Stochastic Systems

Minimizing the Maximum Expected Waiting Time in a Periodic Single-Server Queue with a Service-Rate Control

Ni Ma, Ward Whitt

Minimizing the Maximum Expected Waiting Time in a Periodic Single-Server Queue with a Service-Rate Control

Ni Ma,[a] Ward Whitt[a]

[a] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
Contact: nm2692@columbia.edu, http://orcid.org/0000-0003-2058-0498 (NM); ww2040@columbia.edu, http://orcid.org/0000-0003-4298-9964 (WW)

Abstract. We consider a single-server queue with unlimited waiting space, the first-come, first-served discipline, a periodic arrival-rate function, and independent and identically distributed service requirements, where the service-rate function is subject to control. We previously showed that a rate-matching control, whereby the service rate is made proportional to the arrival rate, stabilizes the queue-length process but not the (virtual) waiting-time process. To minimize the maximum expected waiting time (and stabilize the expected waiting time), we now consider a modification of the service-rate control involving two parameters: a time lag and a damping factor. We develop an efficient simulation search algorithm to find the best time lag and damping factor. That simulation algorithm is an extension of our recent rare-event simulation algorithm for the $GI_t/GI/1$ queue to the $GI_t/GI_t/1$ queue, allowing the time-varying service rate. To gain insight into these controls, we establish a heavy-traffic limit with periodicity in the fluid scale. This produces a diffusion control problem for the stabilization, which we solve numerically by the simulation search in the scaled family of systems with $\rho \uparrow 1$. The state space collapse in that theorem shows that there is a time-varying Little's law in heavy traffic, implying that the queue length and waiting time cannot be simultaneously stabilized in this limit. We conduct simulation experiments showing that the new control is effective for stabilizing the expected waiting time for a wide range of model parameters, but we also show that it cannot stabilize the expected waiting time perfectly.

1. Introduction
1.1. A Nonstationary Stochastic Design Problem

In this paper, we address an open problem from Whitt (2015), which considered the problem of stabilizing performance over time, that is, making a specified time-dependent performance measure a target constant function, in a single-server queue with unlimited waiting space, the first-come, first-served (FCFS) discipline, and a time-varying arrival-rate function. The stabilization is to be achieved with a deterministic service-rate function, under the assumption that the customer service requirements are specified independently of the service-rate control. This is a stochastic design problem instead of a real-time stochastic control problem; that is, the service-rate control is to be determined in advance, assuming full knowledge of the model, but without knowledge of the system state (e.g., the value of the stochastic queue-length process) that will actually prevail at any time.

As explained in section 1 of Whitt (2015), variants of this service-rate control are performed in response to time-varying demand, in many service operations, such as hospital surgery rooms and airport inspection lines, but little is known about the ideal timing and extent of service-rate changes. Service-rate controls for single-server queues are also of current interest within more complex systems, such as in energy-efficient data centers in cloud computing (Kwon and Gautam 2016) and in business process management (Suriadi et al. 2017).

In Whitt (2015) it was shown that a rate-matching control, whereby the service rate is made proportional to the arrival rate, stabilizes the queue-length process but not the (virtual) waiting-time process. In this paper we

develop an algorithm to approximately stabilize the expected waiting time at a target level. It uses a modification of the service-rate control involving two parameters: a time lag and a damping factor.

1.2. Related Literature

There is a large literature on similar stochastic design problems involving setting staffing levels (the number of servers) in a multiserver queue to stabilize performance in the face of time-varying demand (e.g., Jennings et al. 1996, Feldman et al. 2008, Stolletz 2008, Liu and Whitt 2012b, Defraeye and van Nieuwenhuyse 2013, He et al. 2016, Pender and Massey 2017, Liu 2018, Whitt 2018). For a single-server queue, the direct analog would be turning on and off the server, which is a restrictive extreme version of the service-rate control we consider.

The dynamic control problem of turning on and off the server in specified system states has received considerable attention in the stationary setting, starting with Yadin and Naor (1963) and Heyman (1968). Similar dynamic control problems for single-server queues including service-rate controls have been analyzed as Markov decision processes in George and Harrison (2001) and Adusumilli and Hasenbein (2010) and references therein. We emphasize that our design problem is different; our service-rate control is for a nonstationary model and must be set in advance, without knowledge of the system state. In many cases, our new problem is more realistic because arrival rates are often strongly time varying and can be reasonably well estimated in advance, whereas changes to the service rate may be difficult to implement without advance planning. Of course, in general, both problems are important.

Given the extensive research on the staffing design problem for many-server queues, it is natural to consider variants of the successful staffing algorithms, but it is now well known that the behavior of many-server queues tends to be dramatically different from that of single-server queues. That difference can be seen by comparing the many-server fluid models in Liu and Whitt (2012a) with the single-server fluid models in Chen and Mandelbaum (1991), as discussed in Liu and Whitt (2011, p. 836). A simple fluid model supporting the rate-matching control in Whitt (2015) is supported by our heavy-traffic weak law of large numbers in Theorem 4 (see Corollary 3), but we are working to go beyond that.

Hence, it should not be surprising that service-rate controls using variants of the established many-server staffing algorithms are no longer effective for single-server queues. For example, a natural analogue of the square-root staffing function from Jennings et al. (1996) was considered as a candidate service-rate control in equation 2.3 of Whitt (2015) but was found to be ineffective, as illustrated by figure 2 of Whitt (2015). Additionally, variants of the iterated staffing algorithm in Feldman et al. (2008) and Defraeye and van Nieuwenhuyse (2013) were found to be ineffective, evidently because the controls have impact over greater time intervals (are less "local") with single-server systems.

As indicated in Whitt (2015), controlling the service rate to meet time-varying demand is analogous to Kleinrock's classic service-capacity-allocation problem in a stationary Markovian Jackson network (Kleinrock 1964); we allocate service capacity over a time instead of over space (different queues within the network).

1.3. The Rate-Matching Service-Rate Control

Given that the service requirements are specified independently, the actual service times resulting from a time-varying control are relatively complicated, but a construction is given in section 3.1 of Whitt (2015). In Whitt (2015), several controls were considered, but most attention was given to the *rate-matching control*, which chooses the service rate to be proportional to the arrival rate; that is, for a given target traffic intensity ρ, the service-rate function is

$$\mu(t) \equiv \lambda(t)/\rho, \quad t \geq 0, \tag{1}$$

with $\equiv$ denoting equality by definition. In Whitt (2015), theorem 4.2 shows that the rate-matching control stabilizes the queue-length process, theorem 5.1 gives an expression for the waiting time with the rate-matching control, whereas theorems 5.2 and 5.3 establish heavy-traffic limits showing that the queue length is asymptotically stable, but the waiting time is not, being asymptotically inversely proportional to the arrival-rate function.

1.4. The Open Problem: Stabilizing the Expected Waiting Time

The open problem from Whitt (2015) is developing a service-rate control that can stabilize the expected waiting time. (We only discuss the continuous-time virtual waiting-time process in this paper, which is the waiting time of a potential or hypothetical customer if it were to arrive at that time, and so omit "virtual.")

Toward that end, we now study a modification of the rate-matching control. Without loss of generality, we write the periodic arrival-rate function as

$$\lambda(t) \equiv \rho(1 + s(t)), \quad t \geq 0, \tag{2}$$

where $0 < \rho < 1$, and s is a periodic function with period c satisfying

$$\bar{s} \equiv \frac{1}{c} \int_0^c s(u)\, du \equiv 0. \tag{3}$$

As a regularity condition, we require that

$$s_L \leq s(t) \leq s_U \quad \text{for all} \quad t \quad \text{with} \quad -1 \leq s_L \leq 0 \leq s_U < \infty. \tag{4}$$

Most of our numerical examples will be for a sinusoidal function, where $s(t) = \beta \sin(\gamma t)$ for $s(t)$ in (2), so that

$$\lambda(t) \equiv \rho(1 + \beta \sin(\gamma t)), \quad t \geq 0, \tag{5}$$

and β is the relative amplitude, with $0 \leq \beta \leq 1$, and the period is $c = 2\pi/\gamma$.

In the periodic setting of Equations (2)–(4), we consider the rate-matching control in (1) modified by a *time lag* η and *damping factor* ξ; in particular,

$$\mu(t) \equiv 1 + \xi s(t - \eta), \quad t \geq 0, \tag{6}$$

for $0 < \xi \leq 1$ and $\eta > 0$. Thus, the average arrival rate and service rate are $\bar{\lambda} = \rho$ and $\bar{\mu} = 1$, so the long-run traffic intensity is $\bar{\rho} \equiv \bar{\lambda}/\bar{\mu} = \rho$. However, the instantaneous traffic intensity $\rho(t) \equiv \lambda(t)/\mu(t)$ can satisfy $\rho(t) > 1$ for some t in each periodic cycle, for example, if $\rho(1 + \beta) > 1 - \beta$ or, equivalently, if $\beta > (1 - \rho)/(1 + \rho)$ in the setting of Equations (5) and (6).

1.5. Formulation of Optimal Control Problems

Because it is directly of interest, and because we want to allow for imperfect stabilization, we formulate our control problem as minimizing the maximum expected waiting time over a periodic cycle $[0, c]$. We formulate the main optimization problem as a min-max problem; that is,

$$w^* \equiv \min_{\mu(t) \in \mathcal{M}(1)} \max_{0 \leq y \leq 1} \{E[W_y]\}, \tag{7}$$

where $E[W_y]$ is the expected (periodic) steady-state (virtual) waiting time starting at time yc within a cycle of length c, $0 \leq y < c$, and $\mathcal{M}(m)$ is the set of all periodic service-rate functions with average rate m, which we take to be $m \equiv 1$.

Given that the average arrival rate is $\rho < 1$, the obvious reference case is the mean waiting time $E[W]$ in the associated stationary model, which for the $M/GI/1$ model is

$$E[W] = \frac{\rho(1 + c_s^2)}{2(1 - \rho)}, \tag{8}$$

and thus $E[W] = \rho/(1 - \rho)$ in the $M/M/1$ model. However, in general, $E[W]$ is not a lower bound for the average of the periodic steady-state mean $E[W_y]$ over a cycle (see Remark 2 and Example 1).

We have not yet solved this general optimization problem in Equation (7). Here are open problems, applying to the Markovian $M_t/M_t/1$ model and generalizations:

1. For the general periodic problem, what is the solution (value of w^* and set of optimal service-rate functions $\mu^*(t)$ as a function of the model)?

2. For the sinusoidal special case in Equation (5), what is the solution?

3. To what extent do the optimal solutions stabilize the expected waiting time $E[W_y]$ over time? In particular, is it possible to stabilize $E[W_y]$ perfectly?

Remark 1 (Stabilizing the Full Waiting-Time Distribution). Theorems 4.1 and 4.2 of Whitt (2015) show that rate-matching control stabilizes the delay probability $P(W_y > 0)$, whereas corollary 5.1 of Whitt (2015) shows that rate-matching control cannot stabilize the mean waiting time. Theorem 5.2 of Whitt (2015) establishes a heavy-traffic limit (with periodicity in the stronger fluid scale in Section 6 here) that shows that it is not possible to stabilize the queue-length and waiting-time processes simultaneously. Thus, we conclude that it is not possible to stabilize the

full waiting-time distribution. Hence, the open problems above are only for the mean. In this paper we primarily focus on the mean, but we also show that it stabilizes the entire distribution to some extent in Section 9.1.

In this paper we only consider the restricted set of controls in Equation (6). Now our goal is

$$w^*(\eta, \xi) \equiv \min_{\eta, \xi} \max_{0 \leq y \leq 1} \{E[W_y]\}. \tag{9}$$

For practical purposes, this two-parameter control is appealing for its simplicity. We also find that it is quite effective, even though it cannot stabilize $E[W_y]$ perfectly.

We also consider the associated stabilization control, whereby Equation (9) is replaced by

$$w^*_{\text{stab}}(\eta, \xi) \equiv \min_{\eta, \xi} \{\max_{0 \leq y \leq 1} \{E[W_y]\} - \min_{0 \leq y \leq 1} \{E[W_y]\}\}. \tag{10}$$

In our sinusoidal examples, in which there is strong symmetry, we find that the solutions to Equations (9) and (10) are the same (but we have no proof), but neither stabilizes perfectly. For more general periodic arrival-rate functions, we detect differences.

1.6. Organization of This Paper

This paper involves some challenging technical methods. Hence, we present the more accessible results first. We start in Section 2 by presenting two simulation examples to illustrate the effectiveness of our new algorithm. Then, in Section 3, we introduce the two technical tools we will apply: (1) an extension of the rare-event simulation algorithm for the $GI_t/GI/1$ model from Ma and Whitt (2018) to the $GI_t/GI_t/1$ model with a general service-rate control and (2) heavy-traffic limits involving scaling of the underlying deterministic periodic arrival-rate function.

We start in earnest in Section 4. We elaborate on the model and key processes representing the workload and the waiting time in Section 4. Theorem 1 shows that the rate-matching control stabilizes the workload process as well as the queue-length process. We discuss the extension of the rare-event simulation algorithm from Ma and Whitt (2018) to our setting and its application to perform simulation search in Section 5. In both Sections 4 and 5, we will be brief because we can draw on Whitt (2015) and Ma and Whitt (2018).

We establish our main heavy-traffic limits with periodicity in the stronger fluid scaling [see Equation (13)] in Section 6. We present the proof of the main heavy-traffic limit (Theorem 3) in Section 7. We establish heavy-traffic limits with periodicity in the weaker diffusion scaling [see Equation (14)] in Section 8.

We give simulation examples in Section 9. In Section 9.1, we present simulation results using the fluid scaling in Section 6; in Section 9.2, we present simulation results using the diffusion scaling in Section 8. We draw conclusions in Section 10.
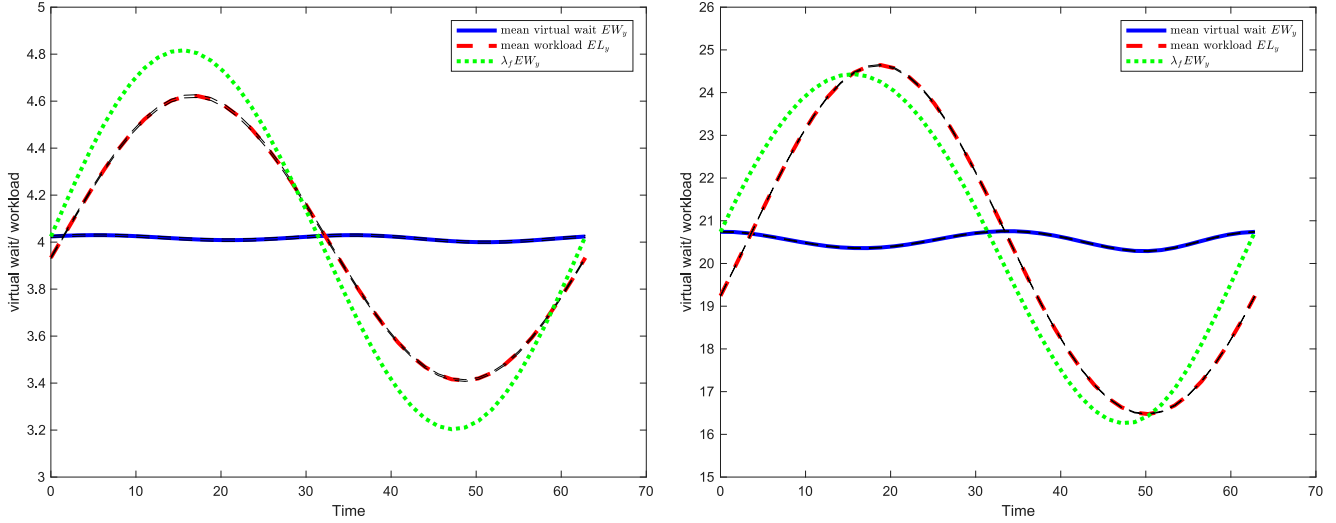
2. Simulation Examples for the $M_t/M_t/1$ Model

To illustrate the effectiveness of our new algorithm, we show results for two simulation examples. We consider the Markovian $M_t/M_t/1$ model with the sinusoidal arrival-rate function in Equations (2)–(5). The first example has model parameters $(\rho, \beta, \gamma) = (0.8, 0.2, 0.1)$, so the average arrival rate is $\bar{\rho} = 0.8$, the average service time is 1, and the cycle length is $c = 2\pi/\gamma = 62.8$. Figure 1 (left) shows the expected steady-state waiting time $E[W_y]$ together with the corresponding expected workload $E[L_y]$ and the product $\lambda(y)E[W_y]$, all for $0 \leq y < 1$. The second example on the right differs only by increasing ρ from 0.8 to 0.95. Figure 1 also shows the upper and lower 95% confidence interval bounds for $E[L_y]$ and $E[W_y]$ with black dashed lines, but these can only be seen by zooming in.

Figure 1 shows that the expected waiting time $E[W_y]$ is well stabilized at a value somewhat higher than the expected steady-state waiting time for the stationary $M/M/1$ model, which is $\rho/(1 - \rho)$ (4 on the left and 19 on the right). The maximum deviation (maximum − minimum) over a cycle is 0.0335 is for $\rho = 0.8$ and 0.4653 for $\rho = 0.95$. Thus, the maximum relative errors are approximately 0.8% for $\rho = 0.8$ and 2.2% for $\rho = 0.95$, clearly adequate for practical applications. Nevertheless, careful simulations and statistical analysis allow us to conclude that it is impossible to stabilize the expected waiting time perfectly with this control. To see the contrasting view with rate-matching control for this same model, see figure 6 of Whitt (2015). (See Ma and Whitt 2015 for more examples.)

It is natural to wonder whether there is any order in the optimal controls found for $\rho = 0.8$ and $\rho = 0.95$ in Figure 1. The dependence on ρ is revealed by the main heavy-traffic limit theorem (Theorem 3).

Figure 1. Estimates of the Periodic Steady-State Values of $E[W_y]$ (Blue Solid Line), $E[L_y]$ (Red Dashed Line), and $\lambda(y)E[W_y]$ (Green Dotted Line) for the Optimal Control (η^*, ξ^*) for the Sinusoidal Example with Parameter Triples $(\rho, \beta, \gamma) = (0.8, 0.2, 0.1)$ (Left) and $(0.95, 0.2, 0.1)$ (Right), So That the Cycle Length Is $c = 2\pi/\gamma = 62.8$

Note. The optimal controls are $(5.84, 0.84)$ for $\rho = 0.8$ and $(15.1, 2.13)$ for $\rho = 0.95$.

Remark 2 (The Cost of Periodicity). The difference between the stable average waiting time in Figure 1 and the value $\rho/(1 - \rho)$ for the stationary model (4 on the left and 19 on the right) might be called the "average cost of periodicity," but we point out that the overall average waiting time with a service-rate control could be much less than in the stationary model. The classic results for the periodic $M_t/GI/1$ queue in Rolski (1981, 1989) do not apply because, in general, the service times are neither independent of the arrival process nor independent and identically distributed (i.i.d.); see Example 1.

Example 1 (Small Expected Waiting Times with Periodicity). To illustrate a nonstationary model with a low average expected waiting time, consider the $M_t/M_t/1$ model with the two-level arrival-rate function with period c:

$$\lambda(t) \equiv \rho b 1_{[(c/2)-\delta, (c/2)+\delta)}(t), \quad 0 \le t < c \quad \text{and} \quad b\delta = c, \tag{11}$$

where $\delta < c/2$, and 1_A is the indicator function of the set A; that is, $1_A(t) = 1$ if $t \in A$ and 0 otherwise. Let the service-rate function be as in Equation (6) with $\eta = 2\delta$ and $\xi = 1$. Then the number of arrivals in the interval $[(c/2) - \delta, (c/2) + \delta)$ has a Poisson distribution with mean ρc, whereas the number of potential departures in the interval $[(c/2) + \delta, (c/2) + 3\delta)$ has a Poisson distribution with mean c. Thus, for $\rho < 1$ and $c = b\delta$ suitably large, the net input over the interval $[(c/2) - \delta, (c/2) + \delta)$ is approximately Gaussian with mean $-(1 - \rho)b\delta$ and variance $(1 + \rho)b\delta$, which is unlikely to be positive. By choosing δ suitably small and $b\delta$ suitably large, subject to specified ρ, we can make the maximum steady-state expected waiting time, and thus the average, approach 0. One way to explain this phenomenon is to observe that the interarrival times and service times will be highly correlated.

Remark 3 (The Single-Parameter Alternative). It is natural to wonder whether we could use only the single control parameter η, fixing $\xi = 1$. If we let $\xi = 1$ and optimize over η in the setting of Figure 1, then, for $\rho = 0.8$ ($\rho = 0.95$), we get $\eta^* = 5.93$ ($\eta^* = 28.3$) and a maximum deviation of 0.4109 (3.034), which yields approximately 10% (14%) relative error instead of 0.8% (2.2%). Hence, we use the two control parameters.

3. The Key Technical Tools
In this section, we discuss the two technical tools that we use.

3.1. The Primary Tool: A Simulation Search Algorithm
Our primary tool for finding good (η, ξ) controls is a simulation search algorithm. For that purpose, we extend the rare-event simulation algorithm for the time-varying workload process in the periodic $GI_t/GI/1$ model in Ma and Whitt (2018) to the $GI_t/GI_t/1$ model, where the service rate is time varying as well. (The notation GI_t means that the process is a deterministic time transformation of a renewal process; see Section 4.) The

workload $L(t)$ represents the amount of work in service time in the system at time t, whereas the waiting time can be represented as the first-passage time

$$W(t) = \inf \left\{ u \geq 0 : \int_t^{t+u} \mu(s)\,ds = L(t) \right\}. \tag{12}$$

The waiting time $W(t)$ coincides with the workload $L(t)$ when $\mu(t) = 1$ for all t, but not otherwise.

As in Ma and Whitt (2018), the rare-event simulation algorithm calculates the periodic steady-state workload L_y and waiting time W_y, starting at time yc within a cycle of length c, $0 \leq y < 1$. We use a search over the parameters (η, ξ), as discussed in Section 5, to solve the optimization problems [Equations (9) and (10)]. The search part is relatively elementary because we have only two control parameters. For background on simulation optimization, see Fu (2015) and Jian and Henderson (2015) and the references therein.

The computational complexity for one control vector (η, ξ) is essentially the same as in Ma and Whitt (2018). In particular, the program running time tends to be proportional to the number of replications and number of y values, which for the case $\rho = 0.8$ in Figure 1 were taken to be 40,000 and 40, respectively. This required approximately 100 minutes on a desktop computer. As indicated in section 4.7 of Ma and Whitt (2018), the run time tends to be of order $(1 - \rho)^{-1}$, so the cases with high traffic intensity are more challenging. The simulation search is performed in stages, with fewer y values and replications in the early stages but the full long run at the end to confirm performance.

3.2. Gaining Additional Insight: Heavy-Traffic Limits

To better understand how the control parameters and performance depend on the model parameters, we establish heavy-traffic (HT) limits, which involve considering a family of models indexed by ρ and letting $\rho \uparrow 1$, drawing on our previous work in Whitt (2014, 2015) and Ma and Witt (2018). That previous work shows that the scaling is very important because there are several possibilities. We use the conventional HT scaling of time by $(1 - \rho)^{-2}$ (usually denoted by n) and space by $1 - \rho$ (usually denoted by $1/\sqrt{n}$), as in chapters 5 and 9 of Whitt (2002), but if we do so without also scaling the arrival-rate function, then the HT limit is easily seen to be the same as if the periodicity were replaced by the constant long-run average, as shown by Falin (1989).

To obtain insight into the periodic dynamics, it is thus important to also scale the arrival-rate function, which is initially specified in Equation (2) with Equations (3) and (4). However, the work of Whitt (2014, 2015) actually uses two different HT scalings of the arrival-rate function. *Our main HT scaling in Section 6 follows Whitt (2015) and has periodicity in the fluid scale*; that is,

$$\lambda_\rho(t) \equiv \rho(1 + s((1 - \rho)^2 t)), \quad t \geq 0, \tag{13}$$

but in Section 8 we also consider the scaling from Whitt (2014) and Ma and Whitt (2018), which has the periodicity in diffusion scale; that is,

$$\lambda_\rho(t) \equiv \rho(1 + (1 - \rho)s((1 - \rho)^2 t)), \quad t \geq 0. \tag{14}$$

The extent of the periodicity is stronger in Equation (13) than in Equation (14) because of the extra factor $(1 - \rho)$ before s in Equation (14). The workload and the waiting time have the same HT limit with the diffusion-scale scaling in Equation (14) but different limits with the fluid-scale scaling in Equation (13). To capture the clear differences shown in Figure 1, we obviously want the stronger fluid scaling in Equation (13). The HT functional central limit theorem (FCLT) in Theorem 3 for the scaling in Equation (13) in Section 6 helps interpret Figure 1.

It is important to note that if we have a constant service rate with this scaling, then the waiting times explode as $\rho \uparrow 1$ because the instantaneous traffic intensity $\rho(t) \equiv \lambda(t) > 1$ over intervals growing as $\rho \uparrow 1$; this case is analyzed in Choudhury et al. (1997).

We also establish an HT functional weak law of large numbers (FWLLN) in Theorem 2, which yields a deterministic fluid approximation. However, it is not very useful because it shows that our proposed control with $\xi = 1$ stabilizes the waiting time perfectly for all η as $\rho \uparrow 1$ (but it helps to see that nothing bad happens).

4. The Model

In this section, we specify the general model, defining the arrival process in Section 4.1 and the basic queuing stochastic processes in Section 4.2. We specialize to the periodic $G_t/G_t/1$ model in Section 4.3. We show that the workload is stabilized by the rate-matching control in Equation (1), extending the results for the queue-length process in Whitt (2015).

4.1. The Arrival Process

We represent the periodic arrival counting process A as a deterministic time transformation of an underlying rate-1 counting process N with associated sequence of interarrival times $\{U_k : k \geq 1\}$ by

$$A(t) \equiv N(\Lambda(t)), \quad \text{where} \quad \Lambda(t) \equiv \int_0^t \lambda(s)\,ds, \quad t \geq 0. \tag{15}$$

where λ is the arrival-rate function. This is a common representation when N is a rate-1 Poisson process; then A is a "nonhomogeneous Poisson process" (NHPP). For the $G_t/G_t/1$ model, N is understood to be a rate-1 stationary point process. Hence, for the $GI_t/GI_t/1$ model, N is an equilibrium renewal process with time between renewals having mean 1, for which the first interrenewal time U_1 has the equilibrium distribution. The representation in Equation (15) has been used frequently for processes N more general than NHPPs, an early source being by Massey and Whitt (1994).

For the sinusoidal arrival-rate function in Equation (5), the associated cumulative arrival-rate function is

$$\Lambda(t) = \rho(t + (\beta/\gamma)(1 - \cos(\gamma t))), \quad t \geq 0. \tag{16}$$

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. For the case of a constant service rate, there are two important cases for the relative amplitude: (1) $0 < \beta < \rho^{-1} - 1$ and (2) $\rho^{-1} - 1 \leq \beta \leq 1$. 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). However, this difficulty can be avoided by a service-rate control.

4.2. The General $G_t/G_t/1$ Model with a Service-Rate Control

We consider a modification of the standard single-server queue with unlimited waiting space where customers are served in order of arrival. Let $\{V_k\}$ be the sequence of service requirements. As in Whitt (2015), we separately define the rate at which service is performed from the service requirement. Given the arrival counting process $A(t)$ defined in Section 4.1, let the total input of work over the interval $[0, t]$ be the random sum

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0. \tag{17}$$

Let service be performed at time t at rate $\mu(t)$ whenever there is work to perform. Paralleling the cumulative arrival rate $\Lambda(t)$ defined in Equation (15), let the cumulative available service rate be

$$M(t) \equiv \int_0^t \mu(s)\,ds, \quad t \geq 0. \tag{18}$$

Let the net-input process of work be $X(t) \equiv Y(t) - M(t)$, $t \geq 0$. Then we can apply the reflection map to the net input process $X(t)$ to represent the workload (the remaining work in service time) at time t, starting empty at time 0, as

$$L(t) = X(t) - \inf\{X(s) : 0 \leq s \leq t\} = \sup\{X(t) - X(s) : 0 \leq s \leq t\}, \quad t \geq 0.$$

In this setting it is elementary that the continuous-time (virtual) waiting time (before starting service) at time t, which we denote by $W(t)$, can be related to $L(t)$.

Lemma 1 (Waiting Time Representation). *The waiting time at time t can be represented as*

$$W(t) = M_t^{-1}(L(t)), \quad t \geq 0, \tag{19}$$

where M_t^{-1} is the inverse of $M_t(u) \equiv M(t + u) - M(t)$ for $M(t)$ in Equation (18).

Proof. By definition,

$$W(t) = \inf\left\{u \geq 0 : \int_t^{t+u} \mu(s)\,ds = L(t)\right\}$$

$$= \inf\{u \geq 0 : M(t + u) - M(t) = L(t)\} = M_t^{-1}(L(t)), \tag{20}$$

for $M_t(u)$ above, as claimed in Equation (19). □

4.3. The Periodic $G_t/G_t/1$ Model

As in Ma and Whitt (2018), we consider the periodic steady state of the periodic $G_t/G_t/1$ model with arrival-rate function in Equation (2). For that purpose, we exploit the arrival-process construction in Equation (15) in terms of the stationary processes $N \equiv \{N(t) : t \geq 0\}$ and $V \equiv \{V_k : k \geq 1\}$ in Equation (15). Let the associated service-rate function $\mu(t)$ also be periodic with cycle length c, average service rate $\bar{\mu} = 1$, and bounds $0 < \mu_L \leq \mu(t) \leq \mu_U < \infty$, for $0 \leq t \leq c$.

As in Ma and Whitt (2018) and earlier in Loynes (1962) and chapter 6 in Sigman (1995), we now convert the standard representation of the workload process in Section 4 to a simple supremum by using a reverse-time construction. To do so, we extend the stationary processes $\{N(t)\}$ and $\{V_k\}$ to the entire real line. We regard the periodic arrival rate and service rate as defined on the entire real line as well, with the functions fixed by their position within the periodic cycle at time 0. With those conditions, the reverse-time construction is achieved by letting the interarrival times and service times be ordered in reverse time going backward from time 0. Then $\tilde{A}(t)$ counts the number of arrivals in $[-t, 0]$, $\tilde{Y}(t)$ is the total input in $[-t, 0]$, and $\tilde{X}(t)$ is the net input in $[-t, 0]$, for $t \geq 0$.

To exploit the reverse-time representation, let

$$\tilde{\Lambda}_y(t) \equiv \Lambda(yc) - \Lambda(yc - t), \quad t \geq 0, \tag{21}$$

be 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.

As an analogue of Equation (21) for the cumulative service rate, let

$$\tilde{M}_y(t) \equiv M(yc) - M(yc - t), \quad t \geq 0. \tag{22}$$

We let the service requirements V_k come from a general stationary sequence with $E[V_k] = 1$.

With this reverse-time representation, the workload at time yc in the system starting empty at time $yc - t$ can be represented as

$$
\begin{aligned}
L_y(t) &= \sup_{0 \leq s \leq t} \{\tilde{X}_y(s)\} \\
&\stackrel{d}{=} \sup_{0 \leq s \leq t} \left\{ \sum_{k=1}^{N(\tilde{\Lambda}_y(s))} V_k - \tilde{M}_y(s) \right\} = \sup_{0 \leq s \leq \tilde{\Lambda}_y(t)} \left\{ \sum_{k=1}^{N(s)} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(s)) \right\},
\end{aligned}
\tag{23}
$$

where $\tilde{X}_y$ is the reverse-time net input of work starting at time yc within the cycle of length c. The other quantities in Equation (23) are the reverse-time cumulative arrival-rate function $\tilde{\Lambda}_y(t)$ in Equation (21) with inverse $\tilde{\Lambda}_y^{-1}(t)$ and the reverse-time cumulative service-rate function $\tilde{M}_y$ in Equation (22) with inverse $\tilde{M}_y^{-1}$.

The equality in distribution in Equation (23) holds because N is a stationary point process, which is a point process with stationary increments and a constant rate.

As $t \to \infty$, $L_y(t) \uparrow L_y(\infty) \equiv L_y$ with probability one (w.p.1) as $t \to \infty$, for

$$L_y \stackrel{d}{=} \sup_{s \geq 0} \left\{ \sum_{k=1}^{N(s)} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(s)) \right\}, \qquad 0 \leq y < 1. \tag{24}$$

Even though Equation (23) is valid for all t, we think of the system starting empty at times $-kc$, for $k \geq 1$, so we let $yc - t = -kc$ or, equivalently, we stipulate that $t = c(k + y)$, $0 \leq y < c$, and consider successive values of k and let $k \to \infty$ to get Equation (24). That makes Equation (23) valid to describe the distribution of $L(c(k + y))$ for all $k \geq 1$.

We now observe that the time transformation in Equation (23) shows that the periodic $G_t/G_t/1$ model is actually equivalent to a $G/G_t/1$ model with a stationary arrival process and a new cumulative service-rate function $\tilde{M}_y(\tilde{\Lambda}_y^{-1}(t))$.

Corollary 1 (Conversion of $G_t/G_t/1$ to an Equivalent $G_t/G/1$). *In addition to representing the periodic steady-state workload L_y in a periodic $G_t/G_t/1$ model as a periodic steady-state workload in a periodic $G/G_t/1$ model, which has a stationary stochastic input and a deterministic service rate, as shown in Equation (24), we can represent it as a periodic steady-state workload in a periodic $G_t/G/1$ model, which has a periodic stochastic input and a constant service rate, via*

$$L_y = \sup \left\{ \sum_{k=1}^{N(\tilde{\Lambda}_y(\tilde{M}_y^{-1}(s)))} V_k - s : s \geq 0 \right\}. \tag{25}$$

Corollary 2 (The Associated Periodic Steady-State Waiting Time). *The periodic steady-state waiting time associated with the periodic steady-state workload in Equation* (24) *is*

$$W_y = \tilde{M}_y^{-1}(L_y), \quad 0 \le y < 1. \tag{26}$$

Proof. Apply the reasoning of Lemma 1. $\square$

In Whitt (2015), we showed that the rate-matching service-rate control in Equation (1) stabilizes the queue-length process. Now we establish the corresponding result for the workload.

Theorem 1 (Stabilizing the Periodic Workload). *If the rate-matching control in Equation* (1) *is used, then* $L_y \stackrel{d}{=} L$ *for* L_y *in Equation* (24), *where* L *is the steady-state workload in the associated (stable) stationary G/G/1 model; that is,*

$$L \stackrel{d}{=} \sup_{s \ge 0} \left\{ \sum_{k=1}^{N(s)} V_k - \rho^{-1}s \right\}, \tag{27}$$

which is independent of y.

Proof. With the rate-matching control, we have $M(t) = c\Lambda(t)$ and $\tilde{M}_y(t) = c\tilde{\Lambda}_y(t)$, $t \ge 0$. As a consequence, $\tilde{M}_y(\tilde{\Lambda}_y^{-1}(t)) = ct$, $t \ge 0$, so

$$L_y \stackrel{d}{=} \sup_{s \ge 0} \left\{ \sum_{k=1}^{N(s)} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(s)) \right\}$$

$$\stackrel{d}{=} \sup_{s \ge 0} \left\{ \sum_{k=1}^{N(s)} V_k - cs \right\} \stackrel{d}{=} L. \quad \square \tag{28}$$

5. The Simulation Search Algorithm

The rare-event simulation algorithm from Ma and Whitt (2018) exploits the classic rare-event simulation algorithm for the $GI/GI/1$ queue, exploiting importance sampling using an exponential change of measure, as in chapter XIII of Asmussen (2003) and chapter VI of Asmussen and Glynn (2007). Hence, our simulation algorithm applies to the $GI_t/GI_t/1$ queue. It was shown in Ma and Whitt (2018) that the algorithm is effective for estimating the mean as well as small tail probabilities (also see Ma and Whitt 2016).

5.1. The $GI_t/GI_t/1$ Model

In the $GI_t/GI_t/1$ setting, the underlying rate-1 process N is an equilibrium-renewal process, which means that U_1 has the stationary-excess or equilibrium distribution U_e, which may be different from the i.i.d. distributions of U_k, $k \ge 2$. Also in the $GI_t/GI_t/1$ setting, the service times V_k are i.i.d. with distribution V and are independent of the arrival process.

The simulation algorithm exploits the discrete-time representation of the workload L_y in Equation (24) and the waiting time W_y; that is,

$$L_y \stackrel{d}{=} \sup_{s \ge 0} \left\{ \sum_{k=1}^{N(s)} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(s)) \right\}$$

$$\stackrel{d}{=} \sup_{n \ge 0} \left\{ \sum_{k=1}^{n} V_k - \tilde{M}_y\left(\tilde{\Lambda}_y^{-1}\left(\sum_{k=1}^{n} U_k \right) \right) \right\},$$

$$W_y \stackrel{d}{=} M_y^{-1}(L_y), \qquad 0 \le y < 1, \tag{29}$$

where M_y is the same as M_t, which is the forward integral of the service rate starting from position y within a cycle.

We exploit the rare-event simulation algorithm in Ma and Whitt (2018), which is based on an exponential change of measure; we refer to Ma and Whitt (2018) for background. In that setting, we use the underlying measure P_{θ^*} determined for $GI/GI/1$ queue. We again use the same notations $X_k(\rho) = V_k - \rho^{-1}U_k$ and partial sum process $S_n \equiv \sum_{k=1}^{n} X_k$ for $GI/GI/1$ and define the new associated process

$$Q_n \equiv \sum_{k=1}^{n} V_k - \tilde{M}_y\left(\tilde{\Lambda}_y^{-1}\left(\sum_{k=1}^{n} U_k \right) \right),$$

which is the process inside the supremum function. To avoid duplication of notation, we let the likelihood function here be denoted by Ψ instead of L. Then the estimator of the rare-event probability for W_y can be derived as follows:

$$
\begin{aligned}
P(W_y > b) &= P(M_y^{-1}(L_y) > b) = P(L_y > M_y(b)) \\
&= P(\tau_{M_y(b)}^Q < \infty) = E_{\theta^*}[\Psi_{\tau_{M_y(b)}^Q}(\theta^*)] \\
&= E_{\theta^*}[m_{X_1}(\theta^*) m_X(\theta^*)^{(\tau_{M_y(b)}^Q - 1)} \exp(-\theta^* S_{\tau_{M_y(b)}^Q})] \\
&= m_{X_1}(\theta^*) E_{\theta^*}[\exp(\theta^* S_{\tau_{M_y(b)}^Q})],
\end{aligned}
\tag{30}
$$

where $\tau_{M_y(b)}^Q$ is the stopping time of process Q_n at level $M_y(b)$, $\Psi_{\tau_{M_y(b)}^Q}(\theta^*)$ is the exponentially tilted likelihood ratio for process Q_n at $n = M_y(b)$, and $m_X(\theta^*)$ is the moment-generating function of X at θ^*. As in Ma and Whitt (2018), the first $X_1(\rho)$ in the partial sum $S_{\tau_{M_y(b)}^Q}$ has a different distribution from $\{X_k, k \geq 2\}$.

5.2. The Extended Algorithm for the $GI_t/GI_t/1$ Model

Here is a summary of the extended algorithm to estimate the tail probabilities in the $GI_t/GI_t/1$ queue with average service rate 1 and average arrival rate ρ:

1. Construct a table of the inverse cumulative arrival-rate function $\rho\tilde{\Lambda}_y^{-1}$ (same as for $GI_t/GI/1$).
2. Determine the required length of partial sums (n_s) needed in each application (same as for $GI_t/GI/1$).
3. For each replication, generate the required vectors of exponentially tilted interarrival times $\rho^{-1}\tilde{U}$ and service times $\tilde{V}$ from $F_{\rho^{-1}U}^{-\theta^*}$ and $F_V^{\theta^*}$, respectively (same as for $GI_t/GI/1$).
4. Calculate the associated vectors of S_n and Q_n, and find out the stopping time $\tau_{M_y(b)}^Q$, which is the hitting time of Q_n at level $M_y(b)$. This step is different from that for $GI_t/GI/1$ in that first we need to calculate $M_y(b)$ as the hitting level instead of b, and second we calculate vector Q_n different from R_n in an additional function $\tilde{M}_y$ in the second term.
5. Use the preceding estimator to calculate the tail probability $P(W_y > b)$ for each replication (same as for $GI_t/GI/1$).
6. Run N i.i.d. replications, and calculate the mean of the estimated values of $P(W_y > b)$ (same as for $GI_t/GI/1$).

5.3. Explicit Representations for the Sinusoidal Case

Here we summarize the expressions for all the basic deterministic rate functions in our sinusoidal examples, extending Equations (5), (6), and (16):
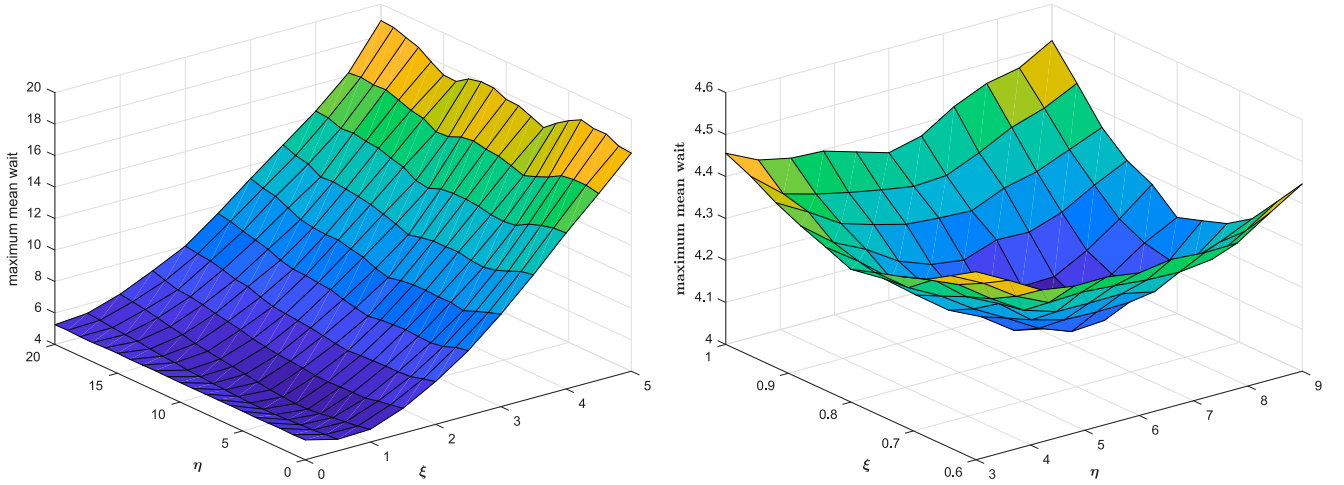
$$
\begin{aligned}
\tilde{\Lambda}_y(t) &= \rho\left(t + \frac{\beta}{\gamma}(\cos(\gamma(t - yc)) - \cos(\gamma yc))\right), \\
M(t) &= t - \xi\frac{\beta}{\gamma}(\cos(\gamma(t - \eta)) - \cos(\gamma\eta)), \\
M_y(t) &= t - \xi\frac{\beta}{\gamma}(\cos(\gamma(t + yc - \eta)) - \cos(\gamma(yc - \eta))), \\
\tilde{M}_y(t) &= t + \xi\frac{\beta}{\gamma}(\cos(\gamma(t + \eta - yc)) - \cos(\gamma(\eta - yc))).
\end{aligned}
\tag{31}
$$

5.4. The Search Algorithm

We use an elementary iterative search algorithm, fixing an initial value of η at the mean for the steady-state model, $\rho/(1 - \rho)$, and searching first over ξ and then over each variable until we get negligible improvement. This simple approach is substantiated by estimating the structure of the objective function. Figure 2 illustrates by showing the maximum waiting time $\max_{0 \leq y \leq c}\{E[W_y]\}$ in the setting of Figure 1 (left). Figure 2 shows estimates of the maximum waiting time $\max_{0 \leq y \leq c}\{E[W_y]\}$ as a function of (η, ξ) in $[0, 20] \times [0, 5]$ (left) and $[3, 9] \times [0.6, 1.0]$ (right) in that setting. Figure 2 shows that the function is not convex as a function of η but suggests that it is unimodal with a unique global minimum, supporting our simple procedure. The plots for the maximum deviation $\max_{0 \leq y \leq c}\{E[W_y]\} - \max_{0 \leq y \leq c}\{E[W_y]\}$ are similar.

We perform the search with fewer points y and replications in the initial stages and then confirm with more points, 40 values of y and 40,000 replications, which yield excellent statistical precision, as can be seen from the narrow confidence interval bands in Figure 1.

Figure 2. Three-Dimensional Plots of Estimates of the Maximum Waiting Time $\max_{0 \leq y \leq c}\{E[W_y]\}$ for (η, ξ) in $[0,20] \times [0,5]$ (Left) $[3,9] \times [0.6, 1.0]$ (Right)

6. Supporting Heavy-Traffic Limits for Periodic Queues

In this section, we obtain an HT FWLLN and an HT FCLT for the periodic $G_t/G_t/1$ model with a general service-rate control of the form in Equation (6). The HT FCLT produces a limit depending on an asymptotic time lag $\hat{\eta}$ and damping factor $\hat{\xi}$, which arise from HT limits; see condition (58) in Theorem 3 and the conclusion in Equation (50). Thus, we reduce the optimization problems over the parameter pairs (η_ρ, ξ_ρ) in Equations (9) and (10), asymptotically as $\rho \uparrow 1$, to diffusion-control problems with the parameter pairs $(\hat{\eta}, \hat{\xi})$.

6.1. The Underlying Rate-1 Processes

As in much of the HT literature, we start by introducing basic rate-1 stochastic processes, but here we consider service requirements instead of service times. We assume that the rate-1 arrival and service-requirements processes N and V specified in Section 4 are independent and each satisfies an FCLT. To state the result, let $\hat{N}_n^a$ and $\hat{S}_n^v$ be the scaled processes defined by

$$\hat{N}_n^a(t) \equiv n^{-1/2}[N^a(nt) - nt] \quad \text{and} \quad \hat{S}_n^v(t) \equiv n^{-1/2}\left[\sum_{i=1}^{\lfloor nt \rfloor} V_k - nt\right], \quad t \geq 0, \tag{32}$$

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

$$\hat{N}_n^a \Rightarrow c_a B_a \quad \text{and} \quad \hat{S}_n^v \Rightarrow c_s B_s \quad \text{in} \quad \mathscr{D} \quad \text{as} \quad n \to \infty, \tag{33}$$

where $\mathscr{D}$ 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), whereas B_a and B_s are independent standard (mean 0, variance 1) Brownian motion processes (BMs). The assumed independence implies joint convergence in Equation (33) by theorem 11.4.4 of Whitt (2002).

We emphasize that *GI* assumptions are not needed, but this is an important special case. If the service times V_k are i.i.d. mean-1 random variables with variance, also the squared coefficient of variation (scv), c_s^2, then the limit in Equation (33) 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 (and scv) c_a^2, then the limit in Equation (33) holds with arrival variability parameter c_a. (See Nieuwenhuis 1989 for theoretical support in the case of an equilibrium renewal process.)

For the queuing HT FCLT, we will apply theorem 9.3.4 of Whitt (2002), which refers to the conditions of theorem 9.3.3. Those conditions require a joint FCLT for the partial sums of the arrival and service processes, notably (3.9) on p. 295. This convergence follows from the FCLTs we assumed for $\hat{N}_n^a$ and $\hat{S}_n^v$ in Equation (33). In particular, the assumed FCLT for N_n^a 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).

6.2. A Family of Models

As a basis for the HT FCLT, we create a model for each ρ, $0 < \rho < 1$. We do so by defining the arrival-rate and service-rate functions.

6.2.1. The Arrival-Rate and Service-Rate Functions.

Let the arrival-rate function in model ρ be as in Equation (13) in the setting of Equations (2)–(4). As a further regularity condition, we also require that the function s be an element of the function space $\mathcal{D}$, as in Whitt (2002). Then the associated cumulative arrival-rate function in model ρ is

$$\Lambda_\rho(t) \equiv \rho\big(t + (1 - \rho)^{-2}S((1 - \rho)^2 t), \quad t \geq 0, \tag{34}$$

where

$$S(t) \equiv \int_0^t s(u)\, du, \tag{35}$$

for s again being the periodic function in Equations (2)–(4). From Equations (34) and (35), we see that the associated arrival-rate function obtained by differentiation in Equation (34) is indeed $\lambda_\rho(t)$ in Equation (33).

The time scaling in Equations (33) and (34) implies that the period in model ρ with arrival-rate function $\lambda_\rho(t)$ in Equation (33) is $c_\rho = c(1 - \rho)^{-2}$, where c is the period of $s(t)$ in Equations (2)–(4). Thus, the period c_ρ in model ρ is growing with ρ. This scaling follows lemma 5.1 and theorem 5.2 of Whitt (2015), with n there replaced by $(1 - \rho)^{-2}$. In particular, the scaling here is in fluid or FWLLN scale and thus is different from the diffusion or FCLT scaling in theorem 3.2 of Whitt (2014) and theorem 2 of Ma and Whitt (2018).

Let $A_\rho(t) \equiv N^a(\Lambda_\rho(t))$ be the arrival process in model ρ, which is obtained by using the cumulative arrival-rate function Λ_ρ in Equation (34) in place of Λ in Equation (15). Given that definition, we see that the cumulative arrival rate is indeed

$$E[A_\rho(t)] = E[N^a(\Lambda_\rho(t))] = \Lambda_\rho(t), \quad t \geq 0. \tag{36}$$

We now define associated scaled time-varying service-rate functions. These are the rate-matching service-rate functions in Whitt (2015) modified by a time lag and a damping factor. In particular,

$$\mu_\rho(t) \equiv 1 + \xi_\rho s((1 - \rho)^2(t - \eta_\rho)) \quad \text{and}$$

$$M_\rho(t) \equiv \int_0^t \mu_\rho(u)\, du = t$$
$$\qquad\qquad + (1 - \rho)^{-2}\xi_\rho S((1 - \rho)^2(t - \eta_\rho)), \quad t \geq 0, \tag{37}$$

where s is the periodic function with period c in Equation (3), whereas η_ρ is the ρ-dependent *time lag* and ξ_ρ is the ρ-dependent *damping factor*. From Equations (37) and (3), we see that the average service rate is $\bar{\mu}_\rho = 1$ for all ρ. As a consequence, the average traffic intensity is $\bar{\lambda}_\rho/\bar{\mu}_\rho = \rho$ for all ρ, whereas the instantaneous traffic intensity at time t is $\lambda_\rho(t)/\mu_\rho(t)$, $t \geq 0$, which is a more complicated periodic function, again with period c.

6.2.2. The Associated Queuing Processes.

Having defined the family of arrival processes $A_\rho(t)$ and deterministic service-rate functions $M_\rho(t)$ above, we define the other queuing processes $Y_\rho(t)$, $X_\rho(t)$, $L_\rho(t)$, and $W_\rho(t)$ as in Section 4.2. Let the completed-work process be defined by

$$C_\rho(t) \equiv Y_\rho(t) - L_\rho(t), \quad t \geq 0. \tag{38}$$

We now can apply Lemma 1 in Section 4 to express the waiting-time process as

$$W_\rho(t) \equiv \inf\{u \geq 0 : M_\rho(t + u) - M_\rho(t) \geq L_\rho(t)\}, \quad t \geq 0. \tag{39}$$

The (virtual) waiting time $W_\rho(t)$ represents the time that a hypothetical arrival at time t would have to wait before starting service.

As in equations (3.7) and (3.8) of Whitt (2015), we can define the queue-length process (number in system) and the departure process in model ρ jointly. We can also express the departure process in terms of the workload process instead of the queue-length process by

$$D_\rho(t) \equiv N^s\left(\int_0^t \mu_\rho(s)1_{\{L_\rho(s) > 0\}}\, ds\right), \quad t \geq 0, \tag{40}$$

but we do not focus on the departure and queue-length processes here.

6.3. The Scaled Queuing Processes

We start with the FWLLN-scaled processes. First, let the scaled deterministic rate functions be

$$\bar{\Lambda}_\rho(t) \equiv (1-\rho)^2 \Lambda_\rho((1-\rho)^{-2}t) \quad \text{and} \quad \bar{M}_\rho(t) \equiv (1-\rho)^2 M_\rho((1-\rho)^{-2}t), \quad t \geq 0, \tag{41}$$

for $\Lambda_\rho(t)$ and $M_\rho(t)$ in Equation (37). We immediately see that

$$\bar{\Lambda}_\rho \to \Lambda_f \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \uparrow 1, \tag{42}$$

where

$$\Lambda_f(t) \equiv t + S(t), \quad t \geq 0, \tag{43}$$

for $S(t)$ in Equation (35).

Let the FWLLN-scaled arrival stochastic process be defined by

$$\bar{A}_\rho(t) \equiv (1-\rho)^2 A_\rho((1-\rho)^{-2}t). \tag{44}$$

Let the input, net-input, workload, completed-work, and waiting-time components of the FWLLN-scaled the vector $(\bar{A}_\rho, \bar{Y}_\rho, \bar{X}_\rho, \bar{L}_\rho, \bar{C}_\rho, \bar{W}_\rho)$ be defined in the same way.

Then let the associated FCLT-scaled deterministic rate functions be defined by

$$\hat{\Lambda}_\rho(t) \equiv (1-\rho)[\Lambda_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)],$$
$$\hat{M}_\rho(t) \equiv (1-\rho)[M_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)] \tag{45}$$

for Λ_f in Equation (43). Let the associated FCLT-scaled stochastic processes be defined by

$$\hat{A}_\rho(t) \equiv (1-\rho)[A_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)],$$
$$\hat{Y}_\rho(t) \equiv (1-\rho)[Y_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)],$$
$$\hat{X}_\rho(t) \equiv (1-\rho)X_\rho((1-\rho)^{-2}t),$$
$$\hat{L}_\rho(t) \equiv (1-\rho)L_\rho((1-\rho)^{-2}t),$$
$$\hat{C}_\rho(t) \equiv (1-\rho)[C_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)],$$
$$\hat{W}_\rho(t) \equiv (1-\rho)W_\rho((1-\rho)^{-2}t), \quad t \geq 0. \tag{46}$$

6.4. The HT FWLLN

We start with the HT FWLLN. The limit provides a deterministic fluid approximation. However, simple fluid approximations evidently are too crude to provide much help. Corollary 3 shows that rate-matching control stabilizes both the workload and the waiting time for the fluid approximation.

Let $\mathcal{D}^k$ be the k-fold product space of $\mathcal{D}$ with itself, let $\Rightarrow$ denote convergence in distribution, and let $x \circ y$ be the composition function defined by $(x \circ y)(t) \equiv x(y(t))$. Let $a \wedge b \equiv \min\{a, b\}$, and let $\psi : D \to D$ be the standard one-dimensional reflection map as in section 13.5 of Whitt (2002); that is,

$$\psi(x)(t) \equiv x(t) - (\inf\{x(s) : 0 \leq s \leq t\} \wedge 0), \quad t \geq 0. \tag{47}$$

Theorem 2 (HT FWLLN). *Under the definitions and assumptions in Section 6 above, if $\xi_\rho \to \xi$ and $\eta_\rho \to \eta$ as $\rho \uparrow 1$, and the system starts empty at time 0, then*

$$\bar{M}_\rho \to M_f \quad \text{in} \quad \mathcal{D}, \quad \text{where} \quad M_f(t) \equiv t + \xi S(t - \eta) \tag{48}$$

and

$$(\bar{A}_\rho, \bar{Y}_\rho, \bar{X}_\rho, \bar{L}_\rho, \bar{C}_\rho, \bar{W}_\rho) \Rightarrow (\bar{A}, \bar{Y}, \bar{X}, \bar{L}, \bar{C}, \bar{W}) \quad \text{in} \quad \mathcal{D}^6 \quad \text{as} \quad \rho \uparrow 1 \tag{49}$$

for $(\bar{A}_\rho, \bar{Y}_\rho, \bar{X}_\rho, \bar{L}_\rho, \bar{C}_\rho, \bar{W}_\rho)$ *defined in Equation* (44), *where*

$$\bar{A}(t) \equiv \bar{Y}(t) \equiv \Lambda_f(t), \quad \bar{X}(t) \equiv S(t) - \xi S(t-\eta), \quad t \geq \eta,$$

$$\bar{L}(t) \equiv \sup_{0 \leq s \leq c} \{X(t) - X(t-s)\}, \quad t \geq c + \eta, \quad \bar{C} \equiv \bar{Y} - \bar{L}, \quad and$$

$$\bar{W}(t) \equiv \inf\{u \geq 0 : M_f(t+u) - M_f(t) \geq \bar{L}(t)\}, \quad t \geq 0, \tag{50}$$

for $\Lambda_f(t)$ *in Equation* (43) *with* $S(t)$ *in Equation* (35), $M_f(t)$ *in Equation* (48), *and* ψ *being the reflection map in Equation* (47).

Proof. We successively apply the continuous mapping theorem (CMT) using the functions in sections 12.7 and 13.2–13.6 of Whitt (2002). First, observe that Equation (48) is a minor modification of Equation (41). Let $\bar{N}_\rho^a$ and $\bar{S}_\rho$ denote $\bar{N}_n^a$ and $\bar{S}_n^v$, respectively, where, paralleling Equation (32), we let $\bar{N}_n^a(t) \equiv n^{-1} N^a(nt)$ and $\bar{S}_n^v \equiv n^{-1} S_{\lfloor nt \rfloor}^v$, $t \geq 0$, and then let $n = (1-\rho)^{-2}$. Then observe that $\bar{A}_\rho = \bar{N}_\rho^a \circ \bar{\Lambda}_\rho$ and $\bar{Y}_\rho = \bar{S}_\rho \circ \bar{A}_\rho$, so we can apply the CMT with the composition map. The limit for $\bar{X}_\rho$ follows from the CMT with addition, and then the limit for $\bar{L}_\rho$ follows from the CMT with the reflection map in Equation (47). To establish the limit for the scaled waiting time $\bar{W}_\rho(t)$ in $\mathscr{D}$, we apply the CMT with the inverse function. Finally, the limit for $\bar{C}_\rho$ again follows from the CMT with addition. □

We obtain stronger results in special cases.

Corollary 3 (FWLLN for the Rate-Matching Service Rate Control). *In addition to the conditions of Theorem* 2, *if* $\eta = 0$ *and* $\xi = 1$, *then* $M_f(t) = \Lambda_f(t)$, $t \geq 0$, *and then* $\bar{X}(t) = \bar{L}(t) = \bar{W}(t) = 0$, *for all* $t \geq 0$, *whereas* $\bar{C} = \bar{Y} = \bar{A} = \Lambda_f$.

Remark 4 (Stabilization Achieved by Many Fluid Models). It is evident that the conclusion of Corollary 3 holds for any single-server fluid model with arrival rate $\lambda(t)$ and service rate $\mu(t)$, provided that $\mu(t) \geq \lambda(t)$ for all t. The (η, ξ) controls are intended to address the time-varying arrival rate in the more general stochastic setting.

As a modification of Corollary 3, we can have all customers wait exactly η if we provide no service until time η.

Corollary 4 (Stabilizing the Waiting Time at Any Positive Value). *In addition to the conditions of Theorem* 2, *if* $\xi = 1$ *and* $M_f(t) = 0$, $0 \leq t < \eta$, *then* $M_f(t) = \Lambda_f(t - \eta)$, $t \geq \eta$, *for a fixed time lag* $\eta > 0$, *so that*

$$\bar{L}(t) = \bar{X}(t) \equiv \bar{X}_\eta(t) = \Lambda_f(t) - \Lambda_f(t-\eta) = \int_{t-\eta}^t \lambda_f(s)\, ds > 0 \tag{51}$$

and

$$\bar{W}(t) = \eta \quad for\ all \quad t \geq \eta. \tag{52}$$

Corollary 4 (Sinusoidal with Damped Time Lag). *In addition to the conditions of Theorem* 2, *suppose that*

$$s(t) \equiv \beta \sin(\gamma t), \quad t \geq 0, \tag{53}$$

for positive constants β *and* γ *with* $\beta < 1$, *so that* $s(t)$ *is periodic with period* $c \equiv c_\gamma = 2\pi/\gamma$. *Then*

$$S(t) = (\beta/\gamma)(1 - \cos(\gamma t)), \quad t \geq \eta, \tag{54}$$

so that

$$\begin{aligned}
\bar{L}(t) &= (\beta/\gamma)([\xi \cos(\gamma(t-\eta)) - \cos(\gamma t)] \\
&\quad + \sup_{0 \leq s \leq c} \{\cos(\gamma(t-s)) - \xi \cos(\gamma(t-\eta-s))\} \\
&= (\beta/\gamma)([\xi \cos(\gamma(t-\eta)) - \cos(\gamma t)] \\
&\quad + \sup_{0 \leq s \leq c} \{\cos(\gamma s) - \xi \cos(\gamma(s-\eta))\}), \quad t \geq c + \eta.
\end{aligned} \tag{55}$$

For the special case $\xi = 1$, $\bar{W}(t) = \eta$. *If in addition, and* $\eta < \pi/\gamma$, *the supremum in* (55) *is attained at* $s^* = (\pi/2\gamma) - \eta/2$, *so that*

$$\bar{L}(t) = (\beta/\gamma)([\cos(\gamma(t-\eta)) - \cos(\gamma t)] + [\cos((\pi/2) - (\gamma\eta/2)) - \cos((\pi/2) + (\gamma\eta/2))]) \tag{56}$$

for $t \geq c + \eta$. *As* $\eta \downarrow 0$,

$$\bar{L}(t)/\eta \to 1 + \beta \sin(\gamma t) = 1 + s(t). \tag{57}$$

Remark 5 (The Impact of High or Low Frequency). Corollary 5 shows the impact of high or low frequency. First, it is well known that high frequency has negligible impact, because performance tends to be determined by the behavior of the cumulative arrival rate function $\Lambda(t)$ in (15) rather than the rate function $\lambda(t)$. From (54) and (55), we see that $S(t) \to 0$ and $\bar{L}(t) \to 0$ as $\gamma \to \infty$. On the other hand, for any fixed t, $s(t) \to 0$ as $\gamma \to 0$.

6.5. The HT FCLT

We now state our main HT result: the HT FCLT with periodicity in fluid scale, as in (13). We present the proof in Section 7 after discussing consequences here.

Theorem 3 (HT FCLT). *In addition to the definitions and assumptions in Section 6 above, including the scaled arrival-rate function in (13), assume that the periodic function $s(t)$ in (3) is continuous and*

$$(1 - \rho)\eta_\rho \to \hat{\eta} \quad and \quad \frac{\xi_\rho - 1}{1 - \rho} \to \hat{\xi} \quad as \quad \rho \uparrow 1, \tag{58}$$

where $0 \leq \hat{\eta} < \infty$ and $0 \leq \hat{\xi} < \infty$. Then there is a limit for the scaled cumulative service-rate functions $\hat{M}_\rho$ in (37) and (45); that is,

$$\hat{M}_\rho(t) \equiv (1 - \rho)[M_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}(t + S(t))]$$
$$\to \hat{M}(t) \equiv -s(t)\hat{\eta} + S(t)\hat{\xi} \quad in \quad \mathcal{D} \quad as \quad \rho \uparrow 1 \tag{59}$$

for $s(t)$ in (3) and $S(t)$ in (35). If, in addition, the system starts empty at time 0, then

$$(\hat{A}_\rho, \hat{Y}_\rho, \hat{X}_\rho, \hat{L}_\rho, \hat{W}_\rho, \hat{C}_\rho) \Rightarrow (\hat{A}, \hat{Y}, \hat{X}, \hat{L}, \hat{W}, \hat{C}) \quad in \quad \mathcal{D}^5 \quad as \quad \rho \uparrow 1 \tag{60}$$

for $(\hat{A}_\rho, \hat{Y}_\rho, \hat{X}_\rho, \hat{L}_\rho, \hat{W}_\rho, \hat{C}_\rho)$ defined in (46), where

$$\hat{A}(t) \equiv (c_a B_a - e) \circ \Lambda_f(t), \quad \hat{Y}(t) \equiv (c_x B - e) \circ \Lambda_f(t), \quad \hat{C}(t) \equiv \hat{Y}(t) - \hat{L}(t),$$
$$\hat{X}(t) \equiv \hat{Y}(t) - \hat{M}(t) = \hat{Y}(t) + s(t)\hat{\eta} - S(t)\hat{\xi}$$
$$= (c_x B \circ \Lambda_f)(t) - \Lambda_f(t) + s(t)\hat{\eta} - S(t)\hat{\xi},$$
$$\hat{L}(t) \equiv \psi(\hat{X})(t) \quad and \quad \hat{W}(t) \equiv \hat{L}(t)/\mu_f(t), \quad t \geq 0. \tag{61}$$

with $c_x \equiv \sqrt{c_a^2 + c_s^2}$, B a BM, ψ the reflection map in (47) and $\mu_f(t) \equiv \lambda_f(t) \equiv 1 + s(t)$, $t \geq 0$, the limiting arrival-rate function, the derivative of Λ_f in (43).

We now draw attention to some important consequences. First, Theorem 3 establishes an HT time-varying (TV) Little's law (LL), paralleling the many-server heavy-traffic (MSHT) TV LL in Sun and Whitt (2018) and exposed for the rate-matching control in theorem 5.2 of Whitt (2015). This is a time-varying version of the familiar state-space collapse, which goes back to the early HT papers (e.g., Whitt 1971). We remark that the relation is different from the time-varying LL discussed in Bertsimas and Mourtzinou (1997), Fralix and Riano (2010), Sigman and Whitt (2019), and Whitt and Zhang (2019).

Corollary 6 (HT Time-Varying Little's Law). *Under the conditions of Theorem 3, the limit processes are related by*

$$\hat{L}(t) = \lambda_f(t)\hat{W}(t), \quad t \geq 0, \quad w.p.1. \tag{62}$$

We now consider an alternative deterministic limit to the HT FWLLN in Theorem 2. Now we assume that the FCLT holds with the variability parameter set equal to 0. For this purpose, we assume that $s(t)$ is differentiable and let $\dot{s}(t)$ be its derivative.

Corollary 7 (The Case of No Variability). *If $c_x = 0$ and $s(t)$ is differentiable in addition to the conditions of Theorem 3, then*

$$\hat{X}(t) = -t + s(t)\hat{\eta} - S(t)(\hat{\xi} + 1), \quad t \geq 0, \tag{63}$$

so that $\hat{L}(t) = \hat{W}(t) = 0$ *for all* $t \geq 0$ *if and only if*

$$\frac{d\hat{X}(t)}{dt} = -1 + \hat{\eta}\dot{s}(t) - (\hat{\xi} + 1)s(t) \leq 0, \quad t \geq 0. \tag{64}$$

In the sinusoidal case with $s(t) \equiv \beta \sin \gamma t$ *in* (5),

$$\frac{d\hat{X}(t)}{dt} = -1 + \hat{\eta}\beta\gamma \cos \gamma t - (\hat{\xi} + 1)\beta \sin \gamma t, \quad t \geq 0. \tag{65}$$

For $\beta = 1$ *and* $\gamma \to 0$,

$$\frac{d\hat{X}(t)}{dt} \to -1 - (\hat{\xi} + 1)\beta \sin \gamma t, \quad t \geq 0, \tag{66}$$

which is strictly positive over subintervals if $\hat{\xi} > 0$.

For the nondegenerate sinusoidal arrival rate function, the derivative in (64) of Corollary 7 implies it is not always possible to stabilize the limiting time-varying diffusion process $\hat{W}$ with $\hat{\xi} > 0$ in Theorem 3. We conjecture that it is never possible to stabilize it perfectly.

We now establish conditions for the optimality of an $(\hat{\eta}^*, \hat{\xi}^*)$ control for the limiting diffusion control problem for either formulation (9) or (10). Our proof will exploit uniform integrability (UI); see p. 31 of Billingsley (1999).

Corollary 8 (Optimality for the Limiting Diffusion Process). *Consider the special case of the* $GI_t/GI_t/1$ *model with* $E[U_k^{2+\epsilon}] < \infty$ *and* $E[V_k^{2+\epsilon}] < \infty$ *for some* $\epsilon > 0$. *If* $(\eta_\rho^*, \xi_\rho^*) \to (\hat{\eta}^*, \hat{\xi}^*)$ *as* $\rho \to 1$, *where* $(\eta_\rho^*, \eta_\rho^*)$ *is the optimal control for problem* (9) *or* (10), *then the limiting control* $(\hat{\eta}^*, \hat{\xi}^*)$ *is optimal for the corresponding diffusion control problem.*

Proof. We let $(\tilde{\eta}, \tilde{\xi})$ be any alternative control for the limiting diffusion process. Then let $(\tilde{\eta}_\rho, \tilde{\xi}_\rho)$ be an associated control for model ρ, $0 < \rho < 1$, where $\tilde{\eta}_\rho \equiv \tilde{\eta}/(1-\rho)$ and $\tilde{\eta}_\rho \equiv 1 + (1-\rho)\tilde{\xi}$. Then, by this construction, condition (58) holds for the family $(\tilde{\eta}_\rho, \tilde{\xi}_\rho)$. We next want to show that the convergence in distribution can be extended to convergence of the means for all t, which requires uniform integrability uniformly in t; see p. 31 of Billingsley (1999). We use the bounds on the second moments to show that it holds.

Toward that end, we exploit the upper bounds for the workload process in the $G_t/G_t/1$ model in terms of the associated workload process in the stationary $G/G/1$ model from section 3 of Ma and Whitt (2018). These bounds extend directly to the $G_t/G_t/1$ model by virtue of Corollary 1. These bounds show that the mean workload is bounded above uniformly in y over the interval $[0, c]$. These bounds also apply to the waiting time process because $W(t) \leq L(t)/\mu_L$, where $\mu_L > 0$ is a lower bound on the service rate, which follows from (4) and (6). For the stationary $GI/GI/1$ model, finite second moments imply the existence of the first moments of the waiting time and uniform integrability needed for convergence; see p. 31 of Billingsley (1999) and sections X.2 and X.7 of Asmussen (2003).

Finally, we observe that our optimal policy $(\eta_\rho^*, \xi_\rho^*)$ has expected value greater than or equal to the alternative policy $(\tilde{\eta}_\rho, \tilde{\xi}_\rho)$ for all ρ, whereas both converge as $\rho \to 1$. Hence, the limit of the optimal policies, $(\hat{\eta}^*, \hat{\xi}^*)$ must be at least as good as $(\tilde{\eta}, \tilde{\xi})$. □

We apply Corollary 8 to support our numerical calculations by observing that $(\eta_\rho^*, \xi_\rho^*)$ when scaled as in (58) converges to a limit. We thus deduce that the limit must be the optimal policy for the diffusion. However, this numerical evidence is not a mathematical proof. Moreover, although the numerical evidence is good, it is not exceptionally good, especially for ξ_ρ^*, as can be seen from Table 1 in Section 9.1.

7. Proof of Theorem 3
To establish (59), apply (37) and (45) to obtain

$$\hat{M}_\rho(t) \equiv (1-\rho)[M_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}(t + S(t))]$$

$$= (1-\rho)^{-1}[\xi_\rho S(t - (1-\rho)^2\eta_\rho) - S(t)]$$

$$= (1-\rho)^{-1}[\xi_\rho S(t - (1-\rho)^2\eta_\rho) - \xi_\rho S(t)] + (1-\rho)^{-1}[\xi_\rho S(t) - S(t)]$$

$$\to -\hat{\eta}\dot{s}(t) + \hat{\xi}S(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \uparrow 1, \tag{67}$$

where on the third line we have subtracted and added the term $\xi_\rho S(t)$ and on the last line we have differentiated using

$$(1-\rho)^2 \eta_\rho / (1-\rho) = (1-\rho)\eta_\rho \to \hat{\eta} \quad \text{as} \quad \rho \uparrow 1$$

by assumption (58). We used the assumed continuity of s to have S be continuously differentiable, so the derivative of $S(t)$ holds uniformly in t over bounded intervals.

We next establish Equation (60). First, the limit for $\hat{A}_\rho$ is given in lemma 5.1 of Whitt (2015), but we need to make an adjustment because the arrival rate in model ρ is chosen to be ρ here as opposed to 1 before. From Equations (43), (42), and (45), we see that

$$\bar{\Lambda}_\rho(t) = \rho \Lambda_f(t) \to \Lambda_f(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \to 1,$$
$$\hat{\Lambda}_\rho(t) = (1-\rho)^{-1}\rho\Lambda_f(t) - (1-\rho)^{-1}\Lambda_f(t) = -\Lambda_f(t) \tag{68}$$

for all ρ, where $\Lambda_f(t)$ is defined in Equation (43). Then the limit for $\hat{A}_\rho$ follows from the standard argument for random sums. The key is to observe that

$$\hat{A}_\rho = \hat{N}_\rho \circ \bar{\Lambda}_\rho + \hat{\Lambda}_\rho, \tag{69}$$

where $\hat{N}_\rho$ is defined to be $\hat{N}_n$ in Equation (32) for $n = (1-\rho)^{-2}$. So we can start with the joint convergence

$$\left(\hat{N}_\rho, \bar{\Lambda}_\rho, \hat{\Lambda}_\rho\right) \Rightarrow (c_a B_a, \Lambda_f, -\Lambda_f) \quad \text{in} \quad \mathcal{D}^3 \quad \text{as} \quad \rho \to 1. \tag{70}$$

We then apply convergence preservation with the map $g(x,y,z) = x \circ y + z$ (composition plus addition) as in section 13.3 of Whitt (2002) to get $\hat{A}_\rho \Rightarrow c_a B_a \circ \Lambda_f - \Lambda_f = (c_a B_a - e) \circ \Lambda_f$ in $\mathcal{D}$.

Similarly, given that $\bar{N}_\rho \equiv (1-\rho)^2 N((1-\rho)^{-2}t) \Rightarrow e$ and $\bar{A}_\rho \equiv (1-\rho)^2 A((1-\rho)^{-2}t)$,

$$\bar{A}_\rho = (1-\rho)^2 N(\Lambda_\rho((1-\rho)^{-2}t)) = (1-\rho)^2 N((1-\rho)^{-2}\rho\Lambda_f(t))$$
$$= \bar{N}_\rho(\rho\Lambda_f(t)) \Rightarrow \Lambda_f \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \to 1. \tag{71}$$

A variant of the random-sum argument holds for $\hat{Y}_\rho$ too. In particular, we start with the joint convergence

$$\left(\hat{S}_\rho, \bar{A}_\rho, \hat{A}_\rho\right) \Rightarrow (c_s B_s, \Lambda_f, c_a B_a \circ \Lambda_f - \Lambda_f) \quad \text{in} \quad \mathcal{D}^3 \quad \text{as} \quad \rho \to 1. \tag{72}$$

The joint convergence holds by virtue of theorems 11.4.4 and 11.4.5 of Whitt (2002). We then apply convergence preservation with the map $g(x,y,z) = x \circ y + z$ (composition plus addition) as in section 13.3 of Whitt (2002) to get

$$\hat{Y}_\rho = \hat{S}_\rho \circ \bar{A}_\rho + \hat{A}_\rho \Rightarrow c_s B_s \circ \Lambda_f + c_a B_a \circ \Lambda_f - \Lambda_f$$
$$\stackrel{\mathrm{d}}{=} c_x B \circ \Lambda_f - \Lambda_f \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \to 1. \tag{73}$$

Then the limits for $\hat{X}_\rho$ and $\hat{L}_\rho$ follow from the continuous mapping theorem with standard reflection map reasoning, for example, as in chapter 9 of Whitt (2002), even though the service-rate function is now more general.

However, the waiting time requires a new treatment. The limit follows from the definition of the scaled service-rate control in Equation (37) and the first-passage-time representation of the waiting time in Equation (39). The structure and result are similar to the Puhalskii (1994) theorem and related results in section 13.7 of Whitt (2002), but they evidently do not apply directly.

We apply Taylor's theorem to a perturbation of S in Equation (35). The essential idea is that

$$(1-\rho)^{-1}[S(t + (1-\rho)u) - S(t)] \to s(t)u \quad \text{as} \quad \rho \to 1 \tag{74}$$

uniformly in t and u over bounded intervals. Just as in Equation (67), we use the assumed continuity of s to have S be continuously differentiable, so the derivative of $S(t)$ holds uniformly in t and u over bounded intervals.

For the specific application, let

$$\tilde{S}_\rho(t, u) \equiv (1-\rho)^{-1}\xi_\rho[S(t + (1-\rho)u - (1-\rho)^2\eta_\rho) - S(t - (1-\rho)^2\eta_\rho)] \tag{75}$$

and

$$\zeta_\rho(t,u) \equiv \tilde{S}_\rho(t,u) - s(t)u. \tag{76}$$

By combining Equation (74) and the two limits in condition (58), we see that $\zeta_\rho(t,u)$ is asymptotically negligible as $\rho \to 1$ uniformly in t and u over bounded intervals. We will use this at the critical final step in the following representation.

To start, let $\tilde{M}_\rho(t,u) \equiv M_\rho((1-\rho)^{-2}t + u)$. Then, from Equations (46) and (39),

$$
\begin{aligned}
\hat{W}_\rho(t) &\equiv (1-\rho)W_\rho((1-\rho)^{-2}t) \\
&= (1-\rho)\inf\{u \geq 0 : \tilde{M}_\rho(t,u) - \tilde{M}_\rho(t,0)) \geq L_\rho((1-\rho)^{-2}t)\} \\
&= \inf\{u \geq 0 : \tilde{M}_\rho(t,(1-\rho)^{-1}u) - \tilde{M}_\rho(t,0) \geq L_\rho((1-\rho)^{-2}t)\} \\
&= \inf\{u \geq 0 : (1-\rho)[\tilde{M}_\rho(t,(1-\rho)^{-1}u) - \tilde{M}_\rho(t,0)] \geq \hat{L}_\rho(t)\} \\
&= \inf\{u \geq 0 : u + \tilde{S}_\rho(t,u) \geq \hat{L}_\rho(t)\} \\
&= \inf\{u \geq 0 : u + s(t)u + \zeta_\rho(t,u) \geq \hat{L}_\rho(t)\} \\
&= \inf\{u \geq 0 : u\lambda_f(t) + \zeta_\rho(t,u) \geq \hat{L}_\rho(t)\}, \quad t \geq 0,
\end{aligned} \tag{77}
$$

where $\lambda_f(t) = t + s(t)$ by Equation (43), and we apply Taylor's theorem with Equations (75) and (76) in line 6 to obtain that $\zeta_\rho(t,u)$ is asymptotically negligible as $\rho \to 1$ uniformly over both t and u over bounded subintervals.

For the final step, to simplify, we make the entire argument deterministic by using the Skorohod representation theorem, as in theorem 3.2.2 of Whitt (2002), to replace the stochastic convergence $\hat{L}_\rho \Rightarrow \hat{L}$ in $\mathscr{D}$ by associated convergence w.p.1. Then we see from line 6 of Equation (77) that in the infimum it suffices to consider u only just beyond $\hat{L}(t)/\lambda_f(t)$, which for t in a bounded interval is bounded for each sample path because $\lambda_f(t)$ has been assumed to be bounded below, whereas $\hat{L}(t)$ is bounded above, for t in a bounded interval. Thus, we can write

$$\frac{\hat{L}_\rho(t) - K\zeta_\rho^\uparrow(t)}{\lambda_f(t)} \leq \hat{W}_\rho(t) \leq \frac{\hat{L}_\rho(t) + K\zeta_\rho^\uparrow(t)}{\lambda_f(t)} \tag{78}$$

for t and u over specified bounded intervals, K an appropriate positive constant, and

$$\zeta_\rho^\uparrow(t) \equiv \sup_{0 \leq u \leq \bar{u}} |\zeta_\rho(t,u)|$$

for an appropriate $\bar{u}$. Given that $\hat{L}_\rho \Rightarrow \hat{L}$ and $\zeta_\rho^\uparrow \to 0$ in $\mathscr{D}$, we can use the standard sandwiching argument (uniformly over bounded time intervals) to obtain convergence $\hat{W}_\rho(t) \Rightarrow \hat{L}(t)/\lambda_f(t) \equiv \hat{W}(t)$ in $\mathscr{D}$, which completes the proof. $\square$

8. An HT FCLT with Periodicity in the Weaker Diffusion Scaling
In this section, we establish an HT FCLT with periodicity holding in the weaker diffusion scale instead of in the fluid scale, as was done in Section 6. The scaling here follows Whitt (2014) and Ma and Whitt (2018) instead of Whitt (2015). In this scaling, the HT limit of the waiting time coincides with the HT limit for the workload process and so does not capture the differences we see in the simulations in preceding sections.

8.1. An Alternative Family of Models
We start with the same basic rate-1 processes in Section 6.1. We then create a model for each ρ, $0 < \rho < 1$, now using Equation (14) instead of Equation (13). This yields the family of cumulative arrival-rate functions

$$\Lambda_\rho(t) \equiv \rho(t + (1-\rho)^{-1}S((1-\rho)^2 t)), \quad t \geq 0, \tag{79}$$

for S in Equation (35). Differentiating in Equation (79) yields the arrival-rate function in Equation (14). Just as before, the time scaling in Equations (14) and (79) implies that the period in model ρ with arrival-rate function $\lambda_\rho(t)$ in Equation (14) is $c_\rho = c(1-\rho)^{-2}$, where c is the period of s in Equations (2)–(4). Thus, the period c_ρ in model ρ is growing with ρ.

8.2. An Associated Family of Service-Rate Controls

Just as in Section 6.2.1, we define associated service-rate controls. Closely paralleling Equations (14) and (79), we define associated scaled time-varying service-rate functions using the control parameters η_ρ and ξ_ρ; that is, for all $t \geq 0$,

$$\mu_\rho(t) \equiv 1 + (1 - \rho)\xi_\rho s(t - \eta_\rho) \quad \text{and}$$

$$M_\rho(t) \equiv \int_0^t \mu_\rho(s)\,ds = t + (1 - \rho)^{-1}\xi_\rho S((1 - \rho)^2(t - \eta_\rho)). \tag{80}$$

Just as in Equation (79), differentiation of $M_\rho(t)$ in Equation (80) shows that it is consistent with $\mu_\rho(t)$. As a consequence of Equation (80), the average service rate is $\bar{\mu}_\rho = 1$, $0 < \rho < 1$.

8.3. The Scaled Queuing Processes

We use the same processes introduced in Section 4 but new scaling. Let the scaled arrival-rate and service-rate functions be defined for $t \geq 0$ by

$$\hat{\Lambda}_\rho(t) \equiv (1 - \rho)[\Lambda_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}t]$$
$$= \rho S(t) - t,$$
$$\hat{M}_\rho(t) \equiv (1 - \rho)[M_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}t]$$
$$= \xi_\rho S(t - (1 - \rho)^2\eta_\rho). \tag{81}$$

Clearly, $\hat{\Lambda}_\rho(t) \to S(t) - t$ as $\rho \to 1$ uniformly over bounded intervals of t. The key is what happens to $\hat{M}_\rho(t)$. From Equation (81), we get the following.

Lemma 2 (HT Limit of $\hat{M}_\rho(t)$). *If $\xi_\rho \to 1$ and $(1 - \rho)^2\eta_\rho \to 0$, then $\hat{M}_\rho(t) \to S(t)$ uniformly over bounded intervals of t.*

Then let associated scaled stochastic processes be defined by

$$\hat{A}_\rho(t) \equiv (1 - \rho)[A_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}t],$$
$$\hat{Y}_\rho(t) \equiv (1 - \rho)[Y_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}t],$$
$$\hat{X}_\rho(t) \equiv (1 - \rho)X_\rho((1 - \rho)^{-2}t),$$
$$\hat{L}_\rho(t) \equiv (1 - \rho)L_\rho((1 - \rho)^{-2}t), \quad \hat{C}_\rho \equiv \hat{Y}_\rho - \hat{L}_\rho,$$
$$\hat{W}_\rho(t) \equiv (1 - \rho)W_\rho((1 - \rho)^{-2}t), \quad t \geq 0. \tag{82}$$

Note that the translation terms in $\hat{\Lambda}_\rho$ and $\hat{M}_\rho$ in Equation (81) are different from the translation terms in Equation (41), whereas the translation terms in $\hat{A}_\rho$ and $\hat{Y}_\rho$ in Equation (82) are different from the translation terms in Equation (46). Thus, the statement of the HT limit that follows is different (and weaker).

8.4. The HT FCLT with Periodicity in Diffusion Scale

Just as in Section 6, the following HT FCLT states that $\hat{A}_\rho$ and $\hat{Y}_\rho$ converge to periodic Brownian motions (PBMs). However, unlike Section 6, $\hat{X}_\rho$ converges to an *ordinary* BM, $\hat{L}_\rho$ and $\hat{W}_\rho$ converge to the *same ordinary* reflected Brownian motion (RBM), whereas $\hat{C}_\rho$ has a complicated limit. We thus show that $\hat{L}_\rho$ and $\hat{W}_\rho$ are asymptotically stable and Markov. Note that the scaling conditions on (η_ρ, ξ_ρ) here are implied by condition (58) in Theorem 3, but as noted earlier, the conclusion is different and weaker because of the different translation terms.

Theorem 4 (Heavy-Traffic Limit Extending Theorem 3.2 of Whitt 2014 and Theorem 2 of Ma and Whitt 2018). *If, in addition to the definitions and assumptions in Equations (79)–(82), $(1 - \rho)^2\eta_\rho \to 0$ and $\xi_\rho \to 1$ as $\rho \to 1$, and the system starts empty at time 0, then*

$$(\hat{\Lambda}_\rho, \hat{M}_\rho, \hat{A}_\rho, \hat{Y}_\rho, \hat{X}_\rho, \hat{L}_\rho, \hat{W}_\rho, \hat{C}_\rho) \Rightarrow (\hat{\Lambda}, \hat{M}, \hat{A}, \hat{Y}, \hat{X}, \hat{L}, \hat{W}, \hat{C}) \tag{83}$$

in $\mathcal{D}^8$ as $\rho \to 1$ for $(\hat{\Lambda}_\rho, \hat{M}_\rho)$ defined in Equation (81) and $(\hat{A}_\rho, \hat{Y}_\rho, \hat{X}_\rho, \hat{L}_\rho, \hat{W}_\rho, \hat{C}_\rho)$ defined in Equation (82), where

$$\hat{\Lambda} \equiv S - e, \quad \hat{A} \equiv c_a B_a + S - e, \quad \hat{M} \equiv S,$$
$$\hat{Y} \equiv \hat{A} + c_s B_s, \quad \hat{X} \equiv \hat{Y} - S \stackrel{d}{=} c_x B - e,$$
$$\hat{L} \equiv \psi(\hat{X}), \quad \hat{W} \equiv \psi(\hat{X}) \quad \text{and} \quad \hat{C} \equiv \hat{Y} - \hat{L}, \tag{84}$$

with B_a and B_s being independent BMs, S in Equation (35), c_a and c_s being the variability parameters in Equation (33), $c_x \equiv \sqrt{c_a^2 + c_s^2}$, and B is a BM.

Proof. We will be brief because most of the argument is essentially the same as in Whitt (2014) and Ma and Whitt (2018). First, the limit for $\hat{A}_\rho$ is given in theorem 3.2 of Whitt (2014). Then the limit for $\hat{Y}_\rho$ follows from theorem 9.3.4 of Whitt (2002), as noted in the proof of theorem 2 in Ma and Whitt (2018). (See $C(t)$ in equation 9.2.4 and C_n in equation 9.3.4 and theorem 9.3.4 of Whitt 2002.) Then the limits for $\hat{X}_\rho$ and $\hat{L}_\rho$ follow from the standard reflection mapping argument even though the service-rate function is now more general. Again, the waiting time requires a new treatment. The limit follows from the first-passage-time representation in Equation (39). In particular, paralleling Equation (77), letting $\tilde{M}_\rho(t, u) \equiv M_\rho((1 - \rho)^{-2}t + u)$, we have

$$
\begin{aligned}
\hat{W}_\rho(t) &\equiv (1 - \rho)W_\rho((1 - \rho)^{-2}t) \\
&= (1 - \rho)\inf\{u \geq 0 : \tilde{M}_\rho(t, u) - \tilde{M}_\rho(t, 0) \geq L_\rho((1 - \rho)^{-2}t)\} \\
&= \inf\{u \geq 0 : \tilde{M}_\rho(t, (1 - \rho)^{-1}u) - \tilde{M}_\rho(t, 0) \geq L_\rho((1 - \rho)^{-2}t)\} \\
&= \inf\{u \geq 0 : (1 - \rho)[\tilde{M}_\rho(t, (1 - \rho)^{-1}u) - \tilde{M}_\rho(t, 0)] \geq \hat{L}_\rho(t)\} \\
&= \inf\{u \geq 0 : u + \zeta_\rho(t, u) \geq \hat{L}_\rho(t)\},
\end{aligned}
\tag{85}
$$

for $t \geq 0$, where

$$
\zeta(t, u) \equiv \xi_\rho\big[S(t + (1 - \rho)u - (1 - \rho)^2\eta_\rho) - S(t - (1 - \rho)^2\eta_\rho)\big],
\tag{86}
$$

which is asympotically negligible as $\rho \to 1$ uniformly in compact intervals, given the conditions on η_ρ and ξ_ρ. As technical support for the last step, note that

$$
S(t + \epsilon) - S(t) \leq s_U\epsilon \quad \text{for all} \quad \epsilon > 0,
\tag{87}
$$

for s_U in Equation (4). Also add and subtract $\xi_\rho S(t)$ and treat the two terms separately; that is,

$$
\begin{aligned}
\xi_\rho S(t + (1 - \rho)u - (1 - \rho)^2\eta_\rho) - S(t) &= \xi_\rho S(t) - S(t) \\
&+ \xi_\rho S(t + (1 - \rho)u - (1 - \rho)^2\eta_\rho) - \xi_\rho S(t).
\end{aligned}
$$

Hence, we can apply the continuous mapping theorem for the inverse in section 13.6 of Whitt (2002) to get $\hat{W}_\rho \Rightarrow \hat{L}$ in $\mathcal{D}$ as $\rho \to 1$, jointly with the other limits. □

9. Simulation Examples
9.1. Simulation Examples in the Setting of Section 6
In this section, we report results of simulation experiments to evaluate the new optimal $(\eta_\rho^*, \xi_\rho^*)$ controls as a function of ρ for models scaled according to Theorem 3, specifically by Equations (13), (34), and (37), so that we can see the systematic behavior.

9.1.1. Sinusoidal Examples. We start with sinusoidal examples and then consider nonsinusoidal examples. Table 1 shows results for four values of the traffic intensity ρ with $\rho \uparrow 1$ for the sinusoidal model in Equations (2)–(6) with HT scaling in Equation (13) with parameters $(\rho, \beta_\rho, \gamma_\rho) = (\rho, 0.2, 2.5(1 - \rho)^2)$. For this case, we found that

Table 1. The (Identical) Solutions to the Minimax and Minimum-Deviation Optimization Problems in (9) and (10) for the Sinusoidal Model in (2)–(6) with HT Scaling in (13) with Parameters $(\rho, \beta_\rho, \gamma_\rho) = (\rho, 0.2, 2.5(1 - \rho)^2)$.

ρ	0.8	0.9	0.95	0.975
$\beta_\rho \equiv \beta$	0.2	0.2	0.2	0.2
γ_ρ	0.1	0.025	0.00625	0.0015625
η_ρ^*	5.80	12.94	27.7	56.6
$\hat{\eta}_\rho^* \equiv (1 - \rho)\eta_\rho^*/\rho$	1.45	1.44	1.46	1.45
ξ_ρ^*	0.842	0.889	0.931	0.960
$\hat{\xi}_\rho^* \equiv (1 - \xi_\rho^*)/(1 - \rho)$	0.79	1.11	1.38	1.60
$\max E[W_y]$	4.03	9.10	19.29	39.61
$(1 - \rho)\max E[W_y]/\rho$	1.008	1.011	1.015	1.016
$min(max - min)$	0.032	0.091	0.143	0.364
Average wait	4.02	9.07	19.21	39.47
$(1 - \rho)\text{ave}\{E[W_y]\}/\rho$	1.005	1.007	1.011	1.012
Relative error (%)	0.8	1.0	0.7	0.9

Notes. The four cases are determined by ρ as specified in the first row. The mean waiting times are reported with and without space scaling in the bottom rows.
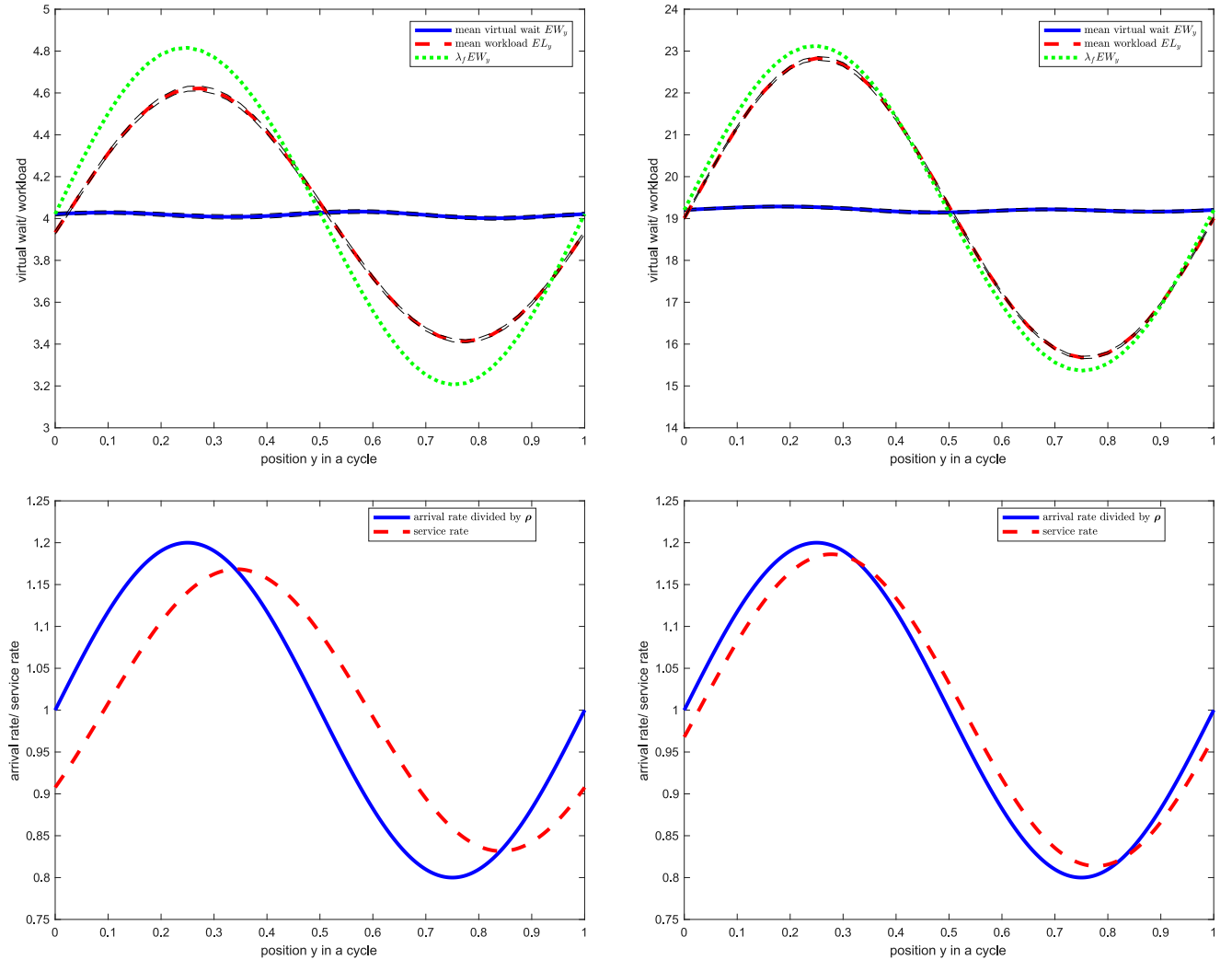
the solutions to optimization problems (9) and (10) are identical, to within our statistical precision. Hence, our solutions are for both problems.

Table 1 shows the estimated optimal controls η_ρ^* and ξ_ρ^* in each case plus scaled versions consistent with condition (58). Table 1 shows that the relative error is roughly independent of ρ, being less than 1% in each case. Table 1 also shows that the limit $\hat{\eta}^* \approx 1.45$ is rapidly approached by $(1-\rho)\eta_\rho^*/\rho$, whereas the limit $\hat{\xi}^* \approx 1.8$ is roughly approached by $(\xi_\rho^* - 1)/(1-\rho)$, both of which are consistent with condition (58). The results support Theorem 3, but unfortunately, the rate of convergence in the control parameters is not fast. Evidently the optimal damping control ξ_ρ^* is more problematic.

For the model in Table 1, Figure 3 shows the expected periodic steady-state virtual waiting time (solid blue line), the expected steady-state workload (the dashed red line), and arrival rate multiplied by the mean waiting time (the dotted green line) for $\rho = 0.8$ (left) and $\rho = 0.95$ (right). As in Figure 1, the 95% confidence interval bands are included, but they can only be seen by zooming in.

We also considered alternative values of the relative amplitude β. Table 2 shows the solutions to the minimum-deviation optimization problem in Equation (10) for the sinusoidal model in Table 1 except that β has been increased to $\beta = 0.8$ from 0.2. Table 2 shows that the relative error is roughly independent of ρ, but the relative error has increased to approximately 10% from approximately 1% in Table 1. Unlike in Figure 3, it

Figure 3. Plots over One Cycle in Table 1: Mean Values (Top) and Arrival Rate (Bottom) for $\rho = 0.8$ (Left) and 0.95 (Right)

Notes. The first row shows the expected periodic steady-state virtual waiting time (solid blue line), the expected steady-state workload (the dashed red line), and arrival rate multiplied by the mean waiting time (the dotted green line) for $\rho = 0.8$ (left) and $\rho = 0.95$ (right) in the base case $(\beta, \gamma) = (0.2, 2.5)$. The second row shows arrival rate divided by ρ and service rate for $\rho = 0.8$ (left) and $\rho = 0.95$ in the same base case. The optimal control parameters are $(\eta_\rho^*, \xi_\rho^*) = (5.80, 0.84)$ for $\rho = 0.8$ and $(27.7, 0.93)$ for $\rho = 0.95$. The maximum minus minimum of EW_y over a cycle equals 0.0321 for $\rho = 0.8$ and 0.1425 for $\rho = 0.95$.

Table 2. Solutions to the Minimum-Deviation Optimization Problem in (10) for the Sinusoidal Model in Table 1, Except β Has Been Increased to $\beta = 0.8$ from 0.2

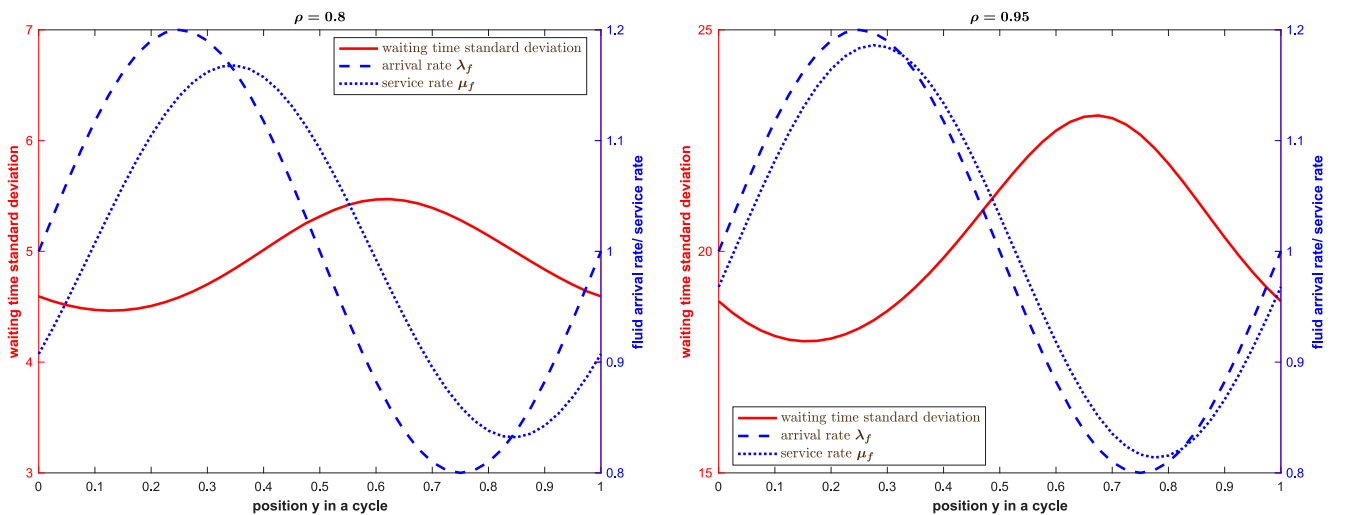| ρ	0.8	0.9	0.95	0.975
$\beta_\rho \equiv \beta$	0.8	0.8	0.8	0.8
γ_ρ	0.1	0.025	0.00625	0.0015625
η_ρ^*	6.08	15.4	33.6	70.3
$\hat{\eta}_\rho^* \equiv (1 - \rho)/\rho\eta_\rho^*$	1.52	1.71	1.77	1.80
ξ_ρ^*	0.874	0.893	0.929	0.960
$\hat{\xi}_\rho^* \equiv (1 - \xi_\rho^*)/(1 - \rho)$	0.63	1.07	1.42	1.60
$\max(E[W_y]) - \min(E[W_y])$	0.54	1.32	2.28	4.55
Average wait	4.33	10.68	23.97	51.76
$(1 - \rho)\text{ave}\{E[W_y]\}/\rho$	1.08	1.19	1.14	1.26
Relative error (%)	12.5	12.4	9.5	8.8

Notes. The four cases are determined by ρ as specified in the first row. The mean waiting times are reported with and without space scaling in the bottom rows.

is evident that the $(\eta_\rho^*, \xi_\rho^*)$ control does not stabilize the expected waiting time perfectly, either for fixed ρ or asymptotically as $\rho \rightarrow 1$.

From cases with $0.2 \leq \beta \leq 0.9$ and $0.8 \leq \rho \leq 0.975$, we conclude that $\hat{\eta}_\rho^* \equiv (1 - \rho)\eta_\rho/\rho$ and $\hat{\xi}_\rho^* \equiv (1 - \xi_\rho)/(1 - \rho)$ are nondecreasing in ρ, whereas $\hat{\eta}_\rho^*$ ($\hat{\xi}_\rho^*$) is nondecreasing (nonincreasing) in β. The relative error tends to be independent of ρ but is increasing in β. The relative error for $\beta = 0.5$ was approximately 4%, whereas the relative error for $\beta = 0.9$ was approximately 22%. The difficulty as $\beta \uparrow 1$ can be partially understood by the rate-matching control, where $E[W_y] \approx c/\lambda_f(t)$ by theorem 5.2 of Whitt (2015), where c is the stable value, which has minimum and maximum values $c/(1 + \beta)$ and $c/(1 - \beta)$ that deviate greatly as $\beta \uparrow 1$. (The constant c is the stable value of the expected queue length.) Tables 1 and 2 also show that the limiting optimal controls $(\hat{\eta}^*, \hat{\xi}^*)$, as well as the relative error, depend on β.
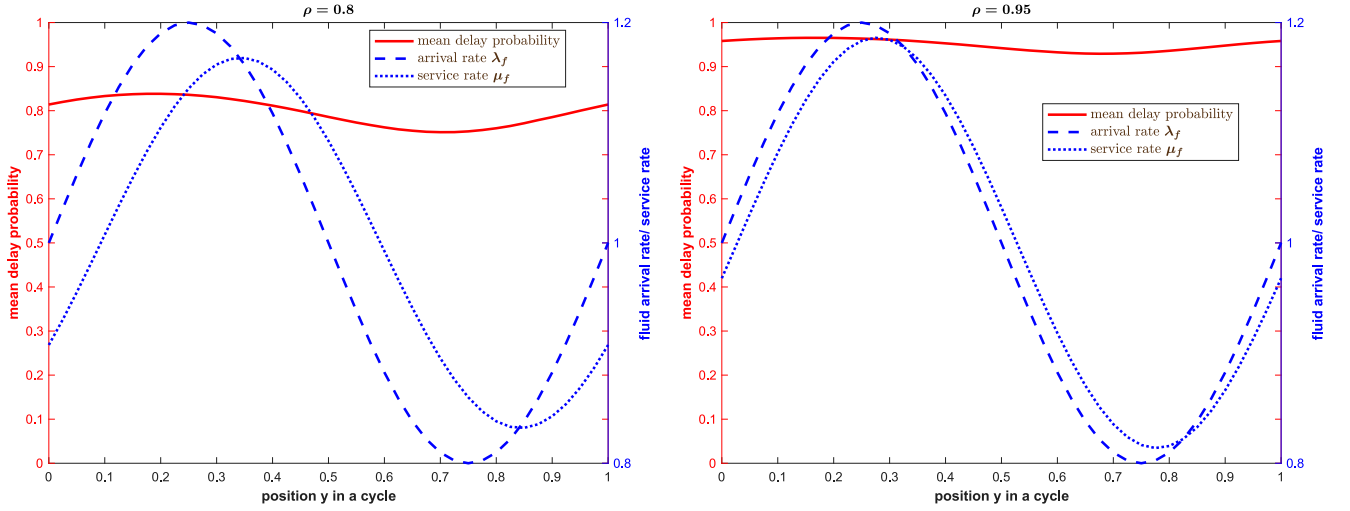
Unlike the rate-matching control in Whitt (2015), which stabilizes the entire queue-length distribution, the optimal modified (η, ξ) control neither stabilizes the mean perfectly nor does it stabilize the entire waiting-time distribution. However, it seems to do a reasonable job of both. Figures 4, 5, and 6 illustrate by showing plots of the time-varying (i) standard deviation $SD[W_y]$, (ii) delay probability $P(W_y > 0)$, and (iii) full complementary cumulative distribution function (CCDF) $\{P(W_y > x) : x \geq 0\}$ for the two cases in Figure 3, that is, for $\rho = 0.8$ (left) and $\rho = 0.95$ (right).

Very roughly, Figures 3, 4, 5, and 6 are consistent with the time-varying waiting-time distribution being exponential, as in the $M/M/1$ stationary model. We should not be surprised that the results look similar for $\rho = 0.8$ and $\rho = 0.95$ because they are scaled to be part of the family of systems satisfying the HT limit. To consider

Figure 4. Plots over One Cycle in Table 1: the Standard Deviation for $\rho = 0.8$ (Left) and 0.95 (Right)

Notes. Estimates of the periodic steady-state standard deviation $SD[W_y]$ for $\rho = 0.8$ (left) and $\rho = 0.95$ (right), shown in red solid line. Also displayed are the fluid arrival rate $\lambda_f = 1 + s(t)$ (blue dashed line) and fluid service rate $\mu_f = 1 + \xi_\rho^* s(t - \eta_\rho^*)$ (blue dotted line).

Figure 5. Plots over One Cycle in Table 1: the Delay Probability for $\rho = 0.8$ (Left) and 0.95 (Right)
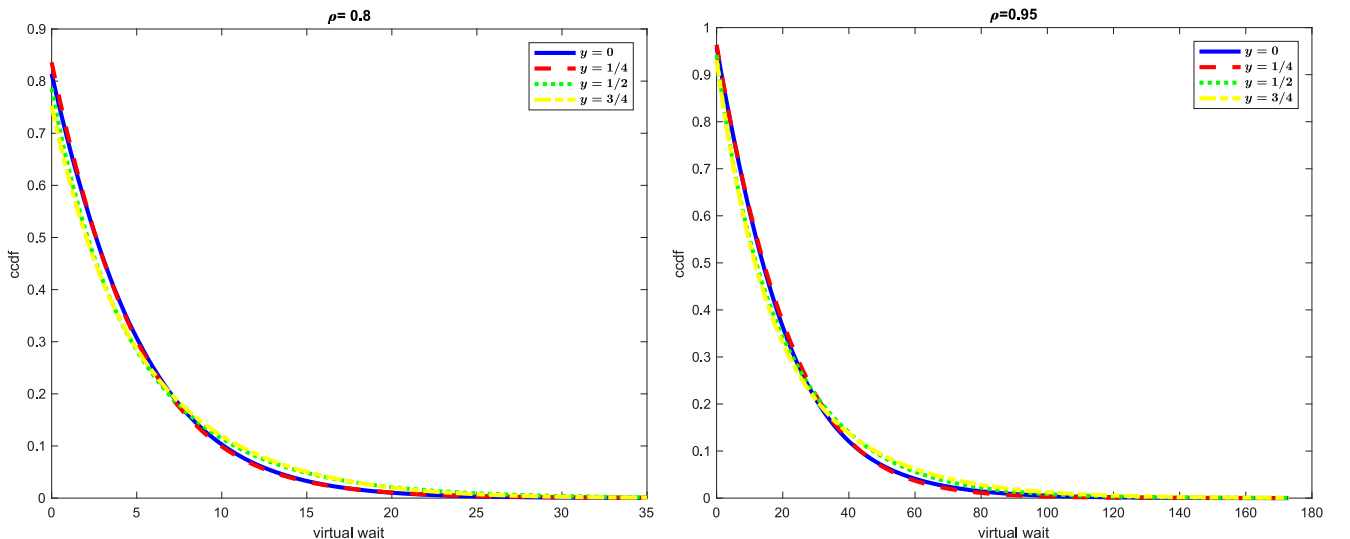

Notes. Estimates of the periodic steady-state probability of delay $P(W_y > 0)$ for $\rho = 0.8$ (left) and $\rho = 0.95$ (right), shown in red solid line. Also displayed are the fluid arrival rate $\lambda_f = 1 + s(t)$ (blue dashed line) and fluid service rate $\mu_f = 1 + \xi_\rho^* s(t - \eta_\rho^*)$ (blue dotted line).

a very different case, Figure 7 shows plots of the estimated waiting-time CCDF (left) and probability density function (PDF, right) for $\rho = 0.95$ in Figure 1, which is not scaled. Without the scaling, the cycles are relatively short. Figure 7, especially the PDF, shows much more varied behavior in this difficult short-cycle setting.

We now show two candidate modifications of the control used in Figure 1. First, Figure 8 shows the analogue of Figure 1 where we fix $\xi = 1$ and use only the single control parameter η. As stated in Remark 2 in Section 2, if we let $\xi = 1$ and optimize over η, then for $\rho = 0.8$ we get $\eta_\rho^* = 5.93$ and a maximum deviation of 0.4109, which yields approximately 10% relative error instead of less than 1%. For $\rho = 0.95$, $\eta_\rho^* = 28.3$, the maximum deviation is 3.034, and the relative error is approximately 14%.
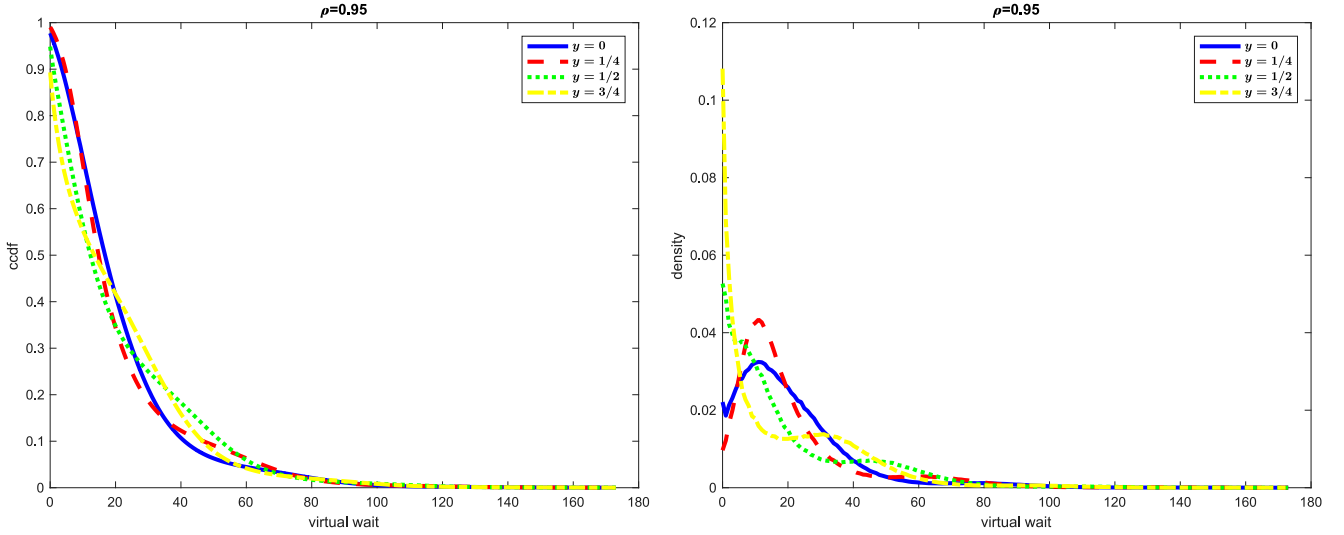
Second, Figure 9 shows the consequences of a direct HT approximation in the setting of Figure 1, obtained by letting $\hat{\eta}^* \approx 1.45$, $\eta_\rho \approx 1.45/(1 - \rho)$, $\hat{\xi}^* \approx 1.80$, and $\xi_\rho \approx 1 - 1.8(1 - \rho)$, on the basis of Table 1. For $\rho = 0.8$, $(\eta_\rho^*, \xi_\rho^*) = (7.25, 0.64)$, and the maximum deviation is 0.6005, yielding approximately 15% relative error. For $\rho = 0.95$, $(\eta_\rho^*, \xi_\rho^*) = (29.0, 0.91)$, and the maximum deviation is 0.9220, yielding approximately 5% relative error. Unlike in Figure 8, we see that the direct HT approximation improves as ρ increases, but the direct two-parameter optimal control is better.

Figure 6. Plots of the Periodic Waiting Time CCDF for 4 places in the Cycle for the Example in Table 1: $\rho = 0.8$ (Left) and 0.95 (Right)

Note. Estimates of the periodic steady-state ccdf $P(W_y > x)$ for four values of y: $0, 1/4, 1/2, 3/4$ for $\rho = 0.8$ (left) and $\rho = 0.95$ (right).

Figure 7. Plots of the Waiting Time CCDF (Left) and PDF (Right) for Four Places in the Cycle for the Example in Figure 1 with ρ 0.95 but Without Heavy-Traffic Scaling
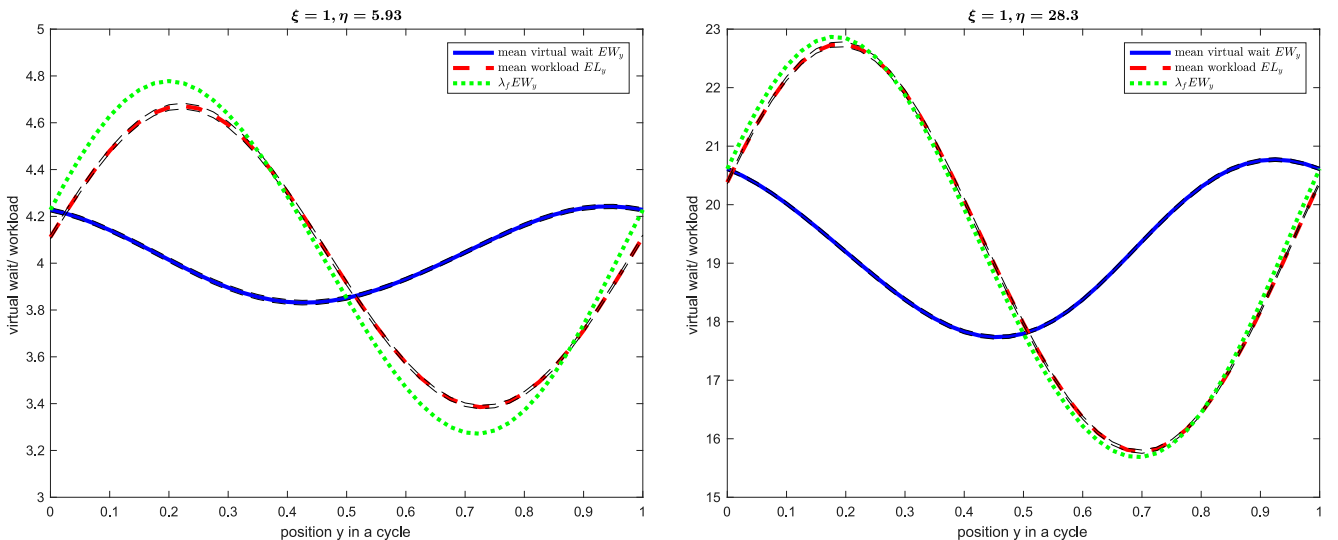

Note. Estimates of the CCDF (left) and PDF (right) of W_y for four values of y: $0, 1/4, 1/2, 3/4$.

Finally, Figure 10 plots two deterministic functions associated with the diffusion limit for the case $\beta = 0.2$, $\gamma = 2.5$, $\hat{\eta} = 1.45$, and $\hat{\xi} = -1.8$. On the left appears $\hat{M}(t) = -\hat{\eta}s(t) + \hat{\xi}S(t) = -1.5\beta\sin(\gamma t) - 1.5(\beta/\gamma)(1 - \cos(\gamma t))$ together with $s(t) = \beta\sin(\gamma t)$ and $S(t) = (\beta/\gamma)(1 - \cos(\gamma t))$. On the right appears the diffusion limit for the net input $\hat{X}(t) = -t - M(t)$ when $c_x = 0$. The plot on the right is consistent with condition (65) for no workload or waiting when $c_x = 0$ in Corollary 7.

9.1.2. Nonsinusoidal Examples. We now turn to nonsinusoidal examples. We consider the piecewise-linear single-peak periodic arrival-rate function
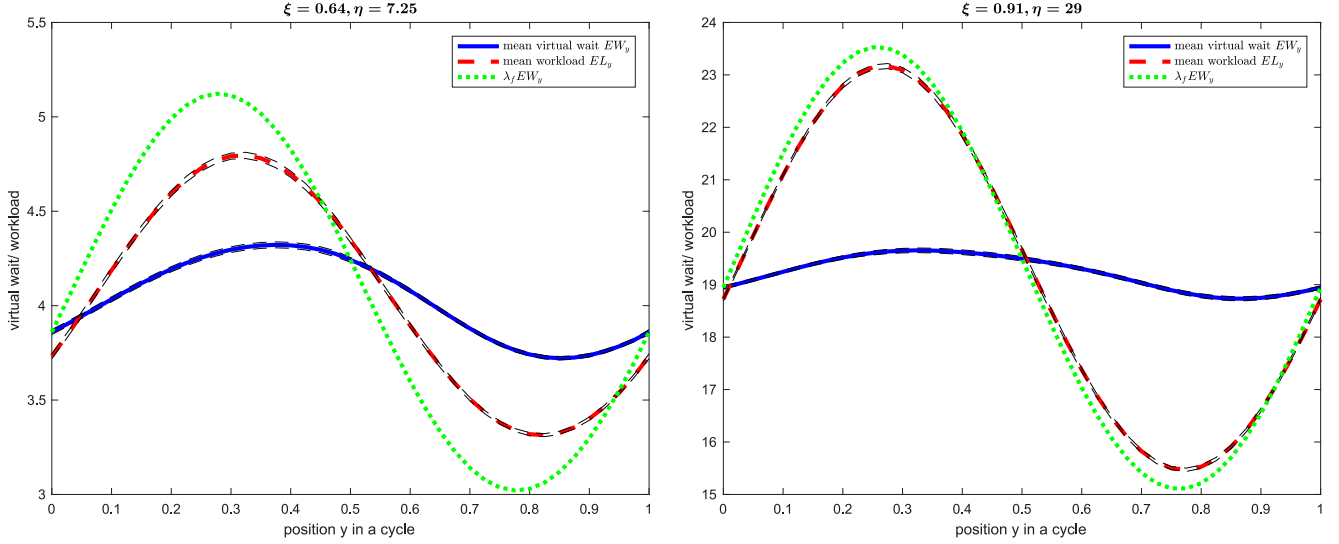
$$\lambda_f(y) = 1 - \beta + \frac{2\beta}{pc}y, \quad 0 \le y < pc \quad \text{and}$$

$$\lambda_f(y) = 1 + \left(\frac{1+p}{1-p}\right)\beta - \frac{2\beta}{c(1-p)}y, \quad pc \le y < c, \tag{88}$$

Figure 8. Estimates of the Expected Waiting Time $E[W_y]$ for the One-Parameter η Control with $\xi \equiv 1$, for the Sinusoidal Example in Figure 1 with Parameter Triple $(\rho, \beta, \gamma) = (0.8, 0.2, 0.1)$ and $\rho = 0.8$ (Left) and $\rho = 0.95$ (Right)

Note. For $\rho = 0.8$, $\eta_\rho^* = 5.93$, the maximum deviation is 0.4109 and the relative error is approximately 10%; for $\rho = 0.95$, $\eta_\rho^* = 28.3$, the maximum deviation is 3.034 and the relative error is approximately 14%.

Figure 9. Estimates of the Expected Waiting Time $E[W_y]$ (Solid Red Line) with the Heavy-Traffic Control Exploiting the Estimated Limiting Controls $\hat{\eta}^* \approx 1.45$ and $\hat{\xi}^* = 1.8$, So That $\eta_\rho^* \approx 1.45/(1-\rho)$ and $\xi_\rho^* \approx 1 - 1.8(1-\rho)$
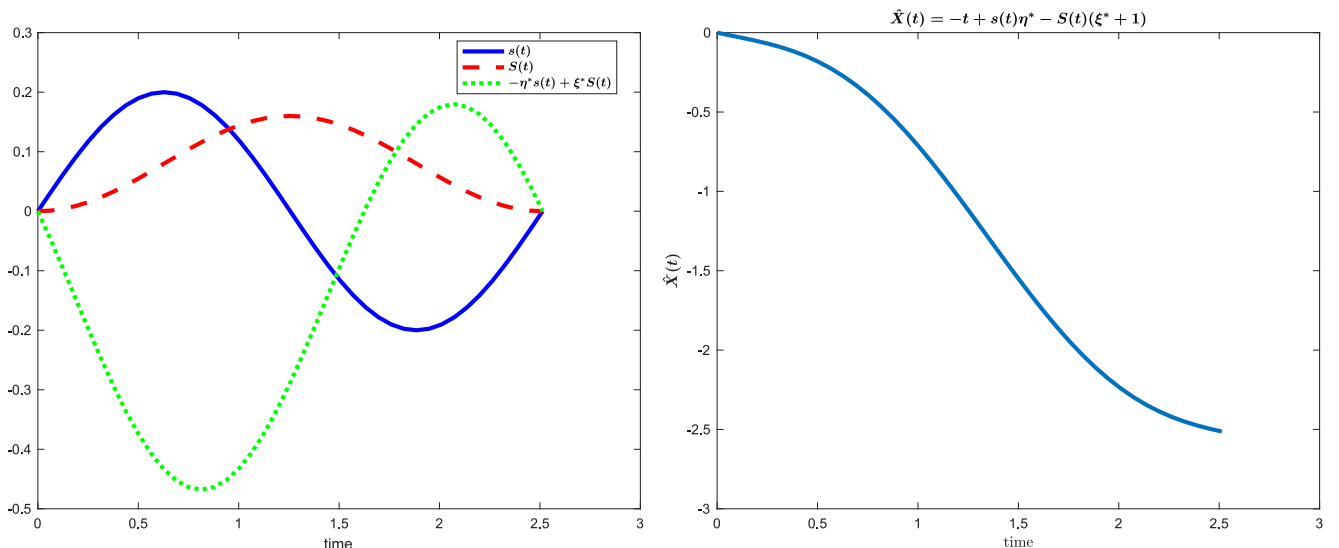

Notes. The plots are for the sinusoidal example in Figure 1 with parameter triple $(\rho, \beta, \gamma) = (0.8, 0.2, 0.1)$ and $\rho = 0.8$ (left) and $\rho = 0.95$ (right). Also displayed are $E[L_y]$, $\lambda_f E[W_y]$, and 95% confidence interval bands, which require zooming in to see.

where $p \in [0, 1)$. This arrival-rate function increases linearly from $1 - \beta$ to $1 + \beta$ on $[0, pc]$ and decreases linearly from $1 + \beta$ to $1 - \beta$ on $[pc, c]$. The periodic arrival-rate function with traffic intensity ρ within one cycle is then $\rho \lambda_f(y)$.

Table 3 reports the optimal η and ξ for $p = 1/2$ (symmetric) and $1/3$ (asymmetric) for $\rho = 0.8$, $\beta = 0.5$ and cycle length $c = 60$. We consider the two objective functions: the maximum expected waiting time and the maximum expected waiting time deviation. Then Figures 11 and 12 plot the expected waiting time under the optimal control for the cases $p = 1/2$ and $p = 1/3$. Table 3 shows a clear difference in the objective functions in the asymmetric case but not in the symmetric case.

Just as for the sinusoidal examples (compare Tables 1 and 2), stabilizing the mean waiting time becomes more difficult as β increases toward the upper limit 1. The most difficult case is $\beta = 1$, where the arrival-rate function is 0 at the end points 0 and c. Table 4 shows the severe performance degradation in this case. Insight

Figure 10. Deterministic Functions Associated with the Diffusion Limit for the Case $\beta = 0.2$, $\gamma = 2.5$, $\hat{\eta} = 1.45$, and $\hat{\xi} = -1.8$.

Notes. On the left appears $\hat{M}(t) = -\hat{\eta}s(t) + \hat{\xi}S(t) = -1.5\beta \sin(\gamma t) - 1.5(\beta/\gamma)(1 - \cos(\gamma t))$ together with $s(t) = \beta \sin(\gamma t)$ and $S(t) = (\beta/\gamma)(1 - \cos(\gamma t))$. On the right appears the diffusion limit for the net input $\hat{X}(t) = -t - M(t)$ when $c_x = 0$, showing that condition (65) holds.

Table 3. Optimal η and ξ for $\rho = 0.8$, $\beta = 0.5$, Cycle Length $c = 60$

p	1/2	1/2	1/3	1/3
Objective	Max EW	Max deviation	Max EW	Max deviation
ρ	0.8	0.8	0.8	0.8
c	60	60	60	60
η_ρ	5.9	5.9	5.8	5.8
$\frac{1-\rho}{\rho}\eta_\rho$	1.48	1.48	1.45	1.45
ξ_ρ	0.86	0.86	0.84	0.87
$\frac{1-\xi_\rho}{(1-\rho)}$	0.7	0.7	0.8	0.65
$\max E[W_y]$	4.1787	4.1787	4.2160	4.2175
$\max(EW_y) - \min(EW_y)$	0.3081	0.3081	0.4200	0.3468
Average wait	4.08	4.08	4.08	4.09
Relative error (%)	7.6	7.6	10.3	8.5

Note. The four cases are determined by the value of p in (88) and the objective for the time-varying waiting time in (9) or (10) in the first two rows of the table.

into the difficult cases with zero or near-zero $\lambda_f(t)$ can be gained from the time-varying Little's law in Corollary 6 and steps (77) and (78) in the proof of Theorem 3.

9.2. Simulation Examples for the Alternative Scaling in Section 8

We now consider four simulation examples in the alternative HT scaling in Section 8. This is the same HT scaling as in Ma and Whitt (2018). We consider the base case of $\beta = 1$, $\gamma = 2.5$, and use

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

Specifically, we consider cases with $\rho = 0.84, 0.92, 0.96, 0.98$. Here we use the lags $\eta_\rho = 5.25, 11.5, 24, 49$ calculated by $\rho/(1-\rho)$, the scaler $\xi_\rho = \rho$. (These are consistent with Theorem 4.)

Figures 13 and 14 show the expected periodic steady-state waiting time (the solid blue line) and the expected steady-state workload (the dashed red line). Figures 13 and 14 show that stabilization is not achieved
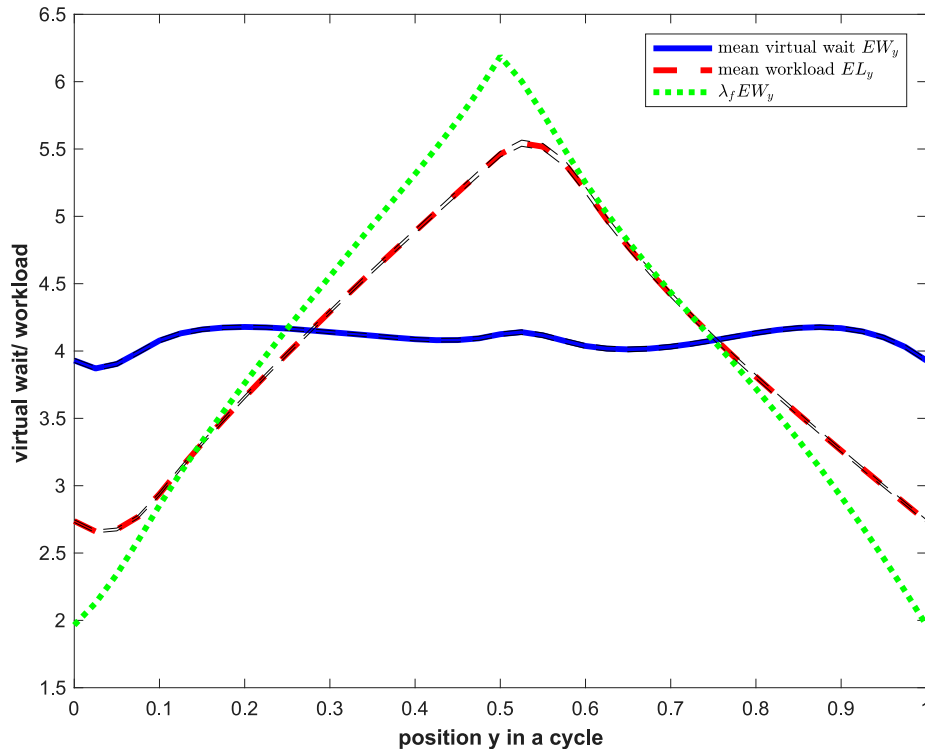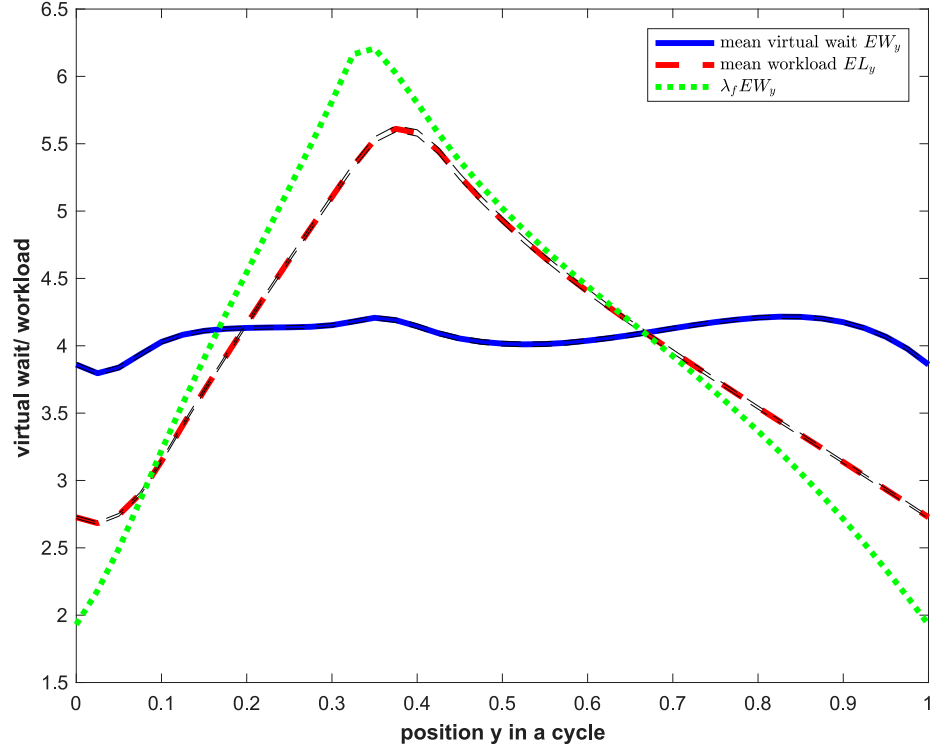
Figure 11. Estimated Optimal Expected Waiting Time (Minimizing the Maximum) for the Non-Sinusoidal Example in (88): for $\rho = 0.8$, $p = 1/2$, and $\beta = 0.5$

Figure 12. Estimated Optimal Expected Waiting Time (Minimizing the Maximum) for the Non-Sinusoidal Example in (88): for $\rho = 0.8, p = 1/3$, and $\beta = 0.5$

10. Conclusions

In this paper, we extended the rare-event simulation algorithm for the periodic $GI_t/GI/1$ model in Ma and Whitt (2018) to the periodic $GI_t/GI_t/1$ model and applied the new algorithm to study methods to stabilize the expected (virtual) waiting time over time. We studied the modification in Equation (6) of the rate-matching service-rate control in Equation (1a) to include a time lag η and a damping factor ξ. We developed and applied a simulation search algorithm to find optimal pairs of control parameters (η, ξ) for the control problems in Equations (9) and (10). Thus, we obtained a practical solution to the open problem in Whitt (2015) of developing an effective way to stabilize the expected waiting time in the periodic single-server model.
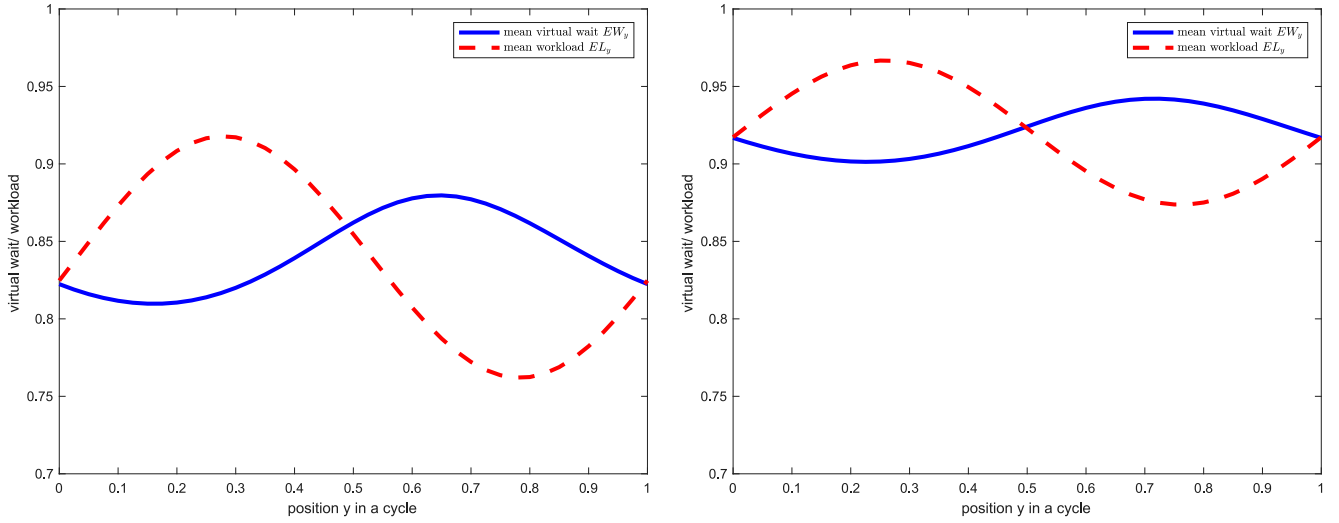
We also established supporting HT limits for the general periodic $G_t/G_t/1$ model and showed that the control problems in Equations (9) and (10) converge to associated diffusion control parameters with

Table 4. Optimal η and ξ for $\rho = 0.8$, $\beta = 1$, Cycle Length $c = 60$

p	1/2	1/2	1/3	1/3
Objective	Max EW	Max deviation	Max EW	Max deviation
ρ	0.8	0.8	0.8	0.8
c	60	60	60	60
η_ρ	5.9	5.9	6.3	6.3
$\frac{1-\rho}{\rho}\eta_\rho$	1.48	1.48	1.58	1.58
ξ_ρ	0.89	0.92	0.84	0.88
$\frac{1-\xi_\rho}{(1-\rho)}$	0.55	0.4	0.8	0.6
$\max E[W_y]$	4.8087	4.8812	4.7656	4.9072
$\max(E[W_y]) - \min(E[W_y])$	1.3715	1.2726	1.8288	1.7775
Average wait	4.32	4.38	4.35	4.40
Relative error (%)	31.7	29.0	42.02	40.40

Note. The four cases are determined by the value of p in (88) and the objective for the time-varying waiting time in (9) or (10) in the first two rows of the table.

Figure 13. The Expected Periodic Steady-State Virtual Waiting Time (Blue Line) and the Expected Steady-State Workload (Red Line) for the Alternative Scaling in Section 9.2 for $\rho = 0.84$, $\beta = 0.16$, $\gamma = 0.064$, $\eta_\rho = 5.25$, $\xi_\rho = 0.84$, Yielding a Maximum Deviation 0.0699 (Left) and $\rho = 0.92$, $\beta = 0.08$, $\gamma = 0.016$, $\eta_\rho = 11.5$, $\xi_\rho = 0.92$, Yielding a Maximum Deviation 0.0408 (Right)
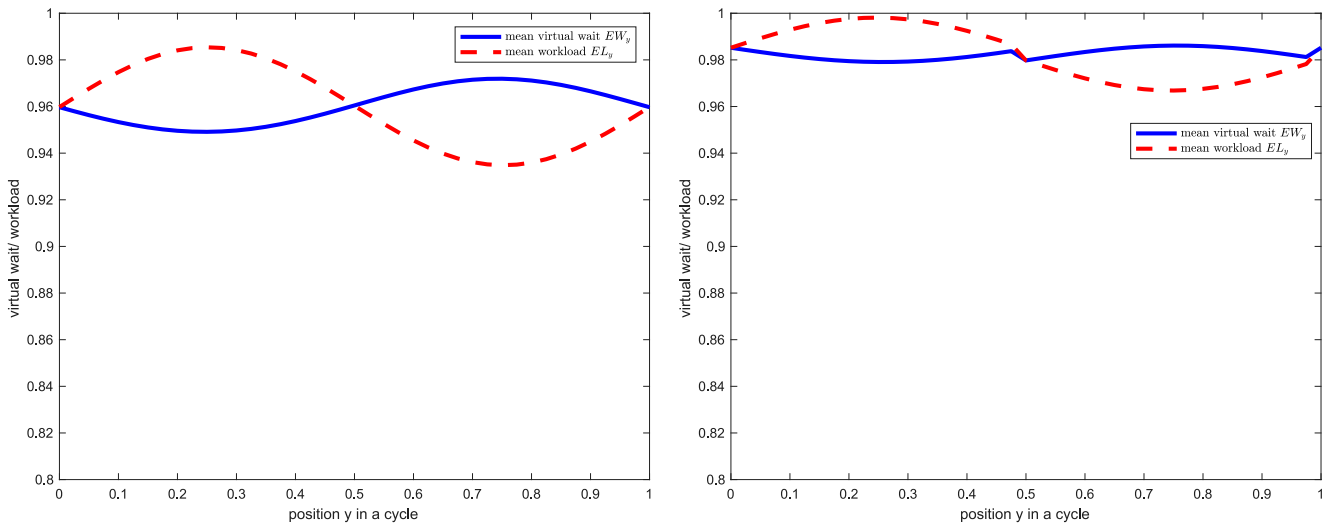

appropriate scaling. The scaling involves the conventional HT scaling associated with $\rho \uparrow 1$, so time is scaled by $(1 - \rho)^{-2}$, whereas space is scaled by $1 - \rho$, but in addition, *to gain insight into the time-varying behavior, we identify and study three different scalings of the arrival-rate function.*

As observed by Falin (1989), if the arrival-rate function is left unscaled, then the HT limit is the same as if the periodicity were not present at all. A major conclusion is that important insight into the time-varying performance can be gained by scaling the arrival-rate function as well. Moreover, as illustrated by Sections 6 and 8, there are two different natural scalings: first, there is the stronger scaling in the fluid scale in Section 6 as in Whitt (2015), and second, there is the weaker scaling in the diffusion scale in Section 8 as in Whitt (2014) and Ma and Whitt (2018). In the weaker scaling, the rate-matching control from Whitt (2015) stabilizes both the queue length and the waiting time, but in the stronger fluid scaling we see significant differences, consistent with the simulation results in Figure 1. This insightful scaling in the fluid scale also yields a limiting diffusion control problem.

We conducted extensive simulation algorithms showing that the new (η, ξ) control is effective in stabilizing the expected waiting time. However, unlike the rate-matching control for the queue length process in

Figure 14. The Expected Periodic Steady-State Virtual Waiting Time (Blue Line) and the Expected Steady-State Workload (Red Line) for the Alternative Scaling in Section 9.2 for $\rho = 0.96$, $\beta = 0.04$, $\gamma = 0.004$, $\eta_\rho = 24$, $\xi_\rho = 0.96$, Yielding a Maximum Deviation 0.0228 (Left) and $\rho = 0.98$, $\beta = 0.02$, $\gamma = 0.001$, $\eta_\rho = 49$, $\xi_\rho = 0.98$, Yielding a Maximum Deviation 0.0070 (Right)

Whitt (2015), the new modified rate-matching control does not stabilize the expected waiting time perfectly, either for fixed ρ or in the HT limit. However, Figures 1 and 3 show that it stabilizes it remarkably well, whereas Figures 4, 5, and 6 show that it stabilizes the full waiting-time distribution quite well too.

It is interesting to consider the performance impact of time-varying arrivals. In Section 1, we observed that the difference between the stable average waiting time in Figure 1 and the value $\rho/(1 - \rho)$ for the stationary model (4 on the left and 19 on the right) might be called the "average cost of periodicity," but Example 1 showed that the overall average expected waiting time with a service-rate control could be much less than in the stationary model. It remains to investigate more carefully.

Indeed, there remain many opportunities for future research, including the open problems mentioned in Section 1.5. It also remains to directly solve the diffusion-control problems with objectives (9) and (10) resulting from Theorem 4, and there are other methods worth carefully studying, such as modifications of the iterated staffing algorithm from Feldman et al. (2008) for single-server models.

Finally, we mention that the methods in this paper generalize and can be applied to other problems. First, the rare-event simulation algorithm in Section 5 applies to any $GI_t/GI_t/1$ model with other service-rate controls. Second, the heavy-traffic limits in Section 6 and 8 evidently extend to general $G_t/G_t/1$ models with other service-rate controls. More generally, simulation of converging stochastic processes is a promising way to numerically solve complex diffusion-control problems.

References

Adusumilli KM, Hasenbein JJ (2010) Dynamic admission and service rate control of a queue. *Queueing Systems* 66(2):131–154.
Asmussen S (2003) *Applied Probability and Queues*, 2nd ed. (Springer, New York).
Asmussen S, Glynn PW (2007) *Stochastic Simulation*, 2nd ed. (Springer, New York).
Bertsimas D, Mourtzinou G (1997) Transient laws of nonstationary queueing systems and their applications. *Queueing Systems* 25(1–4):115–155.
Billingsley P (1999) *Convergence of Probability Measures* (John Wiley & Sons, New York).
Chen H, Mandelbaum A (1991) Discrete flow networks: Bottleneck analysis and fluid approximations. *Math. Oper. Res.* 16(2):408–446.
Choudhury GL, Mandelbaum A, Reiman MI, Whitt W (1997) Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models* 13(1):121–146.
Defraeye M, van Nieuwenhuyse I (2013) Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decision Support Systems* 54(4):1558–1567.
Falin GI (1989) Periodic queues in heavy traffic. *Adv. Appl. Probab.* 21(2):485–487.
Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* 54(2):324–338.
Fralix BH, Riano G (2010) A new look at transient versions of Little's law. *J. Appl. Prob.* 47(02):459–473.
Fu MC, ed. (2015) *Handbook of Simulation Optimization*, International Series in Operations Research and Management Science, vol. 216 (Springer, New York).
George JM, Harrison JM (2001) Dynamic control of a queue with adjustable service rate. *Oper. Res.* 49(5):720–731.
He B, Liu Y, Whitt W (2016) Staffing a service system with non-Poisson nonstationary arrivals. *Probab. Engrg. Inform. Sci.* 30(1):593–621.
Heyman DP (1968) Optimal operating policies for $M/G/1$ queueing systems. *Oper. Res.* 16(2):362–382.
Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42(10):1383–1394.
Jian N, Henderson SG (2015) An introduction to simulation optimization. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds. *Proc. 2015 Winter Simulation Conf.* (ACM, New York), 1780–1794.
Kleinrock L (1964) *Communication Nets: Stochastic Message Flow and Delay* (Dover, New York).
Kwon S, Gautam N (2016) Guaranteeing performance based on time stability for energy-efficient data centers. *IIE Trans.* 48(9):812–825.
Liu Y (2018) Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Oper. Res.* 66(2):514–534.
Liu Y, Whitt W (2011) A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.* 59(4):835–846.
Liu Y, Whitt W (2012a) The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* 71(4):405–444.
Liu Y, Whitt W (2012b) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6):1551–1564.
Loynes RM (1962) The stability of a queue with non-independent inter-arrival and service times. *Math. Proc. Cambridge Philos. Soc.* 58(3):497–520.
Ma N, Whitt W (2015) Using simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rate. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds. *Proc. 2015 Winter Simulation Conf.* (ACM, New York), 1–12.
Ma N, Whitt W (2016) Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statist. Probab. Lett.* 102:202–207.
Ma N, Whitt W (2018) A rare-event simulation algorithm for periodic single-server queues. *INFORMS J. Comput.* 30(1):71–89.
Massey WA, Whitt W (1994) Unstable asymptotics for nonstationary queues. *Math. Oper. Res.* 19(2):267–291.
Nieuwenhuis G (1989) Equivalence of functional limit theorems for stationary point processes and their Palm distributions. *Probab. Theoret. Related Fields* 81(4):593–608.
Pender J, Massey WA (2017) Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probab. Engrg. Inform. Sci.* 31(01):1–42.
Puhalskii A (1994) On the invariance principle for the first passage time. *Math. Oper. Res.* 19(4):946–954.
Rolski T (1981) Queues with nonstationary input stream: Ross's conjecture. *Adv. Appl. Probab.* 13(03):603–618.
Rolski T (1989) Queues with nonstationary inputs. *Queueing Systems* 5(1–3):113–129.
Sigman K (1995) *Stationary Marked Point Processes: An Intuitive Approach* (Chapman and Hall/CRC, New York).
Sigman K, Whitt W (2019) Marked point processes in discrete time. *Queuing Systems* 92(1–2):47–81.

Stolletz R (2008) Approximation of the nonstationary $M(t)/M(t)/c(t)$-queue using stationary models: The stationary backlog-carryover approach. *Eur. J. Oper. Res.* 190(2):478–493.

Sun X, Whitt W (2018) Delay-based service differentiation with many servers and time-varying arrival rates. *Stochastic Systems* 8(3):230–263.

Suriadi S, Wynn MT, Xu J, van der Aalst WMP, ter Hofstede AHM (2017) Discovering work prioritization patterns from event logs. *Decision Support Systems* 100:77–92.

Whitt W (1971) Weak convergence theorems for priority queues: Preemptive-resume discipline. *J. Appl. Probab.* 8(1):74–94.

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

Whitt W (2014) Heavy-traffic limits for queues with periodic arrival processes. *Oper. Res. Lett.* 42(6–7):458–461.

Whitt W (2015) Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems* 81(4):341–378.

Whitt W (2018) Time-varying queues. *Queueing Models Service Management* 1(2):79–164.

Whitt W, Zhang X (2019) Periodic Little's law. *Oper. Res.* 67(1):267–280.

Yadin M, Naor P (1963) Queueing systems with a removable service station. *Oper. Res. Quart.* 14(4):393–405.