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 and asymptotic methods 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: Submitted, August 27, 2016; Revision: June 30, 2017

1. Introduction

Queueing has long played a prominent role in operations research applications. For example, early OR studies include traffic delays at tool 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 queue-

ing 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 employed 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 (ODE's) 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 (NHPP, the 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 ODE's, 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 ODE's 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 employing 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.

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 mean 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). We establish asymptotic results for PRQ and make connections to corresponding asymptotic results for the original stochastic model. In particular, we establish a long-cycle fluid limit (Theorem 2) and a heavy-traffic limit (Theorem 4) for PRQ and compare them to the associated limits for the original queueing model.

Given that PRQ deviates from the original stochastic model, it should not be surprising that the asymptotic behavior of PRQ and the original model do not always coincide, but we find that they sometimes do. In particular, we show that the long-cycle fluid limit for PRQ and the original model coincide. We also find that the HT limits of PRQ and the original model do coincide in some cases.

Finally, we report results of extensive simulation experiments to evaluate the performance of PRQ. These show that PRQ yields useful approximations.

1.1. 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), Heyman and Whitt (1984), Lemoine (1981, 1989) and Rolski (1989), (ii) numerical algorithms, as discussed above, and (iii) asymptotic methods and approximations by Newell (1968a,b,c), Massey (1981), 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 Bertsimas et al. (2011a), 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 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 (2016), 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. (2014) 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).

As in our previous paper, Whitt and You (2016), we not only formulate the new PRQ, but we also study its performance compared to the original stochastic model. For that purpose, we establish new asymptotic results, including a HT limit. For that HT limit, we use the new HT scaling introduced in Whitt (2014, 2016).

1.2. Main Contributions

1. In this paper, we present what we believe is the first application of robust optimization to study the performance of a queueing model with time-varying arrival rates. We focus especially on the periodic steady-state performance of a single-server queue with a periodic arrival-rate function, yielding what we call periodic robust queueing (PRQ).

2. In contrast to the prevalent ODE methods, our TVRQ and PRQ apply to the non-Markovian $G_t/G/1$ model as well as the Markovian special cases, with extension to $G_t/G_t/1$ models, see Remark 1. Both the TVRQ and PRQ formulations and the HT limits provide remarkably tractable approximations; see (6) and (16) for TVRQ and (21), (56) and (EC.14) for PRQ.

3. As in Whitt and You (2016), we exploit 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.

4. For periodic queues, we establish long-cycle fluid limits for both the original queueing system and the PRQ approximation and we prove that those limits coincide (Theorems 1 and 2).

5. We establish new HT limits for PRQ in the $G_t/G/1$ model, which generalize nicely to cover the $G_t/G_t/1$ model, see Remark 7 and §EC.5. These new HT limits exploit the new 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 special cases.

6. We show that the HT limits can be usefully combined with the long-cycle perspective to obtain further insight into the TV behavior of periodic queues. We show that it is important to consider three cases for the instantaneous arrival-rate function $\rho(t)$. Letting $\rho^{\uparrow} \equiv \sup_{t\geq 0} \rho(t)$. We identify three important cases, called: underloaded ($\rho^{\uparrow} < 1$), overloaded ($\rho^{\uparrow} > 1$) and critically loaded ($\rho^{\uparrow} = 1$). We establish results for each case (with slightly different notation) in §4. We find that the HT limits for PRQ coincide with the pointwise-stationary approximation for the original model in both the underloaded case (Theorem 5) and overloaded case (Theorem 6). We find that the scaling in the critically loaded case agrees with the scaling in Whitt (2016).

7. We show that the detailed structure of the objective function in the PRQ provides insight into the structure of the mean and quantiles of the periodic workload; e.g., see §5.1 and §EC.6.3. Thus, we expose a promising way to obtain new insight into the "physics" of TV single-server queues, paralleling Eick et al. (1993) for many-server queues.

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 performance measure of interest as a supremum, thus providing a basis for the RQ optimization.

2.1. A Reverse-Time Construction 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, but that condition can easily be relaxed to consider other initial conditions; e.g., see §13.5

of Whitt (2002b). 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 stochastic process $W(t)$ in (2) into a deterministic approximation as the solution of a deterministic optimization problem. In our simulation experiments we will compare this deterministic approximation to the mean $E[W(t)]$ estimated from multiple independent replications of the model.

In particular, to formulate the deterministic TVRQ approximation for the time-varying workload $W(t)$ for any $t > 0$, we let

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

where $\mathcal{U}_t$ is the deterministic uncertainty set, i.e., the set of allowed sample paths $\{X_t(s) : 0 \le s \le t\}$, 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. This uncertainty set requires that the sample path of the reverse-time net-input process $X_t(s) \equiv Y_t(s) - s$ remain within b standard deviations of its mean at all times s, $0 \le s \le t$.

The uncertainty set in (5) is a natural time-varying generalization of the uncertainty sets in our previous paper, which are similar to the ones used in Bandi et al. (2015). Nevertheless, providing convincing support for this uncertainty set, even in the stationary setting, is somewhat complicated. Thus the choice may ultimately be justified by its utility, which is demonstrated by establishing connections to the performance of the original queueing model. We refer readers to §EC.3 and §EC.4. in Whitt and You (2016) for further discussion of the motivation and justification.

The general idea is that (5) can be based on a Gaussian approximation, which in turn is supported by central limit theorem (CLT) for $Y_t(s)$ under customary regularity conditions. (The basis for the CLT in the non-stationary setting is more obvious in the setting of the next section, which we follow for the rest of this paper.) Of course, the CLT only supports a Gaussian approximation for suitably large t. Thus, it is significant that the deterministic optimization tends to attain its supremum at an intermediate value, which is neither extremely small not extremely large. The CLT for the input process is also the theoretical basis for heavy-traffic limits, so that heavy-traffic limits provide support for the uncertainty set as well. For additional discussion of the CLT in the stationary and nonstationary settings, see §EC.5 of Whitt and You (2016) and Whitt (2014), respectively.

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 (6) is tractable, whereas the original stochastic model is complicated.

2.3. An Alternative Representation of the Workload Process for the $G_t/G/1$ Special Case

For the rest of this paper, we focus on a large class of $G_t/G/1$ models in which the arrival process takes the form of

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

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 function 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

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

where λ_{bd} is a bound. 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 $\sigma_s^2 \equiv Var(V)$. When we introduce additional scaling, it is in the context of this model.

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

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

where $\Lambda_t(s)$ is defined in (11), with $\Lambda_t(t) = \Lambda(t) - \Lambda(0) = \Lambda(t)$ and $\Lambda_t(0) = 0$. The second expression in (12) was first discovered by Lemoine (1981) for the case of the $M_t/G/1$ model. The extended Lemoine representation of the workload for the $G_t/G/1$ queue here is appealing because it has all the stochastic variability in the first term within the supremum but all determinstic variability in the arrival-rate function in the second term within the supremum. It was also exploited to develop a rare-event simulation algorithm for periodic $GI_t/GI/1$ queues in Ma and Whitt (2016). For the $M_t/GI/1$ model, the underlying total input process $\left\{\sum_{k=1}^{N(s)} V_k : s \ge 0\right\}$ is a stationary compound Poisson process whose variance is readily available.

2.4. An Alternative Uncertainty Set Using the Index of Disperion for Work

In this section, drawing on §4.1 of Whitt and You (2016), we give alternative representations for the uncertainty set in (5) and the final TVRQ algorithm in (6) using the index of dispersion for work (IDW). The IDW, denoted by $I_w(t)$, was introduced in Fendick and Whitt (1989) to characterize the variability of the total input of work $Y(t)$ over the time interval $[0,t]$, independent of its mean. The idea is the same as the squared coefficient of variance (scv, variance divided by the square of the mean), which represents the variability of a single random variable independent of scale.

For the base total input process $\tilde{Y}(t) \equiv \sum_{k=1}^{N(s)} V_k$, the IDW is defined as

$$I_w(t) \equiv \frac{Var(\tilde{Y}(t))}{E[V]E[\tilde{Y}(t)]}, \quad t \ge 0; \tag{13}$$

see (1) of Fendick and Whitt (1989). In our setting with mean-1 service times and a rate-1 base process N, the IDW becomes

$$I_w(t) \equiv \frac{Var(\tilde{Y}(t))}{E[\tilde{Y}(t)]} = \frac{Var(\tilde{Y}(t))}{t}, \quad t \ge 0, \tag{14}$$

which is just a scaled version of the variance function. For the $M/GI/1$ model, we have $I_w(t) = c_a^2 + c_s^2 = 1 + c_s^2$ with c_a^2 and c_s^2 being the coefficient of variation of the interarrival and service

distribution, respectively. We assume that IDW as a function of time is bounded, which is to be anticipated.

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

$$\begin{aligned}\mathcal{U}_t &= \left\{ X(t) : X(s) \le E\left[\sum_{k=1}^{N(s)} V_k\right] - \Lambda_t^{-1}(s) + b\sqrt{\mathrm{Var}\left(\sum_{k=1}^{N(s)} V_k\right)}, 0 \le s \le \Lambda(t) \right\} \\ &= \left\{ X(t) : X(s) \le s - \Lambda_t^{-1}(s) + b\sqrt{sI_w(s)}, 0 \le s \le \Lambda(t) \right\} \\ &= \left\{ X(t) : X(s) \le \Lambda_t(s) - s + b\sqrt{\Lambda_t(s) I_w\left(\Lambda_t(s)\right)}, 0 \le s \le t \right\}. \end{aligned} \tag{15}$$

Combining (4) and (15), we have the tractable TVRQ optimization for the $G_t/G/1$ model

$$W^*(t) = \sup_{0 \le s \le t} \left\{ \Lambda_t(s) - s + b\sqrt{\Lambda_t(s) I_w\left(\Lambda_t(s)\right)} \right\}, \quad t \ge 0, \tag{16}$$

with the final expression in (16) 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. We emphasize that (15) and (16) are just alternative expressions for the uncertainty set in (5) and the final expression in (6), but they are convenient for separately characterizing and understanding the performance impact of the stochastic dependence and time-varying arrival rate function.

Remark 1. (a time-varying service rate) Our TVRQ is easily extended to cover a time-varying service rate, where we separate the service requirements V_k from the rate that service is provided, as in Whitt (2015). To treat the familiar M_t service model, let the service requirements be i.i.d. mean-1 exponential random variables. In general, suppose that the server is operating at a deterministic time-varying rate $\mu(t)$ at time t instead of the constant rate of 1. Let $M(t) = \int_0^t \mu(u)du$ be the cumulative service rate. Define $M_t(s) \equiv M(t) - M(t-s)$ as in (11). The extended Lemoine representations of the workload for variable service rate model is obtained by replacing s in (12) with $M_t(s)$, yielding

$$W(t) = \sup_{0 \le s \le t} \left\{ \sum_{k=1}^{N(\Lambda_t(s))} V_k - M_t(s) \right\} = \sup_{0 \le s \le \Lambda_t(t)} \left\{ \sum_{k=1}^{N(s)} V_k - M_t\left(\Lambda_t^{-1}(s)\right) \right\}. \tag{17}$$

Following the logic of (14)-(16), from the first expression we see that the TVRQ optimization for the $G_t/G_t/1$ model is

$$W^*(t) = \sup_{0\le s\le t}\left\{\Lambda_t(s) - M_t(s) + b\sqrt{\Lambda_t(s)I_w\left(\Lambda_t(s)\right)}\right\}, \quad t \ge 0. \tag{18}$$

In §EC.6.4, we report a simulation study to show the PRQ performance for $G_t/G_t/1$ models.

2.5. 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

$$\rho = \bar{\lambda} \equiv \Lambda(c)/c < 1, \tag{19}$$

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

$$W_y = \sup_{s\geq 0}\{Y_y(s) - s\}, \quad 0 \leq y < c, \tag{20}$$

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

$$\begin{aligned} W_y^* &= \sup_{s\geq 0}\{E[Y_y(s)] - s + bSD(Y_y(s))\} \\ &= \sup_{s\geq 0}\left\{\Lambda_y(s) - s + b\sqrt{\Lambda_y(s) I_w\left(\Lambda_y(s)\right)}\right\}, \quad 0 \leq y < c. \end{aligned} \tag{21}$$

3.2. A Periodic Deterministic Fluid Model

In §2.5 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 (19). In order for the fluid model to be interesting, we also assume that

$$\lambda_f^{\uparrow} \equiv \sup_{0\leq s<c}\{\lambda_f(s)\} > 1. \tag{22}$$

Condition (22) ensures that there will be positive workload at some time. If condition (22) 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

$$Y_f(t) \equiv \Lambda_f(t), \quad t \geq 0. \tag{23}$$

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

$$Y_{f,y}(s) = \Lambda_{f,y}(s) \equiv \Lambda_f(y) - \Lambda_f(y-s), \quad s \geq 0, \quad 0 \leq y < c. \tag{24}$$

We now observe that the fluid workload at time y is determined by the input over the cycle ending at time y. The proof appears in §EC.2 along with the other proofs of key results in this section.

PROPOSITION 1. *For the deterministic fluid model, the workload at time y within the cycle $[0,c)$ defined in* (20) *with $Y_{f,y}$ in* (24) *reduces to the supremum over one cycle, i.e.,*

$$W_{f,y} = \sup_{0 \leq u \leq c} \{Y_{f,y}(u) - u\}, \quad 0 \leq y < c. \tag{25}$$

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

$$\lambda_f(t) - 1 \geq 0, \quad 0 \leq t < \delta, \quad \text{and} \quad \lambda_f(t) - 1 \leq 0, \quad \delta \leq t < 1. \tag{26}$$

Often we may require a time shift to satisfy condition (26). In this setting it is easy to determine the periodic fluid W_y, $0 \leq y \leq c$.

PROPOSITION 2. *If conditions* (19), (22) *and* (26) *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*

$$W_{f,y} = \Lambda_f(y) - y, \quad 0 \leq y \leq \delta^*, \quad \textit{and} \quad W_{f,y} = 0, \quad \delta^* \leq y \leq c, \tag{27}$$

and

$$W_f^{\uparrow} \equiv \sup_{0 \leq y \leq c} \{W_{f,y}\} = W_{f,\delta} = \Lambda_f(\delta) - \delta > 0. \tag{28}$$

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*

$$W_{f,y} = (a-1)y, \quad 0 \le y \le \delta, \quad W_f^{\uparrow} = W_{f,\delta} = (a-1)\delta, \tag{29}$$

and

$$W_{f,y} = (a-1)\delta - (1-b)(y-\delta), \quad \delta \le y \le \delta^* \equiv (a-b)\delta/(1-b) \quad \textit{and} \quad W_{f,y} = 0, \quad \delta^* \le y \le c. \tag{30}$$

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,* $\lambda_f(t) = \rho + \beta \sin(2\pi t)$ *and* $t_0 = \arcsin((1-\rho)/\beta)/2\pi$, *then* $\lambda_f(t_0 + t))$ *satisfies condition* (26) *and* $\delta = 0.5 - 2t_0$, *so that* (*in terms of the original* Λ_f)

$$W_f^{\uparrow} = \Lambda_f(0.5 - t_0) - \Lambda_f(t_0) - 0.5 + 2t_0. \tag{31}$$

As $\rho \uparrow 1$, $t_0 \equiv t_0(\rho) \downarrow 0$, $\delta(\rho) \uparrow 0.5$ *and* $W_f^{\uparrow} \to \Lambda(0.5) - 0.5$.

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

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

for the base arrival-rate function λ_f in the fluid model, satisfying (19) and (22). 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

$$\Lambda_\gamma(t) \equiv \gamma^{-1}\Lambda_f(\gamma t) \quad \text{and} \quad \Lambda_{\gamma,y}(t) \equiv \Lambda_\gamma(\gamma^{-1}y) - \Lambda_\gamma(\gamma^{-1}y - t), \quad 0 \le y < c, \tag{33}$$

so that the associated arrival-rate function is as in (32). The periodic structure with (19) and (22) implies the following bound.

LEMMA 1. *In the setting above with* (19) *and* (22)*,*

$$\max\{\Lambda_f(t), \Lambda_{f,y}(t)\} \le \rho t + \lambda^\uparrow c \quad \textit{and} \quad \max\{\Lambda_\gamma(t), \Lambda_{\gamma,y}(t)\} \le \rho t + \lambda^\uparrow c/\gamma \quad \textit{for all} \quad t \ge 0. \tag{34}$$

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

$$A_\gamma(t) \equiv N(\Lambda_\gamma(t)) \quad \text{and} \quad Y_\gamma(t) \equiv \sum_{k=1}^{A_\gamma(t)} V_k, \quad t \ge 0, \tag{35}$$

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

$$t^{-1}N(t) \to 1 \quad \text{as} \quad t \to \infty \quad \text{w.p.1} \tag{36}$$

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

$$|t^{-1}N(t) - 1| < \epsilon \quad \text{for all} \quad t \ge t_0 \quad \text{w.p.1.} \tag{37}$$

Condition (36) 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 (37) is implied by refinements such as the law of the iterated logarithm. Condition (37), 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. We give the proof in §EC.2.

LEMMA 2. *For the periodic $G_t/GI/1$ model under condition* (36)*,*

$$\gamma A_\gamma(\gamma^{-1}(t)) \to \Lambda_f(t) \quad \textit{and} \quad \gamma Y_\gamma(\gamma^{-1}(t)) \to \Lambda_f(t) \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1} \tag{38}$$

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 (20), i.e.,

$$W_{\gamma,y} = \sup_{s \ge 0} \{Y_{\gamma,y}(s) - s\}, \tag{39}$$

where

$$Y_{\gamma,y}(t) \equiv Y_\gamma(y\gamma^{-1}) - Y_\gamma(y\gamma^{-1} - t), \quad t \ge 0, \quad 0 \le y < c, \tag{40}$$

for Y_γ in (35). We get a fluid limit for $W_{\gamma,y}$, again after scaling. The proof appears in §EC.2.

THEOREM 1. (*long-cycle fluid limit*) *For the periodic $G_t/GI/1$ model under conditions* (36) *and* (37)*,*

$$\gamma W_{\gamma,y} \to W_{f,y} \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1}, \tag{41}$$

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

Let $W^*_{\gamma,y}$ be the PRQ workload at time y/γ for $0 \le y < c$, i.e., $W^*_{\gamma,y}$ is the solution to the PRQ problem (21) at time y/γ with $Y_\gamma(t)$ defined in (35).

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

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

4. Heavy-Traffic Limits for Periodic Robust Queueing

We now consider a family of periodic $G_t/G/1$ single-server models indexed by the traffic intensity ρ defined in (19) together with the specified time-scaling factor γ. Before introducing this new scaling, we start with the model defined in §2.3. 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 review the heavy-traffic limit for the workload process itself.

4.1. 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 in (19). To avoid notational confusion, we add a subscript d to the diffusion quantities. We let the cumulative arrival-rate function in model ρ be

$$\Lambda_{\gamma,\rho}(t) \equiv \rho t + (1-\rho)^{-1}\Lambda_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0, \tag{44}$$

so that the associated arrival-rate function is

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

where

$$\Lambda_{d,\gamma}(t) \equiv \int_0^t \lambda_{d,\gamma}(s)\, ds, \quad \lambda_{d,\gamma}(t) \equiv h(\gamma t), \quad \text{and} \quad \int_0^1 h(t)\, dt = 0 \tag{46}$$

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

$$h(t) \geq -\rho/(1-\rho), \quad 0 \leq t < 1, \tag{47}$$

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

$$-\infty < h^\downarrow \equiv \inf_{0\leq t\leq 1}\{h(t)\} < \sup_{0\leq t\leq 1}\{h(t)\} \equiv h^\uparrow < \infty. \tag{48}$$

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 (2016) 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 (44) and let

$$A_{\gamma,\rho}(t) \equiv N\left(\Lambda_{\gamma,\rho}(t)\right), \quad Y_{\gamma,\rho}(t) \equiv \sum_{k=1}^{A_{\gamma,\rho}(t)} V_k, \quad \text{and} \quad X_{\gamma,\rho}(t) \equiv Y_{\gamma,\rho}(t) - t, \quad t \geq 0. \tag{49}$$

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.,

$$W_{\gamma,\rho}(t) = \Psi(X_{\gamma,\rho})(t) = \sup_{0 \leq s \leq t} \{X_{\gamma,\rho}(t) - X_{\gamma,\rho}(t-s)\}. \tag{50}$$

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

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt], \quad \hat{A}_{\gamma,\rho}(t) \equiv (1-\rho)[A_{\gamma,\rho}\left((1-\rho)^{-2}t\right) - (1-\rho)^2 t],$$
$$\hat{X}_{\gamma,\rho}(t) \equiv (1-\rho)X_{\gamma,\rho}\left((1-\rho)^{-2}t\right) \quad \text{and} \quad \hat{W}_{\gamma,\rho}(t) \equiv (1-\rho)W_{\gamma,\rho}\left((1-\rho)^{-2}t\right), \quad t \geq 0. \tag{51}$$

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., $e(t) \equiv t$, $t \geq 0$. Recall that $g(x) = o(x)$ as $x \to 0$ if $g(x)/x \to 0$ as $x \to 0$.

THEOREM 3. (*heavy-traffic FCLT, Theorem 3.2 of Whitt (2014) and Theorem 2 of Ma and Whitt (2016)*) *For the family of $G_t/GI/1$ models indexed by (γ, ρ) with cumulative arrival-rate functions in (44), if $\hat{N}_n \Rightarrow c_a B_a$ as $n \to \infty$, where B_a is a standard Brownian motion, then*

$$(\hat{A}_{\gamma,\rho}, \hat{X}_{\gamma,\rho}, \hat{W}_{\gamma,\rho}) \Rightarrow (\hat{A}_\gamma, \hat{X}_\gamma, \hat{W}_\gamma) \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad \rho \uparrow 1, \tag{52}$$

where

$$(\hat{A}_\gamma, \hat{X}_\gamma, \hat{W}_\gamma) \equiv (c_a B_a + \Lambda_{d,\gamma} - e, \hat{A}_\gamma + c_s B_s, \Psi(\hat{X}_\gamma)), \tag{53}$$

Ψ is the reflection map in (50), $\Lambda_{d,\gamma}$ is defined in (46), 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

$$\hat{W}_\gamma = \Psi(c_a B_a + c_s B_s + \Lambda_{d,\gamma} - e) \stackrel{\mathrm{d}}{=} \Psi(c_x B + \Lambda_{d,\gamma} - e), \tag{54}$$

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 (44).

REMARK 2. (need for an approximation of the periodic heavy-traffic limit) We emphasize that the limit $\hat{W}_\gamma$ for the workload process in Theorem 3 is a RPBM, which is much less tractable than the familiar RBM that arises in the stationary case. Thus, PRQ can be helpful to approximate RPBM as well as the queue.

REMARK 3. (a parametric PRQ stemming from the diffusion approximation) We can obtain an alternative simplified parametric PRQ if we apply the PRQ logic after approximating the net-input process by the net-input of the diffusion process arising in the heavy-traffic limit. We explain in detail in §EC.4.

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

$$W_{\gamma,\rho,y}(t) \stackrel{\mathrm{d}}{=} \sup_{0\le s\le t}\left\{\sum_{k=1}^{A_{\gamma,\rho,y}(t)} V_k - s\right\}, \tag{55}$$

where $A_{\gamma,\rho,y}(t) \equiv A_{\gamma,\rho}(y) - A_{\gamma,\rho}(y-t)$, $A_{\gamma,\rho}(t)$ is defined in (49) and V_k is a generic service time. Under the $G_t/G/1$ setting in §2.3, we immediately get the PRQ optimization problem from (16) by replacing $\Lambda_t(s)$ with $\Lambda_{\gamma,y,\rho}(s)$

$$W^*_{\gamma,\rho,y} = \sup_{s\ge 0}\left\{\Lambda_{\gamma,\rho,y}(s) - s + b\sqrt{\Lambda_{\gamma,\rho,y}(s) I_w\left(\Lambda_{\gamma,\rho,y}(s)\right)}\right\}. \tag{56}$$

For the convenience of further analysis, we note that

$$\begin{aligned}\Lambda_{\gamma,\rho,y}(s) &\equiv \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho}) - \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho} - s) = \Lambda_{\gamma,\rho}(yc_{\gamma,\rho}) - \Lambda_{\gamma,\rho}(yc_{\gamma,\rho} - s)\\ &= \rho s + (1-\rho)^{-1}\int_{y/\gamma-(1-\rho)^2 s}^{y/\gamma} h(\gamma t)dt = \rho s + \frac{1}{\gamma(1-\rho)}\int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt\\ &= \rho s + \frac{1}{\gamma(1-\rho)} H_{\gamma,\rho,y}(s),\end{aligned} \tag{57}$$

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

$$H_{\gamma,\rho,y}(s) \equiv \int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt. \tag{58}$$

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

$$f(t) \equiv -t + 2\sqrt{t} \tag{59}$$

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

$$g_{\gamma,\rho,y}(t) \equiv \frac{4}{b^2 c_x^2 \gamma \rho^2} H_{\gamma,\rho,y}\left(\frac{b^2 c_x^2 \rho}{4(1-\rho)^2} t\right) = \frac{4}{b^2 c_x^2 \gamma \rho^2} \int_{y-\frac{b^2 c_x^2 \gamma \rho}{4} t}^{y} h(s) ds \tag{60}$$

is a periodic function that captures the time-varying part of the arrival rate function. The period of $g_{\gamma,\rho,y}(t)$ is $4/b^2 c_x^2 \gamma \rho$. 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 $g_{\gamma,\rho,y}(t)$. Almost all proofs in this section appear in §EC.3.

Lemma 3. *Let h be a differentiable 1-periodic function whose integral over one period is 0. Assume that h satisfies (48), 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.*

With the two functions defined above, we present a more tractable and intuitive alternate representation to (56), which exposes the three components of the function to be optimized. We remark that the expression is inspired by the classical Kingman's bound and approximation of the relaxation time for $GI/G/1$ model, which brings both the relaxation time and workload back to the scale of $O(1)$.

Lemma 4. *With f and $g_{\gamma,\rho,y}$ defined in (59) and (60), we have*

$$W^*_{\gamma,\rho,y} = \frac{b^2}{2} \cdot \frac{\rho c_x^2}{2(1-\rho)} \cdot \sup_{t\geq 0}\left\{ f(t) + \rho g_{\gamma,\rho,y}(t) + 2\left(\sqrt{(t + (1-\rho) g_{\gamma,\rho,y}(t))\, C_{\gamma,\rho,y}(t)} - \sqrt{t}\right)\right\}, \tag{61}$$

where

$$C_{\gamma,\rho,y}(t) \equiv \frac{1}{c_x^2} \cdot I_w\left(\frac{b^2 c_x^2 \rho^2}{4(1-\rho)^2}\left(t + (1-\rho) g_{\gamma,\rho,y}(t)\right)\right).$$

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 Lemma 3 and that the third component in (EC.14) is typically small, especially when $\rho \approx 1$. Furthermore, we have

$$\lim_{\rho\uparrow 1} C_{\gamma,\rho,y}(t) = \lim_{t\to\infty} I_w(t)/c_x^2 = 1$$

uniformly for t bounded away from 0, where the second equation holds under regularity conditions, see §IV.A of Fendick and Whitt (1989).

Now, we present the heavy traffic limit for PRQ; the proof appears in §EC.3.

THEOREM 4. (*heavy traffic limit for PRQ*) *The heavy traffic limit of the PRQ problem in* (56) *for the* $G_t/G/1$ *model is*

$$\lim_{\rho\uparrow 1} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0} \{f(t) + g_{\gamma,1,y}(t)\}. \tag{62}$$

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*

$$\lim_{\rho\uparrow 1} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} W^*_{\gamma,\rho,y} \leq \lim_{\rho\uparrow 1} f(t) + \lim_{\rho\uparrow 1} g_{\gamma,\rho,y}(t) \leq 1 + \frac{2\beta}{\pi b^2 c_x^2 \gamma}(1-\cos(2\pi y)), \quad 0 \leq y < 1. \tag{63}$$

REMARK 4. (The heavy traffic limits do not coincide in this case.) From the statements of Theorems 3 and 4, it is not obvious if the deterministic heavy-traffic limit for PRQ agrees with the mean value of the steady-state distribution of the heavy-traffic limit of the original stochastic model, paralleling the strong results we obtained for the stationary model in Whitt and You (2016). However, our numerical experiments show that these two do not coincide in general.

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 (19), 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. (1997b). Finally, there is the more complicated critically loaded case. We consider these cases in turns.

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*) *Assume that h is continuously differentiable with $h^{\uparrow} < 1$, then the PRQ problem in* (56) *for the $G_t/G/1$ model admits the double limit*

$$\lim_{\substack{\gamma \downarrow 0 \\ \rho \uparrow 1}} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \frac{1}{1-h(y)}, \tag{64}$$

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

$$W^*_y = \frac{b^2}{2} \cdot \frac{\rho(y) c_x^2}{2(1-\rho(y))} + o(1-\rho). \tag{65}$$

REMARK 5. (asymptotic validity of PSA) We remark that the double limit in Theorem 5 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 (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 limit for PRQ in an overloaded queue*) *The PRQ problem in* (56) *for the* $G_t/G/1$ *model with the heavy-traffic scaling in* (44) *and* $h^{\uparrow} > 1$ *admits the long-cycle limit*

$$(1-\rho)\lim_{\gamma\downarrow 0}\gamma \cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0}\left\{-t + \int_{y-t}^{y} h(s)ds\right\}, \quad 0 \leq \rho < 1. \tag{66}$$

REMARK 6. (the space scaling) When the queue is not overloaded, Theorem 6 yields the trivial limit 0, as does Theorem 2. That implies that the scaling constant γ in (66) then becomes too much to generate an interesting limit. For underloaded queues, we saw in §4.3.1 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 next in §4.3.3.

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 a critically loaded queue*) *Assume that* $h(t)$ *satisfies*

$$h(t) = 1 - ct^p + o(t^p), \ \ as\ t \to 0, \tag{67}$$

for some positive real numbers c and p. Then the long-cycle heavy-traffic limit of the PRQ solution for the $G_t/G/1$ model at the critical point $y = 0$ is in the order of $O(\gamma^{-p/(2p+1)}(1-\rho)^{-1})$ as $(\rho,\gamma) \to (1,0)$.

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. It is shown there that the space scaling of the heavy traffic limit depends on the detailed structure of the arrival-rate function, which turned out to be the same as the PRQ version considered here. Hence, the scaling in PRQ is asymptotically correct in this regime. This is confirmed in §5.1.3, where we present comparisons between simulation and PRQ values for critically loaded queue.

Remark 7. (Heavy-Traffic and long-cycle limits in the $G_t/G_t/1$ case) From (17), we have the following equivalent representation of the steady-state workload

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

which can be viewed as an equivalent system with alternative arrival-rate function $\Lambda_t\left(M_t^{-1}(s)\right)$. One would expect that all the heavy-traffic and long-cycle results for both the stochastic model and the PRQ would generalize to the case where service rate is variable. This is indeed true, which we shall discuss in §EC.5.

5. Simulation Comparisons for the Sinusoidal $G_t/GI/1$ 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 moderate to heavy loads.

For these simulations, we consider the $G_t/GI/1$ model with various rate-1 base processes N in (7), all having sinusoidal arrival-rate functions, and various mean-1 service-time distributions. The arrival processes are all time-transformed renewal processes with mean-1 interarrival times. The

arrival processes and service-time distributions are further specified by the scv of the mean-1 random variables. Our examples use Erlang (E_k), hyperexponential (H_2, mixture of two exponentials with balanced means, p. 137 of Whitt (1982)) and lognormal distributions, with the scv appearing in parentheses. Our base case is $(H_2(4))_t/LN(1)/1$. We vary the arrival process and service-time distribution in order to demonstrate the robustness of our PRQ algorithm and to show how it can be used to expose and separate the impact of the stochastic variability and the impact of the deterministic time-variability.

For the PRQ algorithm applied to the $G_t/G/1$ model in (56), we primarily use the parameter $b = \sqrt{2}$. As explained in Whitt and You (2016), this choice of b makes RQ exact for the mean steady-state workload in the $M/GI/1$ special case. For numerical calculation of the PRQ solution, we create a finite mesh over $[0, T]$, where

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

and $M = \sup_{t\geq 0} I_w(t)$. 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 choose a mesh fine enough such that the error is negligible.

5.1. Heavy Traffic and Long Cycles limits

5.1.1. Underloaded Queues We start with the double limit for underloaded queues in §4.3.1. We consider a special case of arrival-rate functions in (45),

$$\lambda_{\gamma,\rho}(t) = \rho + (1-\rho)\beta \sin\left(2\pi\gamma(1-\rho)^2 t\right). \tag{68}$$

In particular, Figure 1 compares the solution to the PRQ problem in (56) as a function of the position y within a cycle to simulation estimates of the normalized mean workload $2(1-\rho)E\left[W_{\gamma,\rho,y}\right]/\rho$ for $W_{\gamma,\rho,y}$ in (55) and the limit in Theorem 5 in the underloaded $(H_2(4)_t/LN(1)/1$ model with arrival-rate function in (68), in various cases.

Figure 1 confirms Theorem 5 for the underloaded case, where $h^\uparrow = \beta = 0.8 < 1$. We observe that (i) both the simulation values and the PRQ solutions converge to the theoretical limit as $(\gamma, \rho) \to$

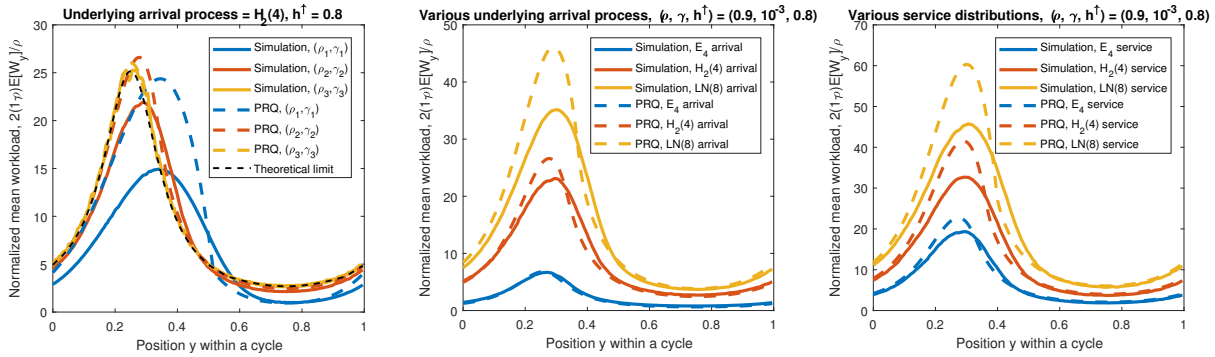

Figure 1 A comparison of the solution to the PRQ problem in (56) as a function of the position y within a cycle to simulation estimations of the normalized mean workload $2(1-\rho)E\left[W_{\gamma,\rho,y}\right]/\rho$ for $W_{\gamma,\rho,y}$ in (55) and the limit in Theorem 5 in the underloaded $(H_2(4)_t/LN(1)/1$ model with arrival-rate function in (68) for $(\gamma,\rho)\in\{(0.7,10^{-2}),(0.9,10^{-3}),(0.95,10^{-4})\}$ (left), for three different arrival processes (middle) and for three different service-time distributions (right).

$(0,1)$; (ii) PRQ captures the essential shape of the simulated mean workload; (iii) PRQ serves as a good approximation for the steady-state mean workload even for moderate traffic intensity and moderate cycle length and across various time-varying arrival processes and service distributions.

Figure 1 (middle) and (right) show the impact of changing variability in the arrival process and the service-time distribution. Consistent with the stationary model, increased variability in either the arrival process or the service process tends to increase congestion. We remark that the story is different from the impact of the service-time distribution on the blocking in the time-varying $M_t/GI/n/0$ loss model; see Davis et al. (1995).

5.1.2. Overloaded Queues Next, we keep the same arrival rate function in (68), but raise the $h^{\uparrow}$ above the critical point of 1, which yields an overloaded queue. Theorem 6 shows that there is a proper long-cycle limit for PRQ, which depends on ρ through a simple scaling of $(1-\rho)$.

Figure 2 (left) shows that both simulated values and PRQ approximations converge to the theoretical limit calculated from Proposition 1, confirming Theorem 2 and 6, while Figure 2 (right) demonstrates that the scaling constant $(1-\rho)$ also appears in the simulated mean workload. Overall, Figure 2 shows that PRQ serves as a reasonable approximation for the overloaded queues even in moderate cycle length and traffic intensities.

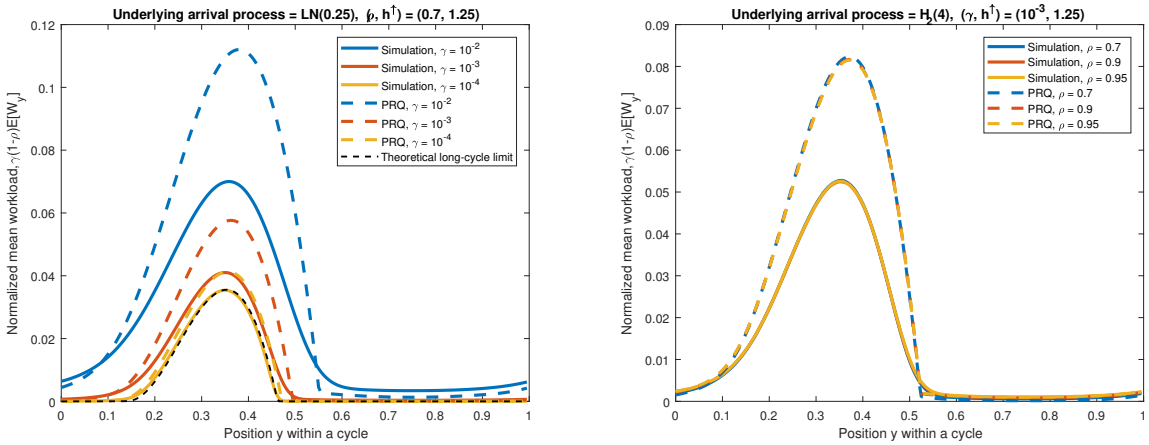

Figure 2 A comparison of the solution to the PRQ problem in (56) as a function of the position y within a cycle to simulation estimations of the normalized mean workload $\gamma(1-\rho)E\left[W_{\gamma,\rho,y}\right]$ for $W_{\gamma,\rho,y}$ in (55) and the limit in Theorem 6 in the overloaded $G_t/LN(1)/1$ model with arrival-rate function in (68) for three values of γ (left) and three values of ρ (right).

5.1.3. Critically Loaded Queues In this experiment, we look at the critically loaded cases, where the arrival rate function (68) with $h^{\uparrow}=1$ is used. To relate to Theorem 7, we perform Taylor's expansion to the sinusoidal arrival rate function around the critical point of $y=0.25$. The relevant parameter in (67) is then $p=2$, resulting in a space scaling proportional to $\gamma^{2/5}$. As discussed in §4.3.3, both the stochastic queues and PRQ approximations exhibit a space scaling of $\gamma^{-2/5}$.

Theorem 7 is confirmed by Figure 3 (left), where we scaled both the estimated mean workload and PRQ approximation with $\gamma^{-2/5}(1-\rho)$. The extra scaling of $(1-\rho)$ is often seen in classical results for stationary queueing models, e.g. in the Kingman's bound. Although the PRQ approximation does not always converge to the mean workload in long-cycle or heavy traffic limits, we can see that it is still a useful approximation. We observe that PRQ predicts the timing of the peak congestion and the sudden drop remarkably well.

Figure 3 (middle) shows that the simulated mean workload depends on the traffic intensity approximately through constant $(1-\rho)$. This is not exact in contrast to the overloaded case. This shows that the critically loaded queue is more complicated. Still, the PRQ algorithm captures this feature.

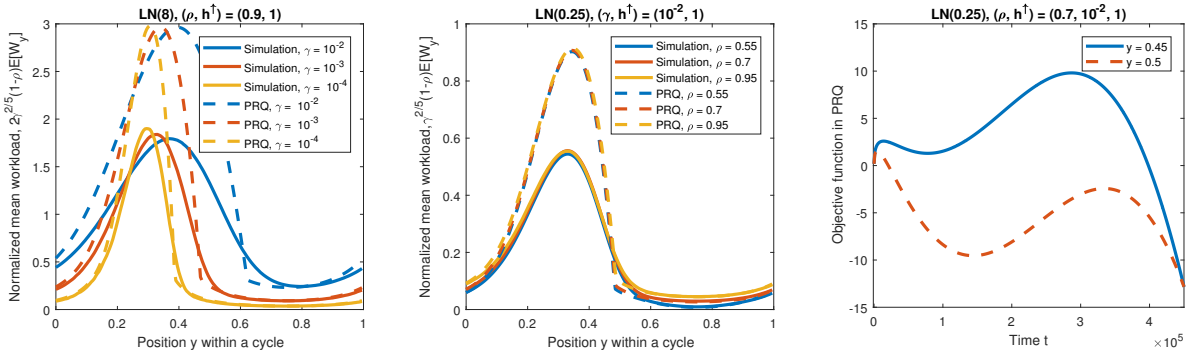

Figure 3 A comparison of the solution to the PRQ problem in (56) as a function of the position y within a cycle to simulation estimations of the normalized mean workload $\gamma^{2/5}(1-\rho)E\left[W_{\gamma,\rho,y}\right]$ for $W_{\gamma,\rho,y}$ in (55) and the limit in Theorem 7 in the critically loaded $G_t/LN(1)/1$ model with arrival-rate function in (68) and underlying stationary point process specified in figure titles, for three choices of γ (left) and for three choices of ρ (middle). The PRQ objective function for the case of $(\rho, h^{\uparrow}) = (0.7, 10^{-2}, 1)$ and $LN(0.25)$ renewal as underlying process is shown on the right.

Figure 3 (right) shows the PRQ objective functions in (56) for $y = 0.45$ and $y = 0.5$ for the case of $(\rho, h^{\uparrow}) = (0.7, 10^{-2}, 1)$ and $LN(0.25)$ renewal as underlying process. Figure 3 (right) explains that the sharp turning is caused by the optimal point switching from one mode to another. This illustrates how plotting the TVRQ objective function can provide useful insight.

6. Conclusions

In this paper, we have developed a time-varying robust queueing (TVRQ) algorithm to approximate the time-varying mean (and quantiles; see §EC.6.3) of the 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 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 reverse time cumulative rate functions Λ_t and M_t for the $G_t/G/1$ queue in (16) and the $G_t/G_t/1$ queue in (18).

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 (21). After developing a deterministic fluid model for the periodic queue in §2.5 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 (19) 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 §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. (1997b). 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 provide helpful insight into 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. 2011a. Theory and applications of robust optimization. *SIAM Review* **53**(3) 464–501.

Bertsimas, D., D. Gamarnik, A. A. Rikun. 2011b. Performance analysis of queueing networks via robust optimization. *Operations Research* **59**(2) 455–466.

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

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., D. L. Lucantoni, W. Whitt. 1997a. Numerical solution of piecewise-stationary $M_t/G_t/1$ queues. *Operatioons Research* **45**(3) 451–463.

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

Davis, J. L., W. A. Massey, W. Whitt. 1995. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Sci.* **41**(6) 1107–1116.

Edie, L. C. 1954. Traffic delays at toll booths. *Operations Research* **2**(2) 107–138.

Eick, S. G., W. A. Massey, W. Whitt. 1993. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** 731–742.

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.

Harrison, J. M., A. J. Lemoine. 1977. Limit theoorems for periodic queues. *Journal of Applied Probability* **14** 566–576.

Heyman, D. P., W. Whitt. 1984. The asymptoic behavior of queues with time-varying arrivals. *Journal of Applied Probability* **21**(1) 143–156.

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

Kolesar, P. J., P. J. Rider, T. B. Craybill, W. E. Walker. 1975. A queueing-linear-programming approach to scheduling police patrol cars. *Operations Research* **23** 1045–1062.

Koopman, B. O. 1972. Air-terminal queues under time-dependent conditions. *Operations Research* **20** 1089–1114.

Lemoine, A. J. 1981. On queues with periodic Poisson input. *Journal of Applied Probability* **18** 889–900.

Lemoine, A. J. 1989. Waiting time and workload in queues with periodic Poisson input. *Journal of Applied Probability* **26**(2) 390–397.

Ma, N., W. Whitt. 2016. A rare-event simulation algorithm for periodic single-server queues. To appear in INFORMS Journal on Computing; 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.

Mandelbaum, A., W. A. Massey, M. I. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* **30** 149–201.

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

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.

Oliver, R. M., A. H. Samuel. 1962. Reducing letter delays in post offices. *Operations Research* **10** 839–892.

Ong, K. L., M. R. Taaffe. 1989. Nonstationary queues with interrupted poisson arrivals and unreliable/repairable servers. *Queueing Systems* **4** 27–46.

Pender, J., W. A. Massey. 2017. Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probability in the Engineering and Information Sciences* **31** 1–42.

Rolski, T. 1989. Queues with nonstationary inputs. *Queueing Systems* **5** 113–130.

Rothkopf, M. H., S. S. Oren. 1979. A closure approximation for the nonstationary $M/M/s$ queue. *Management Science* **25**(6) 522–534.

Taaffe, M. R., K. L. Ong. 1987. Approximating $Ph(t)/M(t)/S/C$ queueing systems. *Annals of Operations Research* **8** 103–116.

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. 2015. Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems* **81** 341–378.

Whitt, W. 2016. Heavy-traffic limits for a single-server queue leading up to a critical point. *Operations Research Letters* **44** 796–800.

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

e-companion

EC.1. Overview

This is an online e-companion to the main paper. It has six sections. First, in §EC.2 we provide the proofs for §3. In §EC.3 we provide the proofs for §4. In §EC.4 we describe the simplified parametric PRQ that follows from applying the heavy-traffic limit to approximate the net-input process by a diffusion process, as mentioned in Remark 3. In §EC.5 we elaborate on Remark 1 about TVRQ for a time-varying service rate. In particular, we extend some of the asymptotic results to the $G_t/G_t/1$ model with TV service as well as TV arrival-rate function Finally, in §EC.6 we present five more simulation examples that demonstrate the effectiveness of PRQ.

.

EC.2. Proofs for Results on PRQ and the Fluid Approximation in §3

Proof of Proposition 1. Let $s = kc + y$, $0 \le y < c$ and $k \ge 0$. Then

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

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

Proof of Lemma 2. Observe that

$$\begin{aligned} \gamma A_\gamma(\gamma^{-1}t) &= \gamma N(\Lambda_\gamma(\gamma^{-1}t)) = \gamma N(\gamma^{-1}\Lambda_f(\gamma(\gamma^{-1}t))) \\ &= \gamma N(\gamma^{-1}\Lambda_f(t)) \to \Lambda_f(t) \quad \text{as} \quad \gamma \downarrow 0 \quad \text{w.p.1} \end{aligned} \tag{EC.2}$$

because $\gamma N(\gamma^{-1}t) \to t$ uniformly over bounded intervals w.p.1 by the FSLLN in (36). A further application of the composition mapping yields the corresponding limit for Y_γ in (35):

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

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. ∎

Proof of Theorem 1. From (39) and (40),

$$\gamma W_{\gamma,y} = \sup_{s\geq 0}\{\gamma Y_{\gamma,y}(\gamma^{-1}s) - s\} \to \sup_{s\geq 0}\{\Lambda_{f,y}(s) - s\} = W_{f,y} \quad \text{as} \quad \gamma \downarrow 0 \quad w.p.1, \tag{EC.3}$$

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 (37) guarantee that it suffices to consider the supremum over a bounded interval, so that the supremum is continuous. ∎

Proof of Theorem 2. Observe that

$$\begin{aligned}
\gamma W^*_{\gamma,y} &= \sup_{s\geq 0}\{\gamma\Lambda_{\gamma,y}(\gamma^{-1}s) - s + \gamma\sqrt{b^2\Lambda_{\gamma,y}(\gamma^{-1}s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s))}\} \\
&= \sup_{s\geq 0}\{\Lambda_{f,y}(s) - s + \sqrt{b^2\gamma\Lambda_{f,y}(s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s))}\} \\
&\to \sup_{s\geq 0}\{\Lambda_{f,y}(s) - s\} = W_{f,y} \quad \text{as} \quad \gamma \downarrow 0,
\end{aligned} \tag{EC.4}$$

where $\Lambda_{\gamma,y}(t)$ is defined in (33) and again $W_{f,y}$ is the workload in the periodic deterministic fluid model. To justify (EC.4), we apply Lemma 1 to see that, $b^2\gamma\Lambda_{f,y}(s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s)) \leq b^2\gamma I^{\uparrow}_w[\rho s + \lambda^{\uparrow}c] \leq \gamma(K_1 s + K_2)$ for constants $I^{\uparrow}_w = \sup_t I_w(t)$, K_1 and K_2 and , 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 (EC.4) 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 $\gamma \downarrow 0$. ∎

EC.3. Proofs of Heavy-Traffic Results from §4

Proof of Lemma 3. (c) and (d) are trivial corollaries of the definition of $g_{\gamma,\rho,y}(\cdot)$. For (a) and (b), note that

$$\begin{aligned}
|g_{\gamma,\rho,y}(t) - h(y)t/\rho| &\leq \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |h(s) - h(y)|ds = \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |h'(\xi)(s-y)|ds \\
&\leq \frac{4M}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |s-y|ds = \frac{4M}{b^2c_x^2\gamma\rho^2}\cdot\frac{1}{2}\left(\frac{b^2c_x^2\gamma\rho}{4}t\right)^2 = N\gamma t^2,
\end{aligned} \tag{EC.5}$$

where $N \equiv Mb^2c_x^2/8$. Note that the second line requires $h(\cdot)$ to be differentiable. (b) follows directly from (EC.5). To prove (a), we note that $|g_{\gamma,\rho,y}(t) - h(y)t| \le |g_{\gamma,\rho,y}(t) - h(y)t/\rho| + |h(y)t|(1-\rho^{-1})$. ∎

Proof of Lemma 4. We write

$$W^*_{\gamma,\rho,y} = \sup_{s\ge 0}\left\{(\rho s - s + bc_x\sqrt{\rho s}) + (\Lambda_{\gamma,y,\rho}(s) - \rho s) + bc_x\left(\sqrt{\Lambda_{\gamma,y,\rho}(s)\frac{I_w\left(\Lambda_{\gamma,\rho,y}(s)\right)}{c_x^2}} - \sqrt{\rho s}\right)\right\}.$$

Together with (59) and (60), the change of variable $s = b^2c_x^2\rho t/4(1-\rho)^2$ yields the desired expression. ∎

Proof of Theorem 4. First, for any small $\varepsilon > 0$, there exist $\delta > 0$ such that

$$\rho g_{\gamma,\rho,y}(t) + 2\left(\sqrt{(t + (1-\rho)g_{\gamma,\rho,y}(t))\, C_{\gamma,\rho,y}(t)} - \sqrt{t}\right) < \varepsilon$$

for all $t < \delta$ and $\rho > \delta$. Recall that $f(t)$ attains its maximum at $t = 1$, it suffices to consider the maximization over interval $t \in [\delta, \infty)$ instead. Since $\lim_{\rho\uparrow 1} C_{\gamma,\rho,y}(t) = 1$ uniformly for all t bounded away from 0, $g_{\gamma,\rho,y}(t)$ and $C_{\gamma,\rho,y}(t)$ are bounded, we have

$$\lim_{\rho\uparrow 1}\sqrt{(t + (1-\rho)g_{\gamma,\rho,y}(t))\, C_{\gamma,\rho,y}(t)} - \sqrt{t} = 0$$

uniformly over $t \in [\delta, \infty)$.

Apply Lemma 4, and note that

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

for any function $f(x)$ and $g(x)$, we have

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

Now, we need only consider a bounded interval of t, becasuse $g_{\gamma,\rho,y}(\cdot)$ is uniformly bounded by definition (60) 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. ∎

Proof of Theorem 5. From Lemma 4, we have

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

Now, let $F_{\gamma,\rho,y}(t)\equiv f(t)+\rho g_{\gamma,\rho,y}(t)+2\left(\sqrt{(t+(1-\rho)g_{\gamma,\rho,y}(t))\,C_{\gamma,\rho,y}(t)}-\sqrt{t}\right)$. For the same reason as discussed in the proof of Theorem 4, we can consider only the t's bounded away from 0. Furthermore, since $F_{\gamma,\rho,y}(\cdot)$ 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

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

where we used the concavity of the square root function. The result then follows from Lemma 3 and the fact that $\lim_{\rho\uparrow 1}C_{\gamma,\rho,y}(t)=1$ uniformly for $t\in[\delta,\infty]$ for any positive δ.

To see that this limit coincides with PSA, note that by (64), 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. ∎

Proof of Theorem 6. Note from (56) that

$$\begin{aligned}W^*_{\gamma,\rho,y}&=\sup_{s\ge 0}\left\{\Lambda_{\gamma,\rho,y}(s)-s+b\sqrt{\Lambda_{\gamma,\rho,y}(s)I_w(\Lambda_{\gamma,\rho,y}(s))}\right\}\\&=\sup_{s\ge 0}\left\{-(1-\rho)\,s+\frac{1}{\gamma(1-\rho)}\int_{y-c^{-1}_{\gamma,\rho}s}^{y}h(u)du+b\sqrt{\Lambda_{\gamma,\rho,y}(s)I_w(\Lambda_{\gamma,\rho,y}(s))}\right\}\\&=\frac{1}{\gamma(1-\rho)}\cdot\sup_{t\ge 0}\left\{-t+\int_{y-t}^{y}h(u)du+\gamma(1-\rho)bc_x\sqrt{\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t)I_w(\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t))}\right\},\end{aligned}\tag{EC.6}$$

where we applied a change of variable $c_{\gamma,\rho}t=s$ in the third line. The result follows from the fact that $I_w(t)$ is bounded and 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. ∎

Proof of Theorem 7. By (67), we have

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

as $\gamma \downarrow 0$ for fixed t, where $M = c\left(b^2c_x^2\rho\right)^p/(4^p(p+1))$. Applying Theorem 4 yields

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

where the t^{p+1} is removed from the little-o expression by noting that it suffices to consider a bounded interval of t from the proof of Theorem 4. The supremum is then achieved at

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

with maximum value

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

as $\gamma\downarrow 0$. ∎

EC.4. A Parametric PRQ Based on the Heavy-Traffic Diffusion Approximation

We now provide the details supporting Remark 3. In particular, we now show that a simplified version of the PRQ can be derived from the heavy-traffic FCLT in Theorem 3. With the heavy-traffic scaling in (51) and the heavy-traffic limit in (52), we have the diffusion approximation of the original stochastic process. In particular, we can start with the following heavy-traffic approximation of the net-input process of the original queueing model

$$(1-\rho)^{-1}\hat{X}_{\gamma,\rho}((1-\rho)^2t) \approx (1-\rho)^{-1}\left(c_xB((1-\rho)^2t) + \Lambda_{d,\gamma}((1-\rho)^2t) - (1-\rho)^2t\right)$$
$$\stackrel{\mathrm{d}}{=} c_x\tilde{B}(t) + (1-\rho)^{-1}\Lambda_{d,\gamma}((1-\rho)^2t) - (1-\rho)t$$
$$= c_x\tilde{B}(t) + \Lambda_{\gamma,\rho}(t) - t$$

where $\Lambda_{\gamma,\rho}$ is defined in (44) and $\tilde{B}$ is again a standard Brownian motion. For slightly more generality, we assume that the periodic arrival-rate function start from a position $y \in [0,1)$ within

a cycle, i.e., move the origin to $yc_{\gamma,\rho} = y/\gamma(1-\rho)^2$ in (44). Applying the reverse-time construction of the workload in §2.1, we have

$$\tilde{W}_{\gamma,\rho,y} \approx \sup_{s\geq 0}\left\{c_x\tilde{B}(s) + \Lambda_{\gamma,\rho,y}(s) - s\right\} \tag{EC.7}$$

as an diffusion approximation of the steady-state mean workload in the system parametrized by (γ,ρ,y), where $\Lambda_{\gamma,\rho,y}(s) \equiv \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho}) - \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho} - s)$. By replacing $\tilde{B}$ by its standard deviation, we immediately obtain a robust queueing approximation of (EC.7)

$$\tilde{W}^*_{\gamma,\rho,y} \equiv \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - s + c_x\sqrt{s}\right\}. \tag{EC.8}$$

As in Whitt and You (2016), we call the formulation in (EC.8) a *parametric PRQ* because we quantify the level of variability by the single parameter c_x^2, which is based on the heavy-traffic limit for the net input process. The parametric PRQ in (EC.8) is in contrast to the *functional PRQ* using I_w as in (16) and (21). In §EC.6.2, we compare the diffusion approximation in (EC.7), the parametric PRQ in (EC.8) and the functional PRQ in (21).

EC.5. Heavy-Traffic and Long-Cycle Limits in the $G_t/G_t/1$ model

In this section, we elaborate on Remark 1 by presenting heavy-traffic and long-cycle limits for the periodic $G_t/G_t/1$ model with sketches of the proofs. We follow the framework for variable service rate introduced in Remark 1, the heavy-traffic scaling in §4.1 and the periodic queueing setup in §4.2. In particular, we focus on the the steady-state workload at a fixed location y within a cycle

$$W_{\gamma,\rho,y} = \sup_{s\geq 0}\left\{\sum_{k=1}^{A_{\gamma,\rho,y}(s)} V_k - M_{\gamma,\rho,y}(s)\right\}$$

as in (17), where $A_{\gamma,\rho,y}(s) \equiv N\left(\Lambda_{\gamma,\rho,y}(s)\right)$. The corresponding PRQ problem is

$$W^*_{\gamma,\rho,y} = \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - M_{\gamma,\rho,y}(s) + b\sqrt{\Lambda_{\gamma,\rho,y}(s) I_w\left(\Lambda_{\gamma,\rho,y}(s)\right)}\right\} \tag{EC.9}$$

as in (18). Here, we keep the same reverse-time cumulative arrival-rate function

$$\Lambda_{\gamma,\rho,y}(s) \equiv \Lambda_{\gamma,\rho}(yc_{\gamma,\rho}) - \Lambda_{\gamma,\rho}(yc_{\gamma,\rho} - s)$$

for $\Lambda_{\gamma,\rho}$ in (44) and $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$. Similarly, we define

$$M_{\gamma,\rho,y}(s) \equiv M_{\gamma,\rho}(yc_{\gamma,\rho}) - M_{\gamma,\rho}(yc_{\gamma,\rho} - s)$$

with

$$M_{\gamma,\rho}(t) \equiv t + (1-\rho)^{-1} M_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0 \tag{EC.10}$$

so that the associated service-rate function is

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

where

$$M_{d,\gamma}(t) \equiv \int_0^t \mu_{d,\gamma}(s)\, ds, \quad \mu_{d,\gamma}(t) \equiv r(\gamma t), \quad \text{and} \quad \int_0^1 r(t)\, dt = 0 \tag{EC.11}$$

for a continuous function r with a cycle length of 1.

With the same heavy-traffic scalings as in (49), we generalize Theorem 3 as follows.

THEOREM EC.1. (*heavy-traffic FCLT for the $G_t/GI_t/1$ model*) *For the family of $G_t/GI_t/1$ models indexed by (γ, ρ) with cumulative arrival-rate functions in* (44) *and cumulative service-rate function in (EC.10), if $\hat{N}_n \Rightarrow c_a B_a$ as $n \to \infty$, where B_a is a standard Brownian motion, then*

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

where

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

Ψ is the reflection map in (50)*, and B_a and B_s are two independent standard* (*mean 0 variance 1*) *Brownian motions.*

Proof. By definition, we have

$$\begin{aligned}
\hat{X}_{\gamma,\rho}(t) &= (1-\rho) X_{\gamma,\rho}\left((1-\rho)^{-2} t\right) \\
&= (1-\rho) \sum_{k=1}^{A_{\gamma,\rho}\left((1-\rho)^{-2} t\right)} V_k - (1-\rho) M_{\gamma,\rho}\left((1-\rho)^{-2} t\right) \\
&= (1-\rho) \sum_{k=1}^{A_{\gamma,\rho}\left((1-\rho)^{-2} t\right)} V_k - (1-\rho)^{-1} t - M_{d,\gamma}(t) \\
&\equiv \Xi_{\gamma,\rho}(t) - M_{d,\gamma}(t)\,.
\end{aligned}$$

where $\Xi_{\gamma,\rho}(t)$ denotes the quantity $\hat{X}_{\gamma,\rho}(t)$ exacly as it appears in Theorem 3, so the result follows. ∎

We remark that this generalized FCLT can be viewed as if we replace $\Lambda_{d,\gamma}$ by $\tilde{\Lambda}_{d,\gamma} \equiv \Lambda_{d,\gamma} - M_{d,\gamma}$ in a $G_t/GI/1$ model, or equivalently, replace h by $\tilde{h} \equiv h - r$ for h in (47) and r in (EC.11).

Next, we generalize the limit theorems for the PRQ problem in (EC.9). As preparation, we re-write $M_{\gamma,\rho,y}$ exactly the same as (57)

$$\begin{aligned} M_{\gamma,\rho,y}(s) &\equiv M_{\gamma,\rho}((k+y)c_{\gamma,\rho}) - M_{\gamma,\rho}((k+y)c_{\gamma,\rho} - s) = M_{\gamma,\rho}(yc_{\gamma,\rho}) - M_{\gamma,\rho}(yc_{\gamma,\rho} - s) \\ &= s + (1-\rho)^{-1} \int_{y/\gamma-(1-\rho)^2 s}^{y/\gamma} r(\gamma t)dt = s + \frac{1}{\gamma(1-\rho)} \int_{y-c_{\gamma,\rho}^{-1}s}^{y} r(t)dt \\ &= s + \frac{1}{\gamma(1-\rho)} R_{\gamma,\rho,y}(s), \end{aligned} \tag{EC.12}$$

where $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$ is the cycle length of $M_{\gamma,\rho,y}$ and $R_{\gamma,\rho,y}(s) \equiv \int_{y-c_{\gamma,\rho}^{-1}s}^{y} r(t)dt$. Similar to (60), we define

$$\tilde{g}_{\gamma,\rho,y}(t) \equiv \frac{4}{b^2 c_x^2 \gamma \rho^2} \int_{y-\frac{b^2 c_x^2 \gamma \rho}{4} t}^{y} (h(s) - r(s))ds \tag{EC.13}$$

All generalizations are trivial in the way that we need only replace $g_{\gamma,\rho,y}$ in the original limits by $\tilde{g}_{\gamma,\rho,y}$ here in appropriate places. Equivalently, this can be done by replacing h by $\tilde{h} \equiv h - r$ appropriately as we observed in the generalized FCLT. We demonstrate this idea by proving a generalized version of Lemma 4.

LEMMA EC.1. *With f, $g_{\gamma,\rho,y}$ and $\tilde{g}_{\gamma,\rho,y}$ defined in* (59), (60) *and* (EC.13), *we have*

$$W^*_{\gamma,\rho,y} = \frac{b^2}{2} \cdot \frac{\rho c_x^2}{2(1-\rho)} \cdot \sup_{t\geq 0} \left\{ f(t) + \rho \tilde{g}_{\gamma,\rho,y}(t) + 2\left(\sqrt{(t + (1-\rho) g_{\gamma,\rho,y}(t))\, C_{\gamma,\rho,y}(t)} - \sqrt{t} \right) \right\}, \tag{EC.14}$$

where

$$C_{\gamma,\rho,y}(t) \equiv \frac{1}{c_x^2} \cdot I_w \left(\frac{b^2 c_x^2 \rho^2}{4(1-\rho)^2} \left(t + (1-\rho) g_{\gamma,\rho,y}(t) \right) \right).$$

Proof. From (EC.9), we write

$$\begin{aligned} W^*_{\gamma,\rho,y} = \sup_{s\geq 0} \{ &(\rho s - s + bc_x\sqrt{\rho s}) + ((\Lambda_{\gamma,y,\rho}(s) - M_{\gamma,y,\rho}(s) + s) - \rho s) \\ &+ bc_x \left(\sqrt{\Lambda_{\gamma,y,\rho}(s) I_w \left(\Lambda_{\gamma,\rho,y}(s) \right) / c_x^2} - \sqrt{\rho s} \right) \Big\}. \end{aligned}$$

Together with (59), (60) and (EC.13), the change of variable $s = b^2 c_x^2 \rho t/4(1-\rho)^2$ yields the desired expression. ∎

Hence, we immediately obtain

THEOREM EC.2. (*heavy traffic limit for PRQ*) *The heavy traffic limit of the PRQ problem in* (EC.9) *for the* $G_t/G_t/1$ *model is*

$$\lim_{\rho\uparrow 1} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0} \{f(t) + \tilde{g}_{\gamma,1,y}(t)\}. \tag{EC.15}$$

Before presenting the long-cycle heavy-traffic limits, we need to adjust the concept of underloaded, critically loaded and overloaded queues. In the case of a $G_t/G_t/1$ queue, the instantaneous traffic intensity becomes

$$\tilde{\rho}(y) = \frac{\rho + (1-\rho)h(y)}{1 + (1-\rho)r(y)} \tag{EC.16}$$

We now distinguish the three cases by the value of $\tilde{\rho}^{\uparrow} \equiv \sup_y\{\tilde{\rho}(y)\}$. So $\tilde{\rho}^{\uparrow} < 1$, $\tilde{\rho}^{\uparrow} = 1$ and $\tilde{\rho}^{\uparrow} > 1$ corresponds to the underloaded, critically loaded and overloaded case, separately. Equivalently, we can also use $\tilde{h}^{\uparrow}$ as the criteria, where $\tilde{h} = h - r$. Using $\tilde{h}^{\uparrow}$ is preferred because (i) it is more consistent with the notation in §4.3; (ii) it is consistent with our observation of replacing h by $\tilde{h}$ when generalizing to the case of $G_t/G_t/1$ models, as we discussed above.

The rest of the generalizations share the similar idea, and only minor adjustments are needed for the proofs. We list them below.

THEOREM EC.3. (*long-cycle heavy-traffic limit for PRQ in an underloaded queue*) *Assume that* h *is continuously differentiable with* $\tilde{h}^{\uparrow} < 1$, *then the PRQ problem in* (EC.9) *for the* $G_t/G_t/1$ *model admits the double limit*

$$\lim_{\substack{\gamma\downarrow 0\\ \rho\uparrow 1}} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \frac{1}{1-\tilde{h}(y)}, \tag{EC.17}$$

so that PRQ is asymptotically consistent with PSA, i.e.,

$$W^*_y = \frac{b^2}{2} \cdot \frac{\tilde{\rho}(y) c_x^2}{2(1-\tilde{\rho}(y))} + o(1-\rho). \tag{EC.18}$$

where $\tilde{\rho}(y)$ *is the instantaneous traffic intensity in (EC.16)*

THEOREM EC.4. (*long-cycle limit for PRQ in an overloaded queue*) *The PRQ problem in* (EC.9) *for the $G_t/G_t/1$ model with the heavy-traffic scaling in* (44) *and $\tilde{h}^{\uparrow} > 1$ admits the long-cycle limit*

$$(1-\rho)\lim_{\gamma\downarrow 0}\gamma\cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0}\left\{-t+\int_{y-t}^{y}\tilde{h}(s)ds\right\}, \quad 0\leq\rho<1. \tag{EC.19}$$

THEOREM EC.5. (*long-cycle heavy-traffic limit for PRQ in a critically loaded queue*) *Assume that $\tilde{h}(t)$ satisfies*

$$\tilde{h}(t) = 1 - ct^p + o(t^p), \ \text{as } t\to 0, \tag{EC.20}$$

for some positive real numbers c and p. Then the long-cycle heavy-traffic limit of the PRQ solution for the $G_t/G_t/1$ model at the critical point $y=0$ is in the order of $O(\gamma^{-p/(2p+1)}(1-\rho)^{-1})$ as $(\rho,\gamma)\to(1,0)$.

EC.6. Additional Simulation Examples

EC.6.1. 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

$$\lambda_\gamma(t) = \rho + \beta\sin(2\pi\gamma t), \tag{EC.21}$$

which is a special case of (32). 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 (22).

Figure EC.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 EC.1 shows that both the simulation results and the solutions to PRQ problem (21) converge to the fluid limit calculated from Proposition 1, confirming Theorem 2.

The estimation in Figure EC.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 EC.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.

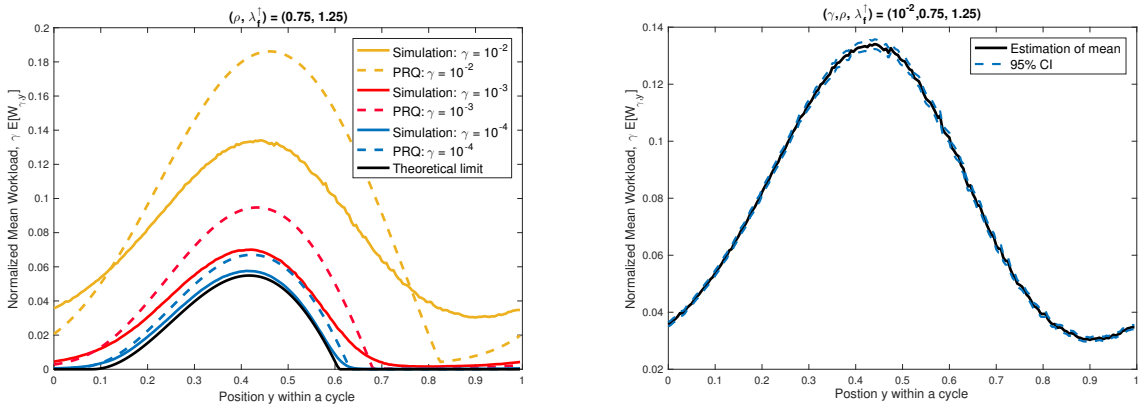

Figure EC.1 A comparison of the solution to the PRQ problem in (56) as a function of the position y within a cycle to simulation estimates of the normalized mean workload $\gamma E[W_{\gamma,y}]$ in (41) and the theoretical limit in Theorem 1 in the $M_t/H_2/1$ model with arrival-rate function in (EC.21) 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 $\gamma = 10^{-2}$.

EC.6.2. Comparing the diffusion approximation and the PRQ

In Remark 3, we compared the diffusion approximation (EC.7) derived from the heavy-traffic FCLT with the parametric and functional PRQ in (EC.8) and (56), separately. In this section, we conduct simulation study on various stochastic models to demonstrate the performance of all three methods and discuss their differences.

From the discussion in §4.2 and §5, we know that the (functional) PRQ performs well under all traffic intensities, especially for system with long arrival cycles, as in Figure EC.2 (left). As a simplified version, the parametric PRQ utilizes only a single parameter to capture the dependence structure of the original system. Under heavy-traffic, the parametric PRQ is asymptotically the same as the functional version, since $I_w(\infty) = c_x^2$ under mild conditions. But for moderate to low traffic intensities, we do not expect the parametric PRQ to outperform the functional one, since the former one uses partial information to capture the transient dependence structure. We shall use the functional PRQ over its parametric version whenever possible.

On the other hand, the diffusion approximation will have excellent performance under heavy traffic by design. In fact, it works very well for systems with simple input such as renewal processes

even for moderate traffic intensities, where the IDW converges to the asymptotic variability parameter c_x^2 faster. This is demonstrated in Figure EC.2 (right), where the diffusion approximation outperforms both parametric and functional PRQ.

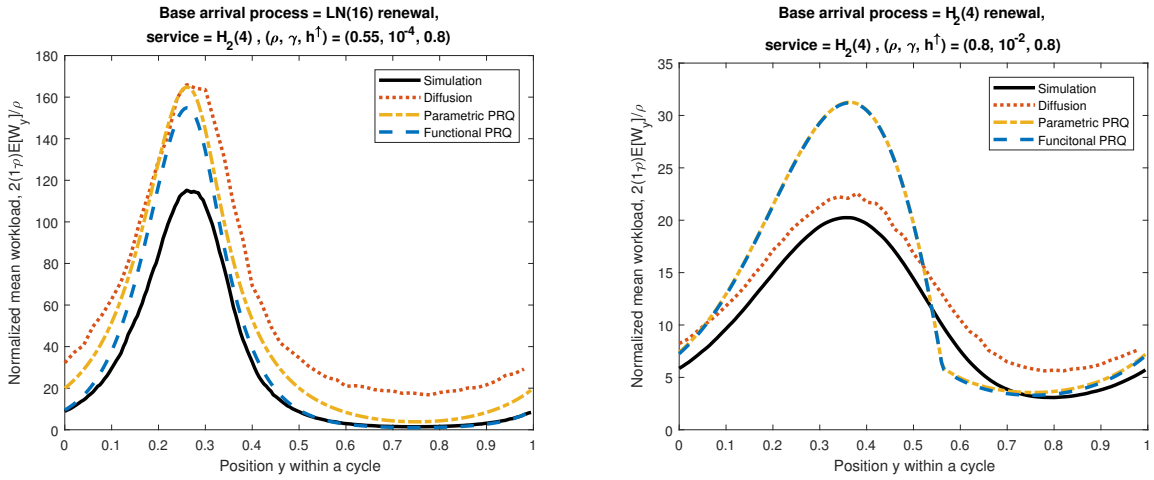

Figure EC.2 A comparison of the diffusion approximation in (EC.7), the solution to the parametric PRQ problem in (EC.8) and the functional PRQ in (56), as functions of the position y within a cycle to simulation estimates of the normalized mean workload $2(1-\rho)E[W_y]/\rho$ for three models.

However, the diffusion approximation may break down in cases with more complex arrival processes such as superposition arrivals. The reason is that it lacks the flexibility to distinguish systems with the same c_x^2 but different transient variability, or the cases where the workload depends heavily on the transient variability. In contrast, the functional PRQ has a more robust performance in complex situations.

In Figure EC.3, we present such a case where PRQ is more helpful. Consider a $(\sum_i G_i)_t/GI/1$ queue with the time-varying superposition arrival process of 10 i.i.d. $LN(16)$ renewal process and $H_2(4)$ service distribution. We consider a underloaded queue with moderate traffic intensity of $\rho = 0.6$ and a cycle length parameter $\gamma = 0.01$. For the superposition of n i.i.d. point processes, the IDC turned out to be a simple time scaling version of the IDC of a single stream, i.e., $I_{a,n}(t) = I_a(t/n)$. Recall that $I_a(0) = 1$, the superposition process acts like a Poisson process in the short run, whose variability is very low. But in the long run, the variability converges to that of a single stream. In

the case here, the asymptotic variability of the superposition arrival process is $c_a^2 = 16 \gg 1$. Figure EC.3 (left) shows the IDC for the superposition arrival process as well as the $LN(16)$ renewal process.

In this case, both the parametric PRQ and the diffusion approximation uses $c_x^2 = c_a^2 + c_s^2 = 20$ to characterize such a complex arrival process. Figure EC.3 (right) shows that both of them overestimate the congestion in the system, while the functional PRQ still provides a reasonable prediction of the mean workload.

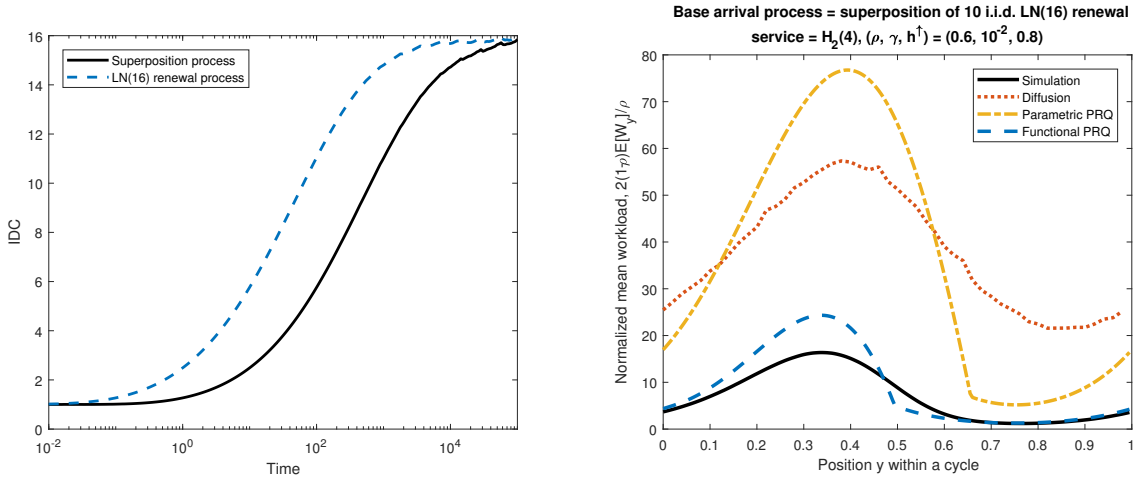

Figure EC.3 A comparison of the diffusion approximation in (EC.7), the solution to the parametric PRQ problem in (EC.8) and the functional PRQ in (56), as functions of the position y within a cycle to simulation estimates of the normalized mean workload $2(1-\rho)E[W_y]/\rho$ for three models. On the left is shown the IDC of the superposition process and a individual stream.

EC.6.3. TVRQ Approximation of the Quantiles

Looking back at the intuition behind the uncertainty set defined in (15), one can easily obtain approximations of the quantile process of the workload by defining the uncertainty set as if we replace the net input process by its quantile process. Assuming a Normal approximation of the net input process, we may use the TVRQ solution with $b = \Phi(p)$ as approximation of the p-th quantile of the workload, where Φ is the Normal cumulative distribution function.

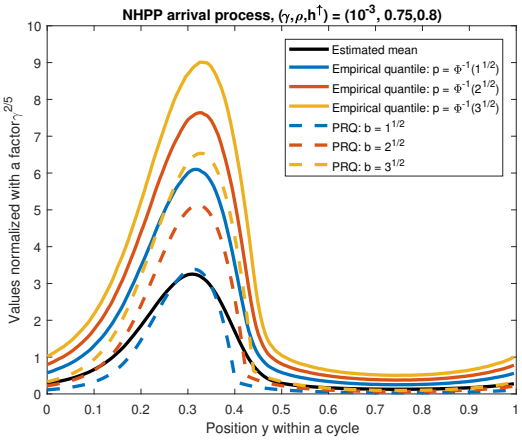

Figure EC.4 A comparison of the PRQ solutions for different values of the parameter b and the corresponding empirical quantile of the simulated values for the critically loaded $M_t/LN(1)/1$ model.

Figure EC.4 shows the impact of the parameter b on the PRQ solution, together with the corresponding quantiles of the simulated values for the critically loaded $M_t/LN(1)/1$ model. The figure shows that the shape of the PRQ solution matches the shape of the quantile process very well, and serves as an reasonable approximation.

EC.6.4. Time-Varying Service Rate

In Remark 1, we introduced the generalization of our TVRQ algorithm for the $G_t/G_t/1$ general model, where the service is delivered at a time-varying rate $\mu(t)$ at time t. In particular, we obtained the TVRQ problem in (18). In this section, we present a simulation experiment on a specific $G_t/G_t/1$ model, where we have the service rate function

$$\mu(t) = 1 + (1-\rho)\alpha \sin(6\pi\gamma(1-\rho)^2 t). \tag{EC.22}$$

We keep the same arrival rate function as in (68), so the cycle length of the arrival rate function is 3 times that of the service rate function. For the base arrival process N, we use the interarrival times with $H_2(4)$ distribution, while the service requiremnts are i.i.d. $LN(2)$ random variables, so we have the $(H_2(4)_t/(LN(2))_t/1$ model with TV arrival rate and service rate as specified above.

Figure EC.5 shows simulation comparisons for different set of parameters. The instantaneous traffic intensity $\rho(t) \equiv \lambda(t)/\mu(t)$ is also displayed in Figure EC.5 (left). Figure EC.5 (left) shows

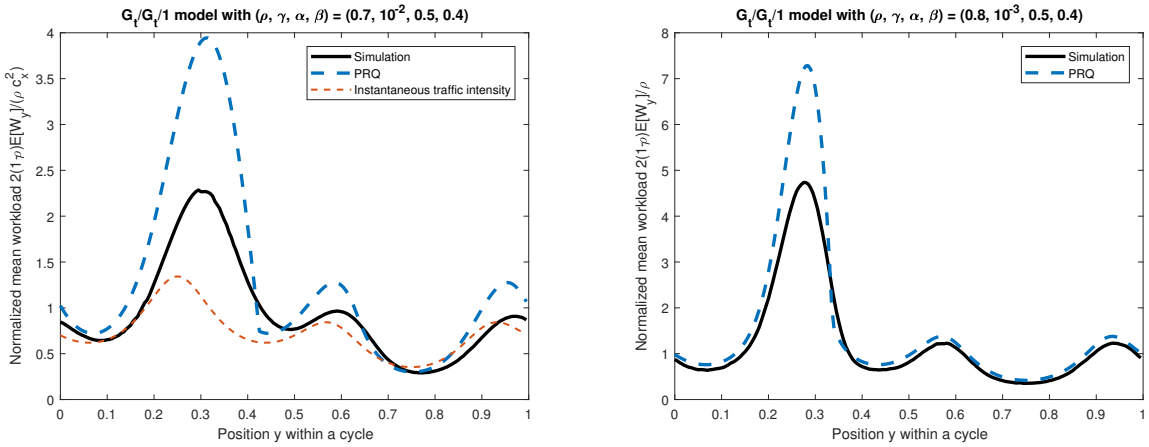

Figure EC.5 A comparison of the PRQ solutions for the $(H_2(4)_t/(LN(2))_t/1$ model and the simulated mean workload for different set of parameters.

that (i) the PRQ approximation performs very well even for moderate traffic intensity $\rho = 0.7$ and cycle length parameter $\gamma = 10^{-2}$, which gives a practical cycle length of $1/(\gamma(1-\rho)^2) \approx 10^3$, and (ii) PRQ accurately predicted the timing of three peak congestion, which are delayed in compare to the peak traffic intensity. Figure EC.5 (right) shows that the accuracy improves when we consider a slightly more congested system with longer cycle length.