (Right)



Note. The arrival rate function is (52) with the parameters specified in each plot.

observe that the same scaling constant in the simulated mean workload.

Remark 4 (The Space Scaling). When the queue is not overloaded, Theorem 5 yields the trivial limit 0, as does Theorem EC.2. That implies that the scaling constant γ in (62) then becomes too much to generate an interesting limit. For underloaded queues, we saw in Section 6.4.1 that the long-cycle scaling constant γ is not needed. For critically loaded queues, the long-cycle scaling is much more interesting; we discuss that case in Section EC.6.

To illustrate, Figure 9 compares PRQ in (20) with parameter $b = 0.5$ as a function of the position y within a cycle to simulation estimates of the normalized mean workload $\gamma(1-\rho)E[W_{\gamma,\rho,y}]$ for $W_{\gamma,\rho,y}$ in (48) and the limit in Theorem 5 in the overloaded $G_t/LN(1)/1$ model with arrival-rate function in (24) and (38) for three values of γ (left) and three values of ρ (right). Figure 9 (left) shows that both simulated values and PRQ approximations converge to the theoretical limit calculated from Proposition EC.1, confirming Theorem 1 and Corollary EC.2, whereas Figure 9 (right) demonstrates that the scaling constant $(1-\rho)$ also appears in the simulated mean workload. Overall, Figure 9 shows that PRQ serves as a reasonable approximation for the overloaded queues even in moderate cycle length and traffic intensities.

7. Conclusions

In this paper, we have developed a time-varying robust queueing algorithm to approximate the time-varying workload in a general $G_t/G_t/1$ single-server queue with time-varying arrival-rate and service-rate functions. Exploiting a reverse-time construction of the steady-state workload at time t in Section 2.1, in Section 2.2 we developed a general TVRQ representation of the steady-state workload at time t as the supremum over an uncertainty set. In (17) in Section 2.3 we expressed it in terms of the index of dispersion for work.

The rest of the paper focused on the special case of periodic RQ with unit service rate. In that case we consider the periodic steady-state workload at place yc, $0 \le y \le 1$, within a periodic cycle of length c, focusing especially on high-volume systems (reflected by long cycles) with heavy loading (associated with high traffic intensities). The general representation of the PRQ workload as a function of y appears in (20). We found that the control parameter b can be used to approximate different quantiles of the workload distribution, as indicated in (22). We also found that the function Π in (22) and the performance of the queue depends on the loading $\rho^{\uparrow}$ as defined in (23).

In Section 4 we found that Π in (28) is effective for underloaded models with $\rho^{\uparrow} < 1$ and is consistent with RQ for the stationary model in Whitt and You (2018b). In contrast, in Section 5 for overloaded models with $\rho^{\uparrow} > 1$, we found that the Gaussian approximation for Π in (36) performs remarkably well. Both PRQ and the more elementary deterministic approximation approximate the location of the peak remarkably well, as illustrated in Figure 4. Overall, the figures in Sections 4, 5, and the EC provide strong support for PRQ.

In Section 6 we established heavy-traffic limits as the long-run average traffic intensity ρ increases toward 1 for both the actual periodic workload and the PRQ, using the scaling in Whitt (2014), but in general these limits do not agree. In Section 6.4 we established double limits as the traffic intensity increases and the cycle length increases. These limits expose three

important cases: First, for underloaded models in which the maximum instantaneous traffic intensity remains less than 1, the limit for PRQ is the same as the pointwise stationary approximation version of the heavy-traffic limit for the stationary model, which has been shown to be asymptotically correct in Whitt and You (2018b). Second, for the overloaded case, we obtain limits with very different scaling that captures the long periods of overloading, just as in Choudhury et al. (1997b). Third, for critically loaded cases, we obtained the limit for PRQ in Theorem EC.3, consistent with Whitt (2016). In each case, we reported results of simulation experiments that confirm the limit theorems and show that PRQ is remarkably effective. Overall, we conclude that TVRQ can provide helpful insight into complex time-varying queueing models.

We regard this paper is an exploration, opening a promising new line of research. There are many directions for further research. For example, it remains to develop theoretical explanations for the function Π in (36) yielding $b = 0.5$ for OL models and the choice $b = 1$ for CL models in Section EC.6. There are opportunities for new insightful asymptotics. It also remains to explore various applications and consider extensions to networks of queues, paralleling Whitt and You (2018a), and queues with multiple servers.

References

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

Bandi C, Bertsimas D, Youssef N (2018) Robust transient analysis of multi-server queueing systems and feed-forward networks. *Queueing Systems* 89(3–4):351–413.

Ben-Tal A, El-Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press, Princeton, NJ).

Bertsimas D, Thiele A (2006) A robust optimization approach to inventory theory. *Oper. Res.* 54(1):150–168.

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

Bertsimas D, Gamarnik D, Rikun AA (2011b) Performance analysis of queueing networks via robust optimization. *Oper. Res.* 59(2): 455–466.

Beyer HG, Sendhoff B (2007) Robust optimzation: A comprehensive survey. *Comput. Methods Appl. Mech. Engrg.* 196(33–34): 3190–3218.

Choudhury GL, Lucantoni DL, Whitt W (1997a) Numerical solution of piecewise-stationary $M_t/G_t/1$queues. *Oper. Res.* 45(3):451–463.

Choudhury GL, Mandelbaum A, Reiman MI, Whitt W (1997b) Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models* 13(1):121–146.

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

Edie LC (1954) Traffic delays at toll booths. *Oper. Res.* 2(2):107–138.

Green LV, Kolesar PJ (1991) The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* 37(1): 84–97.

Harrison JM, Lemoine AJ (1977) Limit theoorems for periodic queues. *J. Appl. Probab.* 14(3):566–576.

Heyman DP, Whitt W (1984) The asymptoic behavior of queues with time-varying arrivals. *J. Appl. Probab.* 21(1):143–156.

Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42(10): 1383–1394.

Keller J (1982) Time-dependent queues. *SIAM Rev.* 24(4):401–412.

Kolesar PJ, Rider PJ, Craybill TB, Walker WE (1975) A queueing-linear-programming approach to scheduling police patrol cars. *Oper. Res.* 23(6):1045–1062.

Koopman BO (1972) Air-terminal queues under time-dependent conditions. *Oper. Res.* 20(6):1089–1114.

Lemoine AJ (1981) On queues with periodic Poisson input. *J. Appl. Probab.* 18(4):889–900.

Lemoine AJ (1989) Waiting time and workload in queues with periodic Poisson input. *J. Appl. Probab.* 26(2):390–397.

Ma N, Whitt W (2018a) A rare-event simulation algorithm for periodic single-server queues. *INFORMS J. Comput.* 30(1):71–89.

Ma N, Whitt W (2019) Minimizing the maximum expected waiting time in a periodic single-server queue with a service-rate control. *Stochastic Systems.* Forthcoming.

Mandelbaum A, Massey WA (1995) Strong approximations for time-dependent queues. *Math. Oper. Res.* 20(1):33–64.

Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1): 149–201.

Massey WA (1985) Asymptotic analysis of the time-varying $M/M/1$ queue. *Math. Oper. Res.* 10(2):305–327.

Massey WA, Pender J (2013) Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* 75(2–4):243–277.

Massey WA, Whitt W (1998) Uniform acceleration expansions for Markov chains with time-varying rates. *Ann. Appl. Probab.* 9(4): 1130–1155.

Newell GF (1968a) Queues with time dependent arrival rates, I. The transition through saturation. *J. Appl. Probab.* 5(2):436–451.

Newell GF (1968b) Queues with time dependent arrival rates, II. The maximum queue and the return to equilibrium. *J. Appl. Probab.* 5(3):579–590.

Newell GF (1968c) Queues with time dependent arrival rates, III. A mild rush hour. *J. Appl. Probab.* 5(3):591–606.

Oliver RM, Samuel AH (1962) Reducing letter delays in post offices. *Oper. Res.* 10(6):839–892.

Ong KL, Taaffe MR (1989) Nonstationary queues with interrupted Poisson arrivals and unreliable/repairable servers. *Queueing Systems* 4(1):27–46.

Pender J, Massey WA (2017) Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probab. Engrg. Inform. Sci.* 31(1):1–42.

Rolski T (1989) Queues with nonstationary inputs. *Queueing Systems* 5(1–3):113–130.

Rothkopf MH, Oren SS (1979) A closure approximation for the nonstationary $M/M/s$ queue. *Management Sci.* 25(6):522–534.

Taaffe MR, Ong KL (1987) Approximating Ph(t)/M(t)/S/C queueing systems. *Ann. Oper. Res.* 8(1):103–116.

Whitt W (1982) Approximating a point process by a renewal process: Two basic methods. *Oper. Res.* 30(1):125–147.

Whitt W (1991) The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Sci.* 37(3):307–314.

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

Whitt W (2014) Heavy-traffic limits for queues with periodic arrival processes. *Oper. Res. Lett.* 42(6–7):458–461.

Whitt W (2015) Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Syst.* 81(4):341–378.

Whitt W (2016) Heavy-traffic limits for a single-server queue leading up to a critical point. *Oper. Res. Lett.* 44(6):796–800.

Whitt W, You W (2018a) A robust queueing network analyzer based on indices of dispersion. Working paper, Columbia University, New York.

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

Ward Whitt is a professor in the Industrial Engineering and Operations Research Department at Columbia University. He was previously at Bell Laboratories and AT&T Labs from 1977 to 2002. He was named an INFORMS Inaugural Fellow in 2002. His research has focused on ways to develop more effective approximations for complex queueing models arising in communication and service systems, including non-Markov models with time-varying arrival rates.

Wei You is a doctoral student in the Industrial Engineering and Operations Research Department at Columbia University. His primary research focus is on queueing theory, applied probability, and their applications to the performance analysis and design of service systems.