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

Time-Varying Robust Queueing

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

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 (i) the time-varying expected workload in models with a general time-varying arrival-rate function and (ii) for the periodic steady-state expected workload in models with a periodic arrival-rate function. We apply simulation to examine the performance of periodic TVRQ (PRQ). We find that PRQ predicts the timing of peak congestion 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.

Key words: robust queueing theory, time-varying arrival rates, nonstationary queues, periodic queues, heavy traffic

History: August 27, 2016

1. Introduction

Many queueing systems in operations research applications have time-varying arrival-rate functions, but this feature is rarely captured in modeling, because there are few helpful analysis tools except simulation for most of the relevant queueing models when we include time-varying arrival-rate functions. Even the Markovian $M_t/M/1$ model with a nonhomgeneous Poisson arrival process (NHPP, the M_t) is relatively hard to analyze. Over the years, supporting asymptotic approximations have been developed for the single-server queue with a time-varying arrival-rate function, e.g.,

see Newell (1968a,b,c), Massey (1981), Keller (1982), Massey (1985), Mandelbaum and Massey (1995) and Whitt (2014), but the limit processes remain difficult to analyze. A way around this impasse has been provided by robust optimization, as in Bertsimas et al. (2011), Ben-Tal et al. (2009), Beyer and Sendhoff (2007); the main idea is to replace a difficult stochastic model by a tractable optimization problem. We apply this approach to develop tractable performance descriptions for the time-varying performance of a single-server queue with a time-varying arrival-rate function.

This paper is a sequel to Bandi et al. (2015) and Whitt and You (2016), which developed robust queueing (RQ) algorithms to approximate the expected steady-state waiting-time and workload in stationary single-server queues. The present paper extends RQ to time-varying RQ (TVRQ) by developing approximations for the time-varying expected steady-state workload in a nonstationary single-server queue, having a time-varying arrival-rate function. This paper is intended to serve as a basis for developing approximations for associated time-varying networks of queues. Bandi et al. (2014) have previously focused on the transient behavior of queues and networks of queues, but they focus on the time-varying behavior of a stationary model, whereas we focus on the time-varying behavior of a nonstationary model. We also focus on the periodic steady state of a periodic model, which can also be viewed as a stationary model under appropriate initial conditions. For the periodic case, we develop a periodic RQ (PRQ). We show that PRQ is effective by establishing heavy-traffic limits and making comparisons with simulations for the special case of the $M_t/GI/1$ model, where the arrival process is a nonhomogeneous Poisson process and the service times are i.i.d. and independent of the arrival process.

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 take a broader view. We think of robust optimization as a useful tool that supplements existing tools in our toolkit. Accordingly, we think that it is important to establish connections between RQ and established queueing theory. Much of this paper and Whitt and You (2016) is devoted to that aim.

2. Time-Varying Robust Queueing (TVRQ): Basic Formulation

Our TVRQ builds on a reverse-time representation of the continuous-time workload process in the single-server queue. Just as in Bandi et al. (2015) and Whitt and You (2016), we use the reverse-time construction to represent the workload as a supremum, thus providing a basis for the RQ optimization.

2.1. A Reverse-Time Contruction of the Workload Process

We consider the standard single-server queue with unlimited waiting space, where customers are served in order of arrival. Throughout this section, we assume that the system starts out empty at time 0. Let $\{(U_k, V_k)\}$ be the sequence of ordered pairs of nonnegative random variables representing the interarrival times and service times. Let an arrival counting process be defined on the positive halfline by $A(t) \equiv \max\{k \geq 1 : U_1 + \cdots + U_k \leq t\}$ for $t \geq U_1$ and $A(t) \equiv 0$ for $0 \leq t < U_1$, and let the total input of work over the interval $[0, t]$ be the random sum

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

Then the workload (the remaining work in service time) at time t, starting empty at time 0, can be represented using the reflection map as $W(t) = \Psi(Y - e)(t)$, where e is the identity map, i.e., $e(t) \equiv t$, $t \geq 0$. Hence,

$$\begin{aligned} W(t) &= \Psi(Y - e)(t) \equiv Y(t) - t - \inf_{0 \leq s \leq t} \{Y(s) - s\} \\ &= \sup_{0 \leq s \leq t} \{Y(t) - Y(s) - (t - s)\} = \sup_{0 \leq s \leq t} \{Y_t(s) - s\}, \quad t \geq 0, \end{aligned} \tag{2}$$

where

$$Y_t(s) \equiv Y(t) - Y(t - s) = \sum_{k=A(t-s)+1}^{A(t)} V_k, \quad 0 \leq s \leq t, \quad t \geq 0, \tag{3}$$

is the cumulative input over the interval $(t - s, t]$.

2.2. A TVRQ Formulation for the Time-Varying Workload

Our RQ optimization problem performs the maximization in (2) subject to deterministic constraints placed on the input process $Y(t)$ in (1). These constraints convert the random variable $W(t)$ in (2) into a deterministic approximaton as the solution of a deterministic optimization problem. In our simulation experiments we will comparing this deterministic approximation to the mean $E[W(t)]$.

In particular, to formulate the deterministic TVRQ approximation for the time-varying workload $W(t)$, we let

$$W^*(t) \equiv \sup_{X_t \in \mathcal{U}_t} \sup_{0 \le s \le t} \{X_t(s)\}, \tag{4}$$

where $\mathcal{U}_t$ is the deterministic time-varying uncertainty set, which we define as

$$\begin{aligned} \mathcal{U}_t &\equiv \{X_t(s) \in \mathbb{R} : X_t(s) \le E[Y_t(s) - s] + bSD(Y_t(s) - s), \quad 0 \le s \le t\} \\ &= \{X_t(s) \in \mathbb{R} : X_t(s) \le E[Y_t(s)] - s + bSD(Y_t(s))), \quad 0 \le s \le t\}, \end{aligned} \tag{5}$$

with SD being the standard deviation. Just as in Bandi et al. (2015) and Whitt and You (2016), the uncertainty set in (5) is based on a Gaussian approximation, which is supported by a central limit theorem (CLT) for $Y_t(s)$ under customary regularity conditions.

To ensure a finite supremum, we assume that $E[Y(t)^2] < \infty$ for all t. Then, since $Y_t(s) \ge 0$ for all t and s, necessarily $0 \le W^*(t) < \infty$ for all t. As a consequence, we have the final TVRQ optimization

$$W^*(t) = \sup_{0 \le s \le t} \sup_{X_t \in \mathcal{U}_t} \{X_t(s)\} = \sup_{0 \le s \le t} \{E[Y_t(s)] - s + bSD(Y_t(s))\}, \quad t \ge 0. \tag{6}$$

It should be noted that the optimization problem in (4) is tractable, whereas the original stochastic model is complicated.

2.3. The $G_t/G/1$ Special Case

We now focus on a large class of $G_t/G/1$ models in which the arrival process takes the form of

$$A(t) = N(\Lambda(t)), \quad t \ge 0, \tag{7}$$

where the underlying process N is assumed to be a general stationary point process, by which we mean that it has stationary increments. We assume that N is a unit rate process with $E[N(t)^2] < \infty$ for all t. The cumulative arrival-rate fucntion is defined as

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0, \tag{8}$$

where the arrival-rate function λ is an element of $\mathcal{D}$ as in Whitt (2002b), i.e., is a right-continuous function with left limits, satisfying

$$0 < \lambda(s) \leq \lambda_{bd}(t) < \infty \quad \text{for all} \quad 0 \leq s \leq t, \quad t > 0. \tag{9}$$

In addition, the service times is a stationary sequence and independent of the arrival process with V_k distributed as V with mean $E[V] = 1$ and finite variance $\sigma_s^2 \equiv Var(V)$.

Given this model structure, we have

$$\{Y_t(s) : 0 \leq s \leq t\} \stackrel{\mathrm{d}}{=} \left\{ \sum_{k=1}^{N(\Lambda_t(s))} V_k : 0 \leq s \leq t \right\} \quad \text{for all} \quad t \geq 0, \tag{10}$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution, which here in (10) we mean as stochastic processes, and

$$\Lambda_t(s) \equiv \Lambda(t) - \Lambda(t-s), \quad 0 \leq s \leq t, \quad t \geq 0. \tag{11}$$

As a consequence of assumption (9), $\Lambda_t(s)$ is strictly increasing and continuous as a function of s with $\Lambda_t(0) = 0$ for each t, so it has a continuous strictly increasing inverse $\Lambda_t^{-1}(s)$ as a function of s with $\Lambda_t(0) = 0$ for each t.

Hence, we can combine (2) and (10) to obtain the alternative representation of the workload as

$$W(t) = \sup_{0 \leq s \leq t} \left\{ \sum_{k=1}^{N(\Lambda_t(s))} V_k - s \right\} = \sup_{0 \leq s \leq \Lambda(t)} \left\{ \sum_{k=1}^{N(s)} V_k - \Lambda_t^{-1}(s) \right\}, \tag{12}$$

where $\Lambda_t(s)$ is defined in (11). The second expression in (12) is appealing because it has all the stochastic variability in the first term inside the supremum. For the $M_t/GI/1$ model, the underlying total input process $\left\{ \sum_{k=1}^{N(s)} V_k : s \geq 0 \right\}$ is a stationary compound Poisson process whose variance is readily avaliable. The second expression in (12) was exploited to develop a rare-event simulation algorithm for periodic queues in Ma and Whitt (2016).

Under this setting, the standard deviation of the random sum $\sum_{k=1}^{N(t)} V_k$ used in (5) is a complicated function of t. Following Fendick and Whitt (1989) and Whitt and You (2016), to approximately characterize the variability of the input of work over the time interval $[0,t]$, independent of its mean, we use the index of dispersion for work (IDW), denoted by $I_w(t)$. In our setting, the IDW is defined by

$$I_w(t) \equiv Var\left(\sum_{k=1}^{N(t)} V_k\right) \Big/ E\left[\sum_{k=1}^{N(t)} V_k\right]. \tag{13}$$

We assume that IDW is finite, which holds true under the setting here and can be anticipated more generally. We remark that under assumptions in Theorem 2 in Whitt and You (2016), the limiting IDW, $I_w(\infty) \equiv \lim_{t\to\infty} I_w(t)$, coincides with the variability parameter in the Brownian limit for the total input process. See Corollary 4 in Whitt and You (2016).

As a consequence, the uncertainty set (5) in the TVRQ can be written as

$$\begin{aligned}
\mathcal{U}_t &= \left\{ X(t) : X(s) \le E\left[\sum_{k=1}^{N(s)} V_k\right] - \Lambda_t^{-1}(s) + b\sqrt{\mathrm{Var}\left(\sum_{k=1}^{N(s)} V_k\right)}, 0 \le s \le \Lambda(t) \right\} \\
&= \left\{ X(t) : X(s) \le s - \Lambda_t^{-1}(s) + b\sqrt{sI_w(s)}, 0 \le s \le \Lambda(t) \right\} \\
&= \left\{ X(t) : X(s) \le \Lambda_t(s) - s + b\sqrt{\Lambda_t(s) I_w\left(\Lambda_t(s)\right)}, 0 \le s \le t \right\}.
\end{aligned} \tag{14}$$

Combining (4) and (14), we have the tractable TVRQ optimization for $G_t/G/1$ model

$$W^*(t) = \sup_{0\le s\le t} \left\{ \Lambda_t(s) - s + b\sqrt{\Lambda_t(s) I_w\left(\Lambda_t(s)\right)} \right\}, \quad t \ge 0, \tag{15}$$

with the final expression in (15) providing a convenient expression for a computational algorithm because $\Lambda_t(s)$ is usually readily available, whereas $\Lambda_t^{-1}(s)$ in the first expression may not be. This extension provides a way to consider models with stochastic dependence as in Whitt and You (2016) as well as time-varying arrival rate function.

To conclude, we remark that the $M_t/GI/1$ is a special case where $I_w(t) \equiv c_x^2 = c_a^2 + c_s^2 = 1 + c_s^2$ with c_s and c_a being the coefficient of variation of the interarrival and service distribution, respectively.

2.4. A Deterministic Fluid Model

A deterministic fluid model is obtained by assuming that the cumulative input of work $Y(t)$ in (1) is a deterministic nondecreasing function, which we take to be $E[Y(t)]$ in an associated stochastic model. To capture the usual idea of a fluid, we also assume that $Y(t)$ is a continuous function of t, which we take to coincide with $\Lambda(t)$. The workload at time t is then defined just as in (2). In this case, the associated TVRQ is defined just as in §2.2, but now we have $SD(Y_t(s)) = 0$ in (5) because of the deterministic property. Thus, for the deterministic fluid model, the TVRQ is necessarily exact. Moreover, when we formulate a fluid limit for the stochastic model, where the stochastic workload process converges to a deterministic fluid model, then the TVRQ will be asymptotically correct, under regularity conditions. We will illustrate for periodic TVRQ in §3.

3. Periodic Robust Queueing (PRQ) and the Fluid Approximation

We now consider periodic models and look at the periodic steady state workload as a function of the place y within a periodic cycle. We will develop a periodic RQ (PRQ) and show that it is asymptotically correct in a long-cycle limit.

3.1. The Periodic Steady-State Workload

If the arrival process and workload process are periodic over the entire real line with period c, then we can obtain an expression for the periodic steady-state workload at time t within the interval $[0, c)$ by letting the system start empty in the distant past. For this periodic steady-state distribution to be well defined, we require that the average arrival rate satisfy

$$\rho = \bar{\lambda} \equiv \Lambda(c)/c < 1, \tag{16}$$

to ensure that the average arrival rate is less than the maximum possible service rate $\mu \equiv 1/E[V] \equiv 1$. We assume that a proper periodic steady-state exists.

Instead of (2) for the transient workload, we have the periodic steady-state workload represented as a supremum over the entire real line. In particular, for a fixed position y within a cycle, we have

$$W_y = \sup_{s \geq 0} \{Y_y(s) - s\}, \quad 0 \leq y < c, \tag{17}$$

where Y_y is defined as in (3).

For the periodic case, starting empty in the distant past, we consider $y \in [0, c)$. Then periodic RQ (PRQ) for the steady-state workload is just TVRQ in (6) except that s is allowed to range over the interval $[0, \infty)$ and that $Y_t(s)$ is replaced by $Y_y(t)$ to emphasize the focus on a fixed location in a cycle. As a consequence, we have the final PRQ optimization

$$W_y^* = \sup_{s \geq 0} \{E[Y_y(s)] - s + bSD(Y_y(s))\}, \quad 0 \leq y < c. \tag{18}$$

3.2. A Periodic Deterministic Fluid Model

In §2.4 we briefly introduced a deterministic fluid model and observed that the workload is defined just as in (2). Now we consider the periodic case. In the next section we will establish a fluid limit for the periodic $G_t/GI/1$ model as the cycle lengths grow. To avoid confusion about notation, we append an extra subscript f for the fluid quantities.

We start with the arrival-rate function $\lambda_f(t)$ satisfying the properties in §2.3, but now we assume as well that the arrival-rate function is periodic with period c and satisfies (16). In order for the fluid model to be interesting, we also assume that

$$\lambda_f^{\uparrow} \equiv \sup_{0 \leq s < c} \{\lambda_f(s)\} > 1. \tag{19}$$

Condition (19) ensures that there will be positive workload at some time. If condition (19) did not hold, then the net rate in at each time would be negative, so that there never would be any workload.

To obtain the deterministic fluid model, we simply let

$$Y_f(t) \equiv \Lambda_f(t), \quad t \geq 0. \tag{20}$$

Notice that $Y_f(t)$ for the fluid model coincide with the expected values of their stochastic counterparts in the $M_t/GI/1$ special case. Now the main quantity we focus on is

$$Y_{f,y}(s) = \Lambda_{f,y}(s) \equiv \Lambda_f(y) - \Lambda_f(y - s), \quad s \geq 0, \quad 0 \leq y < c. \tag{21}$$

We now observe that the fluid workload at time y is determined by the input over the cycle ending at time y.

PROPOSITION 1. *For the deterministic fluid model, the workload at time y within the cycle $[0,c)$ defined in* (17) *with $Y_{f,y}$ in* (21) *reduces to the supremum over one cycle, i.e.,*

$$W_{f,y} = \sup_{0\le u\le c}\{Y_{f,y}(u)-u\}, \quad 0\le y<c. \tag{22}$$

Proof. Let $s=kc+y$, $0\le y<c$ and $k\ge 0$. Then

$$\begin{aligned}
W_{f,y} &= \sup_{0\le s\le\infty}\{Y_{f,y}(s)-s\}, \quad 0\le y<c,\\
&= \sup_{0\le u\le c,\, k\ge 0}\{Y_{f,y}(kc+u)-(kc+u)\}, \quad 0\le y<c,\\
&= \sup_{0\le u\le c,\, k\ge 0}\{(Y_{f,y}(kc+u)-Y_{f,y}(u)-kc)+(Y_{f,y}(u)-u)\}, \quad 0\le y<c,\\
&= \sup_{0\le u\le c,\, k\ge 0}\{-(1-\rho)kc+(Y_{f,y}(u)-u)\}, \quad 0\le y<c,\\
&= \sup_{0\le u\le c}\{Y_{f,y}(u)-u\}, \quad 0\le y<c,
\end{aligned} \tag{23}$$

because the function inside the supremum is strictly decreasing in k. ∎

We now consider a common special case in which, if we start the periodic cycle at an appropriate point, then we can express the arrival-rate function so that the net input rate is positive on an initial subinterval and then negative thereafter. That is, we assume that there exists δ, $0<\delta<c$, such that

$$\lambda_f(t)-1\ge 0, \quad 0\le t<\delta, \quad \text{and} \quad \lambda_f(t)-1\le 0, \quad \delta\le t<1. \tag{24}$$

Often we may require a time shift to satisfy condition (24). In this setting it is easy to determine the periodic fluid W_y, $0\le y\le c$.

PROPOSITION 2. *If conditions* (16), (19) *and* (24) *hold, then there exists one and only one δ^* with $0<\delta<\delta^*<c$ such that $\Lambda_f(\delta^*)=\delta^*$. Moreover, $\Lambda_f(y)-y$ is nondecreasing over $[0,\delta]$ and nonincreasing over $[\delta,c]$, so that*

$$W_{f,y}=\Lambda_f(y)-y, \quad 0\le y\le\delta^*, \quad \textit{and} \quad W_{f,y}=0, \quad \delta^*\le y\le c, \tag{25}$$

and

$$W_f^{\uparrow}\equiv\sup_{0\le y\le c}\{W_{f,y}\}=W_{f,\delta}=\Lambda_f(\delta)-\delta>0. \tag{26}$$

We now apply Proposition 2 to two special cases. The easiest case appears to be the piecewise-constant case with two pieces.

COROLLARY 1. (*piecewise-constant case*) *If, in addition to the conditions of Proposition 2,* $\lambda_f(t) = a1_{[0,\delta)}(t) + b1_{[\delta,c)}(t)$, *where* $a > 1 > b > 0$, *then*

$$W_{f,y} = (a-1)y, \quad 0 \le y \le \delta, \quad W_f^{\uparrow} = W_{f,\delta} = (a-1)\delta, \tag{27}$$

and

$$W_{f,y} = (a-1)\delta - (1-b)(y-\delta), \quad 0 \le y \le y^* \equiv (a-b)\delta/(1-b) \quad \textit{and} \quad W_{f,y} = 0, \quad y^* \le y \le c. \tag{28}$$

The following corollary shows that, for a sinusoidal arrival rate function, the maximum workload is attained shortly before the middle of the arrival-rate cycle.

COROLLARY 2. (*sinusoidal case*) *If, in addition to the conditions of Proposition 2,* $\lambda_f(t) = \rho(1 + \beta \sin(t/\gamma))$ *and* $t_0 = \gamma \arcsin((1-\rho)/\rho\beta)$, *then* $\lambda_f(t_0 + t))$ *satisfies condition* (24) *and* $\delta = (\pi/\gamma) - 2t_0$, *so that* (*in terms of the original* Λ)

$$W_f^{\uparrow} = \Lambda((\pi/\gamma) - t_0) - \Lambda(t_0) - (\pi/\gamma) + 2t_0. \tag{29}$$

As $\rho \uparrow 1$, $t_0 \equiv t_0(\rho) \downarrow 0$, $\delta(\rho) \uparrow \pi/\gamma$ *and* $W_f^{\uparrow} \to \Lambda(\pi/\gamma) - \pi/\gamma$.

3.3. A Long-Cycle Fluid Limit

For periodic queues, it is helpful to consider the case of long cycles relative to a fixed service-time distribution. (This case is equivalent to letting the service times become short relative to a fixed arrival rate function.) We now 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

$$\lambda_\gamma(t) \equiv \lambda_f(\gamma t), \quad t \ge 0, \tag{30}$$

for the base arrival-rate function λ_f in the fluid model, satisfying (16) and (19). 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 $c_\gamma \to \infty$.

In the stochastic model we can also let the cumulative arrival-rate function be defined in terms of the base cumulative arrival-rate function Λ_f in the fluid model. In particular, we let

$$\Lambda_\gamma(t) \equiv \gamma^{-1}\Lambda_f(\gamma t) \quad \text{and} \quad \Lambda_{\gamma,y}(t) \equiv \Lambda_\gamma(\gamma^{-1}y) - \Lambda_\gamma(\gamma^{-1}y - t), \quad 0 \le y < c, \tag{31}$$

so that the associated arrival-rate function is as in (30). The periodic structure with (16) and (19) implies the following bound.

LEMMA 1. *In the setting above with* (16) *and* (19),

$$\max\{\Lambda_f(t), \Lambda_{f,y}(t)\} \le \rho t + \lambda^\uparrow c \quad \textit{and} \quad \max\{\Lambda_\gamma(t), \Lambda_{\gamma,y}(t)\} \le \rho t + \lambda^\uparrow c/\gamma \quad \textit{for all} \quad t \ge 0. \tag{32}$$

Let $A_\gamma(t)$ and $Y_\gamma(t)$ be the associated arrival and cumulative input processes in the $G_t/GI/1$ model, defined as in (1) and (7) by

$$A_\gamma(t) \equiv N(\Lambda_\gamma(t)) \quad \text{and} \quad Y_\gamma(t) \equiv \sum_{k=1}^{A_\gamma(t)} V_k, \quad t \ge 0, \tag{33}$$

where N is a rate-1 stochastic process and $\{V_k\}$ is the i.i.d. sequence of service times with $E[V_k] = 1$ independent of N and thus of A_γ.

As regularity conditions for N, we assume that

$$t^{-1}N(t) \to 1 \quad \text{as} \quad t \to \infty \quad \text{w.p.1} \tag{34}$$

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

$$|t^{-1}N(t) - 1| < \epsilon \quad \text{for all} \quad t \ge t_0 \quad \text{w.p.1.} \tag{35}$$

Condition (34) is a strong law of large numbers (SLLN), which is equivalent to the stronger functions SLLN (FSLLN), see §3.2 of Whitt (2002a), while condition (35) is implied by refinements such as the law of the iterated logarithm. Condition (35), together with Lemma 1, is needed for Theorem 1 to guarantee that a supremum over the entire real line is attained over a bounded subinterval, which allows us to apply a continuous mapping argument. Both conditions hold when N is a Poisson process and can be anticipated more generally.

The basis for the fluid limit is a functional law of large numbers for A_γ and Y_γ after introducing extra time and space scaling.

LEMMA 2. *For the periodic $G_t/GI/1$ model under condition* (34)*,*

$$\gamma A_\gamma(\gamma^{-1}(t)) \to \Lambda_f(t) \quad \textit{and} \quad \gamma Y_\gamma(\gamma^{-1}(t)) \to \Lambda_f(t) \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1} \tag{36}$$

Proof. Observe that

$$\begin{aligned} \gamma A_\gamma(\gamma^{-1}t) &= \gamma N(\Lambda_\gamma(\gamma^{-1}t)) = \gamma N(\gamma^{-1}\Lambda_f(\gamma(\gamma^{-1}t))) \\ &= \gamma N(\gamma^{-1}\Lambda_f(t)) \to \Lambda_f(t) \quad \text{as} \quad \gamma \downarrow 0 \quad \text{w.p.1} \end{aligned} \tag{37}$$

because $\gamma N(\gamma^{-1}t) \to t$ uniformly over bounded intervals w.p.1 by the FSLLN in (34). A further application of the composition mapping yields the corresponding limit for Y_γ in (33):

$$\gamma Y_\gamma(\gamma^{-1}t) = \gamma \sum_{k=1}^{\gamma^{-1}(\gamma A_\gamma(\gamma^{-1}t))} V_k \to \Lambda_f(t) \quad \text{as} \quad \gamma \downarrow 0 \quad \text{w.p.1},$$

because

$$\gamma \sum_{k=1}^{\gamma^{-1}t} V_k \to t \quad \text{as} \quad \gamma \downarrow 0 \quad \text{w.p.1}$$

uniformly over bounded intervals w.p.1 by the FSLLN. ∎

Let $W_{\gamma,y}$ be the periodic steady-state workload at time y/γ for $0 \le y < c$ in $G_t/GI/1$ model γ with arrival rate function $\lambda_\gamma(t)$, defined as in (17), i.e.,

$$W_{\gamma,y} = \sup_{s \ge 0} \{Y_{\gamma,y}(s) - s\}, \tag{38}$$

where

$$Y_{\gamma,y}(t) \equiv Y_\gamma(y\gamma^{-1}) - Y_\gamma(y\gamma^{-1} - t), \quad t \ge 0, \quad 0 \le y < c, \tag{39}$$

for Y_γ in (33). We get a fluid limit for $W_{\gamma,y}$, again after scaling.

THEOREM 1. (*long-cycle fluid limit*) *For the periodic $G_t/GI/1$ model under conditions* (34) *and* (35)*,*

$$\gamma W_{\gamma,y} \to W_{f,y} \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1}, \tag{40}$$

where $W_{f,y}$ is the fluid workload at time y within a cycle of length c.

Proof. From (38) and (39),

$$\gamma W_{\gamma,y} = \sup_{s\geq 0}\{\gamma Y_{\gamma,y}(\gamma^{-1}s) - s\} \to \sup_{s\geq 0}\{\Lambda_{f,y}(s) - s\} = W_{f,y} \quad \text{as} \quad \gamma\downarrow 0 \quad w.p.1, \tag{41}$$

where $W_{f,y}$ is the periodic workload in the periodic fluid model by virtue of Lemma 2 and a further continuity argument. Lemma 2 and condition (35) guarantee that it suffices to consider the supremum over a bounded interval, so that the supremum is continuous. ∎

Let $W^*_{\gamma,y}$ be the PRQ workload at time y/γ for $0 \leq y < c$, i.e., $W^*_{\gamma,y}$ is the solution to the PRQ problem (18) at time y/γ with $Y_\gamma(t)$ defined in (33).

THEOREM 2. (*PRQ is asymptotically correct in the long-cycle fluid limit*) *For the periodic $G_t/GI/1$ model, PRQ with any b, $0 < b < \infty$, is asymptotically exact as $\gamma\downarrow 0$, i.e.,*

$$\gamma W^*_{\gamma,y} \to W_{f,y} \quad \textit{as} \quad \gamma\downarrow 0, \tag{42}$$

where $W_{f,y}$ is the fluid workload at time y within a cycle of length c, so that

$$|\gamma W^*_{\gamma,y} - \gamma W_{\gamma,y}| \to 0 \quad \textit{as} \quad \gamma\downarrow 0 \quad \textit{w.p.1.} \tag{43}$$

Proof. Observe that

$$\begin{aligned}
\gamma W^*_{\gamma,y} &= \sup_{s\geq 0}\{\gamma\Lambda_{\gamma,y}(\gamma^{-1}s) - s + \gamma\sqrt{2b^2\Lambda_{\gamma,y}(\gamma^{-1}s)}\} \\
&= \sup_{s\geq 0}\{\Lambda_{f,y}(s) - s + \sqrt{2b^2\gamma\Lambda_{f,y}(s)}\} \\
&\to \sup_{s\geq 0}\{\Lambda_{f,y}(s) - s\} = W_{f,y} \quad \text{as} \quad \gamma\downarrow 0,
\end{aligned} \tag{44}$$

where $\Lambda_{\gamma,y}(t)$ is defined in (31) and again $W_{f,y}$ is the workload in the periodic deterministic fluid model. To justify (44), we apply Lemma 1 to see that, $2b^2\gamma\Lambda_{f,y}(s) \leq 2b^2\gamma[\rho s + \lambda^\uparrow c] \leq \gamma(K_1 s + K_2)$ for constants K_1 and K_2, so that $\sqrt{2b^2\gamma\Lambda_{f,y}(s)} \leq \sqrt{\gamma(K_1 s + K_2)} \to 0$ uniformly over bounded interval as $\gamma\downarrow 0$. Hence, it suffices to consider the supremum in (44) over a bounded interval, because the function is negative outside that interval for all sufficiently small γ. Since the limit $W_{f,y}$ is the same as in Theorem 1, PRQ has been shown to be asymptotically correct as $\gamma\downarrow 0$. ∎

4. Heavy-Traffic Limits for Periodic Robust Queueing

We now consider a family of periodic $G_t/GI/1$ single-server models indexed by the traffic intensity ρ defined in (16) together with the specified time-scaling factor γ. We scale the models consistently with the heavy-traffic scaling in Whitt (2014). In §4.2 we will show that PRQ has a proper heavy-traffic limit in this scaling. First, in §4.1 we establish heavy-traffic limits for the workload process itself.

4.1. Heavy-Traffic Limits for the Workload Process

We consider a family of models indexed by the long-run average traffic intensity ρ in (16). To avoid notational confusion, we add a subscript d to the diffusion quantities. We let the cumulative arrival-rate function in model ρ be

$$\Lambda_{\gamma,\rho}(t) \equiv \rho t + (1-\rho)^{-1}\Lambda_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0, \tag{45}$$

so that the associated arrival-rate function is

$$\lambda_{\gamma,\rho}(t) \equiv \rho + (1-\rho)\lambda_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0, \tag{46}$$

where

$$\Lambda_{d,\gamma}(t) \equiv \int_0^t \lambda_{d,\gamma}(s)\, ds, \quad \lambda_{d,\gamma}(t) \equiv h(\gamma t), \quad \text{and} \quad \int_0^1 h(t)\, dt = 0 \tag{47}$$

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

$$h(t) \geq -\rho/(1-\rho), \quad 0 \leq t < 1, \tag{48}$$

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

$$-\infty < h^\downarrow \equiv \inf_{0\leq t\leq 1}\{h(t)\} < \sup_{0\leq t\leq 1}\{h(t)\} \equiv h^\uparrow < \infty. \tag{49}$$

There are two 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, so that the cycles are very long.

By essentially the same reasoning as for Theorem 3.2 in Whitt (2014), we obtain a heavy-traffic limit as $\rho \uparrow 1$ 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. Theorem 3.2 of Whitt (2014) is expressed for the queue-length process or number in the system, but it is well known that there are corresponding heavy-traffic limits for the workload process, e.g., see Theorem 9.3.4 of Whitt (2002b). This heavy-traffic limit is for the time-varying behavior starting empty. It 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 (45) and let

$$A_{\gamma,\rho}(t) \equiv N\left(\Lambda_{\gamma,\rho}(t)\right), \quad Y_{\gamma,\rho}(t) \equiv \sum_{k=1}^{A_{\gamma,\rho}(t)} V_k, \quad \text{and} \quad X_{\gamma,\rho}(t) \equiv Y_{\gamma,\rho}(t) - t, \quad t \geq 0. \tag{50}$$

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, i.e.,

$$W_{\gamma,\rho}(t) = \Psi(X_{\gamma,\rho})(t) = \sup_{0 \leq s \leq t} \{X_{\gamma,\rho}(t) - X_{\gamma,\rho}(t-s)\}. \tag{51}$$

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

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt], \quad \hat{A}_{\gamma,\rho}(t) \equiv (1-\rho)[A_{\gamma,\rho}\left((1-\rho)^{-2}t\right) - (1-\rho)^2 t],$$
$$\hat{X}_{\gamma,\rho}(t) \equiv (1-\rho)X_{\gamma,\rho}\left((1-\rho)^{-2}t\right) \quad \text{and} \quad \hat{W}_{\gamma,\rho}(t) \equiv (1-\rho)W_{\gamma,\rho}\left((1-\rho)^{-2}t\right), \quad t \geq 0. \tag{52}$$

Let $\mathcal{D}^k$ be the k-fold product space of the function space $\mathcal{D}$. Reccall that $g(x) = o(x)$ as $x \to 0$ if $g(x)/x \to 0$ as $x \to 0$.

THEOREM 3. (*heavy-traffic FCLT for the workload*) *For the family of $G_t/GI/1$ models indexed by (γ,ρ) with cumulative arrival-rate functions in* (45)*, if $\hat{N}_n \Rightarrow c_a B_a$ as $n\to\infty$, where B_a is a standard Brownian motion, then*

$$(\hat{A}_{\gamma,\rho},\hat{X}_{\gamma,\rho},\hat{W}_{\gamma,\rho}) \Rightarrow (\hat{A}_{\gamma},\hat{X}_{\gamma},\hat{W}_{\gamma}) \quad in \quad \mathcal{D} \quad as \quad \rho\uparrow 1, \tag{53}$$

where

$$(\hat{A}_{\gamma},\hat{X}_{\gamma},\hat{W}_{\gamma}) \equiv (c_a B_a + \Lambda_{d,\gamma} - e, \hat{A}_{\gamma} + c_s B_s, \Psi(\hat{X}_{\gamma})), \tag{54}$$

Ψ is the reflection map in (51)*, $\Lambda_{d,\gamma}$ is defined in* (47)*, and B_a and B_s are two independent standard (mean 0 variance 1) Brownian motions; i.e., $\hat{W}_{\gamma}$ is reflected periodic Brownian motion (RPBM) with*

$$\hat{W}_{\gamma} = \Psi(c_a B_a + c_s B_s + \Lambda_{d,\gamma} - e) \stackrel{\mathrm{d}}{=} \Psi(c_x B + \Lambda_{d,\gamma} - e). \tag{55}$$

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 (45)*.*

Proof. The limit for $\hat{A}_{\gamma,\rho}$ is part of Theorem 3.2 of Whitt (2014). The limit for $\hat{X}_{\gamma,\rho}$ is by the preservation of convergence for random sums with centering terms, exploiting the composition map with centering as in §13.3 and Theorem 7.4.1 of Whitt (2002b). (A detailed proof is given in the appendix.) The limit for The limit for $\hat{W}_{\gamma,\rho}$ follows from the continuity of the reflection map; see §14.5 of Whitt (2002b). The final limit in (55) is the same as for the number in system in Theorem 3.2 of Whitt (2014). ∎

We conclude this section by making some remarks. First, simplification occurs in (55) in common special cases: For the $M_t/GI/1$ model, $c_a = 1$; for the $G_t/M/1$ model, $c_s = 1$. However, even for the $M_t/M/1$ model, the limiting RPBM is complicated, so that PRQ can provide valuable assistance. Second, we do not state a heavy-traffic law of large numbers analogous to Theorem 3.1 of Whitt (2014), because it is implied by Theorem 1 here.

4.2. The Heavy-Traffic Limit for PRQ

We now establish a heavy-traffic limit for PRQ. Again, we add a subscript y to indicate the place in the cycle. In particular, the workload at fixed place y within a cycle for a system which started empty and has run for t time units is

$$W_{\gamma,\rho,y}(t) \stackrel{\mathrm{d}}{=} \sup_{0\le s\le t}\left\{\sum_{k=1}^{A_{\gamma,\rho,y}(t)} V_k - s\right\}, \tag{56}$$

where $A_{\gamma,\rho,y}(t) \equiv A_{\gamma,\rho}(y) - A_{\gamma,\rho}(y-t)$, $A_{\gamma,\rho}(t)$ is defined in (50) and V_k is a generic service time. Under the $G_t/GI/1$ setting in §2.3, we immediately get the PRQ optimization problem from (15) by replacing $\Lambda_t(s)$ with $\Lambda_{\gamma,y,\rho}(s)$.

Now, we will focus on the $M_t/GI/1$ model, in which case we have $I_w(s) = 1 + c_s^2 \equiv c_x^2$ with $c_s^2 \equiv Var(V_k)/E[V_k]^2$. Then (15) becomes

$$W^*_{\gamma,\rho,y} = \sup_{s\ge 0}\left\{\Lambda_{\gamma,y,\rho}(s) - s + bc_x\sqrt{\Lambda_{\gamma,\rho,y}(s)}\right\}. \tag{57}$$

For further analysis, we note that

$$\begin{aligned}\Lambda_{\gamma,\rho,y}(s) &\equiv \Lambda_{\gamma,\rho}((k+y)c_\rho) - \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho} - s) = \Lambda_{\gamma,\rho}(yc_{\gamma,\rho}) - \Lambda_{\gamma,\rho}(yc_{\gamma,\rho} - s)\\ &= \rho s + (1-\rho)^{-1}\int_{y/\gamma-(1-\rho)^2 s}^{y/\gamma} h(\gamma t)dt = \rho s + \frac{1}{\gamma(1-\rho)}\int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt\\ &= \rho s + \frac{1}{\gamma(1-\rho)} I_{\gamma,\rho,y}(s),\end{aligned} \tag{58}$$

where $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$ is the cycle length of $\Lambda_{\gamma,\rho,y}(s)$ and

$$I_{\gamma,\rho,y}(s) \equiv \int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt. \tag{59}$$

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

$$f(t) \equiv -t + 2\sqrt{t} \tag{60}$$

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 (2016). The second function

$$g_{\gamma,\rho,y}(t) = \frac{4}{b^2c_x^2\gamma\rho^2} I_{\gamma,\rho,y}\left(\frac{b^2c_x^2\rho}{4(1-\rho)^2}t\right) = \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} h(s)ds \tag{61}$$

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 $h(t) = 0$. The following lemma presents some basic limits for $g_{\gamma,\rho,y}(t)$.

LEMMA 3. *Let $h(\cdot)$ be a differentiable 1-periodic function whose integral over one period is 0. Assume that $h(\cdot)$ satisfies (49), then*

(a). $\lim_{(\gamma,\rho)\to(0,1)} g_{\gamma,\rho,y}(t) = h(y)t$ *uniformly for t in bounded intervals;*

(b). $\lim_{\gamma\to 0} g_{\gamma,\rho,y}(t) = h(y)t/\rho$ *uniformly for t in bounded intervals;*

(c). $\lim_{\gamma\to\infty} g_{\gamma,\rho,y}(t) = 0$ *uniformly for t over $[0,\infty)$;*

(d). $\lim_{\rho\to 1} g_{\gamma,\rho,y}(t) = g_{\gamma,1,y}(t)$ *uniformly for t in bounded intervals.*

Proof. (c) and (d) are trivial corollaries of the definition of $g_{\gamma,\rho,y}(\cdot)$. For (a) and (b), note that

$$\begin{aligned}|g_{\gamma,\rho,y}(t) - h(y)t/\rho| &\le \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y}|h(s)-h(y)|ds = \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y}|h'(\xi)(s-y)|ds \\ &\le \frac{4M}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y}|s-y|ds = \frac{4M}{b^2c_x^2\gamma\rho^2}\cdot\frac{1}{2}\left(\frac{b^2c_x^2\gamma\rho}{4}t\right)^2 = N\gamma t^2, \end{aligned} \tag{62}$$

where $N \equiv Mb^2c_x^2/8$. Note that the second line require $h(\cdot)$ to be differentiable. (b) follows directly from (62). To prove (a), we note that $|g_{\gamma,\rho,y}(t) - h(y)t| \le |g_{\gamma,\rho,y}(t) - h(y)t/\rho| + |h(y)t|(1-\rho^{-1})$. ∎

With the two functions defined above, we have a more tractable and intuitive alternate representation to (57).

LEMMA 4. *With $h(\cdot)$ and $g_{\gamma,\rho,y}(\cdot)$ defined in (60) and (61), we have*

$$W^*_{\gamma,\rho,y} = \frac{b^2}{2}\cdot\frac{\rho c_x^2}{2(1-\rho)}\cdot\sup_{t\ge 0}\left\{f(t) + \rho g_{\gamma,\rho,y}(t) + 2\left(\sqrt{t + (1-\rho)g_{\gamma,\rho,y}(t)} - \sqrt{t}\right)\right\} \tag{63}$$

Proof. To expose the three components of the function to be optimized, we provide an alternate representation to (57), in particular,

$$W^*_{\gamma,\rho,y} = \sup_{s\ge 0}\left\{(\rho s - s + bc_x\sqrt{\rho s}) + (\Lambda_{\gamma,y,\rho}(s) - \rho s) + bc_x\left(\sqrt{\Lambda_{\gamma,y,\rho}(s)} - \sqrt{\rho s}\right)\right\}. \tag{64}$$

Now perform the change of variables

$$\frac{b^2c_x^2\rho}{4(1-\rho)^2}t = s. \tag{65}$$

With (60), (61) and this change of variables, we see that

$$W^*_{\gamma,\rho,y} = \frac{b^2}{2}\cdot\frac{\rho c_x^2}{2(1-\rho)}\cdot\sup_{t\ge 0}\left\{\left(-t+2\sqrt{t}\right)+\frac{4}{b^2c_x^2\gamma\rho}I_{\gamma,\rho,y}\left(\frac{b^2c_x^2\rho}{4(1-\rho)^2}t\right)+ 2\left(\sqrt{t+\frac{4(1-\rho)}{b^2c_x^2\gamma\rho^2}I_{\gamma,\rho,y}\left(\frac{b^2c_x^2\rho}{4(1-\rho)^2}t\right)}-\sqrt{t}\right)\right\}$$
$$= \frac{b^2}{2}\cdot\frac{\rho c_x^2}{2(1-\rho)}\cdot\sup_{t\ge 0}\left\{f(t)+\rho g_{\gamma,\rho,y}(t)+2\left(\sqrt{t+(1-\rho)g_{\gamma,\rho,y}(t)}-\sqrt{t}\right)\right\}. \quad \blacksquare$$

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 $t=1$, $g_{\gamma,\rho,y}$ is a periodic function fluctuating around 0 with limits in Lamma 3 and the third component in (63) is typically small, especially when $\rho\approx 1$.

Now, we present the heavy traffic limit for PRQ.

THEOREM 4. (*heavy traffic limit for PRQ*) *For the PRQ problem in* (57)*, the heavy traffic limit is*

$$\lim_{\rho\uparrow 1}\frac{2}{b^2}\cdot\frac{2(1-\rho)}{\rho c_x^2}\cdot W^*_{\gamma,\rho,y} = \sup_{s\ge 0}\{f(t)+g_{\gamma,1,y}(t)\}. \tag{66}$$

Proof. First, note that

$$\sup_x\{f(x)\}+\inf_x\{g(x)\}\le\sup_x\{f(x)+g(x)\}\le\sup_x\{f(x)\}+\sup_x\{g(x)\},$$

and that $\sqrt{t+(1-\rho)g_{\gamma,\rho,y}(t)}-\sqrt{t}$ converges to 0 uniformly over $[0,\infty)$ as $\rho\uparrow 1$. Apply Lemma 4, we then have

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

Now, we need only consider a bounded interval of t, becasuse $g_{\gamma,\rho,y}(\cdot)$ is uniformly bounded by definition (61) and thus the objective function in the supremum will be negative outside a bounded interval. The result then follows from part (d) of Lemma 3. $\blacksquare$

We immediately obtain an upper bound for the PRQ solution for the system with sinusoidal arrival rate, which reveals the essential shape of the solution as we shall see in §5.

COROLLARY 3. *Suppose* $h(x)=\beta\sin(2\pi x)$*, then*

$$\lim_{\rho\uparrow 1}\frac{2}{b^2}\cdot\frac{2(1-\rho)}{\rho c_x^2}W^*_{\gamma,\rho,y}\le\lim_{\rho\uparrow 1}f(t)+\lim_{\rho\uparrow 1}g_{\gamma,\rho,y}(t)\le 1+\frac{2\beta}{\pi b^2c_x^2\gamma}(1-\cos(2\pi y)),\quad 0\le y<1. \tag{67}$$

4.3. 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 §3.3. When we let the cycles get long in heavy-traffic, we see that there are three very different cases, depending on the instantaneous arrival rate function. Since the average arrival rate satisfies (16), the model is necessarily stable for each $\rho < 1$ with a proper steady-state distribution, but the local behavior depends on the instantaneous traffic intensity $\rho(y)$. In the heavy-traffic setting of §4-4.2, 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 $h^{\uparrow} = 1$.

In the underloaded case, there will be no times at which the net input rate is positive. For fixed ρ, the system will be stochastically bounded above by a system with the maximum arrival rate, for which there will be a proper steady-state distribution. In that setting, the arrival rate will stay flat for long enough that the system will approach the steady state for that approximately fixed arrival rate. Thus, in that situation, it is appropriate to approximate the time-varying distribution at each time by the steady-state distribution of the model with the arrival rate at that time, which is known as a *pointwise stationary approximation* (PSA); see Green and Kolesar (1991), Whitt (1991) and Massey and Whitt (1997). We will show that if we let the cycles get long for PRQ in an underloaded model, PRQ is asymptotically consistent with PSA.

The overloaded case is very different. In the overloaded case, there will be times at which the net input rate is positive. Hence, 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. (1997). Finally, there is the more complicated critically loaded case. We consider these cases in turn.

4.3.1. Underloaded queues For an underloaded queue, we have the following heavy-traffic double limit.

THEOREM 5. (*long-cycle heavy-traffic limit for PRQ in an underloaded queue*) *For the PRQ problem in* (57)*, if h is continuously differentiable with $h^{\uparrow} < 1$, then we have the double limit*

$$\lim_{\substack{\gamma \downarrow 0 \\ \rho \uparrow 1}} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \frac{1}{1-h(y)}, \tag{68}$$

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

$$W_y^* = \frac{b^2}{2} \cdot \frac{\rho(y) c_x^2}{2(1-\rho(y))} + o(1-\rho). \tag{69}$$

Proof. From Lemma 4, we have

$$\begin{aligned} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} &= \sup_{t \ge 0} \left\{ f(t) + \rho g_{\gamma,\rho,y}(t) + 2\left(\sqrt{t + (1-\rho) g_{\gamma,\rho,y}(t)} - \sqrt{t} \right) \right\} \\ &= \sup_{t \ge 0} \{ F_{\gamma,\rho,y}(t) \}, \end{aligned} \tag{70}$$

where $F_{\gamma,\rho,y}(t) \equiv f(t) + \rho g_{\gamma,\rho,y}(t) + 2\left(\sqrt{t + (1-\rho) g_{\gamma,\rho,y}(t)} - \sqrt{t} \right)$. Note that $F_{\gamma,\rho,y}(\cdot)$ is negative outside a bounded interval and that $\sup_{t \ge 0} \{ -(1-h(y))t + 2\sqrt{t} \} = 1/(1-h(y))$, it suffices to prove that $F_{\gamma,\rho,y}(t)$ converges uniformly to $-(1-h(y))t + 2\sqrt{t}$ over all bounded interval of t as $(\gamma, \rho) \to (0,1)$. To this end, we write

$$\begin{aligned} \left| F_{\gamma,\rho,y}(t) - \left(-(1-h(y))t + 2\sqrt{t} \right) \right| &= \left| \rho g_{\gamma,\rho,y}(t) - h(x)t + 2\left(\sqrt{t + (1-\rho) g_{\gamma,\rho,y}(t)} - \sqrt{t} \right) \right| \\ &\le |g_{\gamma,\rho,y}(t) - h(x)t| + (1-\rho)|g_{\gamma,\rho,y}(t)| + 2\left(\sqrt{(1-\rho) g_{\gamma,\rho,y}(t)} \right), \end{aligned}$$

where we used the concavity of the square root function. The result then follows from Lemma 3.

To see that this limit coincides with PSA, note that by (68), we have

$$W_y^* \approx \frac{b^2}{2} \cdot \frac{\rho c_x^2}{2(1-\rho)(1-h(y))} = \frac{b^2}{2} \cdot \frac{\rho c_x^2}{2(1-(\rho + (1-\rho)h(y)))} = \frac{b^2}{2} \cdot \frac{\rho c_x^2}{2(1-\rho(y))}$$

which is asymptotically correct up to $o(1-\rho)$ in the limit. ∎

Corollary 4. *the the* $M_t/M/1$ *model, the double limit of PRQ agrees with the iterated limit of the actual expected workload as first* $\gamma \downarrow 0$ *and then* $\rho \uparrow 1$*.*

Proof. First, Whitt (1991) has proved that PSA is asymptotically correct as $\gamma \downarrow 0$ for the $M_t/M/1$ model. Second, RQ has been shown to be asymptotically correct for the stationary model as $\rho \uparrow 1$ in Whitt and You (2016). ∎

4.3.2. Overloaded queues The overloaded case is quite different from the underloaded because the instantaneous arrival rate will be higher than the service rate at some time. The longer the cycle, the longer the time the system is overloaded, which will lead to a larger workload. The following limit holds more generally, as $\gamma(1-\rho)\downarrow 0$.

THEOREM 6. (*long-cycle heavy-traffic limit for PRQ in an overloaded queue*) *For the PRQ problem in* (57) *with with the heavy-traffic scaling in* (45) *and* $h^{\uparrow} > 1$, *we have the limit*

$$\lim_{\substack{\gamma\downarrow 0\\ \rho\uparrow 1}} \gamma(1-\rho)\cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0}\left\{-t+\int_{y-t}^{y} h(s)ds\right\}, \quad 0\leq \rho < 1. \tag{71}$$

Proof. Note from (57) that

$$\begin{aligned} W^*_{\gamma,\rho,y} &= \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - s + bc_x\sqrt{\Lambda_{\gamma,\rho,y}(s)}\right\} \\ &= \sup_{s\geq 0}\left\{-(1-\rho)\,s + \frac{1}{\gamma(1-\rho)}\int_{y-c^{-1}_{\gamma,\rho}s}^{y} h(u)du + bc_x\sqrt{\Lambda_{\gamma,\rho,y}(s)}\right\} \\ &= \frac{1}{\gamma(1-\rho)}\cdot\sup_{t\geq 0}\left\{-t+\int_{y-t}^{y} h(u)du + \gamma(1-\rho)bc_x\sqrt{\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t)}\right\}, \end{aligned} \tag{72}$$

where we applied a change of variable $c_{\gamma,\rho}t = s$ in the third line. The result follows from the fact that $\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t)$ is in the order of $\rho c_{\gamma,\rho}t = \rho t/(\gamma(1-\rho)^2)$ when $\gamma\to 0$. Then the third term in the curly brace will be $O\left(\gamma^{1/2}\right)$ and converges to 0 uniformly over bounded intervals of t. Note also that the function in the supremum is negative for all t sufficiently large, we need only consider a bounded interval for t. ∎

4.3.3. critically loaded queues The critically loaded case is more complex in terms of space scaling. Though the space scaling does involve the cycle length parameter γ, it will depend on the detailed structure of the arrival rate function instead of a simple γ we see in Theorem 6. The following theorem reveals the relationship between the space scaling and γ.

THEOREM 7. (*long-cycle heavy-traffic limit for PRQ in an critically loaded queue*) *Assume that* $h(t)$ *satisfies*

$$h(t) = 1 - ct^p, \tag{73}$$

for some real numbers $p\geq 0$. *Then the long-cycle heavy-traffic limit of the PRQ solution at the critical point* $y=0$ *is in the order of* $O(\gamma^{-p/(2p+1)})$.

Proof. By (73), we have

$$g_{\gamma,\rho,0}(t) = \frac{4}{b^2 c_x^2 \gamma \rho^2} \int_{-\frac{b^2 c_x^2 \gamma \rho}{4} t}^{0} h(s)ds = \rho^{-1}\left(1 - \frac{c}{p+1}\left(\frac{b^2 c_x^2 \gamma \rho}{4}\right)^p t^{p+1}\right)$$
$$= \rho^{-1}\left(t - M\gamma^p t^{p+1}\right),$$

where $M = c\left(b^2 c_x^2 \rho\right)^p / (4^p(p+1))$. Plugging it into Theorem 4, we have

$$\frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,1,0} = \sup_{t\geq 0}\{f(t) + g_{\gamma,1,0}(t)\} = \sup_{t\geq 0}\left\{2\sqrt{t} - M\gamma^p t^{p+1}\right\},$$

where the supremum is achieved at

$$t^* = \left(\frac{\gamma^{-p}}{M(p+1)}\right)^{2/(2p+1)},$$

with maximum value

$$(2 - 1/(p+1))\left(\frac{1}{M(p+1)}\right)^{1/(2p+1)} \gamma^{-\frac{p}{2p+1}}$$

as $\gamma \to 0$. ∎

The scaling in Theorem 7 coincides with the scaling in the heavy-traffic FCLT in Theorem 4.1 of Whitt (2016), where the space scaling needed at an isolated critical point is investigated. Hence, the scaling in PRQ is asymptotically correct in this regime. This is confirmed in §5.3.1, where we present comparisons between simulation and PRQ values for critically loaded queue.

5. Simulation Comparisons for the Sinusoidal $M_t/GI/1$ Queue

In this section, we present results of several simulation experiments conducted to evaluate the performance of PRQ. These experiments confirm Theorem 2 showing that PRQ is asymptotically correct in long-cycle limit and show that PRQ provides a useful approximation for the mean workload under heavy loads. For these simulations, we consider the $M_t/H_2/1$ model with sinusoidal arrival-rate functions. The service times are i.i.d random variables drawn from a hyperexponential H_2 distribution with mean 1, $c_s^2 = 2$ and balanced means, as on p. 137 of Whitt (1982).

For the PRQ algorithm applied to the $M_t/GI/1$ model in (57), we primarily use the parameter $b = \sqrt{2}$, just as in Whitt and You (2016). For numerical calculation of the PRQ solution, we create a finite mesh over $[0, T]$, where

$$T = \max\left\{c_{\gamma,\rho}, \frac{b^2 c_x^2 \rho}{4(1-\rho)^2}\right\}.$$

This choice of T ensures that the maximum is obtained in $[0,T]$ as we see from Lemma 4 and the fact that $f(t)$ achieves maximum at $t^* = 1$. Then we simply choose a mesh fine enough such that the error is tolerable.

5.1. Confirming the Long-Cycle Limit

For the long-cycle fluid limit, we look at a sequence of models indexed by the cycle-length parameter γ. For model γ, we let the arrival-rate function be

$$\lambda_\gamma(t) = \rho + \beta \sin(2\pi\gamma t), \tag{74}$$

which is a special case of (30). Because the long-cycle limit in Theorem 2 does not require heavy traffic, we let $\rho = 0.75$. For the overloaded case, we choose $\beta = 0.5$ so that $\lambda_f^\uparrow = 1.25$ in (19).

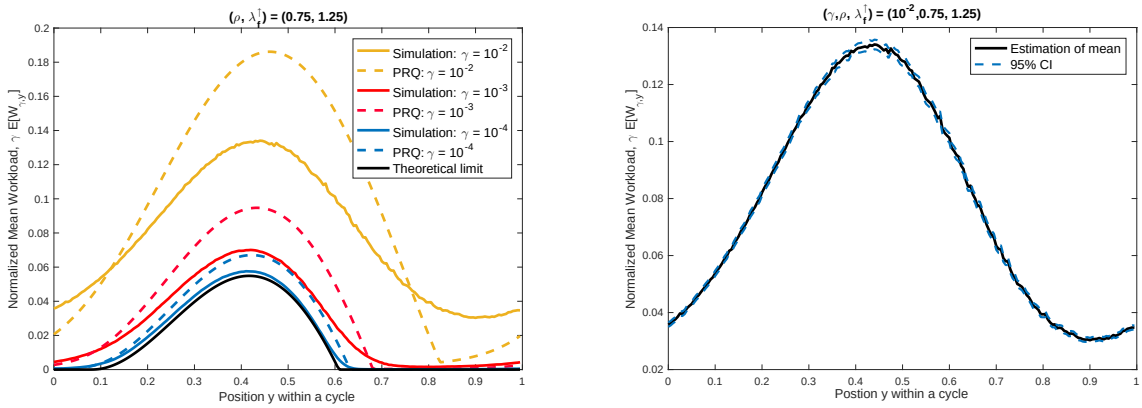

Figure 1 A comparison of the solution to the PRQ problem in (57) as a function of the position y within a cycle to simulation estimates of the normalized mean workload $\gamma E[W_{\gamma,y}]$ in (40) and the theoretical limit in Theorem 1 in the $M_t/H_2/1$ model with arrival-rate function in (74) for $\rho = 0.75$, $\beta = 0.5$ and $\gamma = 10^{-k}$ for $k = 2,3,4$ (on the left). On the right is shown the 95% CI for $\gamma = 10^{-2}$.

Figure 1 compares PRQ to simulation estimates of the mean workload and the heavy-traffic limit for three values of γ: $\gamma = 10^{-k}$ for $k = 2,3,4$. Figure 1 shows that both the simulation results and the solutions to PRQ problem (18) converge to the fluid limit calculated from Proposition 1, confirming Theorem 2.

The estimation in Figure 1 is done over a grid of 100 values, evenly spaced between 0 and 1. For each position y, the expected workload is estimated by the time average of the workload in all intervals of form $[(y+k)\gamma^{-1}, (y+0.01+k)\gamma^{-1})$, where γ^{-1} is the cycle length and k is a positive integer. The statistical precision is shown in Figure 1 (right) in the form of 95% confidence interval. Since the cycle length grows with respect to the decrease of γ, we choose a simulation time proportional to γ^{-1} in order to maintain similar statistical precision.

5.2. Experiments for Heavy Traffic and Long Cycles

We now turn to experiments investigating the heavy traffic limits in §4.2. We consider a special case of arrival-rate functions in (46),

$$\lambda_{\gamma,\rho}(t) = \rho + (1-\rho)\beta \sin\left((1-\rho)^2\gamma t\right). \tag{75}$$

Figure 2 (left) confirms Theorem 5 for the underloaded case, where $h^{\uparrow} = \beta = 0.8 < 1$. We observe that (i) PRQ captures the essential shape of the simulation value and (ii) both simulation values and PRQ solutions converges to the theoretical limit as $(\gamma, \rho) \to (0, 1)$.

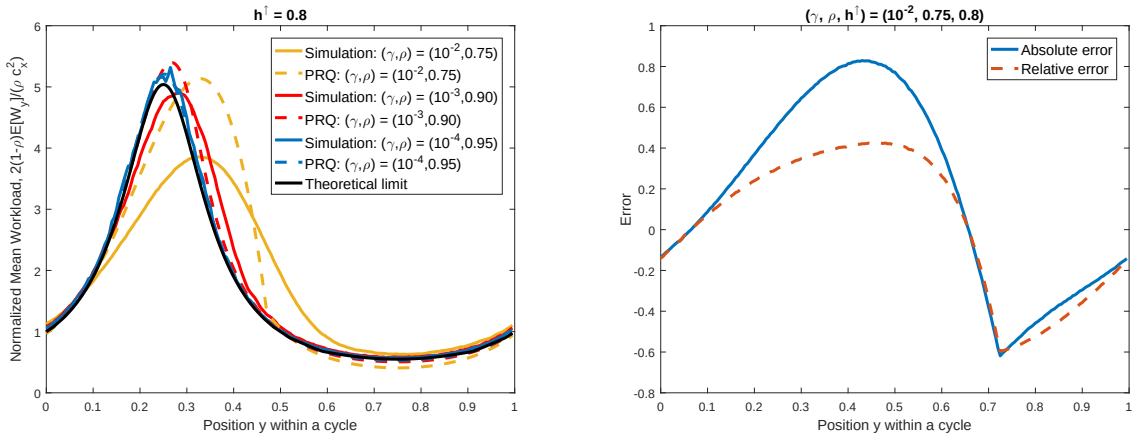

Figure 2 A comparison of the solution to the PRQ problem in (57) as a function of the position y within a cycle to simulation estimates of the mean of the normalized workload $2(1-\rho)E\left[W_{\gamma,\rho,y}\right]/\rho c_x^2$ for $W_{\gamma,\rho,y}$ in (56) and the limit in Theorem 5 in the $M_t/H_2/1$ model with arrival-rate function in (75) for three pairs of (γ, ρ) (left). On the right is shown the absolute and relative errors for $(\gamma, \rho) = (0.01, 0.75)$.

Figure 2 (right) shows the absolute and relative errors, from which we conclude that PRQ serves as a good approximation for the steady-state mean workload even for moderate traffic intensity and moderate cycle length.

5.3. PRQ Capturing Main Features of the Time-Varying Mean

Simulation studies in previous sections confirmed that PRQ is asymptotically correct for both the long-cycle limit and heavy-traffic long-cycle limit in both underloaded and overloaded queues. In this section we look at several features that PRQ captures even for moderate cycle length and traffic intensities.

5.3.1. The space scaling and the quantiles For the critically loaded case, Whitt (2016) showed that the space scaling of the heavy traffic limit depends on the detailed structure of the arrival-rate function and we observed in Theorem 7 that the PRQ solution yields the same space scaling. To show that TVRQ successfully captures this feature, we now present simulation studies for the critically loaded case with arrival rate function in (75), where $h^{\uparrow} = \beta = 1$.

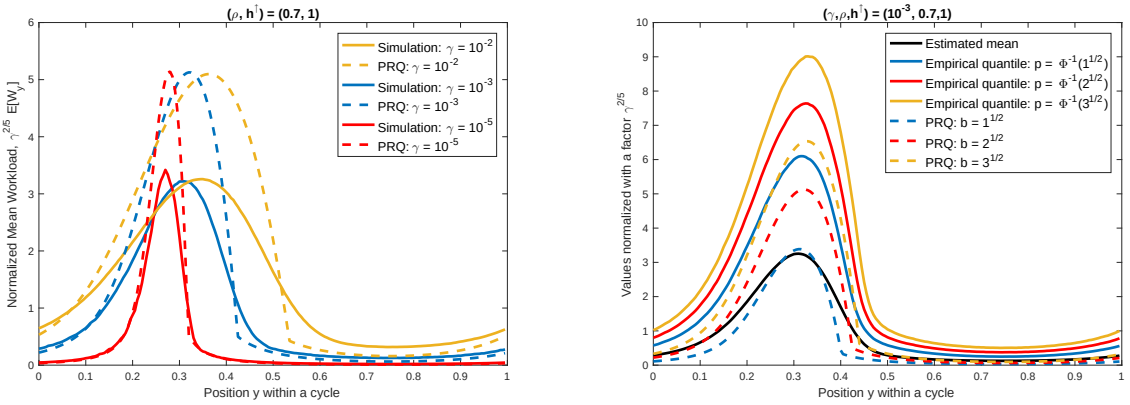

Figure 3 A comparison of the solution to the PRQ problem in (57) as a function of the position y within a cycle to simulation estimates of the mean workload in a critically loaded queue with arrival-rate function in (75) for $\gamma = 10^{-k}$, $k = 2, 3, 5$ (left). The space scaling parameter is $\gamma^{2/5}$ instead of γ for the overloaded case. On the right is shown PRQ solutions for the different parameter b's and the corresponding empirical quantile of the simulated values.

To relate to Theorem 7, we perform Taylor's expansion to the sinusoidal arrival rate function around the critical point of $y = 0.25$. Hence, the relevant parameter in (73) will be $p = 2$, resulting in a space scaling proportional to $\gamma^{2/5}$. Figure 3 (left) confirms Theorem 7; i.e., PRQ successfully captures the space scaling for critically loaded queues, which generally depends on the detailed structure of the arrival-rate function. We observe that PRQ predicts the timing of the peak congestion remarkably well.

Figure 3 (right) shows the impact of the parameter b on the PRQ solution, together with the corresponding quantiles of the simulated values. The correspondence comes from a normal approximation and the intuition that the uncertainty set is defined as if we replace the net input process by its quantile process. The figure shows that the shape of the PRQ solution matches the shape of the quantile very well, which in turns give a good approximation of the mean value.

5.3.2. Location of the peak and the steep drop When the we do not scale the cycle length parameter γ, PRQ may not be asymptotically correct in heavy traffic but still serve as a good approximation, as we see in Figure 2. In fact, for the overloaded queue, traffic intensities almost have no impact on the simulation value beyond a constant scaling of $(1-\rho)$. This feature also appears in the heavy traffic limit theorem for PRQ, i.e., Theorem 6. From (72) there, we see that the long cycle limit for PRQ depends on ρ only through a constant scaling of $(1-\rho)$.

Figure 4 (left) shows PRQ and simulation estimates of the normalized mean workload normalized with a factor of $\gamma(1-\rho)$ for fixed $\gamma = 0.01$ but increasingly larger traffic intensity. We observe that both the PRQ soulution and the simulation values are almost independent of ρ after scaling. In addition, PRQ successfully captures the qualitative changes of the simulation as a function of position y, e.g., the mode around $y = 0.4$ and the sharp turning around $y = 0.7$. Figure 4 (right) shows the PRQ objective functions in (57) for $y = 0.65$ and $y = 0.7$, which explains that the sharp turning is caused by the optimal point switching from one mode to another.

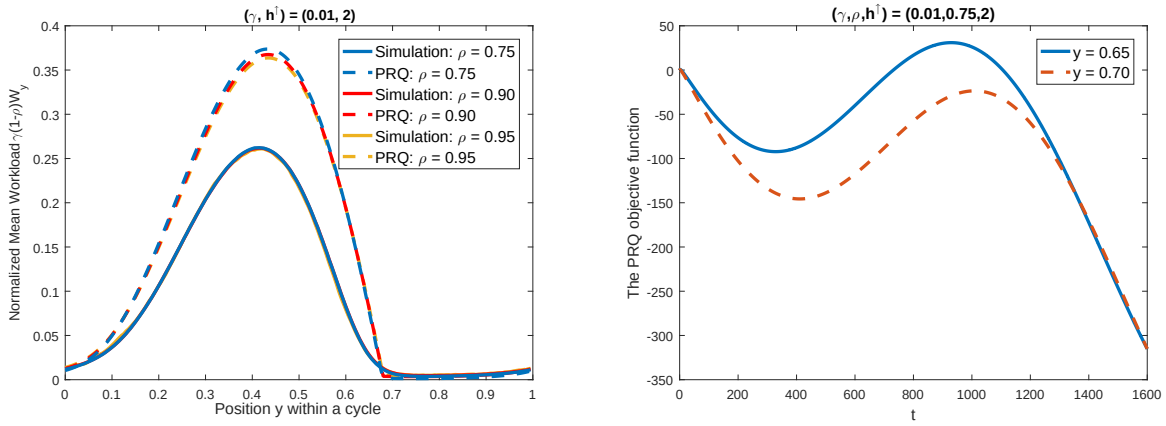

Figure 4 A comparison between the PRQ solutions in (57), as a function of the position y of a cycle, and simulation values for three set of parameters (left). On the right is shown the PRQ objective function over the first cycle $t \in [0, c_{\gamma,\rho}]$ in (57) for two different position parameters y, under parameter triple $(\gamma, \rho, h^{\uparrow}) = (0.01, 0.75, 2)$.

6. Conclusions

In this paper have developed a time-varying robust queueing (TVRQ) algorithm to approximate the time-varying mean (and quantiles; see §5.3.1) of the workload in a general $G_t/G/1$ single-server queue with time-varying arrival rate function. Exploiting a reverse-time construction of the workload process in §2.1, we developed a general TVRQ representation of the workload at time t, starting empty at time 0, as the supremum of an approximating reverse-time net input process in (6). Exploiting the composition representation of the arrival counting process in (7), we obtained the explicit representation in terms of the cumulative arrival rate function Λ for the $G_t/G/1$ queue in (15).

The rest of the paper focused on the special case of periodic RQ (PRQ). In that case we focus on the periodic steady-state workload at place y within a periodic cycle. The general representation of the PRQ workload as a function of y appears in (18). After developing a deterministic fluid model for the periodic queue in §2.4 and §3.2, we established long-cycle limits for both the actual periodic workload and the PRQ that showed the both converge to the same fluid workload, implying that PRQ is asymptotically correct in that limit.

In §4 we established heavy-traffic limits as the long-run average traffic intensity ρ in (16) increases toward 1 for both the for both the actual periodic workload and the PRQ, using the scaling in Whitt (2014), but in general these limits do not agree. In §4.3 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 (PSA) version of the heavy-traffic limit for the stationary model, which has been shown to be asymptotically correct in Whitt and You (2016). 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. (1997). Third, for critically loaded cases, we obtained the limit for PRQ in Theorem 7, consistent with Whitt (2016). Finally, we reported results of simulation experiments that confirm the limit theorems and show that PRQ can be helpful for these complex time-varying models.

Acknowledgments

Support was received from NSF grants CMMI 1265070 and 1634133.

References

Bandi, C., D. Bertsimas, N. Youssef. 2014. Robust transient multi-server queues and feedforward networks. Unpublished manuscript, MIT ORC Center.

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

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

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

Beyer, H. G., B. Sendhoff. 2007. Robust optimzation - a comprehensive survey. *Computer Methods in Applied Mechanics and Engineering* **196**(33-34) 3190–3218.

Choudhury, G. L., A. Mandelbaum, M. I. Reiman, W. Whitt. 1997. Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models* **13**(1) 121–146.

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.

Green, L. V., P. J. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* **37** 84–97.

Keller, J. 1982. Time-dependent queues. *SIAM Review* **24** 401–412.

Ma, N., W. Whitt. 2016. A performance algorithm for periodic queues. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Mandelbaum, A., W. A. Massey. 1995. Strong approximations for time-dependent queues. *Mathematics of Operations Research* **20**(1) 33–64.

Massey, W. A. 1981. Nonstationary queues. Thesis, Stanford University.

Massey, W. A. 1985. Asymptotic analysis of the time-varying $M/M/1$ queue. *Mathematics of Operations Research* **10**(2) 305–327.

Massey, W. A., W. Whitt. 1997. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability* **9**(4) 1130–1155.

Newell, G. F. 1968a. Queues with time dependent arrival rates, I. *Journal of Applied Probability* **5** 436–451.

Newell, G. F. 1968b. Queues with time dependent arrival rates, II. *Journal of Applied Probability* **5** 579–590.

Newell, G. F. 1968c. Queues with time dependent arrival rates, III. *Journal of Applied Probability* **5** 591–606.

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

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

Whitt, W. 2002a. Internet supplement to the book, Stochastic-Process Limits. Available online at: http://www.columbia.edu/∼ww2040.

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

Whitt, W. 2014. Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters* **42** 458–461.

Whitt, W. 2016. Heavy-traffic limits for a single-server queue leading up to a critical point. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Whitt, W., W. You. 2016. Using robust queueing to expose the impact of dependence in single-server queues. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.