e-companion

EC.1. Overview

This is an online e-companion to the main paper. It has five more sections, roughly in order of their appearance in the main paper. First, in §EC.2 we describe the simulation methodology, which is applied throughout the main paper, starting in §4. Second, in §EC.3 we present an alternative framework for the function Π in (22) and (28) based on the $M/M/1$ queue instead of heavy-traffic. Next, in §EC.4 we elaborate on the deterministic model in §5.1 and the long-cycle limit in §5.2. In §EC.5 we provide the proofs for §6. We establish a long-cycle heavy-traffic limit for the critically loaded case in §EC.6. We provide heavy-traffic theory for the $G_t/G_t/1$ model in §EC.7. Finally, in §EC.8 we present additional simulation examples that provide further insight into PRQ.

EC.2. Simulation Methodology

The simulations were conducted with C++ on a personal computer. Each simulation run was for 10^8 time units, but the first 10^7 time units were discarded to allow the system to approach steady-state. Since we have unit-rate service times, this amounts to $10^8\rho$ customers, where ρ is the average arrival rate (and traffic intensity). We then divide the cycle into 100 segments. For each segment, we collect the time average of the workload in that segment for each cycle, so that we have a sample size of $10^6\gamma$ for each segment. The mean of the workload at the start of each segment is then estimated by the sample average while the quantiles are estimated by the sample quantiles. That is, we used one long run instead of independent replications; see Whitt (1991a). The run lengths are long enough to make tight confidence intervals, see Figure EC.2 in §EC.8.

We apply the efficient simulation algorithm proposed in Ma and Whitt (2015). In particular, the inversion table for both the cumulative arrival-rate function and the cumulative service-rate function are generated using Algorithm 1 there. For the time-varying external arrival process, a base renewal arrival process is generated and then converted to its time-varying version by using the inversion table, see Algorithm 2 there. For time-varying service process, we first generate the

(stationary) base service time and record the time that the customer enters service. The base service time is then converted to the service time under the time-varying service rate by using the service-rate inversion table starting from the time that the customer enters service. In the special case of periodic service-rate function, only one inversion table is needed, regardless of the starting time of the service.

A rough estimate of the required run time is 6 minutes to conduct the simulation estimates for one case, while the PRQ calculations are relatively negligible. For example, since the upper left plot in Figure 1 displays 5 percentiles, there are 5 cases, so that it would take about 30 minutes to create that plot. For display, the output is exported to MATLAB.

EC.3. An Alternative $M/M/1$ View of the Function Π

In order to consider possible refinements for the function Π in (28), we now consider a concrete queueing model instead of the HT limit. For the UL stationary $GI/GI/1$ queue, there is an atom at the origin with probability $1-\rho$. In particular, for the $M/M/1$ queue,

$$P(W \le x) = 1 - \rho e^{-\rho x/m}, \quad x > 0, \quad \text{for} \quad m \equiv \rho/(1-\rho). \tag{EC.1}$$

Hence, the p quantile is

$$W(p) = -(m/\rho)\ln((1-p)/\rho). \tag{EC.2}$$

If we apply Theorem 2 and Corollary 3 of Whitt and You (2018b) and apply (27) above, then we can equate (27) and (EC.2) to get the more complex formula for Π:

$$\Pi(b) \equiv \Pi(b,\rho) = 1 - \rho e^{-\rho b^2/2}. \tag{EC.3}$$

Note that Π here is not a surjective mapping. For any $p < 1-\rho$, there is no pre-image b. However, the atom at the origin of the workload W has a probability of $1-\rho$, so that $W(p) = 0$ for any $p < 1-\rho$. In this case, we can set $b = 0$ as the pre-image of any $p < 1-\rho$, so that RQ algorithm gives an approximation of 0 for the quantile, which is exact. Consistent with intuition, formula $\Pi(b,\rho)$ in (EC.3) coincides with (28) when $\rho = 1$. The derivative of $\Pi(b,\rho)$ with respect to ρ is $[(\rho b^2/2) - 1]e^{-\rho b^2/2}$.

EC.4. Supporting Theory for the Periodic Deterministic Model in 5.1

We now elaborate on the periodic deterministic model introduced in §5.1. In particular, we assume that $X(t) = \Lambda(t) - t$, where the arrival rate function λ is periodic and the service rate is constant, so that the TVRQ coincides with the exact workload

$$W_t^* = W_t \equiv W_{det,t} \equiv \sup_{s \geq 0} \{\Lambda_t(s) - s\}, \tag{EC.4}$$

as shown in (30).

EC.4.1. Supremum Over Only One cycle

We start with the arrival-rate function $\lambda(t)$ with period c. In order for the model to be interesting (i.e., for there to be positive workload at some time), we also assume that

$$\lambda^{\uparrow} \equiv \sup_{0 \leq s < c} \{\lambda(s)\} > 1. \tag{EC.5}$$

Now the main quantity we focus on is

$$\Lambda_t(s) \equiv \Lambda(t) - \Lambda(t-s), \quad s \geq 0, \quad 0 \leq t < c. \tag{EC.6}$$

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

PROPOSITION EC.1. *For the deterministic model, the workload at time t within the cycle $[0, c)$ defined in* (EC.4) *reduces to the supremum over one cycle, i.e.,*

$$W_t = \sup_{0 \leq u \leq c} \{\Lambda_t(u) - u\}, \quad 0 \leq t < c. \tag{EC.7}$$

Proof. Let $s = kc + t$, $0 \leq t < c$ and $k \geq 0$. Then

$$\begin{aligned}
W_t &= \sup_{0 \leq s \leq \infty} \{\Lambda_t(s) - s\}, \quad 0 \leq t < c, \\
&= \sup_{0 \leq u \leq c,\, k \geq 0} \{\Lambda_t(kc+u) - (kc+u)\}, \quad 0 \leq t < c, \\
&= \sup_{0 \leq u \leq c,\, k \geq 0} \{(\Lambda_t(kc+u) - \Lambda_t(u) - kc) + (\Lambda_y(u) - u)\}, \quad 0 \leq t < c, \\
&= \sup_{0 \leq u \leq c,\, k \geq 0} \{-(1-\rho)kc + (\Lambda_t(u) - u)\}, \quad 0 \leq t < c, \\
&= \sup_{0 \leq u \leq c} \{\Lambda_t(u) - u\}, \quad 0 \leq t < c,
\end{aligned} \tag{EC.8}$$

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

EC.4.2. Common Special Cases

We now consider a common structural property that holds in many special cases. In many cases, 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

$$\lambda(t) - 1 \geq 0, \quad 0 \leq t < \delta, \quad \text{and} \quad \lambda(t) - 1 \leq 0, \quad \delta \leq t < c. \tag{EC.9}$$

Often we may require a time shift to satisfy condition (EC.9). In this setting it is easy to determine the periodic fluid W_t, $0 \leq t \leq c$.

Proposition EC.2. *If conditions* (EC.5) *and* (EC.9) *hold, then there exists one and only one* δ^* *with* $0 < \delta < \delta^* < c$ *such that* $\Lambda(\delta^*) = \delta^*$. *Moreover,* $\Lambda(t) - t$ *is nondecreasing over* $[0, \delta]$ *and nonincreasing over* $[\delta, c]$, *so that*

$$W_t = \Lambda(t) - t, \quad 0 \leq y \leq \delta^*, \quad \textit{and} \quad W_t = 0, \quad \delta^* \leq t \leq c, \tag{EC.10}$$

and

$$W^{\uparrow} \equiv \sup_{0 \leq t \leq c} \{W_t\} = W_\delta = \Lambda(\delta) - \delta > 0. \tag{EC.11}$$

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

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

$$W_t = (a-1)yt, \quad 0 \leq t \leq \delta, \quad W^{\uparrow} = W_\delta = (a-1)\delta, \tag{EC.12}$$

and

$$W_t = (a-1)\delta - (1-b)(t-\delta), \quad \delta \leq t \leq \delta^* \equiv (a-b)\delta/(1-b) \quad \textit{and} \quad W_t = 0, \quad \delta^* \leq t \leq c. \tag{EC.13}$$

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 EC.2. (*sinusoidal case*) *If, in addition to the conditions of Proposition EC.2,* $\lambda(t) = \rho + \beta \sin(2\pi\gamma t)$*, so that a cycle has period* $c(\gamma) \equiv 1/\gamma$ *and the peak is at* $c(\gamma)/4$ *, and* $t_0 = \arcsin((1-\rho)/\beta)/2\pi\gamma = c(\gamma)\arcsin((1-\rho)/\beta)/2\pi$*, then* $\lambda(t_0 + t))$ *satisfies condition* (EC.9) *and* $\delta = c/2 - 2t_0$*, so that in terms of the original* Λ

$$W^{\uparrow} = W_{c/2-t_0} = \Lambda(c/2 - t_0) - \Lambda(t_0) - (c/2) + 2t_0, \tag{EC.14}$$

so that the time lag in the peak is $(c/4) - t_0 = c(0.25 - \arcsin((1-\rho)/\beta)/2\pi)$*. As* $\rho \uparrow 1$*,* $t_0 \equiv t_0(\rho) \downarrow 0$*,* $\delta(\rho) \uparrow 0.5$ *and* $W^{\uparrow} \to \Lambda(c/2) - 0.5$*.*

Finally, to treat general non-sinusoidal examples it may be useful to consider Taylor series expansions of the arrival rate function in order to obtain simple approximation formulas, as in Remark 10 and §3 of Eick et al. (1993), which focuses on the infinite-server model. If we consider arrival-rate functions with a single peak, then that leads to a quadratic approximation in the neighborhood of the peak.

Thus, we next consider a quadratic function, defined so that the condition in (EC.9) holds. Thus, let

$$\lambda(t) \equiv [a - b(t-p)^2]^+, \quad t \geq 0 \quad \text{for} \quad a > 1 \quad \text{and} \quad b > 0, \tag{EC.15}$$

where $[x]^+ \equiv \max\{x, 0\}$, $a > 1$ because OL and $b > 0$ because the peak is at

$$p \equiv \sqrt{(a-1)/b}. \tag{EC.16}$$

We let the arrival-rate function be periodic with cycle length $c > 2p$, chosen so that the average arrival rate is strictly less than 1.

We have chosen p in (EC.15) and (EC.16) so that the net input rate is initially $\lambda(0) - 1 = 0$, but $\lambda(t) - 1 > 0$ for suitably small t with $t > 0$. This value of p is obtained by solving the equation

$$\lambda(t) - 1 = 0 \quad \text{or} \quad a - 1 = b(t-p)^2, \tag{EC.17}$$

which has solution $t = p$ in (EC.16). Clearly, λ is positive over the interval $(0, 2p)$, symmetric about p with $\lambda(0) = \lambda(2p) = 0$. Thus, we can apply Proposition EC.2 to this quadratic example in (EC.15) for $\delta = 2p$.

COROLLARY EC.3. (*quadratic case*) *If $\lambda(t)$ is a quadratic function as defined in* (EC.15) *with cycle length $c > 2p$ and average arrival rate strictly less than* 1*, then the condition in* (EC.9) *holds for $\delta = 2p$, so that the time lag in the peak is p in* (EC.16) *and*

$$\begin{aligned} W^{\uparrow} &\equiv \sup_{0\le t\le c}\{W_t\} = W_\delta = W_{2p} = \Lambda(2p) - 2p \\ &= 2(\Lambda(p) - p) = 2p\left(\frac{2a+1}{3}\right). \end{aligned} \tag{EC.18}$$

To illustrate how the Taylor series approximatin would work, we consider the sinusoidal example in (24). In the setting of (24), we can move the peak to the origin by replacing sine by cosine. Then using the asymptotic expansion $\cos x = 1 - x^2/2 + O(x^4)$ as $x \downarrow 0$, we get that $a = \rho + \beta$ and $b = 2\beta\pi^2\gamma^2$ in (EC.15). Thus the approximate time lag in the peak of W_0^* in (30) and (EC.4) is

$$\text{time lag} \approx \sqrt{(\rho+\beta-1)/2\beta\pi^2\gamma^2} = \left(\frac{1}{\gamma}\right)\left(\frac{1}{2\pi}\right)\sqrt{2(\rho+\beta-1)/\beta}. \tag{EC.19}$$

The final expression separates out the cycle length γ^{-1} and expresses the time lag relative to 2π, so that $\sqrt{2(\rho+\beta-1)/\beta} = 2\pi/4 \approx 1.56$ means that the time lag would be one quarter of a sine cycle; i.e., $y^* - 0.25 = \sqrt{2(\rho+\beta-1)/2\pi\beta}$.

For example, in the setting of Figure 4, where $\rho = 0.7$ and $\beta = 0.5$, the approximate time lag from the quadratic approximation (EC.19) above is $y^* - 0.25 = 0.4/\pi \approx 0.127$, which indicates the peak congestion should be at about 0.377. This is somewhat smaller than the exact time lag of 0.1475 we obtain from applying Corollary EC.2.

EC.4.3. The Long-Cycle Fluid Limit in §5.2

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

$$\lambda_\gamma(t) \equiv \lambda(\gamma t), \quad t \ge 0, \tag{EC.20}$$

for the base arrival-rate function λ satisfying (EC.5). 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. In particular, we let

$$\Lambda_\gamma(t) \equiv \gamma^{-1}\Lambda(\gamma t) \quad \text{and} \quad \Lambda_{\gamma,y}(t) \equiv \Lambda_\gamma(\gamma^{-1}y) - \Lambda_\gamma(\gamma^{-1}y - t), \quad 0 \le y < c, \tag{EC.21}$$

so that the associated arrival-rate function is as in (EC.20). The periodic structure with (EC.5) implies the following bound.

LEMMA EC.1. *In the setting above with* (EC.5),

$$\max\{\Lambda(t), \Lambda_y(t)\} \le \rho t + \lambda^\uparrow c \quad \textit{and} \quad \max\{\Lambda_\gamma(t), \Lambda_{\gamma,y}(t)\} \le \rho t + \lambda^\uparrow c/\gamma \quad \textit{for all} \quad t \ge 0. \tag{EC.22}$$

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

$$A_\gamma(t) \equiv N(\Lambda_\gamma(t)) \quad \text{and} \quad X_\gamma(t) \equiv \sum_{k=1}^{A_\gamma(t)} V_k - t, \quad t \ge 0, \tag{EC.23}$$

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

$$t^{-1}N(t) \to 1 \quad \text{as} \quad t \to \infty \quad \text{w.p.1} \tag{EC.24}$$

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

$$|t^{-1}N(t) - 1| < \epsilon \quad \text{for all} \quad t \ge t_0 \quad \text{w.p.1.} \tag{EC.25}$$

Condition (EC.24) 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 (EC.25) is implied by refinements such as the law of the iterated logarithm. Condition (EC.25), together with Lemma EC.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 X_γ after introducing extra time and space scaling.

LEMMA EC.2. *For the periodic $G_t/GI/1$ model under condition* (EC.24),

$$\gamma A_\gamma(\gamma^{-1}(t)) \to \Lambda(t) \quad \textit{and} \quad \gamma X_\gamma(\gamma^{-1}(t)) \to \Lambda(t) - t \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1.} \tag{EC.26}$$

Proof. Observe that

$$\begin{aligned} \gamma A_\gamma(\gamma^{-1}t) &= \gamma N(\Lambda_\gamma(\gamma^{-1}t)) = \gamma N(\gamma^{-1}\Lambda(\gamma(\gamma^{-1}t))) \\ &= \gamma N(\gamma^{-1}\Lambda(t)) \to \Lambda(t) \quad \text{as} \quad \gamma \downarrow 0 \quad \text{w.p.1} \end{aligned} \tag{EC.27}$$

because $\gamma N(\gamma^{-1}t) \to t$ uniformly over bounded intervals w.p.1 by the FSLLN in (EC.24). A further application of the composition mapping yields the corresponding limit for X_γ in (EC.23):

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

because

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

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

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

$$W_{\gamma,y} = \sup_{s \ge 0} \{X_{\gamma,y}(s)\}, \tag{EC.28}$$

where

$$X_{\gamma,y}(t) \equiv X_\gamma(y\gamma^{-1}) - X_\gamma(y\gamma^{-1} - t), \quad t \ge 0, \quad 0 \le y < c, \tag{EC.29}$$

for X_γ in (EC.23). We get a fluid limit for $W_{\gamma,y}$, again after scaling.

THEOREM EC.1. (*long-cycle fluid limit*) *For the periodic $G_t/GI/1$ model under conditions* (EC.24) *and* (EC.25),

$$\gamma W_{\gamma,y} \to W_y \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1}, \tag{EC.30}$$

where W_y is the deterministic workload at time y within a cycle of length c.

Proof. From (EC.28) and (EC.29),

$$\gamma W_{\gamma,y} = \sup_{s\ge 0}\left\{\gamma Y_{\gamma,y}(\gamma^{-1}s) - s\right\} \to \sup_{s\ge 0}\left\{\Lambda_y(s) - s\right\} = W_y \quad \text{as} \quad \gamma\downarrow 0 \quad w.p.1, \tag{EC.31}$$

where W_y is the periodic workload in the limiting periodic model by virtue of Lemma EC.2 and a further continuity argument. Lemma EC.2 and condition (EC.25) guarantee that it suffices to consider the supremum over a bounded interval, so that the supremum is continuous. ∎

Let $W^*_{\gamma,y}$ be the PRQ workload at time y/γ for $0\le y<c$.

THEOREM EC.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.,*

$$\gamma W^*_{\gamma,y} \to W_y \quad as \quad \gamma\downarrow 0, \tag{EC.32}$$

where W_y is the deterministic workload at time y within a cycle of length c, so that

$$|\gamma W^*_{\gamma,y} - \gamma W_{\gamma,y}| \to 0 \quad as \quad \gamma\downarrow 0 \quad w.p.1. \tag{EC.33}$$

Proof of Theorem EC.2. Observe that

$$\begin{aligned}
\gamma W^*_{\gamma,y} &= \sup_{s\ge 0}\left\{\gamma\Lambda_{\gamma,y}(\gamma^{-1}s) - s + \gamma\sqrt{b^2\Lambda_{\gamma,y}(\gamma^{-1}s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s))}\right\} \\
&= \sup_{s\ge 0}\left\{\Lambda_y(s) - s + \sqrt{b^2\gamma\Lambda_y(s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s))}\right\} \\
&\to \sup_{s\ge 0}\left\{\Lambda_y(s) - s\right\} = W_y \quad \text{as} \quad \gamma\downarrow 0,
\end{aligned} \tag{EC.34}$$

where $\Lambda_{\gamma,y}(t)$ is defined in (EC.21) and again W_y is the workload in the periodic deterministic fluid model. To justify (EC.34), we apply Lemma EC.1 to see that, $b^2\gamma\Lambda_y(s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s)) \le b^2\gamma I^{\uparrow}_w[\rho s+\lambda^{\uparrow}c] \le \gamma(K_1 s+K_2)$ for constants $I^{\uparrow}_w = \sup_t I_w(t)$, K_1 and K_2 and , so that $\sqrt{2b^2\gamma\Lambda_y(s)} \le \sqrt{\gamma(K_1 s+K_2)} \to 0$ uniformly over bounded interval as $\gamma\downarrow 0$. Hence, it suffices to consider the supremum in (EC.34) over a bounded interval, because the function is negative outside that interval for all sufficiently small γ. Since the limit W_y is the same as in Theorem 1, PRQ has been shown to be asymptotically correct as $\gamma\downarrow 0$. ∎

EC.5. Proofs of Heavy-Traffic Results from §6

Proof of Lemma 1. Observe that

$$\begin{aligned}\Lambda_{\gamma,\rho,y}(s) &\equiv \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho}) - \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho} - s) = \Lambda_{\gamma,\rho}(yc_{\gamma,\rho}) - \Lambda_{\gamma,\rho}(yc_{\gamma,\rho} - s) \\ &= \rho s + (1-\rho)^{-1}\int_{y/\gamma-(1-\rho)^2 s}^{y/\gamma} h(\gamma t)dt = \rho s + \frac{1}{\gamma(1-\rho)}\int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt \\ &= \rho s + \frac{1}{\gamma(1-\rho)}H_{\gamma,\rho,y}(s), \end{aligned} \tag{EC.35}$$

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

$$H_{\gamma,\rho,y}(s) \equiv \int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt. \quad \blacksquare \tag{EC.36}$$

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

LEMMA EC.3. *Let h be a differentiable 1-periodic function whose integral over one period is 0. Assume that h satisfies (41), then*

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

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

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

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

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

$$\begin{aligned}|g_{\gamma,\rho,y}(t) - h(y)t/\rho| &\le \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |h(s)-h(y)|ds = \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |h'(\xi)(s-y)|ds \\ &\le \frac{4M}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |s-y|ds = \frac{4M}{b^2c_x^2\gamma\rho^2}\cdot\frac{1}{2}\left(\frac{b^2c_x^2\gamma\rho}{4}t\right)^2 = N\gamma t^2, \end{aligned} \tag{EC.37}$$

where $N \equiv Mb^2c_x^2/8$. Note that the second line requires $h(\cdot)$ to be differentiable. (b) follows directly from (EC.37). To prove (a), we note that $|g_{\gamma,\rho,y}(t) - h(y)t| \le |g_{\gamma,\rho,y}(t) - h(y)t/\rho| + |h(y)t|(1-\rho^{-1})$. $\blacksquare$

LEMMA EC.4. *With f and $g_{\gamma,\rho,y}$ defined in (55) and (56), we have*

$$W^*_{\gamma,\rho,y} = \frac{b^2}{2}\cdot\frac{\rho c_x^2}{2(1-\rho)}\cdot\sup_{t\ge 0}\left\{f(t) + \rho g_{\gamma,\rho,y}(t) + 2\left(\sqrt{(t+(1-\rho)g_{\gamma,\rho,y}(t))\,C_{\gamma,\rho,y}(t)} - \sqrt{t}\right)\right\}, \tag{EC.38}$$

where

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

Proof. We write

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

Together with (55) and (56), the change of variable $s = b^2c_x^2\rho t/4(1-\rho)^2$ yields the desired expression. ∎

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 EC.3 and that the third component in (EC.38) 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).

Proof of Theorem 3. 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 EC.4, and note that

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

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

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

Now, we need only consider a bounded interval of t, becasuse $g_{\gamma,\rho,y}(\cdot)$ is uniformly bounded by definition (56) and thus the objective function in the supremum will be negative outside a bounded interval. The result then follows from part (d) of Lemma EC.3. ∎

Proof of Theorem 4. From Lemma EC.4, we have

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

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 3, 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 $\sup_{t\geq 0}\{-(1-h(y))t+2\sqrt{t}\}=1/(1-h(y))$, 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

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

where we used the concavity of the square root function. The result then follows from Lemma EC.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 (59), 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 5. Note that

$$\begin{aligned}W^*_{\gamma,\rho,y}&=\sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s)-s+b\sqrt{\Lambda_{\gamma,\rho,y}(s)I_w(\Lambda_{\gamma,\rho,y}(s))}\right\}\\&=\sup_{s\geq 0}\left\{-(1-\rho)\,s+\frac{1}{\gamma(1-\rho)}\int_{y-c^{-1}_{\gamma,\rho}s}^{y}h(u)du+b\sqrt{\Lambda_{\gamma,\rho,y}(s)I_w(\Lambda_{\gamma,\rho,y}(s))}\right\}\\&=\frac{1}{\gamma(1-\rho)}\cdot\sup_{t\geq 0}\left\{-t+\int_{y-t}^{y}h(u)du+\gamma(1-\rho)bc_x\sqrt{\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t)I_w(\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t))}\right\},\end{aligned}\quad\text{(EC.39)}$$

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. ∎

EC.6. Long-Cycle Heavy-Traffic Limits for 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 5. The following theorem reveals the relationship between the space scaling and γ.

THEOREM EC.3. (*long-cycle heavy-traffic limit for PRQ in a critically loaded queue*) *Assume that* $h(t)$ *satisfies*

$$h(t) = 1 - ct^p + o(t^p), \text{ as } t \to 0, \tag{EC.40}$$

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* $(\rho, \gamma) \to (1, 0)$.

Proof. By (EC.40), we have

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

as $\gamma \downarrow 0$ for fixed t, where $M = c\left(b^2 c_x^2 \rho\right)^p / (4^p(p+1))$. Applying Theorem 3 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 3. 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 $\gamma \downarrow 0$. ∎

We remark that the scaling in Theorem EC.3 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 was investigated. It was shown there that the space scaling of the heavy traffic limit depends on the detailed structure of the arrival-rate function.

We conducted simulation experiments to confirm Theorem EC.3. To illustrate, Figure EC.1 (left) shows that PRQ(b) with $b=1$ successfully captured the scaling with respect to γ and ρ, which in this sinusoidal case is $\gamma^{-p/(2p+1)}(1-\rho)^{-1}$ for $p=2$. Figure EC.1 (right) shows that both simulation estimation and the PRQ approximation after scaling is relatively insensitive to the traffic intensity ρ.

We end this section by remarking that, in this simulation example, we consider only the mean and applied the PRQ algorithm with robustness parameter $b=1$. This choice sits between our choice of $b=\sqrt{2}$ for the mean in the underloaded case in §4 and $b=0.5$ for the mean in the overloaded case in §5. Our choice of $b=1$ here was experimental. Finding a suitable function Π in (22) for the critically-loaded models remains to be an important direction for future research.

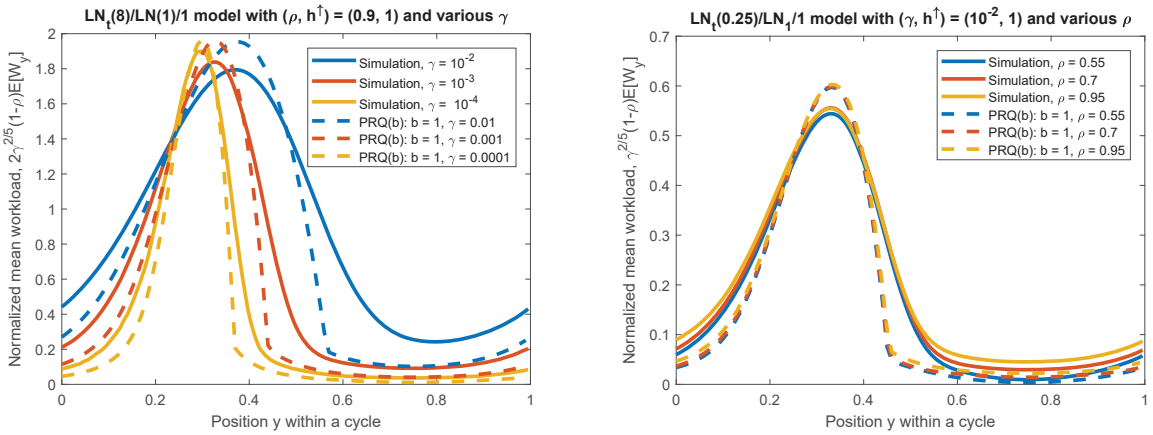

Figure EC.1 Comparing the simulation estimation of the steady-state mean workload to the PRQ(b) approximation in (20) with $b=1$ in two critically-loaded model. The arrival rate function is (52) with the parameters specified in each plot.

EC.7. Heavy-Traffic and Long-Cycle Limits in the $G_t/G_t/1$ model

In this section, we present 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 §6.1 and the periodic queueing setup in §6.3. 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 §6, where $A_{\gamma,\rho,y}(s) \equiv N\left(\Lambda_{\gamma,\rho,y}(s)\right)$. The corresponding PRQ is

$$W^*_{\gamma,\rho,y} = \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - M_{\gamma,\rho,y}(s) + b\sqrt{\Lambda_{\gamma,\rho,y}(s) I_w\left(\Lambda_{\gamma,\rho,y}(s)\right)}\right\} \tag{EC.41}$$

as in (51). 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 (37) 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

$$M_{\gamma,\rho}(t) \equiv t + (1-\rho)^{-1} M_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0 \tag{EC.42}$$

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

$$M_{d,\gamma}(t) \equiv \int_0^t \mu_{d,\gamma}(s)\, ds, \quad \mu_{d,\gamma}(t) \equiv r(\gamma t), \quad \text{and} \quad \int_0^1 r(t)\, dt = 0 \tag{EC.43}$$

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

With the same heavy-traffic scalings as in (42), we generalize Theorem 2 as follows.

THEOREM EC.4. (*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* (37) *and cumulative service-rate function in (EC.42), if $\hat{N}_n \Rightarrow c_a B_a$ as $n\to\infty$, where B_a is a standard Brownian motion, then*

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

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* (43)*, and B_a and B_s are two independent standard* (*mean 0 variance 1*) *Brownian motions.*

Proof. By definition, we have

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

where $\Xi_{\gamma,\rho}(t)$ denotes the quantity $\hat{X}_{\gamma,\rho}(t)$ exacly as it appears in Theorem 2, 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 (40) and r in (EC.43).

Next, we generalize the limit theorems for the PRQ problem in (EC.41). As preparation, we re-write $M_{\gamma,\rho,y}$ exactly the same as (53)

$$\begin{aligned}
M_{\gamma,\rho,y}(s) &\equiv M_{\gamma,\rho}((k+y)c_{\gamma,\rho}) - M_{\gamma,\rho}((k+y)c_{\gamma,\rho} - s) = M_{\gamma,\rho}(yc_{\gamma,\rho}) - M_{\gamma,\rho}(yc_{\gamma,\rho} - s) \\
&= s + (1-\rho)^{-1}\int_{y/\gamma-(1-\rho)^2 s}^{y/\gamma} r(\gamma t)dt = s + \frac{1}{\gamma(1-\rho)}\int_{y-c_{\gamma,\rho}^{-1}s}^{y} r(t)dt \\
&= s + \frac{1}{\gamma(1-\rho)}R_{\gamma,\rho,y}(s), \qquad \text{(EC.44)}
\end{aligned}$$

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 (56), we define

$$\tilde{g}_{\gamma,\rho,y}(t) \equiv \frac{4}{b^2 c_x^2 \gamma \rho^2} \int_{y-\frac{b^2 c_x^2 \gamma \rho}{4} t}^{y} (h(s) - r(s))ds. \tag{EC.45}$$

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 EC.4.

LEMMA EC.5. *With f, $g_{\gamma,\rho,y}$ and $\tilde{g}_{\gamma,\rho,y}$ defined in* (55), (56) *and* (EC.45), *we have*

$$W^*_{\gamma,\rho,y} = \frac{b^2}{2} \cdot \frac{\rho c_x^2}{2(1-\rho)} \cdot \sup_{t\geq 0} \left\{ f(t) + \rho \tilde{g}_{\gamma,\rho,y}(t) + 2\left(\sqrt{(t + (1-\rho) g_{\gamma,\rho,y}(t))\, C_{\gamma,\rho,y}(t)} - \sqrt{t} \right) \right\}, \tag{EC.46}$$

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.41), 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 (55), (56) and (EC.45), 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.5. (*heavy traffic limit for PRQ*) *The heavy traffic limit of the PRQ problem in* (EC.41) *for the $G_t/G_t/1$ model is*

$$\lim_{\rho \uparrow 1} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0} \{ f(t) + \tilde{g}_{\gamma,1,y}(t) \}. \tag{EC.47}$$

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

$$\tilde{\rho}(y) = \frac{\rho + (1-\rho)h(y)}{1 + (1-\rho)r(y)}. \tag{EC.48}$$

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 §5.2; (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.6. (*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.41) *for the $G_t/G_t/1$ model admits the double limit*

$$\lim_{\substack{\gamma\downarrow 0\\ \rho\uparrow 1}} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \frac{1}{1-\tilde{h}(y)}, \tag{EC.49}$$

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

$$W^*_y = \frac{b^2}{2} \cdot \frac{\tilde{\rho}(y) c_x^2}{2(1-\tilde{\rho}(y))} + o(1-\rho), \tag{EC.50}$$

where $\tilde{\rho}(y)$ is the instantaneous traffic intensity in (EC.48).

THEOREM EC.7. (*long-cycle limit for PRQ in an overloaded queue*) *The PRQ problem in* (EC.41) *for the $G_t/G_t/1$ model with the heavy-traffic scaling in* (37) *and $\tilde{h}^{\uparrow} > 1$ admits the long-cycle limit*

$$(1-\rho) \lim_{\gamma\downarrow 0} \gamma \cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0} \left\{ -t + \int_{y-t}^{y} \tilde{h}(s) ds \right\}, \quad 0 \leq \rho < 1. \tag{EC.51}$$

THEOREM EC.8. (*long-cycle heavy-traffic limit for PRQ in a critically loaded queue*) *Assume that $\tilde{h}(t)$ satisfies*

$$\tilde{h}(t) = 1 - ct^p + o(t^p), \text{ as } t \to 0, \tag{EC.52}$$

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.8. Additional Examples

In this final section of the EC we make additonal simulation comparisons to provide further insight into the performance of PRQ.

EC.8.1. Statistical Precision

We start by showing the statistical precision of our estimated steady-state mean workload. Recall that the simulation methodology is described in §3.2. Figure EC.2 shows the estimation of the steady-state mean workload in two cases, together with the 95% confidence interval (CI). We conclude that the run time used here is sufficient to achieve high statistical precision.

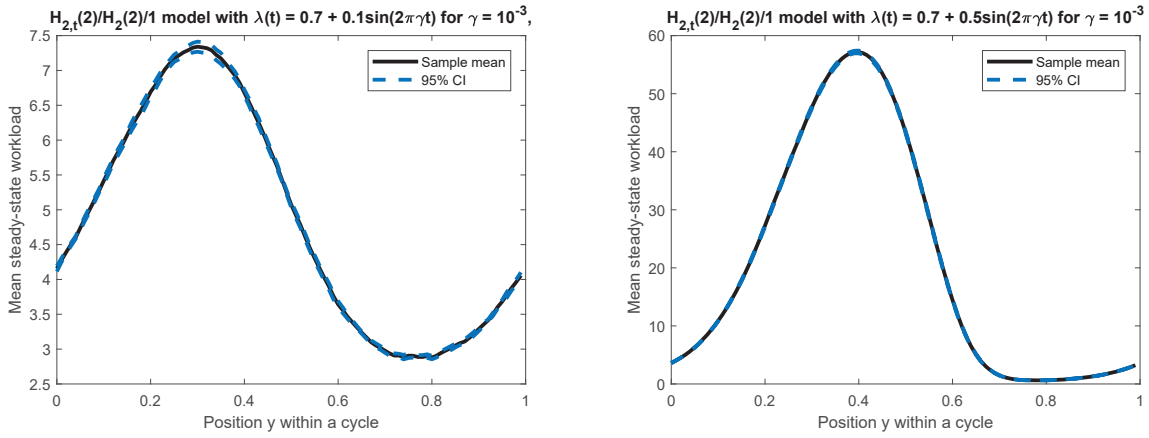

Figure EC.2 The estimated mean of the steady-state mean workload in the $H_{2,t}(2)/H_{2,t}/1$ model with arrival-rate function in (24). Both the UL and OL cases are shown here, with model parameters specified in the titles. The 95% confidence interval is displayed in dashed curves.

EC.8.2. Underloaded Models

For underloaded models, we will compare the simulation estimation of the mean or quantiles of the steady-state workload W_y to the PRQ(b) algorithm specified by (20), (22) and (28). For the mean, we use $b = \sqrt{2}$ as discussed in §4.2. For quantiles, we look at levels $p = 0.95, 0.8, 0.632, 0.4$ and 0.2.

In Figure EC.3, we show the robustness of the PRQ(b) algorithm by presenting four models that share the same arrival rate function but with different interarrival and service time distributions. In

particular, we consider three balanced $GI_t/GI/1$ models with GI being Erlang (E_2) distribution, exponential (M) distribution or hyperexponential ($H_2(2)$) distribution and also one unbalance $LN_t(4)/E_4/1$ model with $LN(4)$ being the Lognormal distirbution with $c_a^2 = 4$. Consistent with our previous observations, PRQ(b) performs very well across different choices of the underlying distribution.

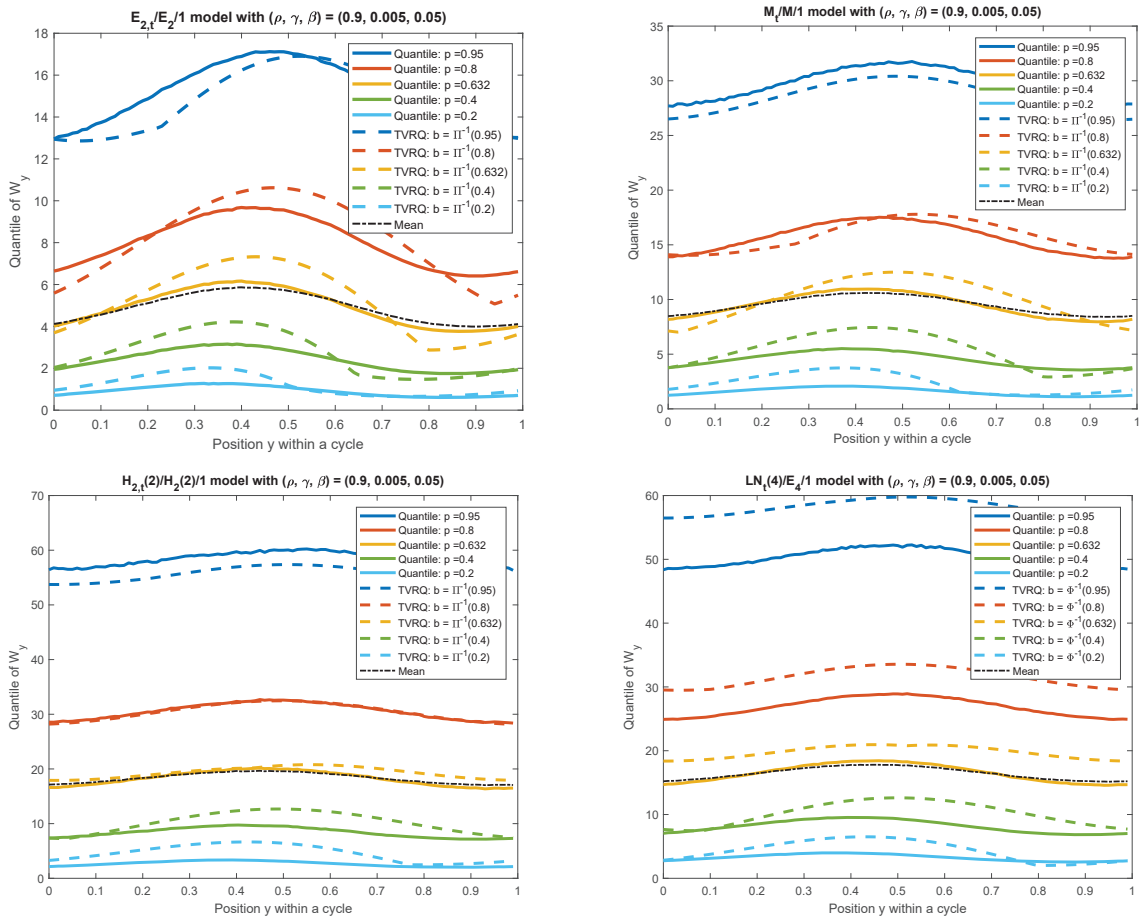

Figure EC.3 The estimated mean and quantiles of the steady-state mean workload in the $GI_t/GI/1$ model with arrival-rate function in (24) and model parameters specified in the titles. Models with four different distributions are displayed to demonstrate the robustness of the PRQ(b) algorithm.

Figure EC.4 supplements Figure 2 by presenting the corresponding long cycle models. In particular, we look at two $M_t/M/1$ models with traffic intensity $\rho = 0.7$ or 0.9. Both examples have a

cycle length of 1000, representing high volumn systems. Both plots compares the TVRQ approximation to the simulation estimation of the quantiles of the stead-state workload, as functions of the position y within a cycle. Again, PRQ(b) performs very well in approximating the full distribution of the steady-state workload.

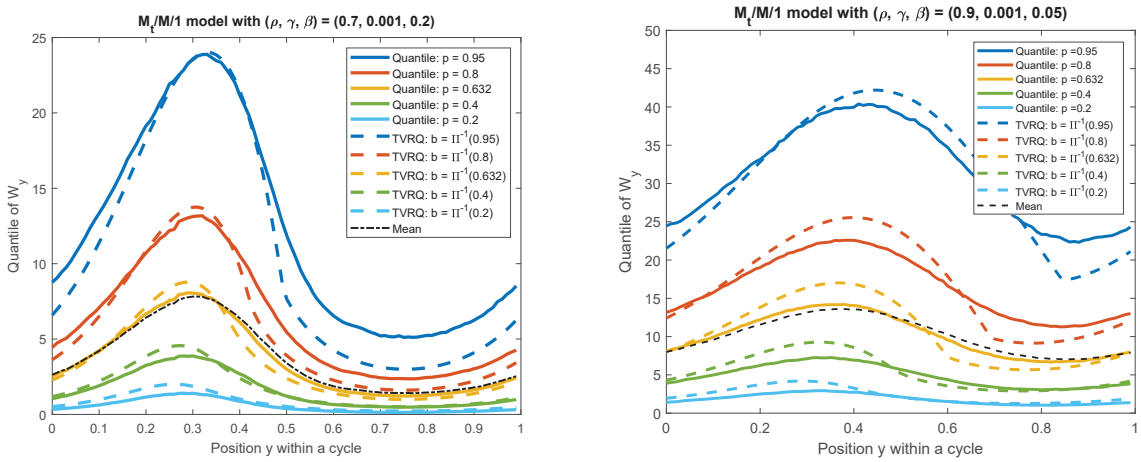

Figure EC.4 The estimated mean and quantiles of the steady-state mean workload in the $M_t/M/1$ underloaded model with arrival-rate function in (24) and model parameters specified in the titles. Left plot shows an example with moderate long-run traffic intensity of $\rho = 0.7$ and a highly variable arrival-rate function $\rho^{\uparrow} = 0.9$. Right plot shows a system with higher traffic intensity of $\rho = 0.9$.

EC.8.3. Overloaded Models

For overloaded models, we will compare the simulation estimation of the mean or quantiles of the steady-state workload W_y to the PRQ(b) algorithm specified by (20), (22) and (36). For the mean, we use $b = 0.5$ as discussed in §5.3. For quantiles, we look at levels $p = 0.9, 0.7, 0.5, 0.3$ and 0.1.

We now present more examples to demonstrate the performance of the PRQ(b) algorithm in the overloaded models. Figure EC.5 present four models with the same arrival-rate function parameters but different underlying distributions for the interarrival and service times. In particular, we consider a wide range of selections in terms of the variability parameter, including a low variability Erlang E_4 distribution and a highly variable hyperexponential $H_2(8)$ distribution. We see that

even the arrival-rate function is fixed, the variability in the underlying distribution can result in different forms of the quantile functions. On the other hand, the PRQ(b) algorithm successfully approximated the distribution of the steady-state workload W_y, as a function of the position y within a cycle.

Figure EC.5 demonstrate that the PRQ(b) algorithm adapts to the changing distribution quite well. However, we can still observe performance degradation as the variability increases, which is caused by our fixed choice of the $\Pi(b)$ function in (36). Further refinements are possible if we allow $\Pi(b)$ to be a function of the variability parameter. But we do not discuss such extensions in this paper.

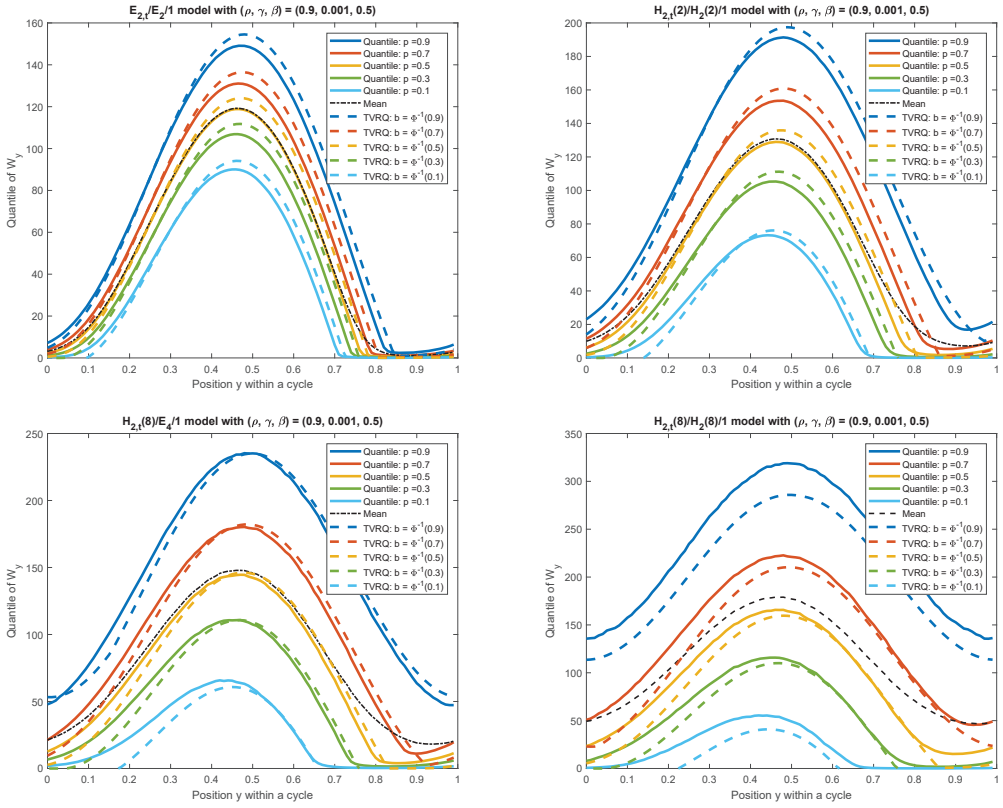

Figure EC.5 The estimated mean and quantiles of the steady-state mean workload in the $GI_t/GI/1$ model with arrival-rate function in (24) and model parameters specified in the titles. Models with four different distributions are displayed to demonstrate the robustness of the PRQ(b) algorithm.

EC.8.4. Long-Cycle Heavy-Traffic Limit

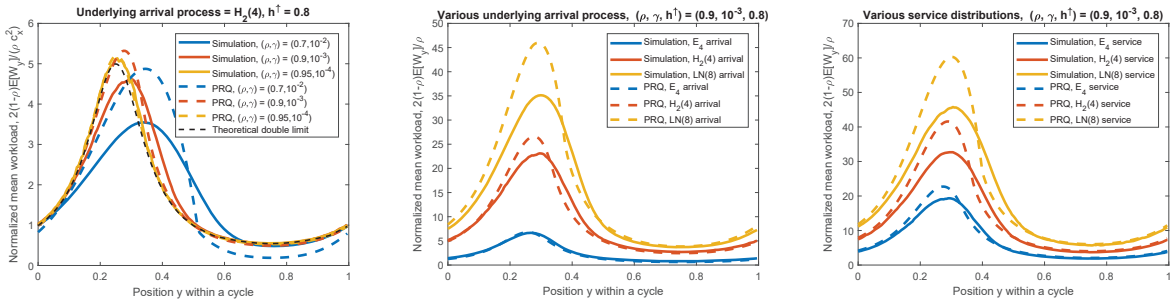

Figure EC.6 A comparison of PRQ in (17) as a function of the position y within a cycle to simulation estimations of the normalized mean workload $2(1-\rho)E\left[W_{\gamma,\rho,y}\right]/\rho$ for $W_{\gamma,\rho,y}$ in (48) and the limit in Theorem 4 in the underloaded $(H_2(4)_t/LN(1)/1$ model with arrival-rate function in (24) and (37) 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).

Figure EC.6 presents simulation comparisons illustrating Theorem 4. In each case, PRQ is compared 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 (48) in the underloaded $H_{2,t}(4)/LN(1)/1$ model with the sinusoidal model in (24) with the scaling in (37)-(39). In particular, the convergence as $\gamma \downarrow 0$ and $\rho \uparrow 1$ is illustrated by considering $(\gamma,\rho) \in \{(0.7,10^{-2}),(0.9,10^{-3}),(0.95,10^{-4})\}$ (left), while the improved performance of PRQ as the level of variability decreases in the arrival and service processes is illustrated in the middle and right.

Figure EC.6 (middle) and (right) show the impact of changing variability in the arrival process and the service-time distribution. Consistent with the stationary model, Figure EC.6 (middle) and (right) show that increased variability in either the arrival process or the service process tends to increase congestion. We remark that the story can be different; e.g., it 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).

References

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

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

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.

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

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.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42** 1383–1394.

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. 2015. Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters* **102** 202–207.

Ma, N., W. Whitt. 2018a. Minimizing the maximum expected waiting time in a periodic single-server queue with a service-rate control. *Stochastic Systems* **8**(4).

Ma, N., W. Whitt. 2018b. A rare-event simulation algorithm for periodic single-server queues. *INFORMS Journal on Computing* **30**(1) 71–89.

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. 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. 1998. 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. 1991a. The efficiency of one long run versus independent replications in steady-state simulation. *Management Science* **37**(6) 645–666.

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

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. 2018a. A robust queueing network analyzer based on indices of dispersion. Under review, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

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