Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

USING SIMULATION TO STUDY SERVICE-RATE CONTROLS TO STABILIZE PERFORMANCE IN A SINGLE-SERVER QUEUE WITH TIME-VARYING ARRIVAL RATE

Ni Ma

Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

Ward Whitt

Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

ABSTRACT

Simulation is used to evaluate the performance of alternative service-rate controls designed to stabilize performance in a queue with time-varying arrival rate, service in order of arrival and unlimited waiting space. Both Markovian and non-Markovian models are considered. Customer service requirements are specified separately from the service rate, which is subject to control. New versions of the inverse method exploiting tables constructed outside the simulation are developed to efficiently generate both the arrival times and service times. The simulation experiments show that a rate-matching service-rate control successfully stabilizes the expected queue length, but not the expected waiting time, while a new square-root service-rate control, based on a assuming that a pointwise-stationary approximation is appropriate, successfully stabilizes the expected waiting time when the arrival rate changes slowly compared to the expected service time.

1 INTRODUCTION

In this paper we study alternative service-rate controls to stabilize performance in a single-server queue with time-varying arrival rate and independent and identically distributed (i.i.d.) service requirements specified separately from the service rate actually provided. Our study parallels Liu and Whitt (2012), He, Liu, and Whitt (2015) and earlier papers cited there that develop time-varying staffing levels (number of servers) to stabilize performance in multi-server queues with flexible staffing.

The present problem of service-rate control is important for systems with only a few servers or with inflexible staffing. In many applications, even though the available service resources are fixed, it is possible to change the processing rate. Two important examples are hospital surgery rooms and airport security lines. The number of hospital surgery rooms often cannot be changed in the short term, but assigning more doctors and nurses can increase the operation completion rate in each surgery room. Similarly, the number of airport security lines may be fixed in the short term, but adding more inspection agents or relaxing inspection requirements can increase the passenger processing rate in each line. In these settings the possible service rates that can be achieved may be limited, but to gain insight into the potential benefits of controlling the service rate, we study the idealized case of a single server where the service rate is fully subject to control.

Specifically, we consider a class of general $GI_t/GI_t/1$ single-server queues with unlimited waiting space, service in order of arrival, a time-varying arrival rate, and a time-varying service rate that is subject to control. Our methods apply to general arrival rate functions, but as in previous work we use stylized sinusoidal arrival rate functions with a range of parameters. We consider arrival processes that are time-transformed stationary renewal processes, with the specified arrival rate function. We assume that the service requirements are i.i.d. random variables with a general distribution, specified independently of

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

the service rate control. The general *GI* arrival and service processes allow different levels of stochastic variability to go with the predictable deterministic variability of the time-varying rates.

We develop new methods to simulate these nonstationary non-Markovian queueing models. As in §7 of Massey and Whitt (1994), Gerhardt and Nelson (2009), Liu, Kuhl, Liu, and Wilson (2014) and He, Liu, and Whitt (2015), we represent the arrival process as the composition of a rate-1 stationary point process and the deterministic cumulative arrival rate function. For this study we use renewal processes for the base rate-1 process, but the method is more general. We efficiently generate both the service times and the arrival times by exploiting tabled inverse functions, as can be done in generating non-uniform random numbers; see §11.2 and §III.2 of Devroye (1986) and §3.8 of L'Ecuyer (2012).

We conduct simulation experiments to study the performance of the service-rate controls to stabilize performance in these systems. We consider four different service-rate controls: a rate-matching control that makes the service rate proportional to the arrival rate and three square-root service-rate controls. The first square-root service-rate control is a natural analog of the offered-load square-root-staffing formula used for many-server queues, where the offered load is the expected number of busy servers in an associated infinite-server system with the same arrival rate and a service-time distribution. Since the service-time distribution is unavailable in advance, we use the service-requirement distribution. The second square-root service-rate control is a variant of the first, in which the arrival rate is used in place of the offered load. The third square-root service-rate control is obtained by solving a quadratic equation, based on a steady-state heavy-traffic approximation assuming that a pointwise-stationary approximation (PSA) is appropriate; see Green, Kolesar, and Whitt (2007).

We show that the rate-matching control stabilizes the expected queue length, but not the expected waiting time, consistent with theoretical results established in Whitt (2014). We show that the expected waiting time tends to be inversely proportional to the arrival rate. We show that the first two square-root service-rate controls that are analogs of the square-root staffing formula for multiple server queues stabilize the mean waiting times to some extent, but not fully. We show that the final square-root control based on the PSA is effective for long cycles, where the PSA is effective, but not more generally.

The remainder of this paper is organized as follows. In §2, we discuss the simulation methodology for generating the nonstationary non-Markovian models. In §3 we define the different service-rate controls; in §4 we describe the simulation experiments; in §5 we show some of the results; and in §6 we draw conclusions.

2 SIMULATION METHODS FOR NONSTATIONARY MODELS

In §2.1 we define the model. Then in §2.2 and §2.3 we describe the new methods to generate the arrival times and the service times, after which the simulation is elementary.

2.1 The Model

We construct the arrival and service processes by using deterministic time-transformations of general rate-1 processes. We first consider the arrival counting process A, where $A(t)$ counts the number of arrivals occurring in the time interval $[0,t]$. We define A using a cumulative arrival rate function

$$\Lambda(t) = \int_0^t \lambda(s)\,ds, \quad t \geq 0, \quad \text{where} \quad 0 < \lambda_L \leq \lambda(t) \leq \lambda_U < \infty, \tag{1}$$

and a general rate-1 counting process N_a with unit jumps. We define A by the composition

$$A(t) \equiv N_a(\Lambda(t)), \quad t \geq 0. \tag{2}$$

Given that $E[N_a(t)] = t$, $t \geq 0$ (the rate-1 property), A defined by (2) has the specified rate: $E[A(t)] = E[N_a(\Lambda(t))] = \Lambda(t)$. The deterministic function $\Lambda(t)$ specifies the predictable variability, while all the unpredictable stochastic variability is specified by the base counting process N_a. This construction is without loss of generality, because given any A with unit jumps and $E[A(t)] = \Lambda(t)$, we can let $N_a = A(\Lambda^{-1}(t))$, $t \geq 0$, where Λ^{-1} is the inverse of Λ, which is well defined. Hence, (2) holds with $E[N_a(t)] = t$, $t \geq 0$.

We now turn to the service process. Paralleling our model of the arrival process, we assume that the service requirements are generated by a counting process N_s with unit jumps, which is independent of N_a. We define the evolution of the queueing model, given the arrival process A, the service requirement process N_s and the time-varying service-rate control $\mu(t)$, by jointly defining the number in system $Q(t)$ and the departure counting process $D(t)$. In particular, we require that these processes satisfy the two equations

$$Q(t) = A(t) - D(t) \quad \text{and} \quad D(t) \equiv N_s(\int_0^t \mu(s) 1_{\{Q(s)>0\}} ds), \quad t \geq 0, \tag{3}$$

The representation (3) can be justified by applying mathematical induction to the successive event changes in $Q(t)$; see §2.1 of Pang, Talreja, and Whitt (2007). Note that the process D has the service rate $\mu(t)$ whenever the system is not empty: $E[D(t)] = \int_0^t \mu(s) 1_{\{Q(s)>0\}} ds$, $t \geq 0$.

In this paper we consider the special case of the model above in which the service requirements S_k are i.i.d random variables with a general cdf G having mean 1 and finite second moment. If the mean were not actually 1 initially, we could rescale both these service requirements and the service-rate control to make it so, so that is without loss of generality. The associated rate-1 counting process is the equilibrium version of the renewal counting process, which differs from the ordinary renewal counting process only by having the first interval having the stationary-excess cdf $G_e(t) = \int_0^t [1 - G(s)] ds$, $t \geq 0$, instead of the cdf G of all other intervals. The same holds for the arrival process. We will generate N_a using i.i.d. random variables with mean 1; then the associated rate-1 process is the equilibrium renewal process.

Often an exceptional first interval is not too important, and can be considered part of the initial conditions, along with starting the queueing system empty. We then can generate both the arrival process and the service process using ordinary renewal processes with mean-1 inter-renewal times. Then the arrival rate is asymptotically correct as $t \to \infty$.

To simulate the model, we first generate the successive arrival times and then the successive service times. It is then straightforward to construct the associated queueing processes. We next describe the two generation steps.

2.2 Generating the Arrival Process

Let A_k and T_k be the arrival times of the k^{th} arrival in the processes A and N_a. The basic construction in (2) implies that $T_k = \Lambda(A_k)$, so that $A_k = \Lambda^{-1}(T_k)$, $k \geq 1$. In our applications using a base renewal process, the times T_k are directly available; e.g., we can generate n i.i.d. uniform random variables on $[0,1]$ and calculate $\widetilde{U}_k = G^{-1}(U_k), 1 \leq k \leq n$, to get random variables with the desired cdf G, where n is large enough to ensure that arrivals cover our considered time interval; see Devroye (1986), L'Ecuyer (2012). Then the renewal times of the base process N_a are their partial sums $T_k = \sum_{i=1}^k \widetilde{U}_i, 1 \leq k \leq n$. If desired, we can obtain an independent random variable approximately distributed as G_e to use as the first time by fixing a large time t, beyond our range of interest, and letting the random variable be the observed excess after t in the ordinary renewal process. The challenge is to evaluate the inverse function Λ^{-1} at each time T_k in order to generate A_k.

In our simulation experiments, the arrival rate function is sinusoidal, with

$$\lambda(t) = 1 + \beta \sin(\gamma t) \quad \text{and} \quad \Lambda(t) = t + (\beta/\gamma)(1 - cos(\gamma t)), \tag{4}$$

but the inverse evidently is not directly available. Hence, we calculate the inverse function outside of the simulation and have it available to apply by table lookup. We take care to do this efficiently.

A major basis for efficiency of a table-lookup scheme is re-using previously constructed tables for new cases. First, we exploit the fact the arrival rate function is periodic. That allows us to only table the inverse over a single cycle. We use the property that $\Lambda(kC) = kC$, $k \geq 1$, where C is the periodic cycle length. As a consequence, $\Lambda^{-1}(kC) = kC$, $k \geq 1$, and

$$\Lambda^{-1}(kC+t) = kC + \Lambda^{-1}(t), \quad 0 \leq t < C. \tag{5}$$

Thus we only need to table the inverse function Λ^{-1} over a single cycle $[0,C]$.

Second, we can also use one constructed inverse function Λ^{-1} to obtain the corresponding inverse functions for scaled versions of the original function Λ. For example, we are interested in the impact of the time-scaling parameter γ in the arrival rate function λ in (4) upon performance. Since $\Lambda(t;\gamma) = \Lambda(\gamma t;1)/\gamma$, we can expreess $\Lambda^{-1}(t;\gamma) = \Lambda^{-1}(\gamma t;1)/\gamma$.

We next develop an efficient way to construct the table of Λ^{-1} over a single cycle $[0,C]$. We specify a large number n_x of equally spaced points of one cycle $[0,C)$, yielding the spacing $\eta = C/n_x$. We then evaluate $\Lambda(t)$ for each of the n_x time points $t = j\eta$.

To have an efficient way of calculating the inverse, we construct an approximation J of the inverse function Λ^{-1} over n_y equally spaced points in the interval $[0,\Lambda(C)] = [0,C]$ with spacing $\delta = C/n_y$. We then approximate the inverse at each point $j\delta$ by the inverse function value $k\eta$, which is the closest point greater equal to the true inverse value, This produces a strictly increasing function b (vector) mapping the finite subset $\{j : 0 \leq j \leq n_y\}$ into the finite subset $\{k : 0 \leq k \leq n_x\}$, so that

$$J(j\delta) = b(j)\eta, \quad 1 \leq j \leq n_y. \tag{6}$$

We extend J to the interval $[0,C]$ by letting $J(t) = J(\lfloor t/\delta \rfloor \delta)$, where $\lfloor x \rfloor$ is the floor function, producing the greatest integer less than or equal to x.

We then specify η and δ to ensure that J is a suitably accurate approximation of Λ^{-1}. By the construction above, it follows that the uniform error bound

$$\|J - \Lambda^{-1}\| \leq \varepsilon \tag{7}$$

is achieved if

$$\eta = \varepsilon/(1+\omega) \quad \text{and} \quad \delta = \lambda_U \eta, \quad \text{where} \quad \omega = \lambda_U/\lambda_L > 1, \tag{8}$$

with λ_L and λ_U being the lower and upper rate bounds in (1). (For additional details, see Ma and Whitt (2015).)

To construct the vector b above and thus the function J in (6), we need two vectors of size n_x and n_y, where $n_x + n_y = C(1+\omega)(1+(1/\lambda_U))/\varepsilon$. Thus, we need storage of order $O(n_x+n_y) = O(C/\varepsilon)$. Since we are able to make a single pass through the data, the computational complexity is of order $O(n_x+n_y) = O(C/\varepsilon)$. With the table, we do not need to do any search for each arrival to get its approximate inverse function value. It only takes two operations for each arrival. Therefore the computation time for calculating arrival times from the table is linear in the number of arrivals.

2.3 Generating the Service Times

We use a similar inverse function method to generate the service times, but the method is more complicated, because to apply (3) we need to keep track of when the server is busy. Thus, we start by developing a recursion.

Let B_k, D_k, V_k and W_k be the times that arrival k who arrives at A_k begins service, departs, spends in service and waits before starting service, respectively. Then we have the basic recursion: $B_k = \max\{D_{k-1}, A_k\}$, $D_k = B_k + V_k$ and $W_k = B_k - A_k$, where the arrival times A_k have been generated already. Given that the system starts empty, we can initialize the recursion with $D_0 = 0$ and $B_1 = A_1$, so that the only variable not formulated in the recursion is the service time V_k.

Since the service requirement S_k is completed by the server busy working from time B_k to time $B_k + V_k$, the service time V_k satisfies the equation

$$S_k = \int_{B_k}^{B_k+V_k} \mu(s)\,ds, \quad k \geq 1. \tag{9}$$

We can solve for service times explicitly by

$$V_k = M^{-1}(S_k + M(B_k)) - B_k, \quad \text{where} \quad M(t) \equiv \int_0^t \mu(s)\,ds \tag{10}$$

and M^{-1} is the inverse of M, which is well defined providing that $0 < \mu_L \leq \mu(t) \leq \mu_U < \infty$, paralleling (1), which we assume to be the case.

Again we work to reduce the computational burden. Just as for the arrival rate function, we see that the function M is typically periodic, so that we only need to compute M^{-1} over a single cycle. We avoid performing the integration in the direct definition of M, we approximate the function M by the piecewise constant function $M(x(i)) = \int_0^{x(i)} \mu(s)\,ds \approx \sum_{j=1}^{i} \mu(x(j))\tau$, implemented with the recursion $M(x(i+1)) = M(x(i)) + \mu(x(i+1))\tau$ for suitably small τ, starting with $M(x(0)) = 0$. To obtain the M^{-1} value for each customer, we table the inverse function much as we did for Λ^{-1}.

3 THE SERVICE-RATE CONTROLS

In this section, we specify the different service-rate controls that we consider.

3.1 The Rate-Matching Control

The first service-rate control is the rate-matching control

$$\mu(t) \equiv \lambda(t)/\rho, \quad t \geq 0, \tag{11}$$

where ρ is the desired traffic intensity. Clearly, the instantaneous traffic intensity is $\rho(t) \equiv \lambda(t)/\mu(t)$ is constant. The simulation experiments show that this control stabilizes the expected queue length (in fact, the entire queue-length distribution), but not the expected waiting time.

3.2 Square-Root Controls Motivated by the Multi-Server Offered-Load Control

We consider two variants of the classical square-root staffing rule for multi-server queues, which lets the time-varying number of servers (staffing) be

$$s(t) \equiv m(t) + \xi\sqrt{m(t)}, \quad t \geq 0, \tag{12}$$

where ξ is a constant and the $m(t)$ is the offered load, which is the expected number of busy servers in the infinite-server system with the same arrival process and service times. The first square-root control is the direct analog

$$\mu(t) \equiv m(t) + \xi\sqrt{m(t)}, \quad t \geq 0, \tag{13}$$

where both $m(t)$ and ξ need to be modified. Since we have time-varying service rates, it is unclear what service times should be used in the infinite-server model. We use the service-requirement distribution directly. For the sinusoidal arrival rate function in (4), explicit formulas for $m(t)$ is given in Eick, Massey, and Whitt (1993); for exponential service times, $m(t) = 1 + (\beta/(1+\gamma^2)(sin\gamma t - \gamma cos\gamma t)$.

The second variant of (12) is (13) with $\lambda(t)$ instead of $m(t)$, i.e.,

$$\mu(t) \equiv \lambda(t) + \xi\sqrt{\lambda(t)}, t \geq 0, \tag{14}$$

Simulations experiments show that these service-rate controls adapted from the multi-server staffing formula stabilize the performance to some extent but are not truly effective for the single-server system.

3.3 The PSA-Based Square-Root Control

To obtain a service-rate control that is effective for stabilizing the mean waiting time, we start by assuming that the PSA approximation is appropriate, so that we can use a time-varying heavy-traffic approximation

$$E[W(t)] \approx \rho(t)V/\mu(t)(1-\rho(t)) = \lambda(t)V/(\mu(t)^2 - \mu(t)\lambda(t)), \quad t \geq 0, \tag{15}$$

where V is a variability parameter, e.g., $V = c_a^2 + c_s^2$; see §5.1 of Whitt (1983). (For $M/GI/1$, this is the exact steady-state formula.) To stabilize, we set $E[W(t)] = w$ and obtain a quadratic equation for $\mu(t)$, yielding

$$\mu(t) \equiv \lambda(t) + (\lambda(t)/2)\left(\sqrt{(\lambda(t)+\zeta)/\lambda(t)} - 1\right), \quad t \geq 0. \tag{16}$$

Simulation experiments verify that this control stabilizes the expected waiting time when the periodic cycles are not too short (when PSA is appropriate), but not when the cycles are short.

4 THE EXPERIMENTS

4.1 Estimating Performance Measures

We mainly evaluate two performance measures for each service-rate control, the expected number of customers in the system, $E[Q(t)]$, and the expected virtual waiting time $E[W(t)]$. The virtual waiting time at time t is defined as the waiting time of a potential or hypothetical arrival (a virtual arrival) at time t, where the waiting time is the time from arrival until starting service.

For each simulation replication, we consider a fixed time interval $[0,T]$ and calculate the performance measures at time points $k\theta, 1 \leq k \leq 1000$, where $\theta = T/1000$. We use $T = 2 \times 10^4$ for $\gamma = 0.001$ and T= 2×10^3 for the other values of γ. We use the longer time interval for very small γ because we want to see the performance over at least several cycles (which each are of length $2\pi/\gamma$). On the other hand, we cannot only fix the number of cycles, because we need enough absolute time to remove the impact of the initial transient.

To calculate these two performance measures, we first derive values of the cumulative arrival function $A(t)$ and the cumulative departure function $D(t)$ at time points $j\theta, 1 \leq j \leq 1000$ from customers' arrival times A_k and departure times D_k. Then we compute $Q(t) = A(t) - D(t)$ and $W(t) = (W_{A(t)} + V_{A(t)} - (t - A_{A(t)}))^+$ at those time points, where the virtual waiting time $W(t)$ actually depends on information after time t, because the service time $V_{A(t)}$ may depend on future service-rate function. But this future effect has been properly accounted for, because the service times have already been generated, according to §2.3.

We generate 10,000 independent replications to estimate mean values of performance measures and to construct their confidence intervals at each of those time points. This sample size is large enough to produce reliable estimation. We estimate the mean values $E[Q(t)]$ and $E[W(t)]$ by taking the average over all replications and construct 95% confidence intervals for these mean values. Since we have a very large sample sizes, z statistics are essentially the same as the natural t statistics.

4.2 The Study Cases

4.2.1 The Rate Functions

We use the sinusoidal arrival rate function in (4) with parameters $\beta = 0.2$ and $\gamma = 10^j$ for $-3 \leq j \leq 2$ to cover a range of different cycle lengths $2\pi/\gamma$. The service rate controls are then as specified in §3. For the infinite-server offered load $m(t)$ with this sinusoidal arrival rate function, formulas are given in Eick, Massey, and Whitt (1993).

4.2.2 Interval Distributions for the Base Renewal Processes

We use renewal processes with i.i.d. interval times having mean 1 for the base processes N_a and N_s used to construct the arrival and service process. We use the squared coefficient of variation (scv, variance divided by the square of the mean), c_a^2 and c_s^2, to characterize the variability. We consider three different distributions: exponential ($c^2 = 1$), hyperexponential (mixture of two exponentials, H_2, $c^2 > 1$) and Erlang (sums of two i.i.d. exponentials, E_2, $c^2 = 0.5$) to represent a range of variability. The H_2 distribution has mean 1 and scv $c^2 = 4$, assuming balanced means $p_1\lambda_1^{-1} = p_2\lambda_2^{-1}$ as in Whitt (1982); it has density $h(x) = p_1\lambda_1 e^{-\lambda_1 x} + p_2\lambda_2 e^{-\lambda_2 x}$, where $p_1 = (5+\sqrt{15})/10$, $p_2 = 1 - p_1$ and $\lambda_i = 2p_i, i = 1,2$. The simulation experiments consider various combinations of these distributions for the arrival and service processes. Some results are for the Markovian $M_t/M_t/1$ model, while others are for the non-Markovian $GI_t/GI_t/1$ systems.

5 SIMULATION RESULTS

In this section, we display simulation results to show the performance of the different service-rate controls.

5.1 The Rate-Matching Control

Figures 1 and 2 show the performance of the rate-matching control for the Markovian $M_t/M_t/1$ system. Figure 1 shows the time-varying means $E[Q(t)]$ and $E[W(t)]$ for three values of γ: 0.001, 0.1 and 10.0. In each case we show the performance over three cycles of length $2\pi/\gamma$, for which the total length is inversely proportional to γ. We show the 95% confidence interval for $E[Q(t)]$ as well as the estimate itself. Figure 1 shows that $E[Q(t)]$ is stabilized in all cases, but $E[W(t)]$ is not. Both means are stabilized for $\gamma = 10.0$ because the cycles are very short, making the arrival process nearly the same as a homogeneous Poisson process (implied by Theorem 1 of Whitt (1984)).

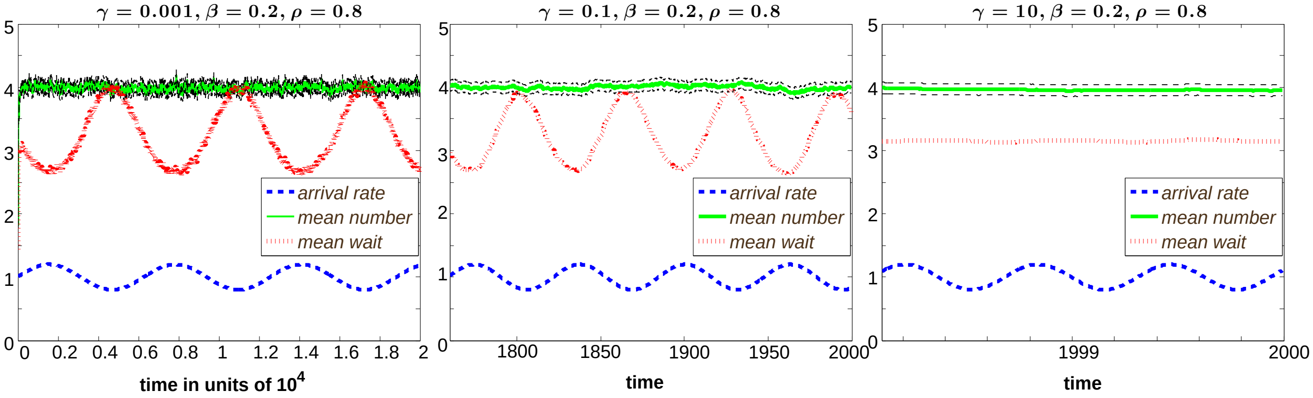


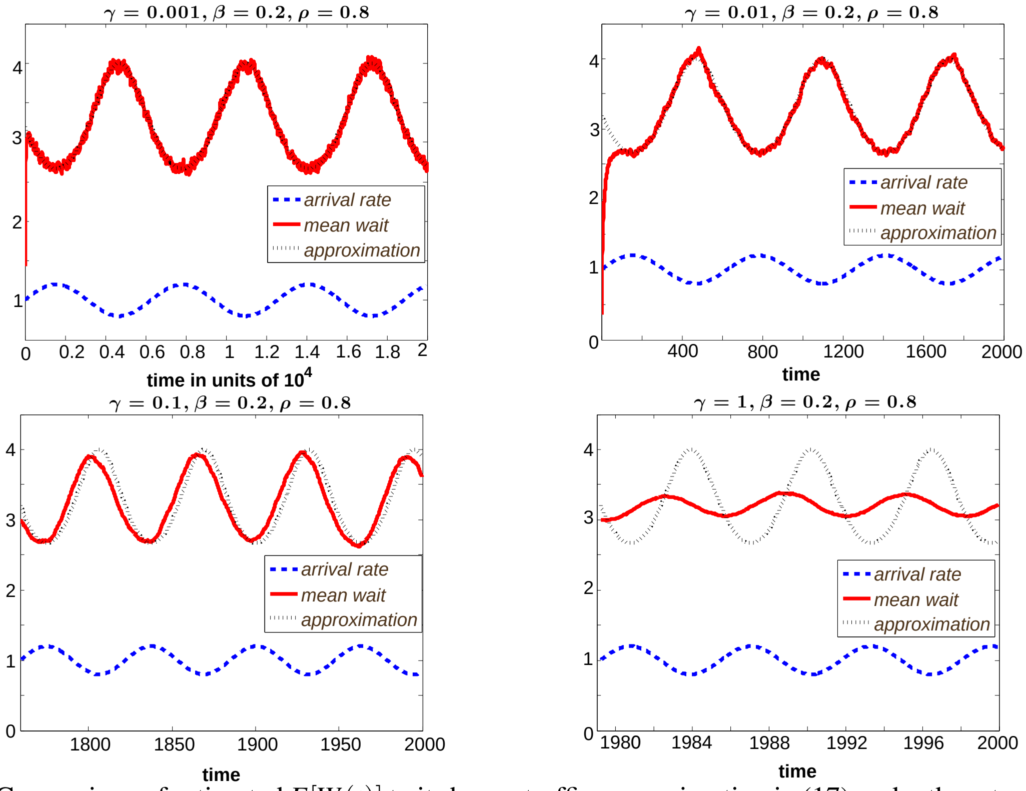
Figure 1: Estimated $E[Q(t)]$ for the rate-matching control in the $M_t/M_t/1$ system with different γ: 0.001 (left), 0.1 (middle) and 10 (right).

Figure 2 compares the estimated $E[W(t)]$ to the heavy-traffic approximation

$$E[W(t)] \approx \frac{E[W(\infty)]}{\lambda(t)} \approx \frac{\rho^2(c_a^2 + c_s^2)}{2(1-\rho)\lambda(t)}, \tag{17}$$

from Whitt (2014). Figure 2 shows that this heavy-traffic approximation works well provided that γ is not too large (the cycles are not too short). A small time-shift error appears at $\gamma = 0.1$ and significant deviation appears for $\gamma \geq 1$. (Above we observed that the rapidly fluctuating arrival rate for the very short cycles makes the model nearly the same as if the arrival rate were constant, equal to its average.)

Figures 3 and 4 present performance results for $E[Q(t)]$ and $E[W(t)]$, respectively, using the rate-matching control applied to three non-Markovian $GI_t/GI_t/1$ systems and three values of γ. (We use (H_2/E_2) to specify that N_a is a H_2 renewal process, while N_s is an E_2 renewal process, and similarly for other cases.) As in stationary models, the performance tends to be proportional to the total variabilty $c_a^2 + c_s^2$. Otherwise, the story is essentially the same as for the $M_t/M_t/1$ model.

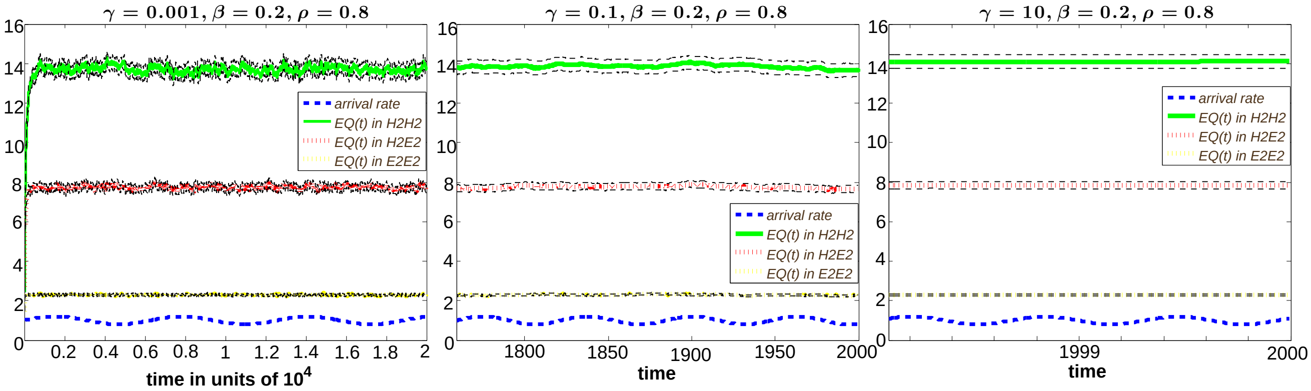

$\gamma = 0.001, \beta = 0.2, \rho = 0.8$

$\gamma = 0.01, \beta = 0.2, \rho = 0.8$

$\gamma = 0.1, \beta = 0.2, \rho = 0.8$

$\gamma = 1, \beta = 0.2, \rho = 0.8$

Figure 2: Comparison of estimated $E[W(t)]$ to its heavy traffic approximation in (17) under the rate-matching control in $M_t/M_t/1$ system with different values of γ: 0.001 (top left), 0.01 (top right), 0.1 (bottom left) and 1 (bottom right).


$\gamma = 0.001, \beta = 0.2, \rho = 0.8$

$\gamma = 0.1, \beta = 0.2, \rho = 0.8$

$\gamma = 10, \beta = 0.2, \rho = 0.8$

Figure 3: Estimated $E[Q(t)]$ for the rate-matching control in three different $G_t/G_t/1$ models, (H_2/H_2), (H_2/E_2) and (E_2/E_2), and three different values of γ: 0.001 (left), 0.1 (middle) and 10 (right).

5.2 The Square-Root Controls Related to the Many-Server Staffing

Figures 5 and 6 presents performance results of the square-root controls related to the many-server staffing formula applied to the Markovian $M_t/M_t/1$ model with different values of γ. The first variant in (13) is shown in Figure 5, while the second variant in (14) is shown in Figure 6. When γ gets larger, $m(t)$ is more different from $\lambda(t)$ and performance is more different for these two controls.

These figures show that neither of these controls is consistently effective. When γ is very small, the offered load $m(t)$ is very close to the arrival rate $\lambda(t)$, explaining why the two left-most plots are very similar.

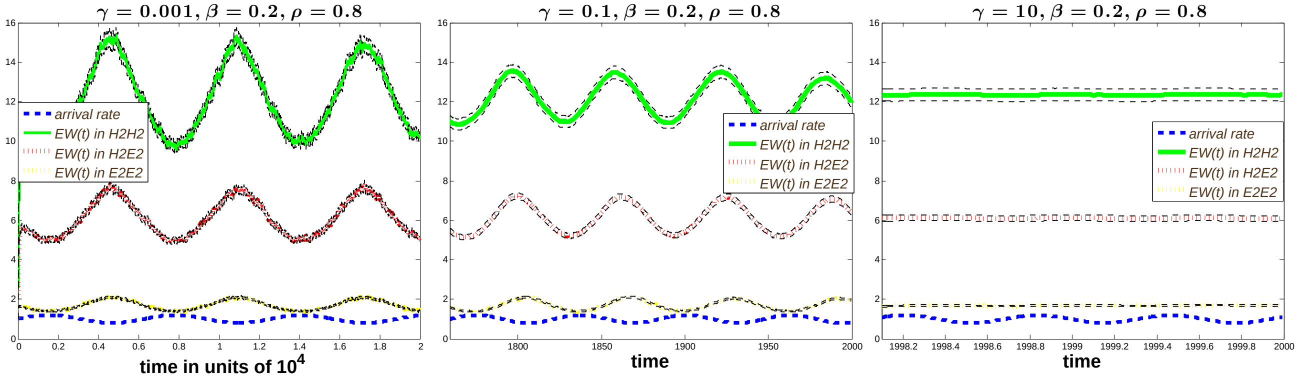


Figure 4: Estimated $E[W(t)]$ for the rate-matching control in three different $G_t/G_t/1$ models, (H_2/H_2), (H_2/E_2) and (E_2/E_2), and three different values of γ: 0.001 (left), 0.1 (middle) and 10 (right)

.

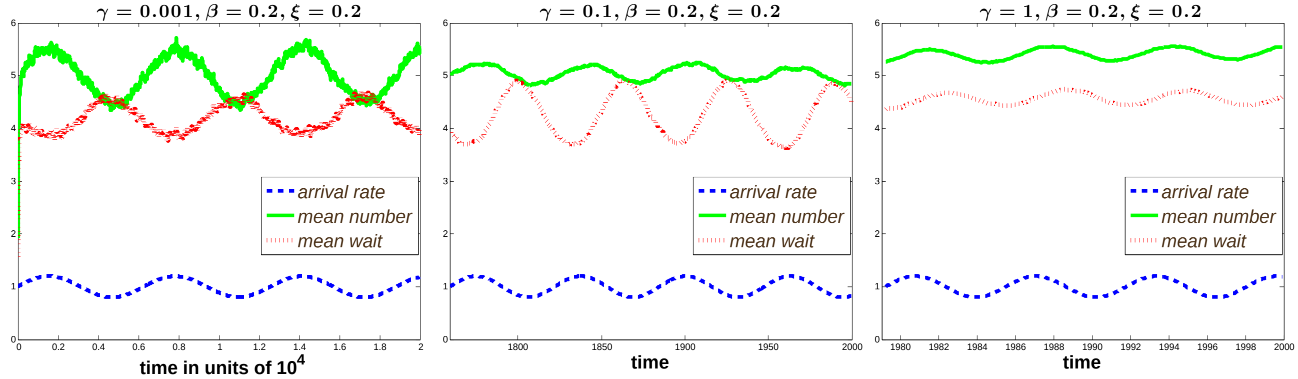


Figure 5: Estimated means $E[Q(t)]$ and $E[W(t)]$ for the square-root control in (13) for the $M_t/M_t/1$ system with different values of γ: 0.001 (left), 0.1 (middle) and 1 (right).

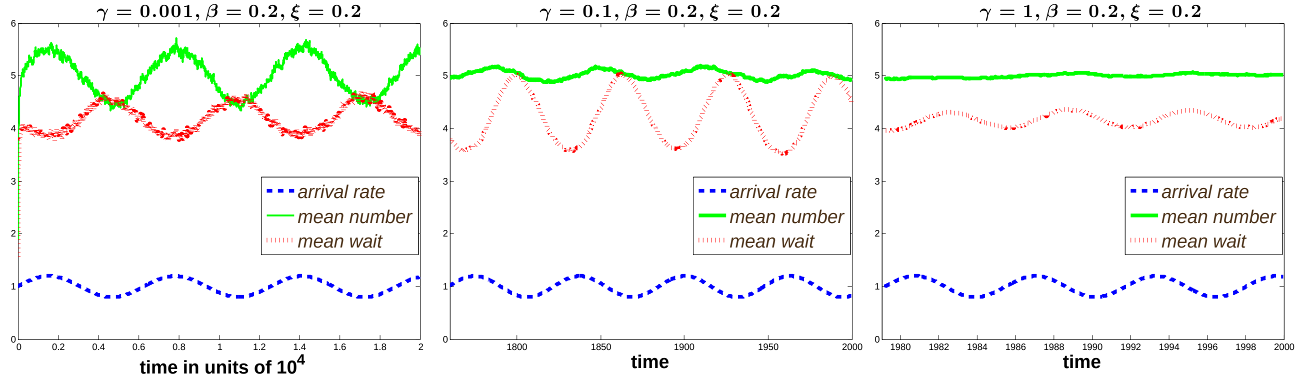


Figure 6: Estimated means $E[Q(t)]$ and $E[W(t)]$ for the square-root control in (14) for the $M_t/M_t/1$ system with different values of γ: 0.001 (left), 0.1 (middle) and 1 (right).

5.3 The PSA Square-Root Control

Figure 7 shows the results of the PSA square-root service-rate control in (16) applied to the $M_t/M_t/1$ system, while Figure 8 shows its application to corresponding (H_2/H_2), (H_2/E_2), and (E_2/E_2), $GI_t/GI_t/1$ systems. When $\gamma = 0.001$, so that the cycles are long and arrival rates change very slowly compared to service times, we see that $E(W(t))$ is stabilized, as intended (while $E(Q(t))$ is not). When $\gamma = 0.1$, so that the cycles are much shorter and PSA is no longer appropriate, $E(W(t))$ becomes periodic.

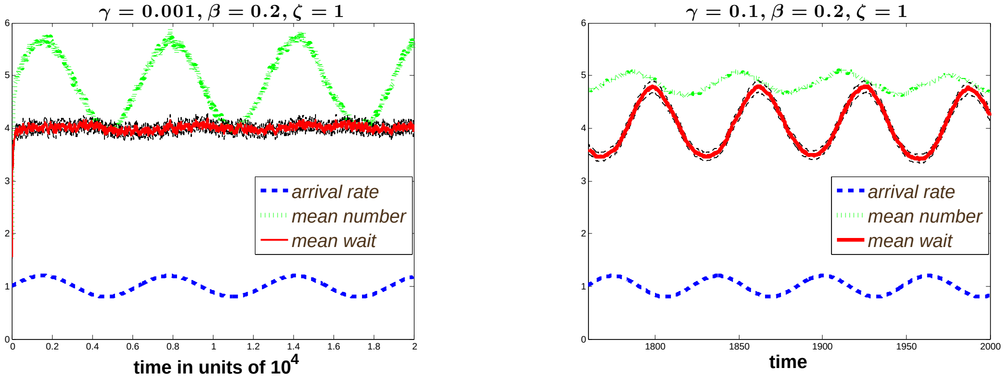


Figure 7: Estimated $E[Q(t)]$ and $E[W(t)]$ for the PSA square-root control in (16) in the $M_t/M_t/1$ model for two values of γ: 0.001 (left) and 0.1 (right).

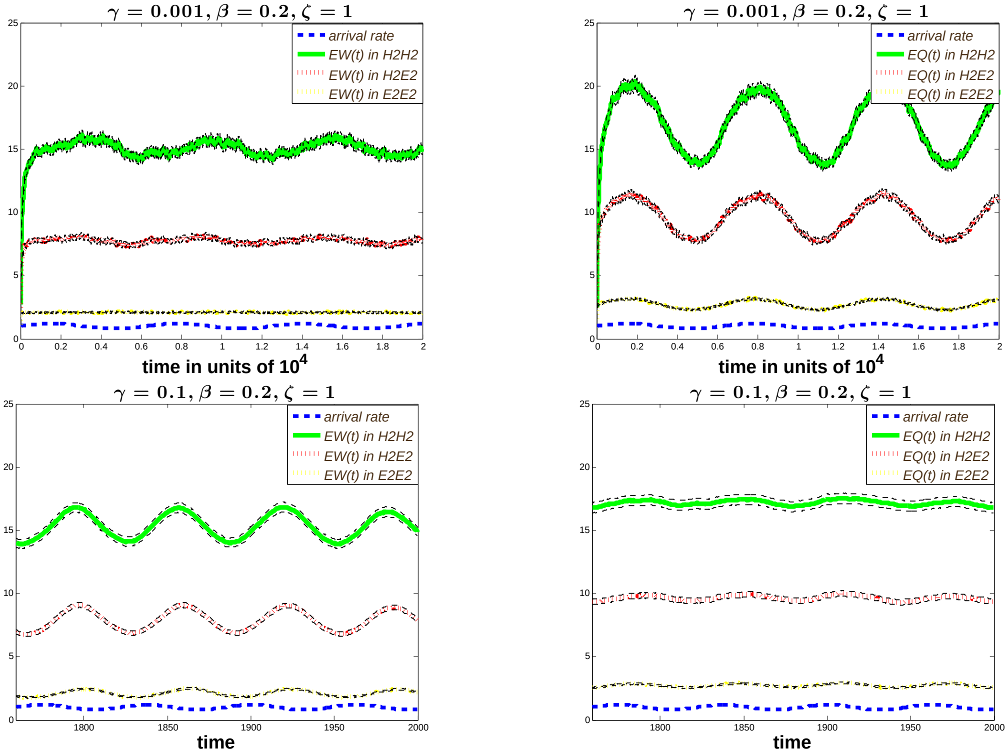


Figure 8: Estimated $E(W(t))$ (left column) and $E(Q(t))$ (right column) for the PSA square-root control in (16) in three $G_t/G_t/1$ models, (H_2/H_2), (H_2/E_2) and (E_2/E_2), for two values of γ: 0.001 (top) and 0.1 (bottom).

6 Conclusions

In this work we have made two contributions: (i) we conducted simulation experiments evaluating the performance of alternative service-rate controls for single-server queues with time-varying arrival rates and

(ii) we developed an efficient algorithm for simulating the model of a time-varying queue with a service-rate control.

In this paper we have described simulation experiments conducted to evaluate the performance of four candidate service-rate controls. The model is a single-server queue with service in order of arrival, unlimited waiting space and a time-varying arrival rate function. The service-rate controls apply to arbitrary arrival rate functions, but for these experiments we used the sinusoidal periodic arrival rate function in (4) with average arrival rate 1, relative amplitude $\beta = 0.2$ and various time-scaling factors γ. The service requirements were i.i.d. random variables specified separately from the service-rate control. The arrival processes were mostly nonhomogeneous Poisson processes, but the method applies to very general arrival processes that can be represented as a deterministic time transformation of a stationary point process as in (2). Experiments were conducted for stationary processes constructed from renewal processes with non-exponential as well as exponential distributions. This allows representing different levels of stochastic variability.

The simulation experiments confirmed theoretical results in Whitt (2014) showing that the rate-matching control in (11) stabilizes the expected queue length $E[Q(t)]$ after an initial transient period, but not the expected waiting time, and that the square-root-service-rate control in (16) stabilizes the mean waiting time when the arrival rate function changes slowly (for long cycles relative to the mean service time). The simulation experiments also showed that the other two service-rate controls in (13) and (14) that are modifications of the classical square-root staffing formula for many-server queues in (12) are not so effective in the present context.

Conducting the simulations for these nonstationary queues turned out to be quite challenging. Thus a significant contribution was developing an efficient simulation algorithm. An important component was exploiting table lookup to calculate the arrival times and service times. The use of tables for a periodic arrival rate function is appealing because the table for one cycle can be used for other cycles and for scaled versions of the original arrival rate function, as shown in §2.2.

ACKNOWLEDGMENTS

Support was received from NSF grant CMMI 1265070.

REFERENCES

Devroye, L. 1986. *Non-Uniform Random Variate Generation*. New York: Springer.

Eick, S. G., W. A. Massey, and W. Whitt. 1993. “$M_t/G/\infty$ Queues with Sinusoidal Arrival Rates”. *Management Science* 39:241–252.

Gerhardt, I., and B. L. Nelson. 2009. “Transforming Renewal Processes for Simulation of Nonstationary Arrival Processes”. *INFORMS Journal on Computing* 21:630–640.

Green, L. V., P. J. Kolesar, and W. Whitt. 2007. “Coping with time-varying demand when setting staffing requirements for a service system”. *Production and Operations Management* 16:13–29.

He, B., Y. Liu, and W. Whitt. 2015. “Staffing a service system with non-Poisson nonstationary arrivals”. working paper, Department of Industrial and Systems Engineering, North Carolina State University.

L’Ecuyer, P. 2012. “Random Number Generation”. In *Handbook of Computational Statistics*, edited by J. E. Gentle, J. K. Hardle, and Y. Mori, Chapter 3, 35–71. New York: Springer.

Liu, R., M. E. Kuhl, Y. Liu, and J. R. Wilson. 2014. “Modeling and simulation of nonstationary non-Poisson arrival processes”. Working paper, North Carolina State University, Raleigh, NC.

Liu, Y., and W. Whitt. 2012. “Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals”. *Oper. Res.* 60 (6): 1551–1564.

Ma, N., and W. Whitt. 2015. “Efficient Simulation of Non-Poisson Non-Stationary Point Processes to Study Queueing Approximations”. Working paper, Columbia University, available at: www.columbia.edu.

Massey, W. A., and W. Whitt. 1994. “Unstable asymptotics for nonstationary queues”. *Math. Oper. Res.* 1:267–291.

Pang, G., R. Talreja, and W. Whitt. 2007. “Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues”. *Probability Surveys* 4:193–267.

Whitt, W. 1982. “Approximating a Point Process by a Renewal Process: Two Basic Methods”. *Operations Research* 30:125–147.

Whitt, W. 1983. “The Queueing Network Analyzer”. *Bell System Technical Journal* 62 (9): 2779–2815.

Whitt, W. 1984. “Departures From a Queue with Many Busy Servers”. *Mathematics of Operations Research* 9:534–544.

Whitt, W. 2014. “Stabilizing Performance in a Single-Server Queue with Time-Varying Arrival Rate”. Working paper, Columbia University, available at: www.columbia.edu, to appear in *Queueing Systems*.

AUTHOR BIOGRAPHIES

NI MA is a doctoral student in the Department of Industrial Engineering and Operations Research at Columbia University. She is currently working with Professor Ward Whitt on service engineering, using methodology in stochastic modeling, queueing theory, simulation and statistical analysis. Her email address is: nm2692@columbia.edu.

WARD WHITT is a Professor in the Department of Industrial Engineering and Operations Research at Columbia University. He joined the faculty there in 2002 after spending 25 years in research at AT&T. He received his Ph.D. from Cornell University in 1969. His recent research has focused on stochastic models of service systems, using both queueing theory and simulation. His e-mail address is: ww2040@columbia.edu.