Stochastic Systems

arXiv: arXiv:0000.0000

MINIMIZING THE MAXIMUM EXPECTED WAITING TIME IN A PERIODIC SINGLE-SERVER QUEUE WITH A SERVICE-RATE CONTROL

By Ni Ma*, and Ward Whitt*

June 20, 2018

We consider a single-server queue with unlimited waiting space, the FCFS discipline, a periodic arrival-rate function and i.i.d. service requirements, where the service-rate function is subject to control. We previously showed that a rate-matching control, where the service rate is made proportional to the arrival rate, stabilizes the queue length process, but not the (virtual) waiting time process. In order 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. That 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 [41], which considered the problem of stabilizing performance over time, i.e., 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

*Department of Industrial Engineering and Operations Research, Columbia University

Keywords and phrases: nonstationary queues, queues with time-varying arrival rate, stabilizing performance, heavy traffic limits, service rate controls

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; i.e., 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 §1 of [41], 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 [19] and in business process management [37].

In [41] it was shown that a rate-matching control, where 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 multi-server queue to stabilize performance in face of time-varying demand, e.g., [8, 10, 14, 16, 20, 23, 30, 35, 42]. 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 [44, 15]. Similar dynamic control problems for single-server queues including service-rate controls have been analyzed as Markov decision processes in [1, 13] 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, while 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 single-server queues. That difference can be seen by comparing the many-server fluid models in [22] to the single-server fluid models in [6], as discussed on p. 836 of [21]. A simple fluid model supporting the rate-matching control in [41] is supported by our heavy-traffic weak law of large numbers in Theorem 6.1, see Corollary 6.1, 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 analog of the square-root staffing function from [16] was considered as a candidate service-rate control in (2.3) of [41], but was found to be ineffective, as illustrated by Figure 2 of [41]. Also variants of the iterated staffing algorithm (ISA) in [10] and [8] 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 [41], 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 [18]; we allocate service capacity over time instead 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 §3.1 of [41]. In [41], 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; i.e., for a given target traffic intensity ρ, the service-rate function is

$$\mu(t) \equiv \lambda(t)/\rho, \quad t \geq 0, \tag{1.1}$$

with $\equiv$ denoting equality by definition. In [41], 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, while 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 [41] 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

$$\lambda(t) \equiv \rho(1 + s(t)), \quad t \geq 0, \tag{1.2}$$

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

$$\bar{s} \equiv \frac{1}{c} \int_0^c s(u)\, du \equiv 0. \tag{1.3}$$

As a regularity condition, we require that

$$s_L \leq s(t) \leq s_U \quad \text{for all} \quad t \quad \text{with} \quad -1 < s_L \leq 0 \leq s_U < \infty. \tag{1.4}$$

Most of our numerical examples will be for a sinusoidal function, where $s(t) = \beta \sin(\gamma t)$ for $s(t)$ in (1.2), so that

$$\lambda(t) \equiv \rho(1 + \beta \sin(\gamma t)), \quad t \geq 0, \tag{1.5}$$

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

In the periodic setting of (1.2)-(1.4), we consider the rate-matching control in (1.1) modified by a *time lag* η and *damping factor* ξ; in particular,

$$\mu(t) \equiv 1 + \xi s(t - \eta), \quad t \geq 0, \tag{1.6}$$

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 that 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 if $\beta > (1 - \rho)/\rho$.

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, i.e.,

$$w^* \equiv \min_{\mu(t) \in \mathcal{M}(1)} \max_{0 \leq y \leq 1} \{E[W_y]\}, \tag{1.7}$$

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

$$E[W] = \frac{\rho(1+c_s^2)}{2(1-\rho)} \tag{1.8}$$

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 2.1 and Example 2.1.

We have not yet solved this general optimization problem in (1.7). 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 (1.5), 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 1.1. (*stabilizing the full waiting time distribution*) Theorems 4.1 and 4.2 of [41] show that the rate-matching control stabilizes the delay probability $P(W_y > 0)$, while Corollary 5.1 of [41] shows that the rate-matching control cannot stabilize the mean waiting time. Theorem 5.2 of [41] establishes a heavy-traffic limit (with periodicity in the stronger fluid scale in §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 show significantly degraded stabilization for $P(W_y > 0)$ and the variance $Var(W_y)$ in Figure 4 in §9.1.

In this paper, we only consider the restricted set of controls in (1.6). Now our goal is

$$w^*(\eta, \xi) \equiv \min_{\eta,\xi} \max_{0 \le y \le 1} \{E[W_y]\}. \tag{1.9}$$

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, where (1.9) is replaced by

$$w^*_{stab}(\eta,\xi) \equiv \min_{\eta,\xi}\{\max_{0\le y\le 1}\{E[W_y]\} - \min_{0\le y\le 1}\{E[W_y]\}\}. \tag{1.10}$$

In our sinusoidal examples, where there is strong symmetry, we find that the solutions to (1.9) and (1.10) 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 the Paper.* Our paper involves some challenging technical methods. Hence, we present the more accessible results first. We start in §2 by presenting two simulation examples to illustrate the effectiveness of our new algorithm. Then in §3 we introduce the two technical tools we will apply: (i) an extension of the rare-event simulation algorithm for the $GI_t/GI/1$ model from [27] to the $GI_t/GI_t/1$ model with a general service-rate control and (ii) heavy-traffic limits involving scaling of the underlying deterministic periodic arrival-rate function.

We start in earnest in §4. We elaborate on the model and key processes representing the workload and the waiting time in §4. Theorem 4.1 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 [27] to our setting and its application to perform simulation search in §5. In both §4 and §5 we will be brief because we can draw upon [41] and [27].

We establish our main heavy-traffic limits with the periodicity in the stronger fluid scaling (see (3.2)) in §6. We present the proof of the main heavy-traffic limit, Theorem 6.2, in §7. We establish heavy-traffic limits with the periodicity in the weaker diffusion scaling (see (3.3)) in §8.

We give simulation examples in §9. In §9.1 we present simulation results using the fluid scaling in §6; in §9.2 we present simulation results using the diffusion scaling in §8. We draw conclusions in §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 (1.2)-(1.5). The first example has model parameters $(\rho,\beta,\gamma) = (0.8, 0.2, 0.1)$, so that 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

$0 \le y < 1$. 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.

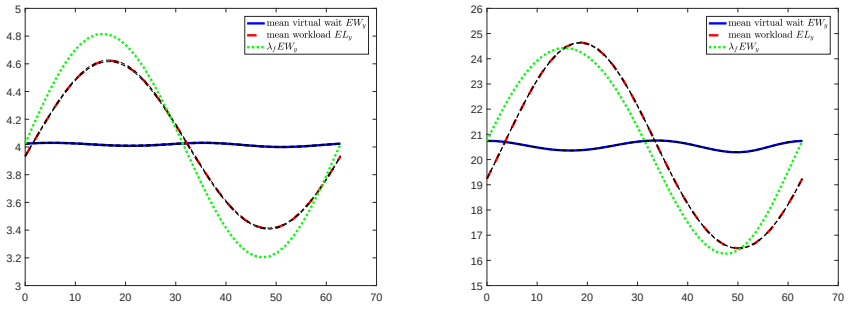


FIG 1. *Estimates of the periodic steady-state values of* $E[W_y]$ *(blue solid line),* $E[L_y]$ *(red dashed line) and* $\lambda(y)E[W_y]$ *(green dotted line) for the optimal control* (η^*, ξ^*) *for the sinusoidal example with parameter triples* $(\rho, \beta, \gamma) = (0.8, 0.2, 0.1)$ *(left) and* $(0.95, 0.2, 0.1)$ *(right), so that the cycle length is* $c = 2\pi/\gamma = 62.8$. *The optimal controls are* $(5.84, 0.84)$ *for* $\rho = 0.8$ *and* $(15.1, 2.13)$ *for* $\rho = 0.95$.

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 about 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 the rate-matching control for this same model, see Figure 6 of [41]. (See [26] for more examples.)

It is natural to wonder if 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 6.2.

REMARK 2.1. (*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

classical results for the periodic $M_t/GI/1$ queue in [32, 33] so not apply because, in general, the service times are neither independent of the arrival process nor i.i.d.; See Example 2.1.

EXAMPLE 2.1. (*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:

$$\lambda(t) \equiv \rho b 1_{[(c/2)-\delta,(c/2)+\delta)}(t), \quad 0 \le t < c \quad \text{and} \quad b\delta = c, \tag{2.1}$$

where $\delta < c/2$ and 1_A is the indicator function of the set A, i.e., $1_A(t) = 1$ if $t \in A$ and 0 otherwise. Let the service-rate function be as in (1.6) 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, while 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 2.2. (*the single-parameter alternative*) It is natural to wonder if 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 about 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 [27] 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 §4.) The workload $L(t)$ represents the amount of work in service

time in the system at time t, while the waiting time can be represented as the first-passage time

$$W(t) = \inf\{u \geq 0 : \int_t^{t+u} \mu(s)\, ds = L(t)\}. \tag{3.1}$$

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

As in [27], 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 employ a search over the parameters (η, ξ), as discussed in §5, in order to solve the optimization problems (1.9) and (1.10). The search part is relatively elementary because we have only two control parameters. For background on simulation optimization, see [12, 17] and the references there.

The computational complexity for one control vector (η, ξ) is essentially the same as in [27]. 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. That required about 100 minutes on a desktop computer. As indicated in §4.7 of [27], the run time tends to be of order $(1-\rho)^{-1}$, so that 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 depends 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 [40, 41, 27]. 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 [39], 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 [9].

To obtain insight into the periodic dynamics, it is thus important to also scale the arrival-rate function, which is ininitially specified in (1.2) with (1.3) and (1.4). However, the papers [40] and [41] actually use two different HT scalings of the arrival-rate function. *Our main HT scaling in §6 follows [41] and has periodicity in the fluid scale*, i.e.,

$$\lambda_\rho(t) \equiv \rho(1 + s((1-\rho)^2 t)), \quad t \geq 0, \tag{3.2}$$

but in §8 we also consider the scaling from [40] and [27], which has the periodicity in diffusion scale, i.e.,

$$\lambda_\rho(t) \equiv \rho(1 + (1-\rho)s((1-\rho)^2 t)), \quad t \geq 0. \tag{3.3}$$

The extent of the periodicity is stronger in (3.2) than in (3.3), because of the extra factor $(1-\rho)$ before s in (3.3). The workload and the waiting time have the same HT limit with the diffusion-scale scaling in (3.3), but different limits with the fluid-scale scaling in (3.2). To capture the clear differences shown in Figure 1, we obviously want the stronger fluid scaling in (3.2). The HT functional central limit theorem (FCLT) in Theorem 6.2 for the scaling in (3.2) in §6 helps interpret Figure 1.

It is important to note that if we have 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 [7].

We also establish a HT functional weak law of large numbers (FWLLN) in Theorem 6.1, 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 §4.1 and the basic queueing stochastic processes in §4.2. We specialize to the periodic $G_t/G_t/1$ model in §4.3. We show that the workload is stabilized by the rate-matching control in (1.1), extending the results for the queue-length process in [41].

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

$$A(t) \equiv N(\Lambda(t)), \quad \text{where} \quad \Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0. \tag{4.1}$$

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 (4.1)

has been used frequently for processes N more general than NHPP's, an early source being by [28].

For the sinusoidal arrival-rate function in (1.5), the associated cumulative arrival-rate function is

$$\Lambda(t) = \rho(t + (\beta/\gamma)(1 - \cos(\gamma t))), \quad t \geq 0. \tag{4.2}$$

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: (i) $0 < \beta < \rho^{-1} - 1$ and (ii) $\rho^{-1} - 1 \leq \beta \leq 1$. In the first case, we have $\rho(t) < 1$ for all t, where $\rho(t) \equiv \lambda(t)$ is the instantaneous traffic intensity, but in the second case we have intervals with $\rho(t) \geq 1$, where significant congestion can build up. If there is a long cycle as well, the system may be better understood from fluid and diffusion limits, as in [7]. However, that 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 [41], we separately define the rate at which service is performed from the service requirement. Given the arrival counting process $A(t)$ defined in §4.1, let the total input of work over the interval $[0, t]$ be the random sum

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0. \tag{4.3}$$

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 (4.1), let the cumulative available service rate be

$$M(t) \equiv \int_0^t \mu(s)\, ds, \quad t \geq 0. \tag{4.4}$$

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 4.1. (*waiting time representation*) *The waiting time at time t can be represented as*

$$W(t) = M_t^{-1}(L(t)), \quad t \geq 0, \tag{4.5}$$

where M_t^{-1} is the inverse of $M_t(u) \equiv M(t+u) - M(t)$ for $M(t)$ in (4.4).

Proof. By definition,

$$\begin{aligned} W(t) &= \inf\{u \geq 0 : \int_t^{t+u} \mu(s)\, ds = L(t)\} \\ &= \inf\{u \geq 0 : M(t+u) - M(t) = L(t)\} = M_t^{-1}(L(t)), \end{aligned} \tag{4.6}$$

for $M_t(u)$ above, as claimed in (4.5). ∎

4.3. *The Periodic $G_t/G_t/1$ Model.* As in [27], we consider the periodic steady state of the periodic $G_t/G_t/1$ model with arrival-rate function in (1.2). For that purpose, we exploit the arrival process construction in (4.1) in terms of the stationary processes $N \equiv \{N(t) : t \geq 0\}$ and $V \equiv \{V_k : k \geq 1\}$ in (4.1). Let the associated service-rate function $\mu(t)$ also be periodic with cycle length c, with average service rate be $\bar{\mu} = 1$, and bounds $0 < \mu_L \leq \mu(t) \leq \mu_U < \infty$, for $0 \leq t \leq c$.

As in [27] and earlier in [24] and Chapter 6 in [34], We now convert the standard representation of the workload process in §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 backwards 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

$$\tilde{\Lambda}_y(t) \equiv \Lambda(yc) - \Lambda(yc - t), \quad t \geq 0, \tag{4.7}$$

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 analog of (4.7) for the cumulative service rate, let

$$\tilde{M}_y(t) \equiv M(yc) - M(yc-t), \quad t \geq 0, \tag{4.8}$$

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

$$\begin{aligned} L_y(t) &= \sup_{0\leq s\leq t} \{\tilde{X}_y(s)\} \\ &\stackrel{d}{=} \sup_{0\leq s\leq t} \Big\{ \sum_{k=1}^{N(\tilde{\Lambda}_y(s))} V_k - \tilde{M}_y(s) \Big\} = \sup_{0\leq s\leq \tilde{\Lambda}_y(t)} \Big\{ \sum_{k=1}^{N(s)} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(s)) \Big\}, \end{aligned} \tag{4.9}$$

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 (4.9) are the reverse-time cumulative arrival-rate function $\tilde{\Lambda}_y(t)$ in (4.7) with inverse $\tilde{\Lambda}_y^{-1}(t)$ and the reverse-time cumulative service-rate function $\tilde{M}_y$ in (4.8) with inverse $\tilde{M}_y^{-1}$.

The equality in distribution in (4.9) holds because N is a stationary point process, which is a point process with stationary increments and a constant rate.

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

$$L_y \stackrel{\mathrm{d}}{=} \sup_{s\geq 0} \Big\{ \sum_{k=1}^{N(s)} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(s)) \Big\}, \qquad 0 \leq y < 1. \tag{4.10}$$

Even though (4.9) is valid for all t, we think of the system starting empty at times $-kc$, for $k \geq 1$, so that 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 (4.10). That makes (4.9) valid to describe the distribution of $L(c(k+y))$ for all $k \geq 1$.

We now observe that the time transformation in (4.9) 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 4.1. (*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* (4.10) *above, 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

$$L_y \;=\; \sup\{\sum_{k=1}^{N(\tilde{\Lambda}_y(\tilde{M}_y^{-1}(s)))} V_k - s : s \geq 0\}. \tag{4.11}$$

COROLLARY 4.2. (*the associated periodic steady-state waiting time*) *The periodic steady-state waiting time associated with the periodic steady-state workload in* (4.10) *is*

$$W_y \;=\; \tilde{M}_y^{-1}(L_y), \quad 0 \leq y < 1. \tag{4.12}$$

Proof. Apply the reasoning of Lemma 4.1. ∎

In [41] we showed that the rate-matching service-rate control in (1.1) stabilizes the queue-length process. Now we establish the corresponding result for the workload.

THEOREM 4.1. (*stabilizing the periodic workload*) *If the rate-matching control in* (1.1) *is used, then $L_y \stackrel{\mathrm{d}}{=} L$ for L_y in* (4.10)*, where L is the steady-state workload in the associated (stable) stationary $G/G/1$ model, i.e.,*

$$L \stackrel{\mathrm{d}}{=} \sup_{s\geq 0}\Big\{\sum_{k=1}^{N(s)} V_k - \rho^{-1}s\Big\}, \tag{4.13}$$

which is independent of y.

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

$$\begin{aligned} L_y \;&\stackrel{\mathrm{d}}{=}\; \sup_{s\geq 0}\Big\{\sum_{k=1}^{N(s)} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(s)\Big\} \\ &\stackrel{\mathrm{d}}{=} \sup_{s\geq 0}\Big\{\sum_{k=1}^{N(s)} V_k - cs\Big\} \stackrel{\mathrm{d}}{=} L. \end{aligned} \tag{4.14}$$

∎

5. The Simulation Search Algorithm. The rare-event simulation algorithm from [27] exploits the classic rare-event simulation algorithm for the $GI/GI/1$ queue, exploiting importance sampling using an exponential change of measure, as in Ch. XIII of [2] and Ch. VI of [3]. Hence our simulation algorithm applies to the $GI_t/GI_t/1$ queue. It was shown in [27] that the algorithm is effective for estimating the mean as well as small tail probabilities. (Also see [25].)

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, $k \geq 2$. Also in the $GI_t/GI_t/1$ setting, the service times V_k's 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 (4.10) and the waiting time W_y, i.e.,

$$\begin{aligned} L_y &\stackrel{\mathrm{d}}{=} \sup_{s\geq 0}\Big\{\sum_{k=1}^{N(s)} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(s))\Big\} \\ &\stackrel{\mathrm{d}}{=} \sup_{n\geq 0}\Big\{\sum_{k=1}^{n} V_k - \tilde{M}_y(\tilde{\Lambda}_y^{-1}(\sum_{k=1}^{n} U_k))\Big\}, \\ W_y &\stackrel{\mathrm{d}}{=} M_y^{-1}(L_y), \qquad 0 \leq y < 1. \end{aligned} \tag{5.1}$$

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 [27], which is based on an exponential change of measure; we refer to [27] 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

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

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 below:

$$\begin{aligned} P(W_y > b) &= P(M_y^{-1}(L_y) > b) = P(L_y > M_y(b)) \\ &= P(\tau^Q_{M_y(b)} < \infty) = E_{\theta^*}[\Psi_{\tau^Q_{M_y(b)}}(\theta^*)] \\ &= E_{\theta^*}[m_{X_1}(\theta^*) m_X(\theta^*)^{(\tau^Q_{M_y(b)}-1)} e^{-\theta^* S_{\tau^Q_{M_y(b)}}}] \\ &= m_{X_1}(\theta^*) E_{\theta^*}[e^{\theta^* S_{\tau^Q_{M_y(b)}}}]. \end{aligned} \tag{5.2}$$

Again the first $X_1(\rho)$ in the partial sum $S_{\tau^Q_{M_y(b)}}$ 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, we 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 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 above 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 (1.5), (1.6) and (4.2):

$$\begin{aligned}
\tilde{\Lambda}_y(t) &= \rho(t + \frac{\beta}{\gamma}(\cos(\gamma(t - yc)) - \cos(\gamma yc))) \\
M(t) &= t - \xi\frac{\beta}{\gamma}(\cos(\gamma(t - \eta)) - \cos(\gamma\eta)) \\
M_y(t) &= t - \xi\frac{\beta}{\gamma}(\cos(\gamma(t + yc - \eta)) - \cos(\gamma(yc - \eta))) \\
\tilde{M}_y(t) &= t + \xi\frac{\beta}{\gamma}(\cos(\gamma(t + \eta - yc)) - \cos(\gamma(\eta - yc))).
\end{aligned} \tag{5.3}$$

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. That simple approach is substantiated by estimating the structure of the objective function. Figure 2 illustrates by showing the maximum waiting time $\max_{0\le y\le c}\{E[W_y]\}$ in the setting of Figure 1 (left). Figure 2 shows estimates of the maximum waiting time

$\max_{0\le y\le c}\{E[W_y]\}$ as a function of (η,ξ) in $[0,20]\times[0,5]$ (left) $[3,9]\times[0.6,1.0]$ (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 $\max_{0\le y\le c}\{E[W_y]\} - \max_{0\le y\le c}\{E[W_y]\}$ are similar.

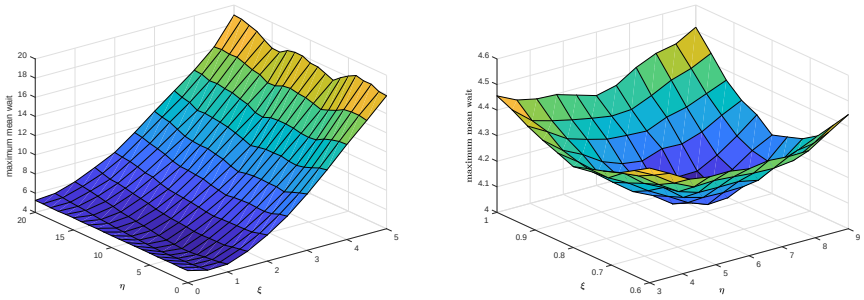


FIG 2. *Three-dimensional plots of estimates of the maximum waiting time* $\max_{0\le y\le c}\{E[W_y]\}$ *for* (η,ξ) *in* $[0,20]\times[0,5]$ *(left)* $[3,9]\times[0.6,1.0]$ *(right).*

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 yields excellent statistical precision, as can be seen from the narrow confidence interval bands in Figure 1.

6. Supporting Heavy-Traffic Limits for Periodic Queues. In this section we obtain a heavy-traffic (HT) functional weak law of large numbers (FWLLN) and a HT functional central limit theorem (FCLT) for the periodic $G_t/G_t/1$ model with a general service-rate control of the form in (1.6). 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 (6.27) in Theorem 6.2 and the conclusion in (6.19). Thus we reduce the optimization problems over the parameter pairs (η_ρ,ξ_ρ) in (1.9) and (1.10), 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 §4 are independent and each satisfies a FCLT. To state the result, let $\hat{N}_n^a$ and $\hat{S}_n^v$ be

the scaled processes defined by
(6.1)

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

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

$$\hat{N}_n^a \Rightarrow c_a B_a \quad \text{and} \quad \hat{S}_n^v \Rightarrow c_s B_s \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{6.2}$$

where $\mathcal{D}$ is the usual function space of right-continuous real-valued functions on $[0, \infty)$ with left limits and $\Rightarrow$ denotes convergence in distribution, as in [39], while B_a and B_s are independent standard (mean 0, variance 1) Brownian motion processes (BM's). The assumed independence implies joint convergence in (6.2) by Theorem 11.4.4 of [39].

We emphasize that GI assumptions are not needed, but that is an important special case. If the service times V_k are i.i.d. mean-1 random variables with variance, also the squared coefficient of variation (scv), c_s^2, then the limit in (6.2) 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 (6.2) holds with arrival variability parameter c_a. (See [29] for theoretical support in the case of an equilibrium renewal process.)

For the queueing HT FCLT, we will apply Theorem 9.3.4 of [39], 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. That convergence follows from the FCLT's we assumed for $\hat{N}_n^a$ and $\hat{S}_n^v$ in (6.2) above. 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 [39].

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 (3.2) in the setting of (1.2)-(1.4). As a further regularity condition, we also require that the function s be an element of the function space $\mathcal{D}$, as in [39]. Then the associated cumulative arrival-rate function in model ρ be

$$\Lambda_\rho(t) \equiv \rho(t + (1-\rho)^{-2} S((1-\rho)^2 t), \quad t \geq 0, \tag{6.3}$$

where

$$S(t) \equiv \int_0^t s(u)\,du, \tag{6.4}$$

for s again being the periodic function in (1.2)-(1.4). From (6.3)-(6.4), we see that the associated arrival-rate function obtained by differentiation in (6.3) is indeed $\lambda_\rho(t)$ in (3.2).

The time scaling in (3.2) and (6.3) implies that the period in model ρ with arrival-rate function $\lambda_\rho(t)$ in (3.2) is $c_\rho = c(1-\rho)^{-2}$, where c is the period of $s(t)$ in (1.2)-(1.4). Thus the period c_ρ in model ρ is growing with ρ. This scaling follows Lemma 5.1 and Theorem 5.2 of [41], 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 [40] and Theorem 2 of [27].

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 (6.3) in place of Λ in (4.1). Given that definition, we see that the cumulative arrival rate is indeed

$$E[A_\rho(t)] = E[N^a(\Lambda_\rho(t))] = \Lambda_\rho(t), \quad t \ge 0. \tag{6.5}$$

We now define associated scaled time-varying service-rate functions. These are the rate-matching service-rate functions in [41] modified by a time lag and a damping factor. In particular,

$$\begin{aligned} \mu_\rho(t) &\equiv 1 + \xi_\rho s((1-\rho)^2(t-\eta_\rho)) \quad \text{and} \\ M_\rho(t) &\equiv \int_0^t \mu_\rho(u)\,du = t \\ &\qquad +(1-\rho)^{-2}\xi_\rho S((1-\rho)^2(t-\eta_\rho)), \quad t \ge 0, \end{aligned} \tag{6.6}$$

where s is the periodic function with period c in (1.3), while η_ρ is the ρ-dependent *time lag* and ξ_ρ is the ρ-dependent *damping factor*. From (6.6) and (1.3), 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, while the instantaneous traffic intensity at time t is $\lambda_\rho(t)/\mu_\rho(t)$, $t \ge 0$, which is a more complicated periodic function, again with period c.

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

$$C_\rho(t) \equiv Y_\rho(t) - L_\rho(t), \quad t \ge 0. \tag{6.7}$$

We now can apply Lemma 4.1 in §4 to express the waiting time process as

$$W_\rho(t) \equiv \inf\{u \geq 0 : M_\rho(t+u) - M_\rho(t) \geq L_\rho(t)\}, \quad t \geq 0. \tag{6.8}$$

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 (3.7) and (3.8) of [41], 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

$$D_\rho(t) \equiv N^s\left(\int_0^t \mu_\rho(s) 1_{\{L_\rho(s)>0\}}\, ds\right), \quad t \geq 0, \tag{6.9}$$

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

6.3. *The Scaled Queueing Processes.* We start with the FWLLN-scaled processes. First let the scaled deterministic rate functions be

$$\bar{\Lambda}_\rho(t) \equiv (1-\rho)^2\Lambda_\rho((1-\rho)^{-2}t) \quad \text{and} \quad \bar{M}_\rho(t) \equiv (1-\rho)^2 M_\rho((1-\rho)^{-2}t), \quad t \geq 0, \tag{6.10}$$

for $\Lambda_\rho(t)$ in (6.3) and $M_\rho(t)$ in (6.6). We immediately see that

$$\bar{\Lambda}_\rho \to \Lambda_f \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \uparrow 1, \tag{6.11}$$

where

$$\Lambda_f(t) \equiv t + S(t), \quad t \geq 0, \tag{6.12}$$

for $S(t)$ in (6.4).

Let the FWLLN-scaled arrival arrival stochastic process be defined by

$$\bar{A}_\rho(t) \equiv (1-\rho)^2 A_\rho((1-\rho)^{-2}t), \tag{6.13}$$

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

$$\begin{aligned} \hat{\Lambda}_\rho(t) &\equiv (1-\rho)[\Lambda_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)], \\ \hat{M}_\rho(t) &\equiv (1-\rho)[M_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)] \end{aligned} \tag{6.14}$$

for Λ_f in (6.12). Let the associated FCLT-scaled stochastic processes be defined by

$$\begin{aligned}
\hat{A}_\rho(t) &\equiv (1-\rho)[A_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)],\\
\hat{Y}_\rho(t) &\equiv (1-\rho)[Y_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)],\\
\hat{X}_\rho(t) &\equiv (1-\rho)X_\rho((1-\rho)^{-2}t),\\
\hat{L}_\rho(t) &\equiv (1-\rho)L_\rho((1-\rho)^{-2}t),\\
\hat{C}_\rho(t) &\equiv (1-\rho)[C_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\Lambda_f(t)],\\
\hat{W}_\rho(t) &\equiv (1-\rho)W_\rho((1-\rho)^{-2}t), \quad t \ge 0. \qquad (6.15)
\end{aligned}$$

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 6.1 below shows that the 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 §13.5 of [39], i.e.,

$$\psi(x)(t) \equiv x(t) - (\inf\{x(s) : 0 \le s \le t\} \wedge 0), \quad t \ge 0. \qquad (6.16)$$

THEOREM 6.1. (*HT FWLLN*) *Under the definitions and assumptions in* §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*

$$\bar{M}_\rho \to M_f \quad \text{in} \quad \mathcal{D}, \quad \text{where} \quad M_f(t) \equiv t + \xi S(t-\eta) \qquad (6.17)$$

and

$$(\bar{A}_\rho, \bar{Y}_\rho, \bar{X}_\rho, \bar{L}_\rho, \bar{C}_\rho, \bar{W}_\rho) \Rightarrow (\bar{A}, \bar{Y}, \bar{X}, \bar{L}, \bar{C}, \bar{W}) \quad \text{in} \quad \mathcal{D}^6 \quad \text{as} \quad \rho \uparrow 1 \qquad (6.18)$$

for $(\bar{A}_\rho, \bar{Y}_\rho, \bar{X}_\rho, \bar{L}_\rho, \bar{C}_\rho, \bar{W}_\rho)$ *defined in* (6.13), *where*

$$\begin{aligned}
\bar{A}(t) &\equiv \bar{Y}(t) \equiv \Lambda_f(t), \quad \bar{X}(t) \equiv S(t) - \xi S(t-\eta), \quad t \ge \eta,\\
\bar{L}(t) &\equiv \sup_{0 \le s \le c}\{X(t) - X(t-s)\}, \quad t \ge c + \eta, \quad \bar{C} \equiv \bar{Y} - \bar{L}, \quad \text{and}\\
\bar{W}(t) &\equiv \inf\{u \ge 0 : M_f(t+u) - M_f(t) \ge \bar{L}(t)\}, \quad t \ge 0. \qquad (6.19)
\end{aligned}$$

for $\Lambda_f(t)$ *in* (6.12) *with* $S(t)$ *in* (6.4), $M_f(t)$ *in* (6.17) *and* ψ *being the reflection map in* (6.16).

Proof. We successively apply the continuous mapping theorem (CMT) using the functions in §12.7 and §§13.2-13.6 of [39]. First, observe that (6.17) is a minor modification of (6.10). Let $\bar{N}^a_\rho$ and $\bar{S}_\rho$ denote $\bar{N}^a_n$ and $\bar{S}^v_n$, respectively, where, paralleling (6.1), we let $\bar{N}^a_n(t) \equiv n^{-1}N^a(nt)$ and $\bar{S}^v_n \equiv n^{-1}S^v_{\lfloor nt \rfloor}$, $t \geq 0$, and then let $n = (1-\rho)^{-2}$. Then observe that $\bar{A}_\rho = \bar{N}^a_\rho \circ \bar{\Lambda}_\rho$ and $\bar{Y}_\rho = \bar{S}_\rho \circ \bar{A}_\rho$, so that 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 (6.16). To establish the limit for the scaled waiting time $\bar{W}_\rho(t)$ in $\mathcal{D}$ 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 6.1. (*FWLLN for the rate-matching service rate control*) *In addition to the conditions of Theorem 6.1, 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* $t \geq 0$, *while* $\bar{C} = \bar{Y} = \bar{A} = \Lambda_f$.

REMARK 6.1. (*stabilization achieved by many fluid models*) It is evident that the conclusion of Corollary 6.1 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 6.1, we can have all customers wait exactly η if we provide no service until time η.

COROLLARY 6.2. (*stabilizing the waiting time at any positive value*) *In addition to the conditions of Theorem 6.1, 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*

$$\bar{L}(t) = \bar{X}(t) \equiv \bar{X}_\eta(t) = \Lambda_f(t) - \Lambda_f(t-\eta) = \int_{t-\eta}^{t} \lambda_f(s)\, ds > 0 \tag{6.20}$$

and

$$\bar{W}(t) = \eta \quad \textit{for all} \quad t \geq \eta. \tag{6.21}$$

COROLLARY 6.3. (*sinusoidal with damped time lag*) *In addition to the conditions of Theorem 6.1, suppose that*

$$s(t) \equiv \beta \sin(\gamma t), \quad t \geq 0, \tag{6.22}$$

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

$$S(t) = (\beta/\gamma)(1 - \cos(\gamma t)), \quad t \geq \eta, \tag{6.23}$$

so that

$$\begin{aligned} \bar{L}(t) &= (\beta/\gamma)([\xi \cos(\gamma(t-\eta)) - \cos(\gamma t)] \\ &\quad + \sup_{0\leq s\leq c} \{\cos(\gamma(t-s)) - \xi \cos(\gamma(t-\eta-s))\} \\ &= (\beta/\gamma)([\xi \cos(\gamma(t-\eta)) - \cos(\gamma t)] \\ &\quad + \sup_{0\leq s\leq c} \{\cos(\gamma s) - \xi \cos(\gamma(s-\eta))\}), \quad t \geq c + \eta. \end{aligned} \tag{6.24}$$

For the special case $\xi = 1$, $\bar{W}(t) = \eta$. If in addition, and $\eta < \pi/\gamma$, the supremum in (6.24) is attained at $s^* = (\pi/2\gamma) - \eta/2)$, so that

$$\bar{L}(t) = (\frac{\beta}{\gamma})([\cos(\gamma(t-\eta)) - \cos(\gamma t)] + [\cos((\pi/2) - (\gamma\eta/2)) - \cos((\pi/2) + (\gamma\eta/2))]) \tag{6.25}$$

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

$$\bar{L}(t)/\eta \to 1 + \beta \sin(\gamma t) = 1 + s(t). \tag{6.26}$$

REMARK 6.2. (*the impact of high or low frequency*) Corollary 6.3 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 (4.1) rather than the rate function $\lambda(t)$. From (6.23) and (6.24), 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 (3.2). We present the proof in §7 after discussing consequences here.

THEOREM 6.2. (*HT FCLT*) *In addition to the definitions and assumptions in §6 above, including the scaled arrival-rate function in (3.2), assume that the periodic function $s(t)$ in (1.3) is continuous and*

$$(1-\rho)\eta_\rho \to \hat{\eta} \quad \textit{and} \quad \frac{\xi_\rho - 1}{1-\rho} \to \hat{\xi} \quad \textit{as} \quad \rho \uparrow 1, \tag{6.27}$$

where $0 \le \hat{\eta} < \infty$ *and* $0 \le \hat{\xi} < \infty$. *Then there is a limit for the scaled cumulative service-rate functions* $\hat{M}_\rho$ *in* (6.6) *and* (6.14)*; i.e.,*

$$\begin{aligned} \hat{M}_\rho(t) &\equiv (1-\rho)[M_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}(t+S(t))] \\ &\to \hat{M}(t) \equiv -s(t)\hat{\eta} + S(t)\hat{\xi} \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad \rho \uparrow 1 \end{aligned} \tag{6.28}$$

for $s(t)$ *in* (1.3) *and* $S(t)$ *in* (6.4). *If, in addition, the system starts empty at time* 0, *then*

$$(\hat{A}_\rho, \hat{Y}_\rho, \hat{X}_\rho, \hat{L}_\rho, \hat{W}_\rho, \hat{C}_\rho) \Rightarrow (\hat{A}, \hat{Y}, \hat{X}, \hat{L}, \hat{W}, \hat{C}) \quad \textit{in} \quad \mathcal{D}^5 \quad \textit{as} \quad \rho \uparrow 1 \tag{6.29}$$

for $(\hat{A}_\rho, \hat{Y}_\rho, \hat{X}_\rho, \hat{L}_\rho, \hat{W}_\rho, \hat{C}_\rho)$ *defined in* (6.15), *where*

$$\begin{aligned} \hat{A}(t) &\equiv (c_a B_a - e) \circ \Lambda_f(t), \quad \hat{Y}(t) \equiv (c_x B - e) \circ \Lambda_f(t), \quad \hat{C}(t) \equiv \hat{Y}(t) - \hat{L}(t), \\ \hat{X}(t) &\equiv \hat{Y}(t) - \hat{M}(t) = \hat{Y}(t) + s(t)\hat{\eta} - S(t)\hat{\xi} \\ &= (c_x B \circ \Lambda_f)(t) - \Lambda_f(t) + s(t)\hat{\eta} - S(t)\hat{\xi}, \\ \hat{L}(t) &\equiv \psi(\hat{X})(t) \quad \textit{and} \quad \hat{W}(t) \equiv \hat{L}(t)/\mu_f(t), \quad t \ge 0. \end{aligned} \tag{6.30}$$

with $c_x \equiv \sqrt{c_a^2 + c_s^2}$, B *a BM,* ψ *the reflection map in* (6.16) *and* $\mu_f(t) \equiv \lambda_f(t) \equiv 1 + s(t)$, $t \ge 0$, *the limiting arrival-rate function, the dervative of* Λ_f *in* (6.12).

We now draw attention to some important consequences. First, Theorem 6.2 establishes a HT time-varying (TV) Little's law (LL), paralleling the many-server heavy-traffic (MSHT) TV LL in [36]. This is a time-varying version of the familiar state-space collapse, which goes back to the early HT papers, e.g., [38]. We remark that the relation is different from the time-varying LL discussed in [4, 11] and [43].

COROLLARY 6.4. (*HT time-varying Little's law*) *Under the conditions of Theorem* 6.2, *the limit processes are related by*

$$\hat{L}(t) = \lambda_f(t)\hat{W}(t), \quad t \ge 0, \quad w.p.1. \tag{6.31}$$

We now consider an alternative deterministic limit to the HT FWLLN in Theorem 6.1. Now we assume that the FCLT holds with the variability parameter set equal to 0. For this purpose, we assume that s is differentiable and let $\dot{s}$ be its derivative.

COROLLARY 6.5. (*the case of no variability*) *If* $c_x = 0$ *and* s *differentiable in addition to the conditions of Theorem* 6.2, *then*

$$\hat{X}(t) = -t + s(t)\hat{\eta} - S(t)(\hat{\xi} + 1), \quad t \ge 0, \tag{6.32}$$

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

$$\frac{d\hat{X}(t)}{dt} = -1 + \hat{\eta}\dot{s}(t) - (\hat{\xi} + 1)s(t) \leq 0, \quad t \geq 0. \tag{6.33}$$

In the sinusoidal case with $s(t) \equiv \beta \sin \gamma t$ *in* (1.5),

$$\frac{d\hat{X}(t)}{dt} = -1 + \hat{\eta}\beta\gamma \cos \gamma t - (\hat{\xi} + 1)\beta \sin \gamma t, \quad t \geq 0. \tag{6.34}$$

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

$$\frac{d\hat{X}(t)}{dt} \to -1 - (\hat{\xi} + 1)\beta \sin \gamma t, \quad t \geq 0, \tag{6.35}$$

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

For the nondegenerate sinusoidal arrival rate function, the derivative in (6.33) of Corollary 6.5 implies it is not always possible to stabilize the limiting time-varying diffusion process $\hat{W}$ with $\hat{\xi} > 0$ in Theorem 6.2. 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 (1.9) or (1.10). Our proof will exploit uniform integrability (UI); see p. 31 of [5].

COROLLARY 6.6. (*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* (1.9) *or* (1.10), *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 diffusioon 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 (6.27) 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 [5]. 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 §3 of [27]. These bounds extend directly to the $G_t/G_t/1$ model by virtue of Corollary 4.1. 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 $W(t) \le L(t)/\mu_L$, where $\mu_L > 0$ is a lower bound on the service rate, which follows from (1.4) and (1.6). 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 [5] and §X.2 and X.7 of [2].

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, while 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 6.6 to support our numerical calculations by observing that $(\eta_\rho^*, \xi_\rho^*)$ when scaled as in (6.27) 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, while the numerical evidence is good, it is not exceptionally good, especially for ξ_ρ^* as can be seen from Table 1 in §9.1 below.

7. Proof of Theorem 6.2. To establish (6.28), apply (6.6) and (6.14) to obtain

$$\begin{aligned}
\hat{M}_\rho(t) &\equiv (1-\rho)[M_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}(t + S(t))] \\
&= (1-\rho)^{-1}[\xi_\rho S(t - (1-\rho)^2\eta_\rho) - S(t)] \\
&= (1-\rho)^{-1}[\xi_\rho S(t - (1-\rho)^2\eta_\rho) - \xi_\rho S(t)] + (1-\rho)^{-1}[\xi_\rho S(t) - S(t)] \\
(7.1) \quad &\to -\hat{\eta} s(t) + \hat{\xi} S(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \uparrow 1,
\end{aligned}$$

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

We next establish (6.29). First, the limit for $\hat{A}_\rho$ is given in Lemma 5.1 of [41], 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 (6.3), (6.11) and (6.14), we see that

$$\begin{aligned}
\bar{\Lambda}_\rho(t) &= \rho\Lambda_f(t) \to \Lambda_f(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \to 1 \\
(7.2) \qquad \hat{\Lambda}_\rho(t) &= (1-\rho)^{-1}\rho\Lambda_f(t) - (1-\rho)^{-1}\Lambda_f(t) = -\Lambda_f(t)
\end{aligned}$$

for all ρ, where $\Lambda_f(t)$ is defined in (6.12). Then the limit for $\hat{A}_\rho$ follows from the standard argument for random sums. The key is to observe that

$$\hat{A}_\rho = \hat{N}_\rho \circ \bar{\Lambda}_\rho + \hat{\Lambda}_\rho, \tag{7.3}$$

where $\hat{N}_\rho$ is defined to be $\hat{N}_n$ in (6.1) for $n = (1-\rho)^{-2}$. So we can start with the joint convergence

$$\left(\hat{N}_\rho, \bar{\Lambda}_\rho, \hat{\Lambda}_\rho\right) \Rightarrow (c_a B_a, \Lambda_f, -\Lambda_f) \quad \text{in} \quad \mathcal{D}^3 \quad \text{as} \quad \rho \to 1, \tag{7.4}$$

We then apply convergence preservation with the map $g(x,y,z) = x \circ y + z$ (composition plus addition) as in §13.3 of [39] 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 $\bar{A}_\rho \equiv (1-\rho)^2 A((1-\rho)^{-2}t)$,

$$\begin{aligned} \bar{A}_\rho &= (1-\rho)^2 N(\Lambda_\rho((1-\rho)^{-2}t)) = (1-\rho)^2 N((1-\rho)^{-2}\rho\Lambda_f(t)) \\ &= \bar{N}_\rho(\rho\Lambda_f(t)) \Rightarrow \Lambda_f \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \to 1. \end{aligned} \tag{7.5}$$

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

$$\left(\hat{S}_\rho, \bar{A}_\rho, \hat{A}_\rho\right) \Rightarrow (c_s B_s, \Lambda_f, c_a B_a \circ \Lambda_f - \Lambda_f) \quad \text{in} \quad \mathcal{D}^3 \quad \text{as} \quad \rho \to 1, \tag{7.6}$$

The joint convergence holds by virtue of Theorems 11.4.4 and 11.4.5 of [39]. We then apply convergence preservation with the map $g(x,y,z) = x \circ y + z$ (composition plus addition) as in §13.3 of [39] to get

$$\begin{aligned} \hat{Y}_\rho &= \hat{S}_\rho \circ \bar{A}_\rho + \hat{A}_\rho \Rightarrow c_s B_s \circ \Lambda_f + c_a B_a \circ \Lambda_f - \Lambda_f \\ &\stackrel{\mathrm{d}}{=} c_x B \circ \Lambda_f - \Lambda_f \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \to 1. \end{aligned} \tag{7.7}$$

Then the limits for $\hat{X}_\rho$ and $\hat{L}_\rho$ follow from the continuous mapping theorem with the standard reflection map reasoning, e.g., as in Chapter 9 of [39], 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 (6.6) and the first-passage-time representation of the waiting time in (6.8). The structure and result are similar to the Puhalskii [31] theorem and related results in §13.7 of [39], but they evidently do not apply directly.

We will apply Taylor's theorem to a perturbation of S in (6.4). The essential idea is that

$$(1-\rho)^{-1}[S(t + (1-\rho)u) - S(t)] \to s(t)u \quad \text{as} \quad \rho \to 1 \tag{7.8}$$

uniformly in t and u over bounded intervals. Just as in (7.1), we use the assumed continuity of s to have S be continuously differentiable, so that the derivative of $S(t)$ holds uniformly in t and u over bounded intervals.

For the specific application, let

$$\tilde{S}_\rho(t,u) \equiv (1-\rho)^{-1}\xi_\rho[S(t+(1-\rho)u-(1-\rho)^2\eta_\rho)-S(t-(1-\rho)^2\eta_\rho)] \tag{7.9}$$

and

$$\zeta_\rho(t,u) \equiv \tilde{S}_\rho(t,u) - s(t)u. \tag{7.10}$$

By combining (7.8) and the two limits in condition (6.27), 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 (6.15) and (6.8),

$$\begin{aligned}
\hat{W}_\rho(t) &\equiv (1-\rho)W_\rho((1-\rho)^{-2}t) \\
&= (1-\rho)\inf\{u \ge 0 : \tilde{M}_\rho(t,u) - \tilde{M}_\rho(t,0)) \ge L_\rho((1-\rho)^{-2}t)\} \\
&= \inf\{u \ge 0 : \tilde{M}_\rho(t,(1-\rho)^{-1}u) - \tilde{M}_\rho(t,0) \ge L_\rho((1-\rho)^{-2}t)\} \\
&= \inf\{u \ge 0 : (1-\rho)[\tilde{M}_\rho(t,(1-\rho)^{-1}u) - \tilde{M}_\rho(t,0)] \ge \hat{L}_\rho(t)\} \\
&= \inf\{u \ge 0 : u + \tilde{S}_\rho(t,u) \ge \hat{L}_\rho(t)\} \\
&= \inf\{u \ge 0 : u + s(t)u + \zeta_\rho(t,u) \ge \hat{L}_\rho(t)\} \\
&= \inf\{u \ge 0 : u\lambda_f(t) + \zeta_\rho(t,u) \ge \hat{L}_\rho(t)\}, \quad t \ge 0, \qquad (7.11)
\end{aligned}$$

where $\lambda_f(t) = t + s(t)$ by (6.12) and we apply Taylor's theorem with (7.9) and (7.10) 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 [39], to replace the stochastic convergence $\hat{L}_\rho \Rightarrow \hat{L}$ in $\mathcal{D}$ by associated convergence w.p.1. Then we see from line 6 of (7.11) that in the infinimum 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, while $\hat{L}(t)$ is bounded above, for t in a bounded interval. Thus, we can write

$$\frac{\hat{L}_\rho(t) - K\gamma_\rho^\uparrow(t)}{\lambda_f(t)} \le \hat{W}_\rho(t) \le \frac{\hat{L}_\rho(t) + K\gamma_\rho^\uparrow(t)}{\lambda_f(t)} \tag{7.12}$$

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

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

for an appropriate $\bar{u}$. Given that $\hat{L}_\rho \Rightarrow \hat{L}$ and $\zeta_\rho^\uparrow \to 0$ in $\mathcal{D}$, 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 $\mathcal{D}$, which completes the proof. ∎

8. A HT FCLT with Periodicity in the Weaker Diffusion Scaling. In this section we establish a HT FCLT with periodicity holding in the weaker diffusion scale instead of in the fluid scale, as was done in §6. The scaling here follows [40] and [27] instead of [41]. In this scaling the HT limits 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 previous sections.

8.1. *An Alternative Family of Models.* We start with the same basic rate-1 processes in §6.1. We then create a model for each ρ, $0 < \rho < 1$, now using (3.3) instead of (3.2). That yields the family of cumulative arrival rate functions

$$\Lambda_\rho(t) \equiv \rho(t + (1-\rho)^{-1} S((1-\rho)^2 t)), \quad t \geq 0, \tag{8.1}$$

for S in (6.4). Differentiating in (8.1) yields the arrival-rate function in (3.3). Just as before, the time scaling in (3.3) and (8.1) implies that the period in model ρ with arrival-rate function $\lambda_\rho(t)$ in (3.3) is $c_\rho = c(1-\rho)^{-2}$, where c is the period of s in (1.2)-(1.4). Thus the period c_ρ in model ρ is growing with ρ.

8.2. *An Associated Family of Service-Rate Controls.* Just as in §6.2.1, we define associated service-rate controls. Closely paralleling (3.3) and (8.1), we define associated scaled time-varying service-rate functions using the control parameters η_ρ and ξ_ρ, i.e., for all $t \geq 0$,

$$\begin{aligned} \mu_\rho(t) &\equiv 1 + (1-\rho)\xi_\rho s(t - \eta_\rho) \quad \text{and} \\ M_\rho(t) &\equiv \int_0^t \mu_\rho(s)\, ds = t + (1-\rho)^{-1}\xi_\rho S((1-\rho)^2(t-\eta_\rho)). \end{aligned} \tag{8.2}$$

Just as in (8.1), differentiation of $M_\rho(t)$ in (8.2) shows that it is consistent with $\mu_\rho(t)$. As a consequence of (8.2), the average service rate is $\bar{\mu}_\rho = 1$, $0 < \rho < 1$.

8.3. *The Scaled Queueing Processes.* We use the same processes introduced in §4, but new scaling. Let the scaled arrival-rate and service-rate

functions be defined for $t \geq 0$ by

$$\begin{aligned} \hat{\Lambda}_\rho(t) &\equiv (1-\rho)[\Lambda_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}t] \\ &= \rho S(t) - t \\ \hat{M}_\rho(t) &\equiv (1-\rho)[M_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}t] \\ &= \xi_\rho S(t - (1-\rho)^2\eta_\rho). \end{aligned} \tag{8.3}$$

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 (8.3), we get

LEMMA 8.1. (*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

$$\begin{aligned} \hat{A}_\rho(t) &\equiv (1-\rho)[A_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}t], \\ \hat{Y}_\rho(t) &\equiv (1-\rho)[Y_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}t], \\ \hat{X}_\rho(t) &\equiv (1-\rho)X_\rho((1-\rho)^{-2}t), \\ \hat{L}_\rho(t) &\equiv (1-\rho)L_\rho((1-\rho)^{-2}t), \quad \hat{C}_\rho \equiv \hat{Y}_\rho - \hat{L}_\rho, \\ \hat{W}_\rho(t) &\equiv (1-\rho)W_\rho((1-\rho)^{-2}t), \quad t \geq 0. \end{aligned} \tag{8.4}$$

Note that the translation terms in $\hat{\Lambda}_\rho$ and $\hat{M}_\rho$ in (8.3) are different from the translation terms in (6.10), while the translation terms in $\hat{A}_\rho$ and $\hat{Y}_\rho$ in (8.4) are different from the translation terms in (6.15). Thus, the statement of the heavy-traffic limit below is different (and weaker).

8.4. *The HT FCLT with Periodicity in Diffusion Scale.* Just as in §6, the following heavy-traffic FCLT states that $\hat{A}_\rho$ and $\hat{Y}_\rho$ converge to periodic Brownian motions (PBM's). However, unlike §6, $\hat{X}_\rho$ converges to an *ordinary* Brownian motion (BM), $\hat{L}_\rho$ and $\hat{W}_\rho$ converge to the *same ordinary* reflected Brownian motion (RBM), while $\hat{C}_\rho$ has a complicated limit. We thus show that $\hat{L}_\rho$ and $\hat{W}_\rho$ are asymptotically stable and Markov. Note that the scaling condition on (η_ρ, ξ_ρ) here are implied by condition (6.27) in Theorem 6.2, but as noted above the conclusion is different and weaker, because of the different translation terms.

THEOREM 8.1. (*heavy-traffic limit extending Theorem 3.2 of* [40] *and Theorem 2 of* [27]) *If, in addition to the definitions and assumptions in* (8.1)-(8.4) *above,* $(1-\rho)^2\eta_\rho \to 0$ *and* $\xi_\rho \to 1$ *as* $\rho \to 1$ *and the system starts empty at time* 0, *then*

$$(\hat{\Lambda}_\rho, \hat{M}_\rho, \hat{A}_\rho, \hat{Y}_\rho, \hat{X}_\rho, \hat{L}_\rho, \hat{W}_\rho, \hat{C}_\rho) \Rightarrow (\hat{\Lambda}, \hat{M}, \hat{A}, \hat{Y}, \hat{X}, \hat{L}, \hat{W}, \hat{C}) \tag{8.5}$$

in $\mathcal{D}^8$ *as* $\rho \to 1$ *for* $(\hat{\Lambda}_\rho, \hat{M}_\rho)$ *defined in* (8.3) *and* $(\hat{A}_\rho, \hat{Y}_\rho, \hat{X}_\rho, \hat{L}_\rho, \hat{W}_\rho, \hat{C}_\rho)$ *defined in* (8.4)*, where*

$$\begin{aligned} \hat{\Lambda} &\equiv S - e, \quad \hat{A} \equiv c_a B_a + S - e, \quad \hat{M} \equiv S, \\ \hat{Y} &\equiv \hat{A} + c_s B_s, \quad \hat{X} \equiv \hat{Y} - S \stackrel{\mathrm{d}}{=} c_x B - e, \\ \hat{L} &\equiv \psi(\hat{X}), \quad \hat{W} \equiv \psi(\hat{X}) \quad \textit{and} \quad \hat{C} \equiv \hat{Y} - \hat{L}, \end{aligned} \tag{8.6}$$

with B_a *and* B_s *being independent BM's,* S *in* (6.4)*,* c_a *and* c_s *being the variability parameters in* (6.2)*,* $c_x \equiv \sqrt{c_a^2 + c_s^2}$ *and* B *is a BM.*

Proof. We will be brief because most of the argument is essentially the same as in [40] and [27]. First, the limit for $\hat{A}_\rho$ is given in Theorem 3.2 of [40]. Then the limit for $\hat{Y}_\rho$ follows from Theorem 9.3.4 of [39], as noted in the proof of Theorem 2 in [27]. (See $C(t)$ in (9.2.4) and C_n in (9.3.4) and Theorem 9.3.4 of [39].) Then the limits for $\hat{X}_\rho$ and $\hat{L}_\rho$ follow from the standard reflection mapping argument as in 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 (6.8). In particular, paralleling (7.11), letting $\tilde{M}_\rho(t,u) \equiv M_\rho((1-\rho)^{-2}t + u)$, we have

$$\begin{aligned} \hat{W}_\rho(t) &\equiv (1-\rho) W_\rho((1-\rho)^{-2}t) \\ &= (1-\rho) \inf\{u \ge 0 : \tilde{M}_\rho(t,u) - \tilde{M}_\rho(t,0) \ge L_\rho((1-\rho)^{-2}t)\} \\ &= \inf\{u \ge 0 : \tilde{M}_\rho(t,(1-\rho)^{-1}u) - \tilde{M}_\rho(t,0) \ge L_\rho((1-\rho)^{-2}t)\} \\ &= \inf\{u \ge 0 : (1-\rho)[\tilde{M}_\rho(t,(1-\rho)^{-1}u) - \tilde{M}_\rho(t,0)] \ge \hat{L}_\rho(t)\} \\ &= \inf\{u \ge 0 : u + \zeta_\rho(t,u) \ge \hat{L}_\rho(t)\}, \end{aligned} \tag{8.7}$$

for $t \ge 0$, where

$$\zeta(t,u) \equiv \xi_\rho\left[S(t + (1-\rho)u - (1-\rho)^2\eta_\rho) - S(t - (1-\rho)^2\eta_\rho)\right], \tag{8.8}$$

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

$$S(t+\epsilon) - S(t) \le s_U \epsilon \quad \text{for all} \quad \epsilon > 0, \tag{8.9}$$

for s_U in (1.4). Also add and subtract $\xi_\rho S(t)$ and treat the two terms separately, i.e.,

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

Hence, we can apply the continuous mapping theorem for the inverse in §13.6 of [39] 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 §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 6.2, specifically by (3.2), (6.3) and (6.6), so that we can see the systematic behavior.

Table 1 shows results for four values of the traffic intensity ρ with $\rho \uparrow 1$ for the sinusoidal model in (1.2)-(1.6) with HT scaling in (3.2) with parameters $(\rho, \beta_\rho, \gamma_\rho) = (\rho, 0.2, 2.5(1-\rho)^2)$. For this case, we found that the solutions to optimization problems (1.9) and (1.10) 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 (6.27). 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 $(1-\rho)\eta_\rho^*/\rho$, while the limit $\hat{\xi}^* \approx 1.8$ is roughly approached by $(\xi_\rho^* - 1)/(1-\rho)$, both of which are consistent with condition (6.27). The results support Theorem 6.2, but unfortunately the rate of convergence in the control parameters is not fast. Evidently the optimal damping control ξ_ρ^* is more problematic.

TABLE 1

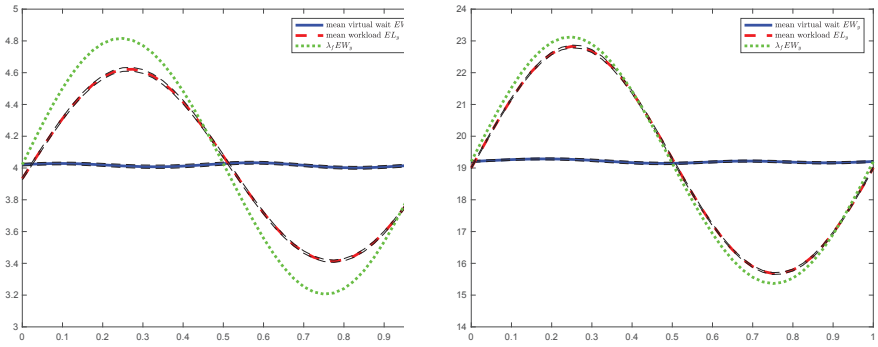
The (identical) solutions to the minimax and minimum-deviation optimization problems in (1.9) and (1.10) for the sinusoidal model in (1.2)-(1.6) with HT scaling in (3.2) with parameters $(\rho, \beta_\rho, \gamma_\rho) = (\rho, 0.2, 2.5(1-\rho)^2)$. The mean waiting times are reported with and without space scaling.

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
$(1-\rho)ave\{E[W_y]\}/\rho$	1.005	1.007	1.011	1.012
relative error	0.8%	1.0%	0.7%	0.9%

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.

FIG 3. *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 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* $\rho = 0.95$.



We also considered alternative values of the relative amplitude β. Table 2 shows the solutions to the minimum-deviation optimization problem in (1.10) for the sinusoidal model in Table 1 except 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 about 10% from about 1% in in Table 1. Unlike in Figure 3, it is evident that the $(\eta_\rho^*, \xi_\rho^*)$ control does not stabilize the expected waiting time perfectly, either for fixed ρ or asymptotically as $\rho \to 1$.

From cases with $0.2 \le \beta \le 0.9$ and $0.8 \le \rho \le 0.975$, 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, while $\hat{\eta}_\rho^*$ ($\hat{\xi}_\rho^*$) 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 about 4%, while the relative error for $\beta = 0.9$ was about 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 [41], where c is the stable value, which has minimum and maximum values $c/(1+\beta)$ and $c/(1-\beta)$, which deviate greatly as $\beta \uparrow 1$. (The constant c is the stable value of the expected queue

TABLE 2

The solutions to the minimum-deviation optimization problem in (1.10) for the sinusoidal model in Table 1 except β has been increased to $\beta = 0.8$ from 0.2. The reported average mean waiting times are reported with and without space scaling.

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
$\max(EW_y) - \min(EW_y)$	0.54	1.32	2.28	4.55
average wait	4.33	10.68	23.97	51.76
$(1-\rho)ave\{E[W_y]\}/\rho$	1.08	1.19	1.14	1.26
relative error	12.5%	12.4%	9.5%	8.8%

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 [41], which stabilizes the entire queue-length distribution, the optimal modified (η, ξ) control does not stabilize the entire waiting time distribution. Figure 4 illustrates by showing plots of the time-varying variance $Var[W_y]$ (left) and delay probability $P(W_y > 0)$ (right) in the setting of Figure 1. We also plot 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). The delay probability tend to reach their upper (lower) bound near the point that the arrival and service rate cross, after the arrival rate has been above (below). The variance tends to be proportional to the sum of the two rates.

We now show two candidate modifications of the control used in Figure 1. First, Figure 5 shows the analog of Figure 1, where we fix $\xi = 1$ and only use the single control parameter η. As we remarked in Remark 2.2 in §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 about 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 about 14%.

Second, Figure 6 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 $\xi_\rho \approx 1 - 1.8(1-\rho)$, based on Table 1. For $\rho = 0.8$, $(\eta_\rho^*, \xi_\rho^*) = (7.25, 0.64)$ and the maximum deviation is 0.6005, yielding about 15% relative error. For $\rho = 0.95$, $(\eta_\rho^*, \xi_\rho^*) = (29.0, 0.91)$ and the maximum deviation is 0.9220 yielding about 5% relative error. Unlike in Figure 5, we see that the direct HT approximation improves as ρ increases, but the direct

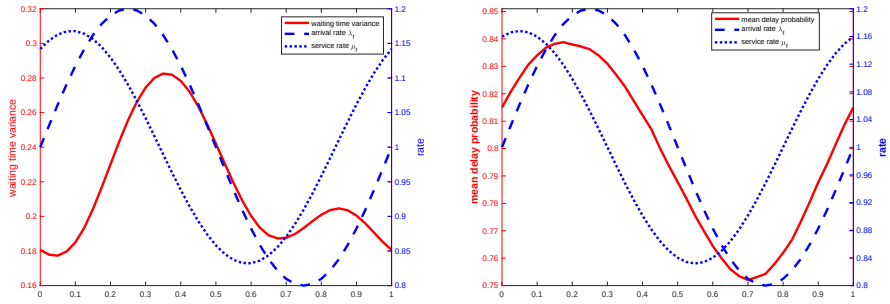


FIG 4. *Estimates of the periodic steady-state variance* $Var[W_y]$ *(left) and probability of delay* $P(W_y > 0)$ *(right), both shown in red solid line, for the sinusoidal example in Figure 1 with parameter triple* $(\rho, \beta, \gamma) = (0.8, 0.2, 0.1)$ *and* $\rho = 0.8$*, using the optimal control* $(\eta_\rho^*, \xi_\rho^*) = (5.84, 0.84)$*. 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).*

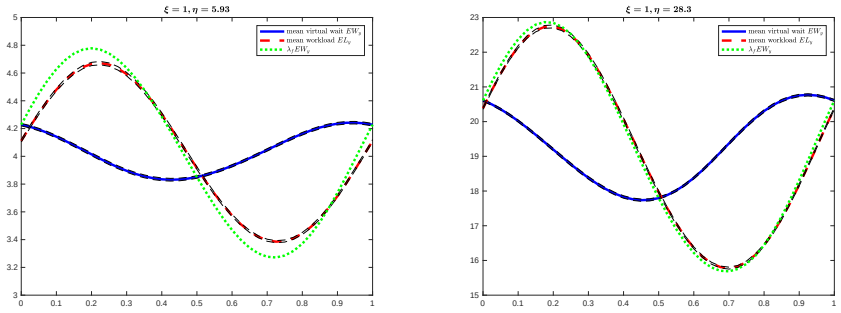


FIG 5. *Estimates of the expected waiting time* $E[W_y]$ *for the one-parameter* η *control with* $\xi \equiv 1$*, for the sinusoidal example in Figure 1 with parameter triple* $(\rho, \beta, \gamma) = (0.8, 0.2, 0.1)$ *and* $\rho = 0.8$ *(left) and* $\rho = 0.95$ *(right). For* $\rho = 0.8$*,* $\eta_\rho^* = 5.93$ *the maximum deviation is* 0.4109 *and the relative error is about* 10%*; for* $\rho = 0.95$*,* $\eta_\rho^* = 28.3$*, the maximum deviation is* 3.034 *and the relative error is about* 14%*.*

two-parameter optimal control is better.

Finally, Figure 7 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 (6.34) for no workload or waiting when $c_x = 0$ in Corollary 6.5.

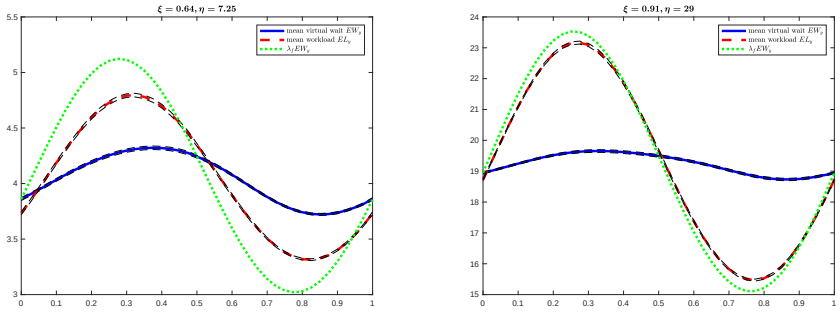


FIG 6. *Estimates of the expected waiting time $E[W_y]$ (solid red line) with the heavy-traffic control exploiting the estimated limiting controls $\hat{\eta}^* \approx 1.45$ and $\hat{\xi}^* = 1.8$, so that $\eta^*_\rho \approx 1.45/(1-\rho)$ and $\xi^*_\rho \approx 1 - 1.8(1-\rho)$. The plots are for 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.*

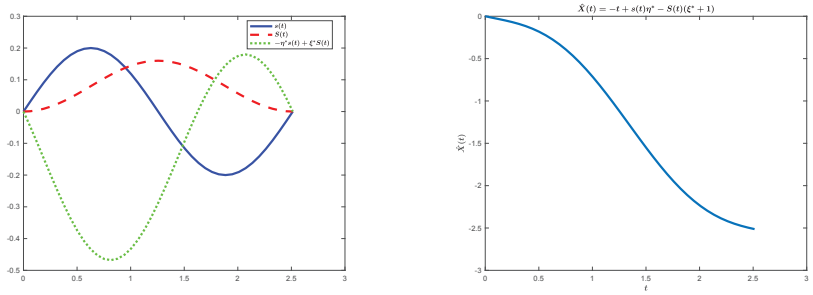


FIG 7. *Deterministic functions associated with the diffusion limit for the case $\beta = 0.2$, $\gamma = 2.5$, $\hat{\eta} = 1.45$ and $\hat{\xi} = -1.8$.. On the left appears $\hat{M}(t) = -\hat{\eta}s(t) + \hat{\xi}S(t) = -1.5\beta\sin(\gamma t) - 1.5(\beta/\gamma)(1-\cos(\gamma t))$ together with $s(t) = \beta\sin(\gamma t)$ 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 $c_x = 0$, showing that condition condition (6.34) holds.*

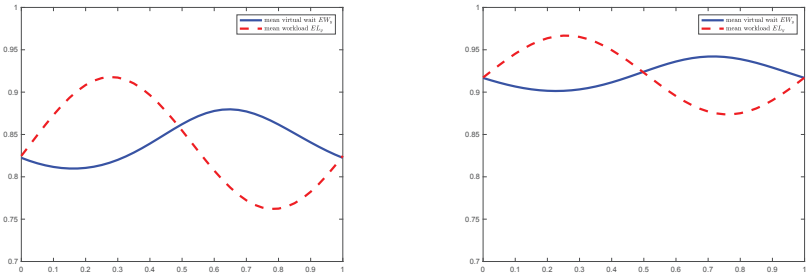
9.2. *Simulation Examples for the Alternative Scaling in §8.* We now consider four simulation examples in the alternative heavy-traffic scaling in §8. This is the same heavy-traffic scaling as in [27]. We consider the base case of $\beta = 1$, $\gamma = 2.5$, and use

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

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 8.1.)

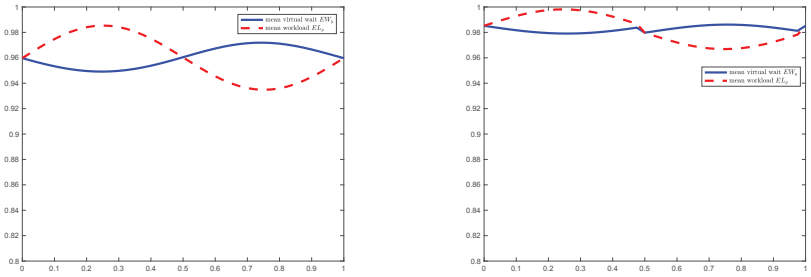
Figures 8-9 show the expected periodic steady-state waiting time (the solid blue line) and the expected steady-state workload (the dashed red line). Figures 8 and 9 show that the stabilization is not achieved well for the lower traffic intensities, but the stabilization improves for both curves as ρ increases. Both processes get quite well stabilized at $\rho = 0.98$, consistent with Theorem 8.1.

FIG 8. *the expected periodic steady-state virtual waiting time (the blue line) and the expected steady-state workload (the red line) 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)*



10. Conclusions. In this paper we extended the rare-event simulation algorithm for the periodic $GI_t/GI/1$ model in [27] 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 a modification of the rate-matching service-rate control in (1.1) to include a time lag η and a damping factor ξ as in (1.6). We developed and applied a simulation search algorithm to find optimal pairs of control parameters (η, ξ) for the control

FIG 9. *the expected periodic steady-state virtual waiting time (the blue line) and the expected steady-state workload (the red line) 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)*



problems in (1.9) and (1.10). Thus, we obtained a practical solution to the open problem in [41] of developing an effective way to stabilize the expected waiting time.

We also established supporting heavy-traffic limits for the general periodic $G_t/G_t/1$ model and showed that the control problems in (1.9) and (1.10) converge to associated diffusion control parameters with appropriate scaling. In Theorem 6.2, the arrival-rate function was scaled as in (3.2), making the periodicity occur in fluid scale, as in [41]. In §8 we also obtained heavy-traffic limits with alternative scaling as in (3.2), making the periodicity occur in diffusion scale, as in [40] and [27]. With scaling in diffusion scale, the workload and waiting time are both asymptotically stabilized as $\rho \uparrow 1$, but that is not consistent with practical examples, as in Figure 1.

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 [41], the new modified rate-matching control does not stabilize the expected waiting time perfectly. Moreover, Figure 4 shows that it does not stabilize the full waiting time distribution. There remain many opportunities for future research, including the open problems mentioned in §1.5. It also remains to directly solve the diffusion control problems with objectives (1.9) and (1.10) resulting from Theorem 6.2

It is interesting to consider the performance impact of time-varying arrivals. In §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 2.1 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.

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

Acknowledgement. Research support was received from NSF (CMMI 1634133).

REFERENCES

[1] K. M. Adusumilli and J. J. Hasenbein. Dynamic admission and service rate control of a queue. *Queueing Systems*, 66:131–154, 2010.

[2] S. Asmussen. *Applied Probability and Queues*. Springer, New York, second edition, 2003.

[3] S. Asmussen and P. W. Glynn. *Stochastic Simulation*. Springer, New York, second edition, 2007.

[4] D. Bertsimas and G. Mourtzinou. Transient laws of nonstationary queueing systems and their applications. *Queueing Systems*, 25:315–359, 1997.

[5] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1999.

[6] H. Chen and A. Mandelbaum. Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.*, 16(2):408–446, 1991.

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

[8] M. Defraeye and I. van Nieuwenhuyse. 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, 2013.

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

[10] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.*, 54(2):324–338, 2008.

[11] B. H. Fralix and G. Riano. A new look at transient versions of Little's law. *J. Appl. Prob.*, 47:459–473, 2010.

[12] M. C. Fu. *Handbook of Simulation Optimization*, volume 216 of *International Series in Operations Research and Management Science*. Springer, New York, 2015.

[13] J. M. George and J. M. Harrison. Dynamic control of a queue with adjustable service rate. *Operations Research*, 49(5):720–731, 2001.

[14] B. He, Y. Liu, and W. Whitt. Staffing a service system with non-Poisson nonstationary arrivals. *Probability in the Engineering and Informational Sciences*, 30(1):593–621, 2016.

[15] D. P. Heyman. Optimal operating policies for $M/G/1$ queueing systems. *Operations Research*, 16(2):362–382, 1968.

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

[17] N. Jian and S. G. Henderson. An introduction to simulation optimization. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, editors, *Proceedings of the 2015 Winter Simulation Conference, Huntington Beach, CA, December 6-9, 2015*, pages 1780–1794. ACM, 2015.

[18] L. Kleinrock. *Communication Nets: Stochastic Message Flow and Delay*. Dover, New York, 1964.

[19] S. Kwon and N. Gautam. Guaranteeing performance based on time stability for energy-efficient data centers. *IIE Transactions*, 48(9):812–825, 2016.

[20] Y. Liu. Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research*, 2018. published online February 8, 2018.

[21] Y. Liu and W. Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.*, 59:835–846, 2011.

[22] Y. Liu and W. Whitt. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems*, 71:405–444, 2012.

[23] Y. Liu and W. Whitt. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.*, 60(6):1551–1564, 2012.

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

[25] N. Ma and W. Whitt. Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters*, 102:202–207, 2015.

[26] N. Ma and W. Whitt. Using simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rate. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, editors, *Proceedings of the 2015 Winter Simulation Conference, Huntington Beach, CA, December 6-9, 2015*, pages 1–12. ACM, 2015.

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

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

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

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

[31] A Puhalskii. On the invariance principle for the first passage time. *Mathematics of Operations Research*, 19(4):946–954, 1994.

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

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

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

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

[36] X. Sun and W. Whitt. Delay-based service differentiation with many servers and time-varying arrival rates. Stochastic Systems, forthcoming; available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2018.

[37] S. Suriadi, M. T. Wynn, J. Xu, W. M. P. van der Aalst, and A. H. M. ter Hofstede. Discovering work prioritization patterns from event logs. *Decision Support Systems*, 100(August):77–92, 2017.

[38] W. Whitt. Weak convergence theorems for priority queues: Preemptive-resume discipline. *Journal of Applied Probability*, 8(1):74–94, 1971.

[39] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.

[40] W. Whitt. Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters*, 42:458–461, 2014.

[41] W. Whitt. Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems*, 81:341–378, 2015.

[42] W. Whitt. Time-varying queues. Queueing Models and Service Management, forthcoming; available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2018.

[43] W. Whitt and X. Zhang. Periodic Little's Law. *Operations Research*, forthcoming; available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2018.

[44] M. Yadin and P. Naor. Queueing systems with a removable service station. *Operational Research Quarterly*, 14:393–405, 1963.

Department of IEOR,
S. W. Mudd Building,
500 West 120th Street,
New York, NY 10027-6699
E-mail: nm2692@columbia.edu
ww2040@columbia.edu