Submitted to *INFORMS Journal on Computing*
manuscript (Please, provide the mansucript number!)

A Rare-Event Simulation Algorithm for Periodic Single-Server Queues

Ni Ma
Industrial Engineering and Operations Research, Columbia University,nm2692@columbia.edu,

Ward Whitt
Industrial Engineering and Operations Research, Columbia University, ww2040@columbia.edu,
http://www.columbia.edu/ ww2040

An efficient algorithm is developed to calculate the periodic steady-state distribution and moments of the remaining workload W_y at time yc within a cycle of length c, $0 \le y < 1$, in a single-server queue with a periodic arrival-rate function. The algorithm applies exactly to the $GI_t/GI/1$ model, where the arrival process is a time-transformation of a renewal process. A new representation of W_y makes it possible to apply a modification of the classic rare-event simulation for the stationary $GI/GI/1$ model exploiting importance sampling using an exponential change of measure. We establish bounds between the periodic workload and the stationary workload with the average arrival rate that enable us to prove that the relative error in estimates of $P(W_y > b)$ is uniformly bounded in b. With the aid of a recent heavy-traffic limit theorem, the algorithm also applies to compute the periodic steady-state distribution of (i) reflected periodic Brownian motion (RPBM) by considering appropriately scaled $GI_t/GI/1$ models and (ii) a large class of general $G_t/G/1$ queues by approximating by $GI_t/GI/1$ models with the same heavy-traffic limit. Simulation examples demonstrate the accuracy and efficiency of the algorithm for both $GI_t/GI/1$ queues and RPBM.

Key words: periodic queues, ruin probabilities, rare-event simulation, exponential change of measure, heavy traffic, reflected periodic Brownian motion

History: January 31, 2017

1. Introduction

For the steady-state performance of the stationary $GI/GI/1$ single-server queue with unlimited waiting room and service in order of arrival, we have effective algorithms, e.g., Abate et al. (1993), Asmussen (2003). We also have exact formulas in special cases and useful general approximation formulas in heavy traffic, e.g., Asmussen (2003), Whitt (2002). For the periodic steady-state performance of associated periodic single-server queues, hav-

ing periodic arrival-rate functions, there is much less available. There is supporting theory in Harrison and Lemoine (1977), Lemoine (1981, 1989), Rolski (1981, 1989). On the algorithm side, there is a recent contribution on perfect sampling in Xiong et al. (2015). Of particular note is the paper on the periodic $M_t/GI/1$ queue by Asmussen and Rolski (1994) that provides a theoretical basis for a rare-event simulation algorithm (although no algorithm is discussed there); also see §VII.6 of Asmussen and Albrecher (2010) and Morales (2004). The goal there was to calculate ruin probabilities, but those are known to be equivalent to waiting-time and workload tail probabilities. A heavy-traffic limit for the periodic $G_t/G/1$ queue, was also established recently by Whitt (2014), which shows that the basic processes can be approximated by reflected periodic Brownian motion (RPBM), but so far there are no algorithms or simple formulas for RPBM.

In this paper, we provide an effective algorithm to calculate the periodic steady-state distribution and moments of the remaining workload W_y at time yc within a cycle of length c, $0 \le y < 1$, in a single-server queue with a periodic arrival-rate function. The algorithm applies exactly to the $M_t/GI/1$ model, where the arrival process is a nonhomogeneous Poisson process (NHPP), and any $GI_t/GI/1$ model, where the arrival process is a time-transformation of an equilibrium renewal process. A new representation of W_y (in (2) below) makes it possible to apply a modification of the classic rare-event simulation for the stationary $GI/GI/1$ model exploiting importance sampling using an exponential change of measure, as in Ch. XIII of Asmussen (2003) and Ch. VI of Asmussen and Glynn (2007). We show that the algorithm is effective for estimating the mean and variance as well as small tail probabilities.

The main example is the periodic $M_t/GI/1$ queue, but our results go well beyond the periodic $M_t/GI/1$ queue. By also treating the more general $GI_t/GI/1$ queue, we are able to apply the algorithm to compute the steady-state distribution of the limiting RPBM in Whitt (2014). To cover the full range of parameters of the RPBM, we need the generalization to $GI_t/GI/1$. (In particular, this enables us to calculate the periodic steady-state distribution of the limiting RPBM for the $GI_t/GI/1$ model in (51) and (55) for any variability parameter c_x.) As we will explain in §6.4, the algorithm for the $GI_t/GI/1$ model can serve as a basis for an approximation algorithm for more general $G_t/G/1$ models, but we do not report simulation results for that extension here.

We report results from extensive simulation experiments for $GI_t/GI/1$ models to demonstrate the effectiveness of the algorithm. Both the convergence to RPBM and the effectiveness of the algorithm for RPBM are demonstrated by displaying the results for a range of traffic intensities ρ approaching 1. This unity in the numerical results requires the non-standard heavy-traffic scaling in Whitt (2014), which we review in §6. (In particular, the deterministic arrival-rate function is scaled as well as space and time; see (41).) The unity in the numerical results provided by the heavy-traffic scaling is in the same spirit as the scaling in the numerical results in Abate and Whitt (1998), Choudhury et al. (1997).

1.1. Using Bounds to Connect to Familiar Rare-Event Simulation Methods

We are able to apply the familiar rare-event simulation for the $GI/GI/1$ model to the periodic $GI_t/GI/1$ model because we can make strong connections between the given periodic $GI_t/GI/1$ model and the associated $GI/GI/1$ model with the constant average arrival rate. In fact, this connection is largely achieved directly by construction, because we represent the periodic arrival counting process A as a deterministic time transformation of an underlying rate-1 counting process N by

$$A(t) \equiv N(\Lambda(t)), \quad \text{where} \quad \Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0, \tag{1}$$

λ is the arrival-rate function, assumed to be positive, and $\equiv$ denotes equality by definition. This is a common representation when N is a rate-1 Poisson process; then A is an NHPP. For the $G_t/G/1$ model, N is understood to be a rate-1 stationary point process. Hence, for the $GI_t/GI/1$ model, N is an equilibrium renewal process with time between renewals having mean 1, which is a renewal process except the first inter-renewal time has the equilibrium distribution. The representation in (1) also has been used for processes N more general than NHPP's by Massey and Whitt (1994), Gerhardt and Nelson (2009), Nelson and Gerhardt (2011), He et al. (2016), Ma and Whitt (2015), Whitt (2015) and Whitt and Zhao (2016).

Given that we use representation (1), we show that it is possible to uniformly bound the difference between the cumulative arrival-rate function Λ and the associated linear cumulative arrival-rate function $\bar{\lambda}e$ of the stationary model, where $\bar{\lambda}$ is the average arrival rate and e is the identity function, $e(t) \equiv t$, $t \geq 0$. Consequently, we are able to bound the difference between the steady-state workloads W in the stationary $G/G/1$ model and W_y in the periodic $G_t/G/1$ model.

1.2. A Convenient Representation for Estimation Efficiency

We exploit the arrival process construction in (1) to obtain a convenient representation of the stationary workload W_y in terms of the underlying stationary process $N \equiv \{N(t) : t \geq 0\}$ in (1) and the associated sequence of service times $V \equiv \{V_k : k \geq 1\}$ via

$$W_y \stackrel{d}{=} \sup_{s \geq 0} \Big\{ \sum_{k=1}^{N(s)} V_k - \tilde{\Lambda}_y^{-1}(s) \Big\}, \quad 0 \leq y < 1, \tag{2}$$

where

$$\tilde{\Lambda}_y(t) \equiv \Lambda(yc) - \Lambda(yc - t), \quad t \geq 0, \tag{3}$$

is the *reverse-time cumulative arrival-rate function* starting at time yc within the periodic cycle $[0, c]$, $0 \leq y < 1$, and $\tilde{\Lambda}_y^{-1}$ is its inverse function, which is well defined because $\tilde{\Lambda}_y(t)$ is continuous and strictly increasing. Representation (2) is convenient because all stochastic dependence is captured by the first term within the supremum, while all deterministic time dependence is captured by the second term.

From the representation in (2), it is evident that from each sample path of the underlying stochastic process (N, V), we can generate a realization of W_y in (2) for each y, $0 \leq y < 1$, by just changing the deterministic function $\tilde{\Lambda}_y^{-1}$. Moreover, from the rare-event construction in §4, we can simultaneously obtain an estimate of $P(W_y > b)$ for all b in the bounded interval $[0, b_0]$ while applying the estimation for the single value b_0. Thus, we can essentially obtain estimates for all *performance parameter pairs* $(y, b) \in [0, 1) \times [0, b_0]$ while doing the estimation for only one pair. This efficiency is very useful to conduct simulation studies to expose the way that $P(W_y > b)$ and the other performance measures depend on (y, b).

1.3. Stylized Sinusoidal Examples

We illustrate the rare-event simulation by showing simulation results for $GI_t/GI/1$ queues with sinusoidal arrival-rate function

$$\lambda(t) \equiv \bar{\lambda}(1 + \beta \sin(\gamma t)), \quad t \geq 0, \tag{4}$$

where β, $0 < \beta < 1$, is the relative amplitude and the cycle length is $c = 2\pi/\gamma$. We let the mean service time be $\mu^{-1} = 1$, so that the average arrival rate is the traffic intensity, i.e., $\bar{\lambda} = \rho$. With this scaling, we see that there is the fundamental *model parameter triple* (ρ, β, γ) or, equivalently, (ρ, β, c). The associated cumulative arrival-rate function is

$$\Lambda(t) = \rho(t + (\beta/\gamma)(1 - \cos(\gamma t)), \quad t \geq 0. \tag{5}$$

and the associated reverse-time cumulative arrival-rate function defined in (3) is

$$\tilde{\Lambda}_y(t) = \rho\,(t + (\beta/\gamma)\,(\cos\,(\gamma(yc - t)) - \cos\,(\gamma yc)))\,, \quad t \geq 0. \tag{6}$$

We only consider the case $\rho < 1$, under which a proper steady-state exists under regularity conditions (which we do not discuss here). Behavior differs for short cycles and long cycles. There are two important cases for the relative amplitude: (i) $0 < \beta < \rho^{-1} - 1$ and (ii) $\rho^{-1} - 1 \leq \beta \leq 1$. In the first case, we have $\rho(t) < 1$ for all t, where $\rho(t) \equiv \lambda(t)$ is the instantaneous traffic intensity, but in the second case we have intervals with $\rho(t) \geq 1$, where significant congestion can build up. If there is a long cycle as well, the system may be better understood from fluid and diffusion limits, as in Choudhury et al. (1997). (Tables 8 and 9 illustrate the significant performance difference for the mean $E[W_y]$.)

1.4. Organization of the Paper

We start in §2 by reviewing the reverse-time representation of the workload process, which leads to representation (2). In §3 we establish the bounds and associated asymptotic and approximations connecting the periodic model to the associated stationary model with the average arrival rate. In §4 we develop the simulation algorithm for the $GI_t/GI/1$ model and establish theoretical results on its efficiency. We also discuss the computational complexity and running times. In §5 we present simulation examples. In §6 we review and extend the heavy-traffic FCLT in Theorem 3.2 of Whitt (2014), which explains the scaling that unifies our numerical results in the simulation experiments. in §6.4 we discuss the approximation for general periodic $G_t/G/1$ models. In §7 we draw conclusions. We present additional material in the online supplement Ma and Whitt (2016), including approximations for the important asymptotic decay rate and more simulation examples.

2. Reverse-Time Representation of the Workload Process

We consider the standard single-server queue with unlimited waiting space where customers are served in order of arrival. Let $\{(U_k, V_k)\}$ be a sequence of ordered pairs of interarrival times and service times. (in §2 and in §3 we do not need to impose any GI conditions.) 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

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0. \tag{7}$$

Then we can apply the reflection map to the net input process $Y(t)-t$ to represent the workload (the remaining work in service time) at time t, starting empty at time 0, as

$$W(t) = Y(t) - t - \inf\{Y(s) - s : 0 \le s \le t\} = \sup\{Y(t) - Y(s) - (t-s) : 0 \le s \le t\}, \quad t \ge 0.$$

We now convert this standard representation to a simple supremum by using a reverse-time construction, as in Loynes (1962) and Chapter 6 in Sigman (1995). This is achieved by letting the interarrival times and service times be ordered in reverse time going backwards from time 0. Then $\tilde{A}(t)$ counts the number of arrivals and $\tilde{Y}(t)$ is the total input over the interval $[-t, 0]$ for $t \ge 0$. With this reverse-time construction (interpretation), we can write

$$W(t) = \sup\{\tilde{Y}(s) - s : 0 \le s \le t\}, \quad t \ge 0, \tag{8}$$

and we have $W(t)$ increasing to $W(\infty) \equiv W$ with probability 1 (w.p.1) as $t \uparrow \infty$. In a stable stationary setting, under regularity conditions, we have $P(W < \infty) = 1$; see §6.3 of Sigman (1995).

We now consider the periodic arrival-rate function $\lambda(t)$ with cycle length c, average arrival rate $\bar{\lambda} = \rho < 1$ and bounds $0 < \lambda_L \le \lambda(t) \le \lambda_U < \infty$ for $0 \le t \le c$. As in (1), we can construct the arrival process A by transforming a general rate-1 stationary process N by the cumulative arrival-rate function. We let the service times V_k be a general stationary sequence with $E[V_k] = 1$.

We now exploit (8) in our more specific periodic $G_t/G/1$ context. The workload at time yc in the system starting empty at time $yc - t$ can be represented as

$$\begin{aligned} W_y(t) &= \sup_{0 \le s \le t} \{\tilde{Y}_y(s) - s\} \\ &\stackrel{d}{=} \sup_{0 \le s \le t} \Big\{ \sum_{k=1}^{N(\tilde{\Lambda}_y(s))} V_k - s \Big\} \\ &= \sup_{0 \le s \le \tilde{\Lambda}_y(t)} \Big\{ \sum_{k=1}^{N(s)} V_k - \tilde{\Lambda}_y^{-1}(s) \Big\}, \end{aligned} \tag{9}$$

where $\tilde{Y}_y$ is the *reverse-time total input of work* starting at time yc within the cycle of length c, $\tilde{\Lambda}_y(t)$ is the reverse-time cumulative arrival-rate function in (3) and $\tilde{\Lambda}_y^{-1}$ is its inverse function, which are defined in terms of the cumulative arrival-rate function $\Lambda(t)$ in (1). The second line equality in distribution holds when N is a stationary point process,

which is a point process with stationary increments and a constant rate. In the $GI_t/GI/1$ setting, N is an equilibrium renewal process and thus this regularity condition is satisfied. Note that in this specific setting, V_k's are i.i.d. with distribution V, but U_1 has equilibrium distribution U_e, which may be different from the i.i.d. distributions of U_k, $k \geq 2$ in (9). Just as $W(t) \uparrow W$ w.p.1 as $t \to \infty$, so $W_y(t) \uparrow W_y$ w.p.1 as $t \to \infty$, for W_y in (2).

Even though (9) is valid for all t, we think of the system starting empty at times $-kc$, for $k \geq 0$, so that we let $yc - t = -kc$ or, equivalently, we stipulate that $t = c(k+y)$, $0 \leq y < 1$, and consider successive values of k and let $k \to \infty$ to get (2). That makes (9) valid to describe the distribution of $W(c(k+y))$ for all $k \geq 0$. We think that (9) and (2) are new representations, but they can be related to various special cases in the literature.

3. Bounds and Approximations for General Periodic $G_t/G/1$ Queues

We first bound the periodic system above and below by modifications of the corresponding stationary system with an arrival process that has the average arrival rate. Then we establish limits and introduce approximations. In doing so, we extend results in Asmussen and Rolski (1994).

3.1. Basic Bounds

We now compare the periodic steady-state workload W_y in (2) and the associated stationary workload W defined as in (2) with $\rho^{-1}s$ replacing $\tilde{\Lambda}_y^{-1}(s)$:

$$W \stackrel{d}{=} \sup_{s \geq 0} \Big\{ \sum_{k=1}^{N(s)} V_k - \rho^{-1} s \Big\}, \tag{10}$$

Note that in both (2) and (10), N is understood to be a stationary point process. In particular, for the $GI_t/GI/1$ model, N is an equilibrium renewal process with the first inter-renewal time having the equilibrum distribution, therefore W is the stationary workload in the associated $GI/GI/1$ model, which may differ from the stationary waiting time in the same model. We now show that we can bound W_y above and below by a constant difference from the stationary workload W by rewriting (2) as

$$W_y = \sup_{s \geq 0} \Big\{ \sum_{k=1}^{N(s)} V_k - \rho^{-1} s - (\tilde{\Lambda}_y^{-1}(s) - \rho^{-1} s) \Big\}. \tag{11}$$

From (11), we immediately obtain the following lemma.

LEMMA 1. (*upper and lower bounds on W_y*) *For W_y in* (2) *and W in* (10),

$$W_y^- \equiv W - \zeta_y^- \le W_y \le W - \zeta_y^+ \equiv W_y^+ \tag{12}$$

where

$$\zeta_y^- \equiv \sup_{0\le s\le \rho c} \{\tilde{\Lambda}_y^{-1}(s) - \rho^{-1}s\} \ge 0 \quad \textit{and} \quad \zeta_y^+ \equiv \inf_{0\le s\le \rho c} \{\tilde{\Lambda}_y^{-1}(s) - \rho^{-1}s\} \le 0. \tag{13}$$

Note that the supremum and infimum in (13) are over the interval $[0, \rho c]$. Because the average arrival rate is ρ, $\tilde{\Lambda}_y(c) = \Lambda(c) = \rho c$ and thus $\tilde{\Lambda}_y^{-1}(\rho c) = c$. Given that Λ is continuous and strictly increasing, we can use properties of the inverse function as in §13.6 of Whitt (2002) to determine an alternative representation of the bounds in terms of the reverse-time cumulative arrival-rate function $\tilde{\Lambda}_y$. We emphasize that these bounds depend on y.

LEMMA 2. (*alternative representation of the bounds*) *The constants ζ_y^- and ζ_y^+ can also be expressed as*

$$\zeta_y^- = -\rho^{-1} \inf_{0\le s\le c} \{\tilde{\Lambda}_y(s) - \rho s\} \ge 0 \quad \textit{and} \quad \zeta_y^+ = -\rho^{-1} \sup_{0\le s\le c} \{\tilde{\Lambda}_y(s) - \rho s\} \le 0. \tag{14}$$

Proof. We use basic properties of inverse functions, as in §13.6 of Whitt (2002). First, note that, for any homeomorphism ϕ on the interval $[0, c]$,

$$\sup_{0\le s\le c} \{\phi(s) - s\} = \sup_{0\le s\le c} \{\phi(\phi^{-1}(s)) - \phi^{-1}(s)\} = \sup_{0\le s\le c} \{s - \phi^{-1}(s)\} = -\inf_{0\le s\le c} \{\phi^{-1}(s) - s\}. \tag{15}$$

To treat ζ_y^- in (13), we apply (15) to $\tilde{\Lambda}_y^{-1}$ after rescaling time to get

$$\begin{aligned} \sup_{0\le s\le \rho c} \{\tilde{\Lambda}_y^{-1}(s) - \rho^{-1}s\} &= \sup_{0\le u\le c} \{\tilde{\Lambda}_y^{-1}(\rho u) - u\} = -\inf_{0\le u\le c} \{\rho^{-1}\tilde{\Lambda}_y(u) - u\} \\ &= -\rho^{-1} \inf_{0\le s\le c} \{\tilde{\Lambda}_y(s) - \rho s\}. \end{aligned} \tag{16}$$

In (16), the first equality is by making the change of variables $u = \rho^{-1}s$; the second equality is by (15) plus Lemma 13.6.6 of Whitt (2002), i.e., $(\tilde{\Lambda}_y^{-1} \circ \rho e)^{-1} = (\rho^{-1} e \circ \tilde{\Lambda}_y) = \rho^{-1}\tilde{\Lambda}_y$; the third equality is obtained by multiplying and dividing by ρ. ∎

We now combine the one-sided extrema into an expression for the absolute value.

COROLLARY 1. (*single bound*) *As a consequence,*

$$\begin{aligned} |W_y - W| \le \zeta &\equiv \max\{\zeta_y^-, -\zeta_y^+\} \\ &= \rho^{-1}\|\tilde{\Lambda}_y - \rho e\|_c \equiv \rho^{-1} \sup_{0\le s\le c} \{|\tilde{\Lambda}_y(s) - \rho s|\} < \infty. \end{aligned} \tag{17}$$

COROLLARY 2. (*bounds in the sinusoidal case*) *For the sinusoidal case in* (4)*, the bounds can be expressed explicitly as*

$$\zeta_y^- = \frac{\beta(\cos(\gamma cy)+1)}{\gamma} \quad \textit{and} \quad \zeta_y^+ = \frac{\beta(\cos(\gamma cy)-1)}{\gamma}. \tag{18}$$

Proof. By (6),

$$\tilde{\Lambda}_y(t) - \rho t = (\rho\beta/\gamma)\left(\cos(\gamma(cy-t)) - \cos(\gamma cy)\right), \quad t \ge 0, \tag{19}$$

from which (18) follows by choosing t to make $\cos(\gamma(cy-t)) = \pm 1$. ∎

3.2. Tail Asymptotics for the Periodic $G_t/G/1$ Model

For many models, it is possible to obtain an approximation for W of the form

$$P(W > b) \approx Ae^{-\theta^* b}, \quad b \ge 0, \tag{20}$$

based on the limit

$$\lim_{b\to\infty} e^{\theta^* b} P(W > b) = A. \tag{21}$$

For the $GI/GI/1$ model, the limit (21) is discussed in §XIII.5 of Asmussen (2003), where the random variable $X_k \equiv V_k - T_k$ is required to have a nonlattice distribution. However, the limit (21) also has been established for much more general models, allowing dependence among the interarrival times and service times; see Abate et al. (1994), Choudhury et al. (1996) and references therein. If indeed, the limit (21) holds for W, then we easily get corresponding bounds for W_y.

We remark that logarithmic asymptotics from Glynn and Whitt (1994) supports the weaker approximation

$$P(W_y > b) \approx P(W > b) \approx e^{-\theta^* b}, \quad b \ge 0. \tag{22}$$

The following corollary draws implications from the limit (20), from the bounds we have established, assuming that the limit (20) is valid.

COROLLARY 3. (*tail-limit bounds*) *If* $e^{\theta^* b} P(W > b) \to A$ *as* $b \to \infty$ *for some* $\theta^* > 0$*, then*

$$\begin{aligned} \limsup_{b\to\infty} e^{\theta^* b} P(W_y > b) &\le \lim_{b\to\infty} e^{\theta^* b} P(W > b + \zeta_y^+) = A_y^+ \equiv Ae^{-\zeta_y^+ \theta^*} \quad \textit{and} \\ \liminf_{b\to\infty} e^{\theta^* b} P(W_y > b) &\ge \lim_{b\to\infty} e^{\theta^* b} P(W > b + \zeta_y^-) = A_y^- \equiv Ae^{-\zeta_y^- \theta^*}. \end{aligned} \tag{23}$$

as $b \to \infty$*. If* $e^{\theta^* b} P(W_y > b) \to A_y$ *as* $b \to \infty$*, then*

$$A_y^- \le A_y \le A_y^+ \quad \textit{and} \quad A_y^- \le A \le A_y^+. \tag{24}$$

For the $GI/GI/1$ model, we have the Cramer-Lundberg inequality for W in Theorem XIII.5.1 of Asmussen (2003), yielding $P(W > b) \le e^{-\theta^* b}$ for all b.

COROLLARY 4. (*periodic Cramer-Lundberg bound*) *For the periodic $GI_t/GI/1$ model,*

$$P(W_y > b) \le e^{-\theta^*(b+\zeta_y^+)} \quad \textit{for all} \quad b > 0.$$

4. Simulation Methodology for the $GI_t/GI/1$ Model

We now apply the representation in (2) and the bounds in §3 to obtain an effective rare-event simulation method for the periodic $GI_t/GI/1$ queueing model. Our approach is to first generate exponentially tilted interarrival times and service times until a process involving them hits a given level b and then to calculate an estimate of tail probability using these generated values for each simulation replication. Hence, the algorithm is primarily deterministic calculations. We obtain estimates of statistical precision by performing a large number of independent replications.

4.1. Exponential Tilting for the $GI/GI/1$ Model

We apply the familiar rare-event simulation method for the stationary $GI/GI/1$ model using importance sampling with an exponential change of measure, as in §XIII of Asmussen (2003) and §§V and VI of Asmussen and Glynn (2007). For the discrete-time waiting times in the $GI/GI/1$ model based on $\{(\rho^{-1}U_k, V_k)\}$, where $\{U_k\}$ and $\{V_k\}$ are independent sequences of i.i.d. nonnegative mean-1 random variables, the key random variables are $X_k(\rho) \equiv V_k - \rho^{-1}U_k$. We assume that U_k, V_k and thus $X_k(\rho)$ have finite moment generating functions (mgf's) $m_U(\theta)$, $m_V(\theta)$, and $m_X(\theta) \equiv m_{X(\rho)}(\theta)$, e.g., $m_V(\theta) \equiv E[e^{\theta V_k}]$, and probability density functions (pdf's) f_U, f_V and $f_X \equiv f_{X(\rho)}$. As usual, we define the twisted pdf $f_{X,\theta}(x) = e^{\theta x} f_X(x)/m_X(\theta)$ and for our simulation use the "optimal value" θ^* such that $m_X(\theta^*) = 1$. That optimal tilting parameter coincides with the asymptotic decay rate θ^* in Corollary 3.

There are several simplifications that facilitate implementation. First, as in Example XIII.1.4 of Asmussen (2003), we can construct the tilted pdf $f_{X,\theta}(x)$ by constructing associated tilted pdf's of f_U and f_V, in particular, because $X_k(\rho) \equiv V_k - \rho^{-1}U_k$, it suffices to let $f_{V,\theta}(x) = e^{\theta x} f_V(x)/m_V(\theta)$ and

$$f_{-U/\rho,\theta}(x) = \frac{e^{\theta x} f_{-U/\rho}(x)}{m_{-U/\rho}(\theta)} \quad \text{or} \quad \frac{e^{-\theta y/\rho}\rho f_U(y)}{m_U(-\theta/\rho)} \tag{25}$$

with the second expression obtained after making a change of variables, so that $m_X(\theta) = m_V(\theta)m_U(-\theta/\rho)$. We thus obtain the i.i.d. tilted random variables with pdf $f_{X,\theta^*}(x)$ by simulating independent sequences of i.i.d. random variables with the pdf's $f_{V,\theta^*}(x)$ and $f_{-U/\rho,\theta^*}(x)$.

Second, for all our examples, we consider common distributions that produce twisted pdf's having the same form as the original pdf's; it is only necessary to change the parameters. In particular, this property holds for the M, H_2, E_k and $M+D$ distributions that we propose to exploit in §6.4. In particular, if V is a rate-μ exponential (M) random variable with pdf $f_V(x) = \mu e^{-\mu x}$, then $f_{V,\theta}(x)$ is again an exponential random variable with parameter $\mu - \theta$, where we are required to have $\mu > \theta > 0$. Moreover, for the $M/M/1$ queue with arrival rate λ and service rate μ, the associated optimal tilted parameters are $\lambda_{\theta^*} = \mu$ and $\mu_{\theta^*} = \lambda$; i.e., the optimal tilting just switches the arrival and service rates; see Example XIII.1.5 of Asmussen (2003).

If V has an H_2 pdf $f_V(x) = p\mu_1 e^{-\mu_1 x} + (1-p)\mu_2 e^{-\mu_2 x}$, having parameter triple (p, μ_1, μ_2), then $f_{V,\theta}(x)$ again has an H_2 distribution, but with a new parameter triple $(p_\theta, \mu_{1,\theta}, \mu_{2,\theta})$, where $\mu_{j,\theta} = \mu_j - \theta$ and $p_\theta = [p\mu_1/(\mu_1 - \theta)/\{[p\mu_1/(\mu_1 - \theta)] + [(1-p)\mu_2/(\mu_2 - \theta)]\}$. We remark that the twisted H_2 pdf does not inherit the balanced-means property of the original H_2 pdf and has a different squared coefficient of variation (scv, variance divided by the square of the mean), but still $c^2 > 1$.

We now turn to the pdf's with scv $c^2 < 1$. First, a twisted E_k distribution is again E_k. More generally (because E_k is a special gamma distribution), if V has a gamma pdf $f_V(x;\alpha,\mu) = \mu^\alpha x^{\alpha-1} e^{-\mu x}/\Gamma(\alpha)$, then $f_{V,\theta}(x)$ has a gamma pdf with parameter pair $(\alpha_\theta, \mu_\theta) = (\alpha, \mu - \theta)$; see §V.1.b of Asmussen and Glynn (2007). Finally, if V is an $M+D$ distribution with parameter pair (d, μ), then the twisted distribution is an $M+D$ distribution with parameter pair $(d, \mu - \theta)$.

As a consequence, we can generate the tilted random variables in the standard way given underlying uniform random variables; e.g., we can apply the function $h(x) = -\log(1-x)/\mu$ to a vector of uniform random variables to obtain the corresponding vector of exponential random variable with mean $1/\mu$. For each H_2 random variable we can use two uniforms, one to select the exponential component and the other to generate the appropriate exponential; i.e., a random variable X with the H_2 distribution having parameter triple (p, μ_1, μ_2) can be expressed in terms of the pair of i.i.d. uniforms (Z_1, Z_2) as

$$X = -((1/\mu_1)1_{\{Z_1 \le p\}} + (1/\mu_2)1_{\{Z_1 > p\}}) \log(Z_2), \tag{26}$$

where 1_A is the indicator variable with $1_A = 1$ on the event A.

4.2. Rare-Event Simulation for Stationary Waiting Time in $GI/GI/1$ Model

Let W^* denote the steady-state discrete-time waiting time, which coincides with the steady-state continuous-time workload W in the $GI/GI/1$ model for Poisson arrivals, but not otherwise. The heavy-traffic limits coincide, as can be seen from Theorem 9.3.4 of Whitt (2002).

The standard simulation for rare-event probability of large waiting times in the $GI/GI/1$ model is achieved by performing the change of measure using the tilted interarrival times and service times, as indicated in §4.1, where the tilting parameter θ^* coincides with the asymptotic decay rate in §3.2, as described in Ch. XIII of Asmussen (2003) and §VI.2a of Asmussen and Glynn (2007).

To implement the simulation, we generate the random variables U_k and V_k from their tilted distributions with θ^*. We estimate the tail probability of stationary waiting time $P(W^* > b)$ by its representation as $P(\tau_b^S < \infty)$, where τ_b^S is the first hitting time of S_n at level b, with $S_n \equiv \sum_{k=1}^n X_k(\rho)$. The tail probability can be expressed in terms of the stopped sum $S_{\tau_b^S}$ using the underlying probability measure P_{θ^*}. Note that $S_{\tau_b^S} = b + Y(b)$, where $Y(b)$ is the overshoot of b by $\{S_n\}$, all under P_{θ^*}. Under the new probability measure P_{θ^*}, S_n hits b with probability 1, so we only need to estimate the likelihood ratio. Thus the tail probability of the $GI/GI/1$ steady-state waiting time W^* can be expressed as

$$\begin{aligned} P(W^* > b) &= P(\tau_b^S < \infty) = E_{\theta^*}[I\{\tau_b^S < \infty\} L_{\tau_b^S}(\theta^*)] = E_{\theta^*}[L_{\tau_b^S}(\theta^*)] \\ &= E_{\theta^*}[m_X(\theta^*)^{\tau_b^S} e^{-\theta^* S_{\tau_b^S}}] = E_{\theta^*}[e^{-\theta^* S_{\tau_b^S}}] = e^{-\theta^* b} E_{\theta^*}[e^{-\theta^* Y_S(b)}], \end{aligned} \tag{27}$$

where $L_{\tau_b^S}(\theta^*)$ is the likelihood ratio of $\{X_k(\rho)\}_{1 \le k \le \tau_b^S}$ with respect to P_{θ^*} . The second moment of this estimator is $E_{\theta^*}[L_{\tau_b^S}(\theta^*)^2] = E_{\theta^*}[e^{-2\theta^* S_{\tau_b^S}}]$. Theorem XIII.7.1 of Asmussen (2003) shows that the rare-event estimator of $P(W > b)$ has relative error that is uniformly bounded in b as $b \to \infty$. (The proof of Theorem XIII.7.1 relies on Theorems XIII.5.1-3; the pdf assumption implies that X has a nonlattice distribution.)

4.3. Rare-Event Simulation for Stationary Workload in $GI/GI/1$ Model

We are interested in the rare-event probability of large stationary workload W as in (10), where arrival process N is an equilibrium renewal process, because this is the process that we used to develop bounds of W_y in section 3. The classical exponential tilting method

applies to simulating the rare-event probability of stationary waiting time W^* as reviewed in §4.2. The stationary waiting time is as in (10) with N being the renewal process without the exceptional first inter-renewal time. To apply this exponential tilting method to stationary workload W, we need to make a slight modification of the algorithm above.

Now the equilibrium renewal process N has the exceptional first interarrival time and a constant rate ρ. We still use the usual partial sum process $S_n \equiv \sum_{k=1}^{n}(V_k - \rho^{-1}U_k)$, where V_k are still i.i.d with distribution V, but U_1 has the equilibrium distribution of U_e and $U_k, k \geq 2$ are i.i.d with distribution U. We do the same tilting for all $X_k(\rho)$'s still using P_{θ^*}, with $dP_{\theta^*}(x) = [e^{\theta^* x}/m_X(\theta^*)]dP(x)$. Note that θ^* is solved from $m_{X_k}(\theta^*) = 1$, where $k \geq 2$ and when $k = 1$, this equation may not hold. Now the likelihood ratio becomes

$$\begin{aligned} L_{\tau_b^S}(\theta^*) &= m_{X_1}(\theta^*) \times m_{X_2}(\theta^*) \times ... \times m_{X_{\tau_b^S}}(\theta^*)/(e^{\theta(X_1+X_2+...+X_{\tau_b^S})}) \\ &= m_{X_1}(\theta^*)e^{-\theta S_{\tau_b^S}}, \end{aligned}$$

where the second line follows because $m_{X_k}(\theta^*) = 1$.

Then we need to add a constant multiplier $m_{X_1}(\theta^*)$ to equation (27):

$$\begin{aligned} P(W > b) &= P(\tau_b^S < \infty) \\ &= E_{\theta^*}[L_{\tau_b^S}(\theta^*)] \\ &= E_{\theta^*}[m_{X_1}(\theta^*)m_X(\theta^*)^{\tau_b^S - 1}e^{-\theta^* S_{\tau_b^S}}] \\ &= E_{\theta^*}[m_{X_1}(\theta^*)e^{-\theta^* S_{\tau_b^S}}] \\ &= m_{X_1}(\theta^*)e^{-\theta^* b}E_{\theta^*}[e^{-\theta^* Y_S(b)}]. \end{aligned} \tag{28}$$

Note that (28) is also different from (27) in that the first $X_1(\rho)$ in the partial sum $S_{\tau_b^S}$ may have a different distribution from $\{X_k(\rho), k \geq 2\}$. The exact form of $m_{X_1}(\theta^*)$ is as below

$$\begin{aligned} m_{X_1}(\theta^*) &= E\{\exp\{\theta^* V - \theta^* \rho^{-1} U_e\}\} \\ &= E\{\exp\{-\theta^* \rho^{-1} U_e\}\}/E\{\exp\{-\theta^* \rho^{-1} U\}\}. \end{aligned}$$

where the second line still follows from $m_{X_k}(\theta^*) = 1$ and thus $E\{\exp\{\theta^* V\}\} = 1/E\{\exp\{-\theta^* \rho^{-1} U\}\}$.

Given that the estimator in (27) has bounded relative error as b goes to infinity, the estimator in (28) has bounded relative error as b goes to infinity as well. This is because

when b is large, the first X_1 does not influence the distribution of the overshoot $Y_S(b)$ and thus $Y_S(b)$ has the same distribution under P_{θ^*} in both estimators.

Table 1 shows simulation estimates for the workload tail probabilities $P(W > b)$ and the associated waiting-time tail probabilities $P(W^* > b)$ using the algorithms in §4.3 and §4.2 respectively. In both cases, we refer to the estimates as $P(W > b) \equiv \hat{p} = Ae^{-\theta^* b}$, where θ^* is common to both. We use a very small $\rho = 0.1$ here so that workload and waiting time probabilities are very different. These numerical results match the exact values of $\hat{p}$ and A calculated from Theorem X.5.1 of Asmussen (2003).

Table 1 Comparison of the steady-state workload and waiting-time tail probabilities for $b = 4, 20$ in the stationary $H_2/M/1$ queue with $\rho = 0.1$. The exact values are calculated from Theorem X.5.1 of Asmussen (2003).

	workload	waiting time	workload	waiting time
ρ	0.1	0.1	0.1	0.1
θ^*	0.8690	0.8690	0.8690	0.8690
exact A	0.1	0.1310	0.1	0.1310
exact p	0.003093	0.004050	2.83E-09	3.70E-09
b	4	4	20	20
$\hat{p}$	0.003104	0.004055	2.84E-09	3.69E-09
$e^{-\theta^* b}$	0.0309	0.0309	2.83E-08	2.83E-08
A	0.1004	0.1311	0.1004	0.1305
s.e.	2.73E-05	3.55E-05	2.49E-11	3.25E-11
%95 CI lb	0.003050	0.003985	2.79E-09	3.63E-09
%95 CI ub	0.003157	0.004125	2.89E-09	3.76E-09
r.e.	0.008788	0.008765	0.008771	0.008792

4.4. Applying the Bounds to Treat the Periodic Case

From (2), we see that any positive b must be hit for the first time at an arrival time. Thus, we have the alternative discrete-time representation

$$W_y = \sup_{n\geq 0}\Big\{\sum_{k=1}^{n} V_k - \tilde{\Lambda}_y^{-1}(N^{-1}(n))\Big\} = \sup_{n\geq 0}\Big\{\sum_{k=1}^{n} V_k - \tilde{\Lambda}_y^{-1}(\sum_{k=1}^{n} U_k)\Big\}, \tag{29}$$

where U_k is the k^{th} interarrival time in the equilibrium renewal process N, i.e. U_1 assumes the equilibrium distribution U_e while $\{U_k, k \geq 2\}$ are i.i.d. with distribution U.

For the periodic $GI_t/GI/1$ model with $\bar{\lambda} = \rho$, we can apply a variant of the exponential change of measure for the waiting times in the $GI/GI/1$ model in §4.1 above. We use the underlying measure P_{θ^*} determined for $GI/GI/1$. we use the usual partial sum process $S_n \equiv \sum_{k=1}^{n} X_k(\rho)$ for $GI/GI/1$ and the associated process

$$R_n \equiv \sum_{k=1}^{n} V_k - \tilde{\Lambda}_y^{-1}(\sum_{k=1}^{n} U_k). \tag{30}$$

We estimate the tail probability $P(W_y > b)$ by its representation as $P(\tau_b^R < \infty)$, where τ_b^R is the first hitting time of R_n at level b. Under the new probability measure, R_n hits b with probability 1, so we only need to estimate the likelihood ratio. We still twist $X_k(\rho) = V_k - \rho^{-1}U_k$ in the same way, which is equivalent to twisting V_k and $\rho^{-1}U_k$ separately, as discussed in §4.1. Then the likelihood ratio for $\{X_k(\rho) : 1 \le k \le n\}$ is the same as before, i.e., $L_n(\theta) = m_{X_1}(\theta)m_X(\theta)^{(n-1)}e^{-S_n}$. As a consequence, we obtain the representation

$$\begin{aligned} P(W_y > b) &= P(\tau_b^R < \infty) = E_{\theta^*}[L_{\tau_b^R}(\theta^*)] \\ &= E_{\theta^*}[m_{X_1}(\theta^*)m_X(\theta^*)^{(\tau_b^R-1)}e^{-\theta^* S_{\tau_b^R}}] = m_{X_1}(\theta^*)E_{\theta^*}[e^{-\theta^* S_{\tau_b^R}}]. \end{aligned} \tag{31}$$

Still note that the first $X_1(\rho)$ in the partial sum $S_{\tau_b^R}$ has a different distribution from $\{X_k, k \ge 2\}$.

At first glance, (31) does not look so useful, because the random sum $S_{\tau_b^R}$ involves the hitting time τ_b^R for $\{R_n\}$ instead of $\{S_n\}$, but we can shift the focus to $R_{\tau_b^R}$ because we can bound the difference between $S_{\tau_b^R}$ and $R_{\tau_b^R}$.

LEMMA 3. (*bound on difference of random sums*) *Under the assumptions above,*

$$|S_{\tau_b^R} - R_{\tau_b^R}| \le \zeta \equiv \max\{|\zeta_y^+|, \zeta_y^-\}, \tag{32}$$

where ζ_y^+ and ζ_y^- are the one-sided bounds in (13) *and* (14)*. In addition, $\tau_{b-\zeta}^S \le \tau_b^R \le \tau_{b+\zeta}^S$.*

Proof. The bound in (32) follows immediately from (13) and (14), because

$$|R_n - S_n| = |\Big(\sum_{k=1}^n V_k - \tilde{\Lambda}_y^{-1}\sum_{k=1}^n U_k\Big) - \Big(\sum_{k=1}^n V_k - \sum_{k=1}^n \rho^{-1}U_k\Big)| \le \zeta \equiv \max\{|\zeta_y^+|, \zeta_y^-\} \tag{33}$$

for all $n \ge 1$, where ζ_y^+ and ζ_y^- are the one-sided bounds in (13) and (14). ∎

Lemma 3 allows us to focus on $R_{\tau_b^R}$, where τ_b^R is the hitting time for $\{R_n\}$. To do so, we impose an additional regularity condition. The regularity condition requires the excess service-time distribution in probability measure P_{θ^*} be bounded above in stochastic order by a proper cdf, i.e.,

$$P_{\theta^*}(V > t + x | V > t) \equiv \frac{P_{\theta^*}(V > t + x)}{P_{\theta^*}(V > t)} \le G^c(x) \quad \text{for all} \quad t \ge 0, \tag{34}$$

where $G^c(x) \equiv 1 - G(x) \to 0$ as $x \to \infty$. For example, it suffices for the service time to be bounded. It also suffices for the service-time distribution to have an exponential tail, which holds if there is a constant $\eta > 0$ such that

$$e^{\eta x} P_{\theta^*}(V > x) \to L, \quad 0 < L < \infty \quad \text{as} \quad x \to \infty. \tag{35}$$

If (35) holds, then

$$\frac{e^{\eta(t+x)}P_{\theta^*}(V>t+x)}{e^{\eta t}P_{\theta^*}(V>t)} \to 1 \quad \text{as} \quad t\to\infty, \tag{36}$$

so that (34) holds asymptotically with $G^c(x)\equiv e^{-\eta x}$. It holds over any bounded interval because the ratio is continuous and bounded, given (35). Of course, condition (34) would not hold if $x^pP_{\theta^*}(V>x)\to L$ as $x\to\infty$ for $0<L<\infty$ and $p>0$.

THEOREM 1. (*bounded relative error*) *The rare-event simulation algorithm for the tail probability* $P(W_y>b)$ *in the periodic* $GI_t/GI/1$ *queue is unbiased and, if the service-time distribution satisfies condition* (34)*, then the rare-event simulation algorithm produces relative error that is uniformly bounded in* b*, just as for the stationary* $GI/GI/1$ *model, provided that the conditions for the rare-event simulation in the* $GI/GI/1$ *model are imposed so that the estimates are unbiased with bounded relative error.*

Proof. The unbiasedness follows from (31). Lemma 3 allows us to focus on $R_{\tau_b^R}$. The remaining result parallels Theorem XIII.7.1 in Asmussen (2003) for the $GI/GI/1$ model, which draws on Theorems XIII.5.1-3. Just as $S_{\tau_b^S}=b+Y_S(b)$, where $Y_S(b)$ is the overshoot of b upon first passage to b in the random walk $\{S_n\}$, so is $R_{\tau_b^R}=b+Y_R(b)$, where $Y_R(b)$ is the overshoot of b upon first passage to b in the sequence $\{R_n\}$. The results for the stationary case are based on the well developed theory for that overshoot, which depend on the random walk structure. In contrast, less is known for $\{R_n\}$. However, we do see from (29) that the overshoot can be regarded as an excess-distribution of the last service time. Thus, under the extra condition (34), we can again apply the proof in Asmussen (2003), using

$$e^{-k\theta^* b}\ge E_{\theta^*}[e^{-k\theta^* R_{\tau_b^R}}]\ge e^{-k\theta^* b}E_{\theta^*}[e^{-k\theta^* Y_R(b)}]\ge ce^{-k\theta^* b}$$

for $0<c<1$, where $c=E[e^{-k\theta^* Z}]$, $P(Z>x)=G^c(x)$, $x\ge 0$, and k is a positive integer. ∎

4.5. The Mean and Variance

We now show how tail-integral representations of the mean and higher moments on p. 150 of Feller (1971) can be exploited to obtain corresponding rare-event simulations of these related quantities. Recall that, for any nonnegative random variable X, the mean can be expressed as

$$E[X]=\int_0^\infty P(X>t)\,dt, \tag{37}$$

while the corresponding representation of the p^{th} moment for any $p > 1$ is

$$E[X^p] = \int_0^\infty pt^{p-1} P(X > t)\, dt. \tag{38}$$

To obtain a finite algorithm, it is natural to approximate the integrals for the mean and the second moment by finite sums plus a tail approximation, i.e.,

$$\begin{aligned} E[W_y] &\approx \sum_{k=0}^{n} (P(W_y > k\delta)\delta) + \frac{P(W_y > n\delta)}{\theta^*} \\ E[W_y^2] &\approx \sum_{k=0}^{n} (2P(W_y > k\delta)k\delta) + 2P(W_y > n\delta)(\frac{n\delta}{\theta^*} + \frac{1}{(\theta^*)^2}). \end{aligned} \tag{39}$$

In each case, the second term is based on applying the tail integral formula over $[n\delta, \infty)$ with the approximation

$$P(W_y > n\delta + x) \approx P(W_y > n\delta) e^{-\theta^* x} \tag{40}$$

and integrating.

To understand how to choose the discretization parameter δ in (39), suppose that $P(W > t) = ae^{-\theta^* t}$. In that case, the infinite sum for the mean can be expressed as

$$\sum_{k=0}^{\infty} a\delta e^{-\theta^* k\delta} = \frac{a}{\theta^*} \left(1 + \theta^* \frac{\delta}{2} + O(\delta^2)\right) \quad \text{as} \quad \delta \downarrow 0,$$

so that the relative error for the mean is $\theta^*(\delta/2) + O(\delta^2)$. Similarly, the corresponding calculation for the second moment indicates an asymptotic relative error proportional to $\theta^*\delta$. The subsequent truncation approximations involving n imposes no additional error, provided that the tail is exponential, which is likely to hold in view of §3.2. Thus, the truncation is good provided that approximation (40) is good, which can be checked with the algorithm.

In closing, we remark that because $\theta^*(\rho)$ tends to be of order $1 - \rho$ as $\rho \uparrow 1$, as explained in §2.2 of Ma and Whitt (2016), we can maintain fixed relative error in the discretization if we let δ be inversely proportional to $1 - \rho$ or $\theta^*(\rho)$ as $\rho \uparrow 1$. That can be useful because otherwise the computational complexity increases as ρ increases, as we show in the next sections. We illustrate letting δ increase with increasing ρ in Table 10.

4.6. The Algorithm

This exponential tilting algorithm to estimate tail probabilities $P(W_y > b)$ in the $GI_t/GI/1$ queue is based on equation (31) with the following steps. (We elaborate on Steps 4 and 5 in Ma and Whitt (2016).) Without loss of generality, we assume service rate is $\mu = 1$ and thus $\bar{\lambda} = \rho$.

Step 1. Before we conduct the simulation, we first **construct a table of the inverse cumulative arrival-rate function** $\rho\tilde{\Lambda}_y^{-1}$, i.e., the inverse of the reverse-time cumulative arrival-rate function $\tilde{\Lambda}_y$ in (3) scaled by ρ, for each time yc in the cycle to be considered. For that purpose, we use Algorithm 1 in Ma and Whitt (2015). That algorithm constructs an approximation J_y to the inverse function $\rho\tilde{\Lambda}_y^{-1}$ for one cycle from the interval $[0, c]$ to the interval $[0, c]$. This table is the same for a fixed y no matter what value ρ takes, which will be used for efficiently calculating $\tilde{\Lambda}_y^{-1}$ later. The computational complexity has shown to be of order $O(c/\epsilon)$, where c is the length of a cycle of the periodic arrival-rate function and ϵ is an allowed error tolerance.

Step 2. Again, before we conduct the simulation, we **determine the required number of partial sums** needed in each replication, which we denote by n_s. Note that we need this step because Matlab is much faster in vector operations than in loops. However, if another software is used to implement this algorithm, we can skip this step and generate exponentially twisted service times and interarrival times one by one in a loop until the hitting time τ_b^R is reached. Given the largest b under consideration, we estimate the expected number by $m_s \equiv b/E_{\theta^*}[V_k - \rho^{-1}U_k]$ by approximating the sum by Brownian motion which is asymptotically correct as b gets large, e.g. by §5.7.5 of Whitt (2002). If we use a Brownian motion approximation for the random walk, then we can get that the approximate mean and variance by applying Theorems 5.7.13 and 5.7.9 of Whitt (2002). For the canonical Brownian motion in Theorem 5.7.13, the variance of the first passage time is equal to the mean, but in general the ratio of the variance to the mean is proportional to the scv $c_X^2 \equiv Var(X)/E[X]^2$. Hence, we use $n_s = \max\{C, Lm_s\}$, where C is a minimum number like 100 and L is a safety-factor multiplier to account for the stochastic variability, which might be taken to be simply 10, but could be constructed more carefully. The largest value of b will depend on the case. If we want to treat multiple cases at once for simulation efficiency, we need to determine the largest required value of n_s. If m_s is large, then it is natural to use $n_s = m_s + 5\sqrt{c_X^2 m_s}$ instead of $n_s = 10m_s$, because then $5\sqrt{c_X^2 m_s}$ is about 5 standard deviations, which should be sufficient, and beneficial if $5\sqrt{c_X^2 m_s} << (L-1)m_s$.

Step 3. As the first part of the actual stochastic simulation, for each replication we now **generate the required random vectors of tilted interarrival times and service times**; For each replication, generate $\tilde{V} \equiv (V_1, ..., V_n)$ and $\rho^{-1}\tilde{U} \equiv (\rho^{-1}U_1, ..., \rho^{-1}U_n)$ where $n = n_s$ from step 2 above, V_k are i.i.d. random variables from $F_V^{\theta^*}$, the exponentially tilted distribution of V_k with parameter θ^* and $\rho^{-1}U_k$ i.i.d. from $F_{\rho^{-1}U}^{-\theta^*}$, the exponentially tilted distribution of $\rho^{-1}U_k$ with parameter $-\theta^*$. The distributions of V_k and U_k under the tilted probability measure P_{θ^*} were discussed in §4.1.

Step 4. Using vector operations, we **calculate the associated vectors of partial sums and transformed partial sums**. Use Algorithm 2 in Ma and Whitt (2015) to calculate the time-transformed arrival times.

Step 5. **Use (31) to calculate the tail probability $\mathbf{P(W_y > b)}$.** If n_s is not large enough to reach hitting times τ_b^R, we repeat Step 3 to generate additional vectors of $\tilde{V}$ and $\rho^{-1}\tilde{U}$ and repeat Step 4 to calculate additional partial sums and transformed partial sums. We treat the cases of the tail probability for a single value of b differently from multiple values of b, as required when we estimate moments. For multiple values of b, we use one loop to find all stopping times at each element of the vector b.

Step 6. **We run the algorithm for N i.i.d. replications.** Estimate $P(W_y > b)$, EW_y and EW_y^2 by the sample averages over the N replications. We estimate the associated confidence intervals in the usual way, using the Gaussian distribution if N is large enough and the Student-t distribution otherwise. ∎

In conclusion, we point out that there is flexibility in the order of the steps specified above. We can re-use random variables if we generate the random vectors in an early step. We can avoid storage problems if we perform calculations for each replication separately. As usual, there is a tradeoff in storage requirements and computation efficiency.

4.7. Computational Complexity and Running Times

We implemented the algorithm using matlab on a desktop computer. All examples were for the sinusoidal arrival-rate function λ in (4) with associated reverse-time cumulative arrival-rate function $\tilde{\Lambda}_y$ in (6). Because we used matlab, it was important to use vector calculations in step 3 to avoid loops.

We now specify the **computational complexity of the algorithm** above. Given the inverse function table for $\tilde{\Lambda}_y^{-1}$ computed in advance using the algorithm in Ma and Whitt (2015), the remaining algorithm has an approximate linear computational complexity of

$O(b/E_{\theta^*}[V_k - \rho^{-1}U_k])$, Specifically for the $M_t/M/1$ model, the computational complexity is $O(b\rho/(1-\rho))$, being directly proportional to b and inversely proportional to $1-\rho$. This can be made precise as $b \uparrow \infty$ or as $\rho \uparrow 1$, and presumably in some joint limit as $b/(1-\rho) \uparrow$, but we do not do that here. For b large or for ρ large, we can perform asymptotics to make the following approximations valid.

The hitting time τ_b of the random walk S_n as defined in (30) has expectation $E(\tau_b) = b/(E_{\theta^*}(V_k - \rho^{-1}U_k))$ by approximating S_n by a Brownian motion, for b that is very large compared to the step size of the random walk. Now consider the hitting time τ_b of R_n as defined in (30). Since the average arrival rate $\bar{\lambda} = \rho$, the expected value of this hitting time is approximately the same as that for S_n.

When both V_k and $\rho^{-1}U_k$ are exponential random variables with rates 1 and ρ respectively, under the new measure θ^*, they are still exponential with rates ρ and 1 respectively. Thus $b/E_{\theta^*}(V_k - \rho^{-1}U_k) = b/(1/\rho - 1) = b\rho/(1-\rho)$.

It can be advantageous to estimate the tail probabilities $P(W_y > b)$ for multiple values of b simultaneously. This can be done for each b by keeping track of the passage times for them while considering the largest value of b. This is very useful when we want to plot the cdf or its probability density function (pdf), or when we want to calculate the mean.

We now describe our experiments with **running times** on a desktop computer. Before conducting the simulation, we did step 1, constructing the table of the inverse function $\rho\tilde{\Lambda}_y^{-1}$ in one cycle, which takes computational time of $O(c/\epsilon) = O(1/\gamma\epsilon)$ by Theorem 3.1 of Ma and Whitt (2015), where c is the cycle length of the arrival rate function, γ is the parameter in the sinusoidal arrival-rate function and ϵ is the error bound we choose for the inverse function table. The longest cycle we consider has $\gamma = 0.00025$ (for (42) with $\rho = 0.99$), or $c = 25,120$. For $\epsilon = 10^{-4}$, it took 0.08 seconds to form the table needed for a single value of y.

In each replication, we can quickly determine the required length of the random variable vector, generate the vectors of random variables and calculate the partial sums, which are steps 2 to 4. The most time is required for step 5, searching for the stopping time for one b, or for all stopping times for a long vector of b. When we do the search for one b, the computational time is $O(b/(E_{\theta^*}[V_k - \rho^{-1}U_k])$, which is the approximate expected stopping time. When we do this for a long vector of b, we use a big loop which takes time linear in the maximum stopping time and the length of vector b, i.e., $O(max(b)/(E_{\theta^*}[V_k - \rho^{-1}U_k] +$

$length(b)$). Specifically, for the $M_t/M/1$ queue, the computational times are $O(b\rho/(1-\rho)$ and $O(max(b)\rho/(1-\rho)+length(b))$ respectively. For example, in $M_t/M/1$ queue, when $\rho = 0.8$, we choose $max(b) = log(1000)/\theta^* = log(1000)/(1-\rho)$, $\delta = 0.0002/(1-\rho)$, then maximum stopping time $O(max(b)\rho/(1-\rho))$ is negligible compared to the length of the vector b. The first part of time increases as ρ increases while the second part does not depend on ρ as both the largest b and δ are inversely proportional to $(1-\rho)$. In this case, when we did $40,000$ replications, the run time was 127 seconds on the desktop to find all stopping times, whereas it took about 10 seconds to find one stopping time for the largest b.

5. Simulation Examples

We now give examples to illustrate the new simulation algorithm. All our examples are for the sinusoidal arrival-rate function in (4) with parameter triple $(\bar{\lambda}, \beta, \gamma)$. More results appear in the online supplement.

5.1. Estimating the Tail Probabilities $P(W_y > b)$

We start by illustrating the efficiency of the rare-event simulation estimator of the tail probability $P(W_y > b)$, which gets exponentially small as b increases, and thus is prohibitively hard to estimate accurately by direct simulation. Table 2 shows that the relative errors of simulation estimates of $P(W_y > b)$ for the $M_t/M/1$ model in several cases are approximately independent of b. That property held in all models considered.

In particular, Table 2 shows estimates of $\hat{p} \equiv P(W_y > b) \equiv A_y e^{-\theta^* b}$ and the components A_y and $e^{-\theta^* b}$ for the special case $y = 0.0$ based on 5000 i.i.d. replications. Table 2 also shows estimates of the standard error (s.e.) of $\hat{p}$, the upper and lower bounds of the 95% confidence interval (CI), and the relative error (r.e.), which is the s.e. divided by the estimate of the mean. For Table 2, we used the arrival-rate function (4) with $\bar{\lambda} = 1$, and $E[V_1] = 0.8$, so that $\rho = 0.8$. We let $\beta = 0.2$ and consider three values of γ: 10, 1 and 0.1, making cycle lengths of 0.628, 6.28 and 62.8. The rapid fluctuation with $\gamma = 10$ makes the arrival process very similar to a homogeneous Poisson process, because the cumulative arrival-rate function approaches a linear function; see Theorem VIII.4.10 in Jacod and Shiryaev (1987), Problem 1 on p. 360 of Ethier and Kurtz (1986) and Whitt (2016). We also simulated the $M/M/1$ model with $\beta = 0$ to verify simulation correctness.

Table 2 shows that the rare-event simulation is effective for estimating $P(W_0 > b)$, because the relative error is approximately independent of b for each γ, ranging from about 0.0029 for $\gamma = 10$ to about 0.0055 for $\gamma = 0.1$.

Table 2 **Estimates of $\hat{p} \equiv P(W_y > b) \equiv A_y e^{-\theta^* b}$ in the $M_t/M/1$ model with sinusoidal arrival-rate function in (4) as a function of γ and b for: $\rho = 0.8, \bar{\lambda} = 1, \mu = 1.25$ and $\beta = 0.2$ based on 5000 replications.**

	b	$\hat{p}$	$exp(-\theta^* b)$	$A_0(b)$	s.e.	95% CI (lb)	(ub)	r.e.
$\gamma = 10$	10	0.0654	0.0821	0.797	1.87E-04	0.0651	0.0658	0.00286
	20	0.00537	0.00674	0.797	1.55E-05	0.00534	0.00540	0.00289
	40	3.61E-05	4.54E-05	0.795	1.05E-07	3.59E-05	3.63E-05	0.00290
	80	1.64E-09	2.06E-09	0.796	4.82E-12	1.63E-09	1.65E-09	0.00294
$\gamma = 1$	10	0.0628	0.0821	0.765	1.87E-04	0.0624	0.0632	0.00298
	20	0.00516	0.00674	0.766	1.51E-05	0.00513	0.00519	0.00292
	40	3.49E-05	4.54E-05	0.769	1.00E-07	3.47E-05	3.51E-05	0.00287
	80	1.58E-09	2.06E-09	0.767	4.65E-12	1.57E-09	1.59E-09	0.00294
$\gamma = 0.1$	10	0.0413	0.0821	0.503	2.33E-04	0.0409	0.0418	0.00565
	20	0.00360	0.00674	0.535	1.98E-05	0.00356	0.00364	0.00550
	40	2.50E-05	4.54E-05	0.551	1.37E-07	2.47E-05	2.53E-05	0.00548
	80	1.12E-09	2.06E-09	0.545	6.20E-12	1.11E-09	1.14E-09	0.00552

5.2. Unified Numerical Results Via Heavy-Traffic Scaling

We produce unified numerical results by exploiting heavy-traffic scaling. In particular, we scale the arrival rate function so that the performance measures have heavy-traffic limits as $\rho \uparrow 1$, which we explain in §6. In the special case of (4), we consider an arrival-rate function scaled by the overall traffic intensity ρ, specifically,

$$\lambda_\rho(t) = \rho + (1-\rho)\rho\beta \sin\left(\gamma(1-\rho)^2 t\right), \quad t \geq 0, \tag{41}$$

so that the cycle length in model ρ is $c_\rho = c^*(1-\rho)^{-2} = 2\pi/(\gamma(1-\rho)^2)$. After scaling, the cycle length is $c^* = 2\pi/\gamma$.

When we consider the periodic steady-state workload, we include spatial scaling by $1-\rho$ Hence, to have asymptotically convergent models, we should choose parameter four-tuples $(\bar{\lambda}_\rho, \beta_\rho, \gamma_\rho, b_\rho)$ indexed by ρ, where

$$(\bar{\lambda}_\rho, \beta_\rho, \gamma_\rho, b_\rho) = (\rho, (1-\rho)\beta, (1-\rho)^2\gamma, (1-\rho)^{-1}b), \tag{42}$$

where (β, γ, b) is a feasible base triple of positive constants with $\beta < 1$. (We must constrain $\beta_\rho \leq 1$ so that $\lambda_\rho(t) \geq 0$ for all t.) Hence, we have the ρ-dependent constraint $\rho_b = (1-\rho)\beta \leq 1$. There is no problem if $\beta \leq 1$, but we may want to consider $\beta > 1$. In that case, β_ρ is only well defined for $\rho \geq 1 - (1/\beta)$. For example, if $\beta = 5.0$, then we require that $\rho \geq 0.8$.

EXAMPLE 1. (Using $M_t/M/1$ to estimate the performance of RPBM)

To illustrate how we can apply simulations of the $M_t/M/1$ model with increasing traffic intensities, let the base parameter triple be $(\beta, \gamma, b) = (1.0, 2.5, 4.0)$. Then the parameter 4-tuple for $\rho = 0.8$ is

$$(\bar{\lambda}_\rho, \beta_\rho, \gamma_\rho, b_\rho) = (0.8, (1-0.8)\beta, (1-0.8)^2\gamma, (1-0.8)^{-1}b) = (0.8, 0.2, 0.1, 20.0). \tag{43}$$

The associated parameter 4-tuple for $\rho = 0.9$ is $(0.90, 0.10, 0.025, 40.00)$.

Let W be the steady-state workload in the stationary $M/M/1$ model with the same scaling, which has an exponential distribution except for an atom $1-\rho$ at the origin. Table 3 shows estimates of the ratio $P(W_y > b_\rho)/P(W > b_\rho)$ for 5 different values of $1-\rho$, where we successively divide $1-\rho$ by 2, and 8 different values of the position y within the cycle in the $M_t/M/1$ model with sinusoidal arrival-rate function in (41) with the parameter 4-tuple in (42) using the base parameter triple $(\beta, \gamma, b) = (1.0, 2.5, 4.0)$. (The paramter 4-tuples for $\rho = 0.8$ and $\rho = 0.9$ are shown above.)

Table 3 Comparison of the ratios $P(W_y > b_\rho)/P(W > b_\rho)$, where W is for the stationary model, for 5 different values of $1-\rho$ and 8 different values of the position y within the cycle in the $M_t/M/1$ model with sinusoidal arrival-rate function in (41) with the parameter 4-tuple in (42) using the base parameter triple $(\beta, \gamma, b) = (1.0, 2.5, 4.0)$.

y	$1-\rho = 0.16$	$1-\rho = 0.08$	$1-\rho = 0.04$	$1-\rho = 0.02$	$1-\rho = 0.01$
0.000	0.96364	0.96523	0.96424	0.96357	0.96344
0.125	0.97619	0.97686	0.97504	0.97493	0.97482
0.250	1.00456	1.00450	1.00255	1.00251	1.00305
0.375	1.03278	1.03264	1.03035	1.03152	1.03152
0.500	1.04565	1.04470	1.04278	1.04346	1.04405
0.625	1.03213	1.03096	1.03230	1.03150	1.03204
0.750	1.00225	1.00404	1.00425	1.00277	1.00241
0.875	0.97371	0.97696	0.97629	0.97457	0.97545
avg diff	0.00037	0.00112	0.00015	-0.00019	
avg. abs. dif	0.00099	0.00121	0.00081	0.00039	
rmse	0.00116	0.00134	0.00096	0.00049	

Table 3 shows that, for each fixed y, all estimates as a function of ρ serve as reasonable practical approximations for the others as well as for the RPBM limit developed in §6. The convergence in Table 3 is summarized by showing the average difference, average absolute difference and root mean square error (rmse) of the entry with the corresponding estimate for $\rho = 0.99$ in the final column, taken over 40 evenly spaced values of y in the interval $[0, 1)$.

5.3. Hyperexponential Examples

We now present results from simulation experiments with nonexponential service times and interarrival times in the base process N. In particular, we work with hyprexponential (H_2) examples.

Tables 4, 5 and 6 show estimates of $P(W_y > b)$ for the $M_t/M/1$, $M_t/H_2/1$ and $(H_2)_t/M/1$ models, respectively. All three tables show results for $y = 0.0$ and $y = 0.5$ as a function of $1-\rho$ with base parameter triple $(\beta, \gamma, b) = (1, 2.5, 4)$ in (42) based on 40,000 replications. The mean service time is fixed at $\mu^{-1} = 1$, so that $\bar{\lambda} = \rho$ in all cases. The scv of the H_2 cdf is always $c^2 = 2$. The scaling in (42) is performed as a function of ρ in order to produce nearly stable results in each row.

We start by showing the estimate of the tail probability $\hat{p} \equiv P(W_y > b) \equiv A_y e^{-\theta^* b}$. Then we show the corresponding estimates for the components $e^{-\theta^* b}$ and $A_y \equiv e^{\theta^* b}\hat{p}$. We then show the lower and upper bounds in (23) of Corollary 3. We then show the s.e., the associated 95% CI bounds (lb and ub), and the r.e. In all cases the relative error is less than 0.0015 or 0.15%.

For the two cases $y = 0.0$ and $y = 0.5$, we also display estimates of scaled tail probabilities, $P(W_y > b)/P(W > b)$, where $P(W > b)$ is the corresponding estimate for the stationary model. We do this because we seek estimates that are more stable as functions of $1-\rho$, and thus support approximations for the limiting RPBM tail probability, which is the scaled limit as $\rho \uparrow 1$. In Tables 5 and 6 for the $M_t/H_2/1$ and $(H_2)_t/M/1$ models we also show the alternative ratios $P(W_y > b)/\rho$; we do not show that for $M_t/M/1$ in Table 4 because the ratios are proportional, because $P(W > b) = \rho e^{-\theta^* b}$ for $M/M/1$ and $\theta^*(\rho) = 1-\rho$. Tables 5 and 6 show that greater stability is achieved with the ratio $P(W_y > b)/(W > b)$.

Tables 4, 5 and 6 strongly support the heavy-traffic limit in Theorem 2, establishing convergence to RPBM as $\rho \uparrow 1$. The stability of the scaled quantities is especially clear through the ratios $P(W_y > b)/P(W > b)$. For the ratios at the bottom of the tables, we also show the difference and absolute difference of the value with value in the final column of the table.

A close examination of Tables 5 and 6 show that there is a consistent sign in the differences in the second-to-last row, being positive for the $M_t/H_2/1$ in Table 5 and negative for the $(H_2)_t/M/1$ model Table 6. These consistent signs in Tables 5 and 6 suggest that the two cases $M_t/H_2/1$ and $(H_2)_t/M/1$ serve as one-sided bounds on RPBM. We provide

Table 4 Simulation estimates of $\hat{p} \equiv P(W_y > b) \equiv A_y e^{-\theta^* b}$ in the $M_t/M/1$ model for $y = 0.0$ and $y = 0.5$ as a function of $1-\rho$ with base parameter triple $(\beta, \gamma, b) = (1, 2.5, 4)$ in (42) based on $40,000$ replications.

$1-\rho$	0.16	0.08	0.04	0.02	0.01
$\hat{p}$ for $y = 0.0$	0.011053	0.012192	0.012814	0.013122	0.013263
$e^{-\theta^* b}$	0.0183	0.0183	0.0183	0.0183	0.0183
A_y	0.604	0.666	0.700	0.716	0.724
A_y^- LB in (23)	0.377	0.413	0.431	0.440	0.445
A_y^+ UB in (23)	0.840	0.920	0.960	0.980	0.990
s.e.	1.75E-05	1.69E-05	1.71E-05	1.73E-05	1.74E-05
95% CI (lb)	0.01102	0.01216	0.01278	0.01309	0.01323
(ub)	0.01109	0.01223	0.01285	0.01316	0.01330
r.e.	0.001582	0.001387	0.001333	0.001319	0.001313
$P(W_y > b)/P(W > b)$	0.71845	0.72356	0.72879	0.73103	0.73144
diff w.r.t. last column	0.01298	0.00788	0.00264	0.00041	0.00000
abs diff	0.01298	0.00788	0.00264	0.00041	0.00000
$\hat{p}$ for $y = 0.5$	0.025888	0.028396	0.029551	0.030110	0.030430
$e^{-\theta^* b}$	0.0183	0.0183	0.0183	0.0183	0.0183
A_y	1.413	1.550	1.613	1.644	1.661
A_y^- LB in (23)	0.840	0.920	0.960	0.980	0.990
A_y^+ UB in (23)	1.869	2.047	2.137	2.181	2.203
s.e.	3.87E-05	3.74E-05	3.80E-05	3.86E-05	3.89E-05
95% CI (lb)	0.02581	0.02832	0.02948	0.03003	0.03035
(ub)	0.02596	0.02847	0.02963	0.03019	0.03051
r.e.	0.001496	0.001318	0.001286	0.001281	0.001279
$P(W_y > b)/P(W > b)$	1.68266	1.68517	1.68068	1.67751	1.67821
diff w.r.t. last column	-0.00445	-0.00696	-0.00247	0.00071	0.00000
abs diff	0.00445	0.00696	0.00247	0.00071	0.00000

Table 5 Simulation estimates of $\hat{p} \equiv P(W_y > b) \equiv A_y e^{-\theta^* b}$ in the $M_t/H_2/1$ model for $y = 0.0$ and $y = 0.5$ as a function of $1-\rho$ with base parameter triple $(\beta, \gamma, b) = (1, 2.5, 4)$ in (42) based on $40,000$ replications.

$1-\rho$	0.16	0.08	0.04	0.02	0.01
$\theta^*(\rho)$	0.101	0.0519	0.0263	0.0132	0.00664
$\hat{p}$ for $y = 0.0$	0.050594	0.052946	0.054024	0.054544	0.054904
$e^{-\theta^* b}$	0.0807	0.0747	0.0720	0.0707	0.0701
A_y	0.627	0.708	0.750	0.771	0.783
A_y^- LB in (23)	0.477	0.532	0.560	0.573	0.580
A_y^+ UB in (23)	0.789	0.894	0.947	0.974	0.987
s.e.	7.49E-05	5.64E-05	5.13E-05	5.03E-05	5.01E-05
95% CI (lb)	0.05045	0.05284	0.05392	0.05445	0.05481
(ub)	0.05074	0.05306	0.05412	0.05464	0.05500
r.e.	0.001480	0.001065	0.000950	0.000923	0.000913
$P(W_y > b)/P(W > b)$	0.79534	0.79246	0.79200	0.79200	0.79377
diff w.r.t. last column	-0.00158	0.00131	0.00177	0.00177	0.00000
abs diff	0.00158	0.00131	0.00177	0.00177	0.00000
A_y/ρ	0.74662	0.76999	0.78125	0.78680	0.79107
diff w.r.t. last column	0.04445	0.02108	0.00982	0.00427	0.00000
abs diff	0.04445	0.02108	0.00982	0.00427	0.00000
$\hat{p}$ for $y = 0.5$	0.086646	0.092721	0.095707	0.096711	0.097186
$e^{-\theta^* b}$	0.0807	0.0747	0.0720	0.0707	0.0701
A_y	1.074	1.241	1.329	1.367	1.386
A_y^- LB in (23)	0.789	0.894	0.947	0.974	0.987
A_y^+ UB in (23)	1.305	1.502	1.603	1.654	1.679
s.e.	1.25E-04	9.42E-05	8.49E-05	8.28E-05	8.28E-05
95% CI (lb)	0.08640	0.09254	0.09554	0.09655	0.09702
(ub)	0.08689	0.09291	0.09587	0.09687	0.09735
r.e.	0.001442	0.001016	0.000887	0.000856	0.000852
$P(W_y > b)/P(W > b)$	1.36208	1.38777	1.40307	1.40428	1.40505
diff w.r.t. last column	0.04297	0.01728	0.00198	0.00077	0.00000
abs diff	0.04297	0.01728	0.00198	0.00077	0.00000
A_y/ρ	1.27865	1.34842	1.38403	1.39507	1.40028
diff w.r.t. last column	0.12163	0.05186	0.01625	0.00521	0.00000
abs diff	0.12163	0.05186	0.01625	0.00521	0.00000

Table 6 Simulation estimates of $\hat{p} \equiv P(W_y > b) \equiv A_y e^{-\theta^* b}$ in the $(H_2)_t/M/1$ model for $y = 0.0$ and $y = 0.5$ as a function of $1 - \rho$ with base parameter triple $(\beta, \gamma, b) = (1, 2.5, 4)$ in (42) based on $40,000$ replications.

$1-\rho$	0.16	0.08	0.04	0.02	0.01
$\theta^*(\rho)$	0.113	0.0548	0.0270	0.0134	0.00669
$\hat{p}$ for $y=0$	0.038876	0.046701	0.050799	0.053020	0.053985
$e^{-\theta^* b}$	0.0593	0.0645	0.0670	0.0682	0.0689
A_y	0.655	0.724	0.758	0.777	0.784
A_y LB	0.477	0.532	0.559	0.573	0.580
A_y UB	0.840	0.920	0.960	0.980	0.990
s.e.	4.36E-05	4.56E-05	4.73E-05	4.88E-05	4.95E-05
95% CI (lb)	0.03879	0.04661	0.05071	0.05292	0.05389
(ub)	0.03896	0.04679	0.05089	0.05312	0.05408
r.e.	0.001123	0.000976	0.000932	0.000920	0.000917
$P(A_y>b)/P(A>b)$	0.78051	0.78763	0.78988	0.79280	0.79187
diff	0.01136	0.00424	0.00199	-0.00093	0.00000
abs diff	0.01136	0.00424	0.00199	0.00093	0.00000
A_y/ρ	0.78015	0.78747	0.78988	0.79279	0.79186
diff	0.01171	0.00439	0.00198	-0.00094	0.00000
abs diff	0.01171	0.00439	0.00198	0.00094	0.00000
$\hat{p}$ for $y=0.5$	0.071241	0.084111	0.090923	0.094201	0.096045
$e^{-\theta^* b}$	0.0593	0.0645	0.0670	0.0682	0.0689
A_y	1.201	1.305	1.357	1.380	1.395
A_y LB	0.840	0.920	0.960	0.980	0.990
A_y UB	1.477	1.592	1.648	1.677	1.691
s.e.	7.61E-05	7.71E-05	7.93E-05	8.13E-05	8.21E-05
95% CI (lb)	0.07109	0.08396	0.09077	0.09404	0.09588
(ub)	0.07139	0.08426	0.09108	0.09436	0.09621
r.e.	0.001068	0.000917	0.000873	0.000863	0.000855
$P(A_y>b)/P(A>b)$	1.43030	1.41856	1.41378	1.40857	1.40881
diff	-0.02149	-0.00975	-0.00497	0.00024	0.00000
abs diff	0.02149	0.00975	0.00497	0.00024	0.00000
A_y/ρ	1.42963	1.41826	1.41378	1.40856	1.40878
diff	-0.02085	-0.00948	-0.00500	0.00023	0.00000
abs diff	0.02085	0.00948	0.00500	0.00023	0.00000

strong theoretical support for this idea in Theorem 1 and Corollary 1 of Ma and Whitt (2016). Those results show that the one-sided bounds apply exactly to the asymptotic decay rates θ^*, which is the dominant part of the actual tail probability. For the cases considered in Table 6, it is natural to wonder if the refinement of the rare-event algorithm for the first non-exponential interarrival time makes much difference. We show that it does not for these cases with higher ρ in §4.6 of Ma and Whitt (2016).

Tables 4, 5 and 6 show that the bounds A_y^- and A_y^+ in (23) are not too close, and thus not good approximations for the actual A_y. Experiments show that the average of the two bounds is not a consistently good approximation for A_y either.

Simulation results over a wide range of y show that $P(W_y > b)$ consistently increases from a minimum at $y = 0$ to a maximum at $y = 0.5$ and then decreases to back to the minimum at $y = 1$, with The values for $y = 1/4$ and $y = 3/4$ being approximately equal to $P(W > b)$. It remains to establish theoretical supporting results.

5.4. Estimating the Moments of W_y

We now apply the extension of the algorithm in §4.5 to estimate the first two moments of W_y, reporting the estimated mean and standard deviation. In Table 7 we first show preliminary results for the stationary $M/M/1$ model, so that we can judge the algorithm

Table 7 **Estimated mean $E[W]$ and standard deviation $SD(W)$ as a function of $1-\rho$ for five cases of the stationary $M/M/1$ queue: $\mu = 1, \bar{\lambda} = \rho$**

$1-\rho$	0.16	0.08	0.04	0.02	0.01
n_s in (39)	40,000	40,000	40,000	40,000	40,000
δ in (39)	0.001	0.001	0.001	0.001	0.001
largest b	41	86	173	345	691
$P(W>0)$	0.8396	0.9201	0.9601	0.9799	0.9900
exact	0.8400	0.9200	0.9600	0.9800	0.9900
s.e. of $P(W>0)$	6.86E-04	3.71E-04	1.93E-04	9.73E-05	4.98E-05
%95 CI of $P(W>0)$	[0.8383, 0.8410]	[0.919, 0.921]	[0.9598, 0.9605]	[0.9797, 0.9801]	[0.9899, 0.9901]
$E[W]$	5.249	11.499	23.999	49.000	99.000
exact	5.250	11.500	24.000	49.000	99.000
s.e. of $E[W]$	1.59E-03	1.27E-03	9.51E-04	6.93E-04	4.94E-04
%95 CI of $E[W]$	[5.246, 5.252]	[11.497, 11.502]	[23.997, 24.001]	[48.999, 49.001]	[98.999, 99.001]
$E[W\|W>0]$	6.251	12.497	24.995	50.003	100.005
%95 CI of $E[W\|W>0]$	[6.238,6.265]	[12.485, 12.510]	[24.983, 25.007]	[49.992, 50.014]	[99.994, 100.015]
$E[W^2]$	65.624	287.494	1199.982	4899.957	19,800.03
exact	65.625	287.500	1200.000	4900.000	19,800.00
s.e. of $E[W^2]$	1.50E-02	2.33E-02	3.40E-02	4.92E-02	7.04E-02
%95 CI of $E[W^2]$	[65.595, 65.654]	[287.449, 287.540]	[1199.92, 1200.05]	[4899.86, 4900.05]	[19,799.89, 19,800.17]
$SD[W]$	6.170	12.460	24.981	49.990	99.995
exact	6.1695	12.450	24.980	49.990	99.995
$P(W>0)/\rho$	0.9995	1.0002	1.0001	0.9999	1.0000
exact	1.0000	1.0000	1.0000	1.0000	1.0000
$(1-\rho)E[W]$	0.8398	0.9200	0.9600	0.9800	0.9900
$(1-\rho)SD[W]$	0.9873	0.9968	0.9992	0.9998	0.9999
$(1-\rho)E[W]/\rho$	0.9998	0.9999	0.9999	1.0000	1.0000
$(1-\rho)SD[W]/\rho$	0.8293	0.9171	0.9593	0.9798	0.9899
$(1-\rho)E[W\|W>0]$	1.0002	0.9998	0.9998	1.0001	1.0000
$(1-\rho)SD[W\|W>0]$	1.0002	1.0000	1.0000	1.0000	1.0000

against known exact results. For ease of comparison, we show the corresponding known exact values for $P(W>0)$, $E[W]$, $E[W^2]$ and $SD(W)$. The first section of Table 7 with three rows shows the algorithm parameters. The final seven rows of Table 7 are included to show alternatives ways of scaling aimed at achieving stable values across all values of $1-\rho$. In this case, knowing that W has an exponential distribution except for an atom of mass $1-\rho$ at the origin, we are not surprised to see that the final two rows provide the best scaling. We will use those rows in the following tables for time-varying arrival-rate functions.

Tables 8 and 9 show corresponding estimates of the time varying mean $E[W_y]$ and standard deviation $SD(W_y)$ for the special case of $y = 0.5$ for associated $M_t/M/1$ model with arrival-rate function in (4) for base parameter pairs $(\beta, \gamma) = (1, 2.5)$ and $(\beta, \gamma) = (4, 2.5)$ using the scaling convention in (42). Both have cycle length $2\pi/\gamma$, which equals

$6.28/0.1 = 62.8$ for $\rho = 0.8$. The higher relative amplitude in Table 9 leads to much larger mean values at $y = 0.5$, which tends to produce the largest values in the cycle. As can be seen from the online supplement, much lower values occur for $y = 0$, which tends to produce the least values.

Table 8 **Estimated mean $E[W_y]$ and standard deviation $SD(W_y)$ as a function of $1-\rho$ for five cases of the $M_t/M/1$ queue at $y = 0.5$: $\mu = 1, \bar{\lambda} = \rho$ and base parameter pair $(\beta, \gamma) = (1, 2.5)$**

$1-\rho$	0.16	0.08	0.04	0.02	0.01
n_s in (39)	40,000	40,000	40,000	40,000	40,000
δ in (39)	0.001	0.001	0.001	0.001	0.001
largest b	41	86	173	345	691
$P(W_y > 0)$	0.8801	0.9411	0.9714	0.9851	0.9930
s.e. of $P(W_y > 0)$	9.85E-04	6.54E-04	4.51E-04	2.92E-04	2.19E-04
%95 CI of $P(W_y > 0)$	[0.8782, 0.8820]	[0.9399, 0.9424]	[0.9705, 0.9723]	[0.9845, 0.9856]	[0.9926, 0.9934]
$E[W_y]$	6.839	14.927	31.194	63.667	128.411
std of $E[W_y]$	6.42E-03	1.20E-02	2.36E-02	4.69E-02	9.30E-02
%95 CI of $E[W_y]$	[6.827, 6.852]	[14.903, 14.950]	[31.147, 31.240]	[63.575, 63.759]	[128.228, 128.593]
$E[W_y \mid W_y > 0]$	7.771	15.860	32.113	64.632	129.315
%95 CI of $E[W_y \mid W_y > 0]$	[7.740, 7.803]	[15.814, 15.907]	[32.036, 32.189]	[64.501, 64.763]	[129.075, 129.554]
$E[W_y^2]$	97.057	427.685	1795.344	7344.665	29,673.77
std of $E[W_y^2]$	7.81E-02	0.302	1.207	4.829	19.314
%95 CI of $E[W_y^2]$	[96.90, 97.21]	[427.09, 428.28]	[1793.0, 1797.7]	[7335.2, 7354.13]	[29,636, 29,712]
$SD[W_y]$	7.091	14.314	28.676	57.369	114.824
$P(W_y > 0)/\rho$	1.0478	1.0230	1.0119	1.0052	1.0030
$(1-\rho)E[W_y \mid W_y > 0]$	1.2434	1.2688	1.2845	1.2926	1.2931
$(1-\rho)SD[W_y \mid W_y > 0]$	1.1301	1.1395	1.1433	1.1452	1.1472

Table 9 **Estimated mean $E[W_y]$ and standard deviation $SD(W_y)$ as a function of $1-\rho$ for five cases of the $M_t/M/1$ queue at $y = 0.5$: $\mu = 1, \bar{\lambda} = \rho$ and base parameter pair $(\beta, \gamma) = (4, 2.5)$ having larger relative amplitude**

$1-\rho$	0.16	0.08	0.04	0.02	0.01
n_s in (39)	40,000	40,000	40,000	40,000	40,000
δ in (39)	0.001	0.001	0.001	0.001	0.001
largest b	41	86	173	345	691
$P(W_y > 0)$	0.9728	0.9883	0.9967	0.9965	0.9993
s.e. of $P(W_y > 0)$	3.61E-03	2.69E-03	2.05E-03	1.16E-03	8.52E-04
%95 CI of $P(W_y > 0)$	[0.9657, 0.9799]	[0.9831, 0.9936]	[0.9927, 1.0000]	[0.9943, 0.9988]	[0.9976, 1.0000]
$E[W_y]$	15.148	33.583	70.677	145.183	294.222
std of $E[W_y]$	5.58E-02	1.13E-01	2.27E-01	4.59E-01	9.15E-01
%95 CI $E[W_y]$	[15.04, 15.26]	[33.36, 33.81]	[70.23, 71.12]	[144.3, 146.1]	[292.4, 296.0]
$E[W_y \mid W_y > 0]$	15.572	33.980	70.909	145.690	294.437
%95 CI of $E[W_y \mid W_y > 0]$	[15.35, 15.80]	[33.58, 34.39]	[70.2, 71.6]	[144.5, 147.0]	[292.4, 296.7]
$E[W_y^2]$	331.868	1528.127	6547.951	27,092.17	110,239.9
std of $E[W_y^2]$	1.023	4.263	17.227	69.632	0.785
%95 CI of $E[W_y^2]$	[329.9, 333.9]	[1519.8, 1536.5]	[6514, 6582]	[26,955, 27,228]	[109,691, 110,787]
$SD[W_y]$	10.119	20.007	39.405	77.551	153.861
$P(W_y > 0)/\rho$	1.1581	1.0743	1.0383	1.0169	1.0094
$(1-\rho)E[W_y \mid W_y > 0]$	2.4915	2.7184	2.8364	2.9138	2.9444
$(1-\rho)SD[W_y \mid W_y > 0]$	1.5892	1.5830	1.5704	1.5442	1.5371

Finally, Table 10 shows estimates of the time varying mean $E[W_y]$ and standard deviation $SD(W_y)$ for the special case of $y = 0.5$ for associated $(H_2)_t/M/1$ model with arrival-rate function in (4) for base parameter pairs $(\beta, \gamma) = (1, 2.5)$, but here we let δ increase as $1-\rho$

decreases. Table 10 shows that the precision remains good for all ρ. (For the cases considered in Table 10, the refinement of the rare-event algorithm for the first non-exponential interarrival time does not make too much difference, but it matters more than for Table 6, as we show in §4.6 of Ma and Whitt (2016).)

Table 10 **Estimated mean $E[W_y]$ and standard deviation $SD(W_y)$ as a function of $1-\rho$ for five cases of the $(H_2)_t/M/1$ queue at $y=0.5$: $\mu=1, \bar{\lambda}=\rho$ and base parameter pair $(\beta,\gamma)=(1,2.5)$.**

$1-\rho$	0.16	0.08	0.04	0.02	0.01
$\theta^*(\rho)$	0.113	0.0548	0.0270	0.0134	0.00669
n_s	40,000	40,000	40,000	40,000	40,000
δ	0.001	0.002	0.004	0.008	0.016
largest b	41	86	173	345	691
$P(W_y>0)$	0.8721	0.9382	0.9691	0.9853	0.9923
s.e. of $P(W_y>0)$	7.36E-04	4.81E-04	3.18E-04	2.34E-04	1.51E-04
%95 CI of $P(W_y>0)$	[0.8707, 0.8736]	[0.9373, 0.9391]	[0.9685, 0.9697]	[0.9848, 0.9857]	[0.9920, 0.9926]
$E[W_y]$	9.125	20.501	43.720	88.613	179.456
std of $E[W_y]$	5.56E-03	1.05E-02	2.07E-02	4.07E-02	8.18E-02
%95 CI of $E[W_y]$	[9.114, 9.135]	[20.480, 20.521]	[43.162, 43.243]	[88.533, 88.693]	[179.296, 179.616]
$E[W_y\|W_y>0]$	10.462	21.851	45.114	89.937	180.845
%95 CI of $E[W_y\|W_y>0]$	[10.432, 10.492]	[21.807, 21.895]	[44.510, 44.651]	[89.814, 90.060]	[180.630, 181.061]
$E[W_y^2]$	175.380	814.768	3489.720	14,425.330	58,633.918
std of $E[W_y^2]$	8.65E-02	0.350	1.424	5.703	23.026
%95 CI of $E[W_y^2]$	[175.210, 175.549]	[814.081, 815.455]	[3,486.928, 3,492.511]	[14,414, 14,436]	[58,588, 58,679]
$SD[W_y]$	9.598	19.862	40.289	81.074	162.571
$P(W_y>0)/\rho$	1.0383	1.0198	1.0095	1.0054	1.0023
$(1-\rho)E[W_y]$	1.4599	1.6401	1.7488	1.7723	1.7946
$(1-\rho)SD[W_y]$	1.5357	1.5889	1.6116	1.6215	1.6257
$(1-\rho)E[W_y]/\rho$	1.7380	1.7827	1.8216	1.8084	1.8127
$(1-\rho)SD[W_y]/\rho$	1.2900	1.4618	1.5471	1.5891	1.6095
$(1-\rho)E[W_y\|W_y>0]$	1.6739	1.7481	1.8045	1.7987	1.8085
$(1-\rho)SD[W_y\|W_y>0]$	1.5316	1.5818	1.5828	1.6189	1.6243

6. The Supporting Heavy-Traffic FCLT for Periodic Queues

To explain the unified numerical results in §5, we now review and extend the heavy-traffic (HT) functional central limit theorem (FCLT) for periodic $G_t/G/1$ queues in Theorem 3.2 of Whitt (2014). An extension of the HT FCLT in Whitt (2014) is needed because that HT FCLT is stated for the scaled arrival process and the scaled queue-length process, but not the scaled workload process that we consider here. A similar argument applies to the workload process, jointly with the other processes, but it is more natural to apply Theorem 9.3.4 of Whitt (2002) than Iglehart and Whitt (1970), because the workload process is defined there in §9.2 essentially the same way as the workload is defined in §2.

The innovative part of Whitt (2014) is the new HT scaling in (41) to capture the impact of the periodicity in an interesting and revealing way, as demonstrated by the tables in §5. As shown in Whitt (2014), the periodicity has no impact on the heavy-traffic limit if this additional scaling is not included. (That elementary observation was made earlier by Falin (1989); the main contribution of Whitt (2014) is the new scaling.)

6.1. The Heavy-Traffic FCLT

We assume that the rate-1 arrival and service processes N and V specified in §2 are independent and each satisfies a FCLT. To state the result, let $\hat{N}_n$ and $\hat{S}_n^v$ be the scaled processes defined by

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt] \quad \text{and} \quad \hat{S}_n^v(t) \equiv n^{-1/2}[\sum_{i=1}^{\lfloor nt \rfloor} V_k - nt], \quad t \geq 0, \tag{44}$$

with $\equiv$ denoting equality in distribution and $\lfloor x \rfloor$ denoting the greatest integer less than or equal to x. We assume that

$$\hat{N}_n \Rightarrow c_a B_a \quad \text{and} \quad \hat{S}_n^v \Rightarrow c_s B_s \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{45}$$

where $\mathcal{D}$ is the usual function space of right-continuous real-valued functions on $[0, \infty)$ with left limits and $\Rightarrow$ denotes convergence in distribution, as in Whitt (2002), while B_a and B_s are independent standard (mean 0, variance 1) Brownian motion processes (BM's). The assumed independence implies joint convergence in (45) by Theorem 11.4.4 of Whitt (2002).

We emphasize that GI assumptions are not needed, but that is an important special case. If the service times V_k are i.i.d. mean-1 random variables with variance = scv c_s^2, then the limit in (45) holds with service variability parameter c_s. Similarly, if the base arrival process is a renewal process or an equilibrium renewal process with times between renewals having mean 1 and variance = scv c_a^2, then the limit in (45) holds with arrival variability parameter c_a. (See Nieuwenhuis (1989) for theoretical support in the case of an equilibrium renewal process.)

Theorem 9.3.4 of Whitt (2002) refers to the conditions of Theorem 9.3.3, which requires a joint FCLT for the partial sums of the arrival and service processes, notably (3.9) on p. 295. That convergence follows from the FCLT's we assumed for N and V in (45) above. In particular, the assumed FCLT for N implies the associated FCLT for the partial sums of the interarrival times by Theorem 7.3.2 and Corollary 7.3.1 of Whitt (2002).

We create a model for each ρ, $0 < \rho < 1$, by defining the arrival-rate function

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

where λ_d is a periodic function with period c^* satisfying

$$\bar{\lambda}_d \equiv \frac{1}{c^*} \int_0^{c^*} \lambda_d(s)\, ds \equiv 0. \tag{47}$$

As a regularity condition, we also require that the function λ_d be an element of $\mathcal{D}$. As a consequence of (46) and (47), the average arrival rate is $\bar{\lambda}_\rho = \rho$, $0 < \rho < 1$. Hence, (41) is a special case of (46); see §6.3 below.

We can also work with cumulative functions and let the cumulative arrival-rate function in model ρ be

$$\Lambda_\rho(t) \equiv \rho t + (1-\rho)^{-1}\Lambda_d((1-\rho)^2 t), \quad t \geq 0, \tag{48}$$

where

$$\Lambda_d(t) \equiv \int_0^t \lambda_d(s)\, ds, \tag{49}$$

for λ_d again being the periodic function in (47). From (48)-(49), we see that the associated arrival-rate function obtained by differentiation in (48) is (46).

The time scaling in (46) and (48) implies that the period in model ρ with arrival-rate function $\lambda_\rho(t)$ in (46) is $c_\rho = c^*(1-\rho)^{-2}$, where c^* is the period of $\lambda_d(t)$ in (47). Thus the period c_ρ in model ρ is growing with ρ.

Now let $A_\rho(t) \equiv N(\Lambda_\rho(t))$ be the arrival process, using the cumulative arrival-rate function Λ_ρ in (48) in place of Λ in (1). Let $Q_\rho(t)$ and $W_\rho(t)$ be the associated queue length process and workload process in the $G_t/G/1$ model with arrival process $A_\rho(t)$ in (46) and service times from the fixed service process V, constructed as in §9.2 of Whitt (2002). Then let associated scaled arrival, queue length and workload processes be defined by

$$\begin{aligned} &\hat{A}_\rho(t) \equiv (1-\rho)[A_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}t], \\ &\hat{Q}_\rho(t) \equiv (1-\rho)Q_\rho((1-\rho)^{-2}t) \quad \text{and} \quad \hat{W}_\rho(t) \equiv (1-\rho)W_\rho((1-\rho)^{-2}t), \quad t \geq 0. \end{aligned} \tag{50}$$

The scaled processes in (50) and the HT limit all have cycle length c^*.

The following heavy-traffic FCLT states that $\hat{A}_\rho$ converges to periodic Brownian motion (PBM), while $\hat{Q}_\rho$ and $\hat{W}_\rho$ converge to a common reflected periodic Brownian motion (RPBM). To explain, let e be the identity function with $e(t) = t$, $t \geq 0$. By a PBM, we mean a process $cB + \Lambda - e \equiv \{cB(t) + \Lambda_d(t) - t : t \geq 0\}$, where B is a BM and Λ_d is of the form (49), so that the process has periodic deterministic drift $\lambda_d(t) - 1$. Let ψ be the usual one-dimensional reflection map as on pp. 87, 290 and 439 of Whitt (2002). Given that $cB + \Lambda - e$ is a PBM, $\psi(cB + \Lambda - e)$ is a RPBM. To state the HT FCLT, let $\mathcal{D}^k$ be the k-fold product space of $\mathcal{D}$ with itself and let $\stackrel{\mathrm{d}}{=}$ denote equality in distribution.

THEOREM 2. (*heavy-traffic limit extending Whitt (2014)*) *If, in addition to the definitions and assumptions in* (44)-(50) *above, the system starts empty at time* 0, *then*

$$(\hat{A}_\rho, \hat{Q}_\rho, \hat{W}_\rho) \Rightarrow (X_a, Z, Z) \quad in \quad \mathcal{D}^3 \quad as \quad \rho \uparrow 1, \tag{51}$$

where

$$X_a \equiv c_a B_a + \Lambda_d - e, \quad X \equiv X_a - c_s B_s \quad and \quad Z \equiv \psi(X), \tag{52}$$

with B_a *and* B_s *being independent BM's,* Λ_d *in* (49) *and* c_a *and* c_s *being the variability parameters in* (45), *so that* $X \stackrel{\mathrm{d}}{=} c_x B$, *where* $c_x \equiv \sqrt{c_a^2 + c_s^2}$ *and* B *is a BM.*

The joint limit for $(\hat{A}_\rho, \hat{Q}_\rho)$ is established in Theorem 3.2 of Whitt (2014), which in turn follows quite directly from Iglehart and Whitt (1970). (We remark that there is a typographical error in the translation term on the first line of (13) in the proof of Theorem 3.2 of Whitt (2014); it should be $-(1-\rho)^{-2}t$ as in equation (11) there instead of $-(1-\rho)^{-2}\rho t$.) To treat the workload, we apply Theorem 9.3.4 of Whitt (2002), which implies that the limit for $\hat{W}_\rho$ is the same as for the limit for $\hat{Q}_\rho$.

Unfortunately, the periodic feature makes the RPBM complicated, so that it remains to derive explicit expressions for its transient and periodic steady-state distributions. The present paper contributes by developing an effective algorithm to calculate the periodic steady-state distribution.

6.2. Approximations for the Periodic Steady State Workload

Our algorithm for the periodic steady-state distribution of RPBM calculates the periodic steady-state distribution of the scaled workload process in a $GI_t/GI/1$ queue for suitably large ρ and uses Theorem 2 for justification. While that approach is intuitively reasonable, there are steps that remain to be justified. Proper justification requires an additional limit interchange argument, which has been done in some contexts, e.g., see Budhiraja and Lee (2009), but here is left for a topic of future research.

Hence, we assume that those steps are justified. In particular, we assume that the workload process and the limiting RPBM have proper periodic steady-state distributions for each ρ and that there is convergence in distribution of the scaled periodic steady state workload to the periodic steady state of RPBM as $\rho \uparrow 1$. In particular, in addition to the limit $\hat{W}_\rho \Rightarrow Z$ in $\mathcal{D}$ as $\rho \uparrow 1$ established in Theorem 2, we assume that

$$W_\rho((k+y)c_\rho) \Rightarrow W_{\rho,y}(\infty) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad k \to \infty, \tag{53}$$

where $P(W_{\rho,y}(\infty) < \infty) = 1$ for all ρ and y, $0 < \rho < 1$ and $0 \le y < 1$, or, equivalently,

$$\hat{W}_\rho((k+y)c^*) \Rightarrow \hat{W}_{\rho,y}(\infty) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad k \to \infty, \tag{54}$$

where $P(\hat{W}_{\rho,y}(\infty) < \infty) = 1$ for all ρ and y, $0 < \rho < 1$ and $0 \le y < 1$, and

$$Z((k+y)c^*) \Rightarrow Z_y(\infty) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad k \to \infty, \tag{55}$$

where $P(Z_y(\infty) < \infty) = 1$ for all y, $0 \le y < 1$. With these assumptions, our algorithm applies to RPBM using the approximation

$$P(Z_y(\infty) > x) \approx P(\hat{W}_{\rho,y}(\infty) > x) \tag{56}$$

where ρ is chosen to be suitably large.

6.3. Application to the Sinusoidal Arrival-Rate Function

For the sinusoidal example in (4), we let

$$\lambda_d(t) \equiv \bar{\lambda}\beta \sin(\gamma t), \quad t \ge 0, \tag{57}$$

for $\lambda_d(t)$ in (47), so that the cycle length is $c^* = 2\pi/\gamma$. With (57) and $\bar{\lambda} \equiv \rho$, (46) becomes (41), so that the cycle length in model ρ is $c_\rho = c^*(1-\rho)^{-2} = 2\pi/(\gamma(1-\rho)^2)$. When we consider the periodic steady-state workload, the time scaling is gone but we still have the spatial scaling. When the traffic intensity is ρ, we multiply by $1-\rho$; i.e., we have

$$\hat{W}_{\rho,y}(\infty) = (1-\rho)W_{\rho,y}(\infty). \tag{58}$$

Hence, to have asymptotically convergent models, we should choose parameter four-tuples $(\bar{\lambda}_\rho, \beta_\rho, \gamma_\rho, b_\rho)$ indexed by ρ as indicated in (42).

6.4. Approximations for the Periodic $G_t/G/1$ Model

To apply the heavy-traffic FCLT to generate approximations for the performance of the periodic steady-state workload in a general periodic $G_t/G/1$ model (without i.i.d. assumptions), we assume that the assumptions in §6.1 are satisfied so that Theorem 2 is valid. We then approximate the model by a $GI_t/GI/1$ model which has the same HT FCLT limit process. In other words, we approximate the underlying rate-1 arrival counting process N by a renewal process with i.i.d. mean-1 times between renewals having scv c_a^2, where c_a is

the arrival process variability parameter in the assumed FCLT (45). Similarly, we approximate the sequence of mean-1 service times $\{V_k\}$ by a sequence of mean-1 i.i.d. random variables with a scv equal to c_s^2, where c_s is the service variability parameter in the assumed FCLT (45). Both approximations are exact for GI.

To construct the specific GI arrival and service processes, we follow the approximation scheme in §3 of Whitt (1982). We apply the same method for the interarrival times U_k of N as we do to the service times V_k, so we only discuss the service times. If $c_s^2 \approx 1$, then we use a mean-1 exponential (M) distribution; if $c_s^2 > 1$, then we use a mean-1 hyperexponential (H_2) distribution with pdf $f_V(x) = p_1\mu_1 e^{-\mu_1 x} + p_2\mu_2 e^{-\mu_2 x}$, with $p_1 + p_2 = 1$, having parameter triple (p_1, μ_1, μ_2). To reduce the parameters to two (the mean and scv), we assume balanced means, i.e., $p_1/\mu_1 = p_2/\mu_2$, as in (3.7) of Whitt (1982). If $c_s^2 < 1$ and if $c_s^2 \approx 1/k$ for some integer k, then we use a mean-1 Erlang (E_k) distribution (sum of k i.i.d. exponential variables), otherwise if $c_s^2 < 1$, then we use the $D + M$ distribution, i.e., a sum of a deterministic constant (D) and an exponential (M) distribution with rate μ, which has pdf $f_V(x) = \mu e^{-\mu(x-d)}$, $x \geq d$, as in (3.11) and (3.12) of Whitt (1982).

7. Conclusions

We have developed a new algorithm to calculate the distribution of the periodic steady-state remaining workload W_y, at time yc within a periodic cycle of length c, $0 \leq y < 1$, in a general $GI_t/GI/1$ single-server queue with periodic arrival-rate function. The key model assumption is the representation in (1) of the arrival process as a time-transformation of a rate-1 process. The algorithm is based on the new representation of W_y in (2) derived in §1.1 and §2. In §4 we developed an algorithm for computing the exact tail probabilities $P(W_y > b)$ in the $GI_t/GI/1$ model based on the established rare-event simulation algorithm for the associated stationary $GI/GI/1$ model. That connection is supported by the close relation between the two models, established in §3.

We also have shown that the algorithm can be applied together with the heavy-traffic FCLT in Whitt (2014) reviewed in §6 to also calculate the periodic steady-state distribution and moments of reflected periodic Brownian motion (RPBM). In addition, the algorithm can be applied to approximate the tail probabilities in the more general $G_t/G/1$ model by choosing special parameters (the squared coefficients of variation (scv) of interrenewal times) in the $GI_t/GI/1$ model to insure that the two systems obey the same heavy-traffic FCLT.

We have verified the effectiveness of the algorithm for $GI_t/GI/1$ queues and $RPBM$ by conducted extensive simulation experiments for the $GI_t/GI/1$ model with sinusoidal arrival rate in §1.3 and a range of traffic intensities. Some of these are reported in §5 and in the online supplement Ma and Whitt (2016). It remains to investigate the algorithm for $G_t/G/1$ queues more general than $GI_t/GI/1$.

Acknowledgments

We thank Wei You and anonymous referees for carefully reading our paper and NSF (CMMI 1265070 and 1634133) for research support.

References

Abate, J., G. L. Choudhury, W. Whitt. 1993. Calculation of the $GI/G/1$ steady-state waiting-time distribution and its cumulants from Pollaczek's formula. *Archiv fur Elektronik und bertragungstechnik* **47**(5) 311–321.

Abate, J., G. L. Choudhury, W. Whitt. 1994. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models* **10**(1) 99–143.

Abate, J., W. Whitt. 1998. Calculating transient characteristics of the Erlang loss model by numerical transform inversion. *Stochastic Models* **15** 223–230.

Asmussen, S. 2003. *Applied Probability and Queues*. 2nd ed. Springer, New York.

Asmussen, S., H. Albrecher. 2010. *Ruin Probabilities*. 2nd ed. World Scientific, Singapore.

Asmussen, S., P. W. Glynn. 2007. *Stochastic Simulation*. 2nd ed. Springer, New York.

Asmussen, S., T. Rolski. 1994. Risk theory in a periodic environment: the Cramer-Lundberg approximation and Lundberg's inequality. *Mathematics of Operations Research* **19**(2) 410–433.

Budhiraja, A., C. Lee. 2009. Stationary distribution convergence for generalized jackson networks in heavy traffic. *Mathematics of Operations Research* **34**(1) 45–56.

Choudhury, G. L., D. M. Lucantoni, W. Whitt. 1996. Squeezing the most out of ATM. *IEEE Transactions on Communications* **44**(2) 203–217.

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

Ethier, S. N., T. G. Kurtz. 1986. *Markov Processes: Characterization and Convergence*. Wiley, New York.

Falin, G. I. 1989. Periodic queues in heavy traffic. *Advances in Applied Probability* **21** 485–487.

Feller, W. 1971. *An Introduction to Probability Theory and its Applications*. Second edition ed. John Wiley, New York.

Gerhardt, I., B. L. Nelson. 2009. Transforming renewal processes for simulation of nonstationary arrival processes. *INFORMS Journal on Computing* **21** 630–640.

Glynn, P. W., W. Whitt. 1994. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Studies in Applied Probability, Papers in Honour of Lajos Takacs*. Applied Probability Trust, Sheffield, England, 131–156.

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

He, B., Y. Liu, W. Whitt. 2016. Staffing a service system with non-Poisson nonstationary arrivals. Probability in the Engineering and Informational Sciences, published online June 2016.

Iglehart, D. L., W. Whitt. 1970. Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Advances in Applied Probability* **2**(2) 355–369.

Jacod, J., A. N. Shiryaev. 1987. *Limit Theorems for Stochastic Processes*. Springer, New York.

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.

Loynes, R.M. 1962. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society* **58**(3) 497–520.

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. 2016. Online supplement to "a rare-event simulation algorithm for periodic single-server queues". Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Massey, W. A., W. Whitt. 1994. Unstable asymptotics for nonstationary queues. *Math. Oper. Res.* **19** 267–291.

Morales, M. 2004. On a surplus process under a periodic environment: a simulation approach. *North American Actuarial Journal* **8**(4) 76–89.

Nelson, B. L., I. Gerhardt. 2011. Modeling and simulating renewal nonstationary arrival processes to facilitate analysis. *Journal of Simulation* **5** 3–8.

Nieuwenhuis, G. 1989. Equivalence of functional limit theorems for stationary point processes and their Palm distributions. *Probabiity and Related Fields* **81** 593–608.

Rolski, T. 1981. Queues with nonstationary input stream: Ross's conjecture. *Advances in Applied Probability* **13** 603–618.

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

Sigman, K. 1995. *Stationary Marked Point Processes: An Intuitive Approach*. Chapman and Hall/CRC, New York.

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

Whitt, W. 2002. *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. A Poisson limit for the departure process from a queue with many slow server. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html.

Whitt, W., J. Zhao. 2016. Staffing to stabilize blocking in loss models with non-Markovian arrivals. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html.

Xiong, Y., D. J. Murdoch, D. A. Stanford. 2015. Perfect sampling of a single server queue with periodic Poisson arrivals. *Queueing Systems* **80** 15–33.