

Stabilizing Performance in a Single-Server Queue with Time-Varying Arrival Rate

Ward Whitt

Department of Industrial Engineering and Operations Research,
Columbia University, New York, NY, 10027 {ww2040@columbia.edu}

January 27, 2015

Abstract

We consider a class of general $G_t/G_t/1$ single-server queues, including the $M_t/M_t/1$ queue, with unlimited waiting space, service in order of arrival and a time-varying arrival rate, where the service rate at each time is subject to control. We study the rate-matching control, where the service rate is made proportional to the arrival rate. We show that the model with the rate-matching control can be regarded as a deterministic time transformation of a stationary $G/G/1$ model, so that the queue length distribution is stabilized as time evolves. However, the time-varying virtual waiting time is not stabilized. We show that the time-varying expected virtual waiting time with the rate-matching service-rate control becomes inversely proportional to the arrival rate in a heavy-traffic limit. We also show that no control that stabilizes the queue length asymptotically in heavy-traffic can also stabilize the virtual waiting time. Then we consider two square-root service-rate controls and show that one of these stabilizes the waiting time when the arrival rate changes very slowly relative to the average service time, so that a pointwise stationary approximation is appropriate.

Keywords: stabilizing performance, queues with time-varying arrival rates, nonstationary queues, heavy-traffic limits, single-server queues with time-varying arrival rates, service-rate controls, heavy-traffic scaling, pointwise stationary approximations.

1 Introduction

In this paper we study controls to stabilize the performance of a single-server queueing system with a time-varying arrival rate function. We assume that there is unlimited waiting space and that service is provided in order of arrival.

It has been shown how server staffing (choosing a time-varying number of servers) can be used to achieve this goal in multi-server systems with fixed service-time distribution for each customer when the required number of servers is not too small and there is flexibility in its assignment; see [18], then [10, 15] and more recently [6, 16, 22, 39]. In contrast, here we consider a single-server queue, in which there is no flexibility in the number of servers. To achieve stabilization, we assume that the service rate of the single server is flexible and subject to control. In doing so, we assume that the service rate can be specified separately from the random service requirements as a deterministic function. For example, a customer service requirement might correspond to the size of a message to be transmitted in a communication network, while the service rate might be the processing rate of the message. Thus a service requirement S with a constant service rate μ would lead to a service time of S/μ . However, here the service rate can change while the customer is in service. With this approach, all randomness appears through the service requirements. We also assume that the service requirements are stochastically independent of the arrival process.

Even though the stabilization problem for multi-server queues has been studied for twenty years, the present paper evidently is the first formulation of an analogous problem for nonstationary single-server queues. Moreover, the previous staffing algorithms for multi-server queues evidently do not apply directly. In fact, even constructing the service times is not entirely straightforward, so that simulation experiments are somewhat challenging. We show how to construct the service times in §3.1 and develop alternative service time models that can serve as useful approximations and may be of direct interest in §3.2.

Having a single-server queue where the the service rate is a continuous deterministic function subject to control is an idealization of what occurs in many service operations, such as hospital surgery rooms and airport security inspection lines. In the short run, there may be a fixed number of service facilities, sometimes only one, but the processing rate can be increased by assigning additional staff or changing procedures, which may occur at some cost. Assigning more doctors and nurses can increase the rate of completed operations; assigning more inspection agents at the airport security line or relaxing the inspection requirements can increase the rate at which passengers are

processed through inspection. In these applications, the possible service rate functions may not actually be continuous, or even fully under control. Nevertheless, to better understand the possible benefits of these practical service-rate controls, it is helpful to understand what controls are desirable in the ideal situation when any deterministic continuous service-rate control function is possible.

There is an important precedent in earlier work. By having the service rate function as the control, our problem is similar to the capacity allocation problem for open Jackson queueing networks in steady state, considered by Kleinrock [19], extended for approximations of generalized Jackson networks in [31] and reviewed in §5.7 of [20], in §7 of [3] and elsewhere. Now, instead of allocating capacity (which corresponds to service rate) to several queues in different locations, we allocate capacity to a single queue at different times. We hope to consider how to allocate capacity over both time and space in future work.

2 Overview

2.1 The Rate-Matching Control

In this paper we primarily consider the simple *rate-matching control*, which chooses the service rate to be proportional to the arrival rate; i.e., for a given target traffic intensity ρ , we let the service rate be

$$\mu(t) \equiv \frac{\lambda(t)}{\rho}, \quad t \geq 0. \quad (2.1)$$

In considering this rate-matching control, we assume that the arrival rate function is deterministic and known. In future work, we intend to consider the cases in which the rate-matching control is used with (i) an estimate of a deterministic arrival rate function obtained from data and (ii) a stochastic arrival rate function, as in a Cox process; see [16].

By definition, the rate-matching control stabilizes the time-varying instantaneous traffic intensity $\rho(t) \equiv \lambda(t)/\mu(t)$ for all $t \geq 0$. We will show that the rate-matching control in (2.1) stabilizes the mean queue length as $t \rightarrow \infty$ (to allow the effect of the initial condition to dissipate), but not the mean waiting time. This is illustrated by Figure 1, which shows simulation estimates of the time-varying mean number in the system, $E[Q(t)]$ (left), and the mean waiting time, $E[W(t)]$ (right), for the $M_t/M_t/1$ model with mean-1 service requirements and sinusoidal arrival rate function $\lambda(t) \equiv 1 + \beta \sin \gamma t$ for $\beta = 0.2$ and $\gamma = 0.001$ (long cycles); see §8. We let the target traffic intensity be $\rho = 0.8$. As for the stationary $M/M/1$ queue, the mean steady-state number in the system should be $\rho/(1 - \rho) = 4.0$. The plot on the left in Figure 1 shows that the mean queue

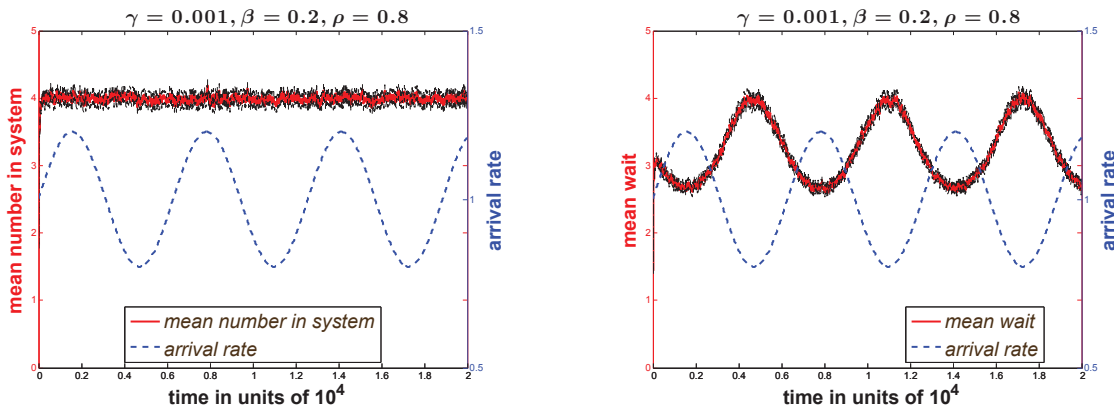


Figure 1: Simulation estimates of the time-varying mean number in the system, $E[Q(t)]$ (left), and the mean waiting time, $E[W(t)]$ (right), for the $M_t/M_t/1$ model with sinusoidal arrival rate function $\lambda(t) \equiv 1 + \beta \sin \gamma t$ with $\beta = 0.2$ and $\gamma = 0.001$ with the rate-matching control in (2.1).

length is indeed stabilized at 4.0, but the plot on the right shows that the time-varying mean virtual waiting time $E[W(t)]$ is periodic. The mean waiting time is stabilized to some extent by the rate-matching control in (2.1), but not nearly as well as the number in system. The 95% confidence intervals are also displayed along with the estimates in both plots. The dashed blue line shows the arrival rate with values on the right vertical axis.

The key idea supporting the positive result for $E[Q(t)]$ in Figure 1 is that, under the rate-matching service-rate control in (2.1), the queue-length process can be represented as a deterministic time transformation of a corresponding queue length process in a stationary model, as shown in §4. That construction directly implies the stabilization. However, the story for the waiting times is more complicated. Theorem 5.2 shows that, with the rate-matching service-rate control, the time-varying expected virtual waiting time is asymptotically inversely proportional to the time-varying arrival rate in a heavy-traffic limit. Paralleling Theorem 2 and Corollary 1 of [22] for staffing multi-server queues, Theorem 5.3 shows that no control that asymptotically stabilizes the queue length in this heavy-traffic regime can simultaneously stabilize the virtual waiting time. Nevertheless, for models with a periodic arrival rate function, Theorem 6.2 establishes that the waiting times of successive arrivals does have a proper limit. That occurs because the arrival time eventually occurs randomly over the periodic cycle.

2.2 Non-Markov Nonstationary $G_t/G_t/1$ Models

This paper makes significant contributions for the nonstationary Markov $M_t/M_t/1$ model, but the results are established in greater generality. In particular, they are established for a large class of

$G_t/G_t/1$ models, defined in §3. In particular, we assume that the arrival counting process A can be represented as the composition of a general counting process N_a and a deterministic cumulative arrival rate function Λ by the composition

$$A(t) \equiv N_a(\Lambda(t)) = N_a\left(\int_0^t \lambda(s) ds\right), \quad t \geq 0, \quad (2.2)$$

where N_a is a rate-1 stochastic counting process with unit jumps (so that arrivals occur one at a time) satisfying a functional strong law of large numbers (FSLLN) and a functional central limit theorem (FCLT). For the rate-matching control, the service times can be defined analogously. This construction is discussed in §7 of [25] and [13]. It has been used in stabilizing performance of many-server queues with non-Poisson arrivals in [16, 23].

It is important to recognize that this is a special construction, treating only a subclass of all non-Poisson nonstationary arrival processes, but we think that it is a useful way to draw conclusions about such complicated models. First, the construction is without loss of generality for the $M_t/M_t/1$ model. Second, the construction applies directly to the service times when the service times are i.i.d. random variables with a general distribution, so that the $M_t/G_t/1$ model is natural and also without loss of generality as well. This is important because in service systems it has been found that service times are often non-exponential [4].

To understand the restriction on the G_t arrival process more generally, it is helpful to consider the special case in which the process N_a is a rate-1 Markov modulated Poisson process (MMPP) with a finite-state continuous-time Markov environment process, yielding an arrival rate of γ_k in state k [12, 16]. The composition construction in (2.2) implies that the arrival rate of A at time t when the environment process is in state k is simply the product $\lambda(t)\gamma_k$. More generally, a nonstationary MMPP with a finite-state Markov environment process could have arrival rate $\gamma_k(t)$, which is a general function of the two variables k and t . Clearly, the construction here yields only a subset of all possible cases, but nevertheless we believe that it usefully goes beyond the $M_t/M_t/1$ model. It allows some characterization of the stochastic variability of the arrival and service processes instead of none at all. It remains to determine how useful is the “one-dimensional” characterization of non-Poisson stochastic variability in the non- M_t G_t arrival process. Since non- M_t properties often arise through structural features such as having arrivals be departures or overflows from another queue, there is good reason to expect that the present approach will prove useful. Moreover, the heavy-traffic limit identifies parsimonious characterizations of the stochastic variability in the arrival and service processes.

We have verified that the rate-matching control actually works in this more general G_t setting by conducting simulation experiments for $G_t/G_t/1$ models with mean-1 service requirements where both the service requirements and the nonstationary arrival process are constructed from renewal counting processes, where the times between renewals are i.i.d. random variables having non-exponential distributions, including Erlang E_2 and hyperexponential H_2 distributions. The plots look just like the plots for the $M_t/M_t/1$ models displayed in this introduction.

2.3 Two Square-Root Service-Rate Controls

As an analog of Kleinrock's [19] square-root capacity allocation formula (appearing in (7.7) here), we also consider the (first) *square-root service-rate control*

$$\mu(t) \equiv \lambda(t) + \xi \sqrt{\lambda(t)}, \quad t \geq 0. \quad (2.3)$$

where ξ is a positive parameter, and a *second square-root service-rate control*

$$\mu(t) \equiv \lambda(t) + \frac{\lambda(t)}{2} \left(\sqrt{1 + \frac{\zeta}{\lambda(t)}} - 1 \right), \quad t \geq 0, \quad (2.4)$$

where ζ is a positive parameter.

The first square-root service-rate control in (2.3) is interesting because it is a natural analog of the square-root staffing formula used for many-server queues. Nevertheless, Figure 2 shows that (2.3) is *not* effective in stabilizing either $E[Q(t)]$ or $E[W(t)]$ for the same arrival process as in Figure 1, again with 95% confidence intervals, even though there are very long cycles. We use $\xi = 2$, which would make $\rho = 0.8333$ and the steady state mean waiting time $EW = \rho/\mu(1 - \rho) = 3.333$ for the constant arrival rate occurring when $\beta = 0$. Even though this service-rate control does stabilize both $E[Q(t)]$ or $E[W(t)]$ to some extent, Figure 2 shows that they are both periodic. That is expected because the control (2.3) is reasonable, but based on a different objective function (following [3, 19, 20, 31]). See §8 for more discussion.

On the other hand, when the arrival rate changes very slowly, so that the pointwise stationary approximation (PSA) is effective [14, 35], so that the queue lengths at the (far apart) times where the arrival rate differs tend to be stochastically independent, the control in (2.4) does stabilize the mean waiting time, as it is designed to do. Figure 3 shows the corresponding performance estimates for the second square-root service-rate control in (2.4) with parameter $\zeta = 1.0$, again with the same arrival process as before.

Here is how the rest of this paper is organized: We start in §3 by defining the specific $G_t/G_t/1$ model, showing how to construct the service times, and showing that the queue length process

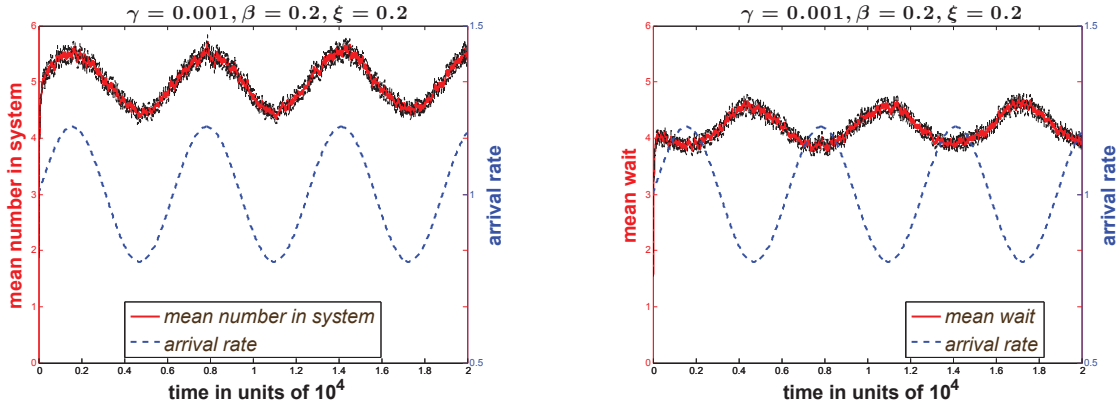


Figure 2: Simulation estimates of the time-varying mean number in the system, $E[Q(t)]$ (left) and the mean waiting time, $E[W(t)]$ (right) for the $M_t/M_t/1$ model, having the same sinusoidal arrival rate with long cycles ($\gamma = 0.001$), with the first square-root service-rate control in (2.3).

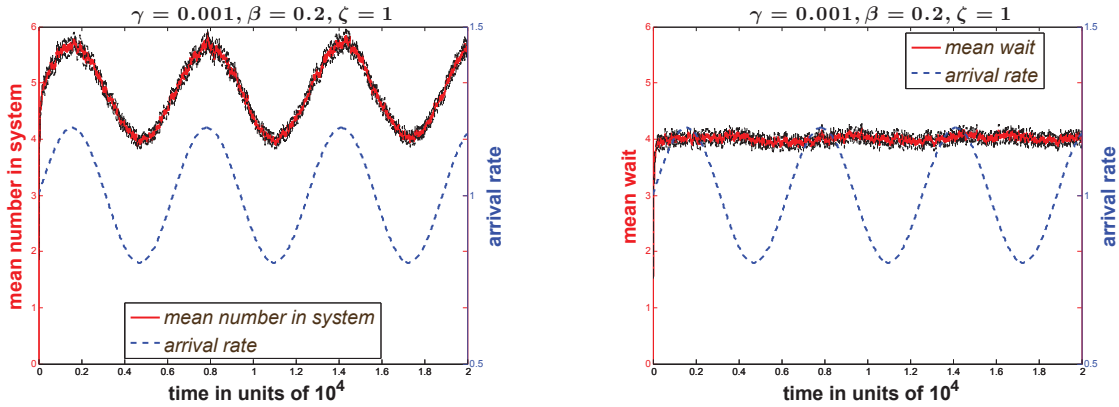


Figure 3: Simulation estimates of the time-varying mean number in the system, $E[Q(t)]$ (left) and the mean waiting time, $E[W(t)]$ (right) for the $M_t/M_t/1$ model, having the same sinusoidal arrival rate with long cycles ($\gamma = 0.001$), with the second square-root service-rate control in (2.4).

in this model is a deterministic time transformation of the queue length process in an associated stationary $G/G/1$ model. In §4 we establish positive stabilization properties of the rate-matching control. In §5 we give an explicit representation of the time-varying waiting time in terms of the waiting time in the corresponding stationary $G/G/1$ model and establish the heavy-traffic limit theorems. In §6 we consider the special case of a periodic arrival rate function in more detail. After Theorem 6.1 formalizes the notion of a periodic steady state, Theorem 6.2 establishes a periodic heavy-traffic limit for the waiting times of successive arrivals. As in [21] for multi-server queues, this illustrates a nearly periodic situation in which the limit depends on the order of the two iterated limits as $n \rightarrow \infty$ and $t \rightarrow \infty$. In §7 we consider the square-root service-rate controls in (2.3) and (2.4). We show that they are optimal with appropriate objective functions when a pointwise

stationary approximation is appropriate, as in [35]. We discuss the simulation experiments further in §8. Alternative controls to systematically stabilize the time-varying waiting time are introduced and evaluated with simulation experiments in [24].

3 The Model

As indicated in §2.2, we exploit a special composition construction of the arrival and service processes in order to obtain a general $G_t/G_t/1$ model. In particular, we assume that the arrival process is defined by the composition in (2.2) where N_a is a rate-1 stochastic counting process with unit jumps (so that arrivals occur one at a time) satisfying a functional strong law of large numbers (FSLLN) and a functional central limit theorem (FCLT), i.e.,

$$\bar{N}_{a,n} \rightarrow e \quad \text{and} \quad \hat{N}_{a,n} \Rightarrow c_a B_a \quad \text{in } \mathcal{D} \quad \text{as } n \rightarrow \infty, \quad (3.1)$$

with

$$\bar{N}_{a,n}(t) \equiv n^{-1}N_a(nt) \quad \text{and} \quad \hat{N}_{a,n}(t) \equiv n^{-1/2}[N_a(nt) - nt], \quad t \geq 0, \quad (3.2)$$

e the identity function, $e(t) = t$, $t \geq 0$, B_a a standard (drift 0, variance 1) Brownian motion (BM), \Rightarrow denoting convergence in distribution and \mathcal{D} denoting the function space of right-continuous real-valued functions on the interval $[0, \infty)$ with left limits, as in [36], while Λ is a deterministic cumulative arrival rate function, satisfying

$$\Lambda(t) \equiv \int_0^t \lambda(s) ds, \quad t \geq 0, \quad (3.3)$$

with λ being the arrival rate function, which is assumed to be strictly positive and continuous with finite long-run average

$$\bar{\lambda} \equiv \lim_{t \rightarrow \infty} t^{-1}\Lambda(t). \quad (3.4)$$

Without loss of generality, we assume that $\bar{\lambda} = 1$. In addition, we assume that $\lambda(t)$ is uniformly bounded above and below.

The composition construction in (2.2) is a standard way to construct a nonhomogeneous Poisson process (NHPP, M_t), which is an important special case; then N_a above is a rate-1 Poisson process. This composition model has all unpredictable stochastic variability in the arrival process associated with the processes N_a and its FCLT behavior characterized by the single variability parameter c_a^2 , while all the predictable deterministic variability associated with the deterministic arrival rate function $\lambda(t)$ and the associated cumulative rate function Λ . If the process N_a is a renewal counting

process, then c_a^2 is the scv of a time between renewals (which requires a finite second moment), but N_a can be more general; e.g., see §4.4 of [36].

We specify the random service requirements of successive customers separately from the service rate, which is deterministic and subject to control. For the first six sections of the paper, we assume that, for each ρ , $0 < \rho < 1$, μ_ρ is defined by the rate-matching policy, as specified in (2.1). We assume that the successive service requirements are generated (in a way to be explained in the next paragraph) from a rate-1 stochastic counting process N_s with unit jumps, independent of N_a , satisfying an FLLN and an FCLT, i.e.,

$$\bar{N}_{s,n} \rightarrow e \quad \text{and} \quad \hat{N}_{s,n} \Rightarrow c_s B_s \quad \text{in } \mathcal{D} \quad \text{as } n \rightarrow \infty, \quad (3.5)$$

where

$$\bar{N}_{s,n}(t) \equiv n^{-1} N_s(nt) \quad \text{and} \quad \hat{N}_{s,n}(t) \equiv n^{-1/2} [N_s(nt) - nt], \quad t \geq 0, \quad (3.6)$$

with B_s being a standard BM, necessarily independent of B_a . The case of i.i.d. service requirements with a general distribution having a finite second moment is a natural special case.

As usual, the queue length process can be defined as

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0, \quad (3.7)$$

where $D(t)$ is the total number of departures in the interval $[0, t]$. We assume that the system starts empty at time 0. We understand $D(t)$ to satisfy

$$D(t) \equiv N_s \left(\int_0^t \mu(s) 1_{\{Q(s) > 0\}} ds \right) = N_s \left(\int_0^t (\lambda(s)/\rho) 1_{\{Q(s) > 0\}} ds \right), \quad t \geq 0, \quad (3.8)$$

where 1_A is the indicator function, equal to 1 on A and 0 otherwise. Note that Q and D in (3.7) and (3.8) are defined in terms of each other. However, as in Lemma 2.1 of [28], there is a unique solution, as can be proved by induction on the successive events in the processes A and S , which necessarily occur one at a time because of the unit-jump assumptions for N_a and N_s .

3.1 Direct Construction of the Service Times

The present paper differs from the majority of the literature on single-server queues by *not* introducing the sequence of successive *service times*, which we denote as $\{V_k : k \geq 1\}$, as a model primitive. Instead, here we have the sequence of successive *service requirements* $\{S_k : k \geq 1\}$ specified as the times between events in the counting process N_s , while the service rate $\mu(t)$ is time-dependent and subject to control. For the rate-matching control in (2.1) and the square-root

service-rate controls in (2.3) and (2.4), the service rate becomes a fully specified function that is continuous and positive.

We now show how to construct the sequence of successive service times, assuming that the sequence $\{S_k : k \geq 1\}$ of service requirements is given and the service rate $\mu(t)$ is a fully specified continuous function, uniformly bounded above and below, just like λ . That condition on μ follows from the assumption about λ with (2.1) or (2.3). This construction is important for computer simulations.

We assume that the system starts empty. Let A_k, B_k, D_k , be the times at which customer k arrives, begins service and departs, respectively. Let V_k and W_k be the durations (length of the time intervals) that customer k spends in service and spends waiting in queue before starting service, respectively. Since the system starts empty, $D_0 = 0, B_1 = A_1 \geq 0$. As usual, we have the basic recursions

$$B_k = D_{k-1} \vee A_k, \quad D_k = B_k + V_k \quad \text{and} \quad W_k = B_k - A_k, \quad k \geq 1, \quad (3.9)$$

where $a \vee b \equiv \max\{a, b\}$. The complication is that V_k is not specified exogenously.

To construct V_k , we need to properly relate rates to requirements and time. When we do so, we see that V_k is specified implicitly via the equation

$$S_k = \int_{B_k}^{B_k + V_k} \mu(s) ds, \quad k \geq 1. \quad (3.10)$$

If we let

$$M(t) \equiv \int_0^t \mu(s) ds, \quad t \geq 0, \quad (3.11)$$

then we see that $M(t)$ is the total amount of service completed in the interval $[0, t]$, assuming that the server is busy continuously. Since M is strictly increasing and continuous, it has an inverse M^{-1} . With that inverse, we obtain an explicit formula for the service times, in particular,

$$V_k = M^{-1}(S_k + M(B_k)) - B_k, \quad k \geq 1. \quad (3.12)$$

For example, if $\mu(t) = \mu, t \geq 0$, then $M(t) = \mu t$ and $M^{-1}(t) = t/\mu, t \geq 0$. Hence, $M(B_k) = \mu B_k$, $M^{-1}(S_k + M(B_k)) = (B_k + S_k/\mu)$ and $V_k = S_k/\mu$ for all k , as it should.

3.2 Alternative Service-Time Models

Since the service-time formula (3.12) is somewhat complicated, it is helpful to have a useful practical approximation. In fact, there are alternative service time models that might be of interest as

models in their own right. The first alternative model assigns each customer a constant service rate determined when the customer begins service. Then the service time of customer k is simply

$$V_k = \frac{S_k}{\mu(B_k)}. \quad (3.13)$$

The second alternative assigns each customer a constant service rate determined when the customer arrives. Then the service time of customer k is simply

$$V_k = \frac{S_k}{\mu(A_k)}. \quad (3.14)$$

The model with service times in (3.13) is a natural approximation for our main model of time-dependent service rate $\mu(t)$ applying at time t if the customer is in service at that time. Indeed, the model with service times in (3.13) can be achieved by employing local linear Taylor approximations

$$M(t+s) \approx M(t) + \mu(t)s \quad \text{and} \quad M^{-1}(t+s) = M^{-1}(t) + \frac{s}{\mu(M^{-1}(t))}, \quad (3.15)$$

assuming that s is relatively small. We obtain the second from the inverse function theorem from calculus. In particular, with an abuse of notation, let $\mu^{-1}(t)$ be the derivative of $M^{-1}(t)$. By the inverse function theorem, $\mu^{-1}(t) = 1/\mu(M^{-1}(t))$. Thus the corresponding Taylor approximation for $M^{-1}(t+s)$ is given in (3.15). When we apply the Taylor approximation in (3.12), regarding S_k as a small perturbation about $M(B_k)$, we get

$$V_k = M^{-1}(M(B_k)) + \frac{S_k}{\mu(M^{-1}(M(B_k)))} - B_k \approx \frac{S_k}{\mu(B_k)}. \quad (3.16)$$

In the actual model, the service rate may keep changing, but this seems to be a reasonable approximation. Under heavy-traffic conditions, the difference will be negligible.

3.3 Time Transformation of Stationary Model

We now show that, with the rate-matching service-rate control in (2.1), we can circumvent the construction of the service times in (3.12) in order to deduce some important structure. (With this approach, we do *not* use the approximation in (3.16).) An important consequence of the composition construction in (2.2)-(3.8) above is that the queue length process $Q(t)$ depending on the arrival rate function $\lambda(t)$ can be related to the associated queue-length process $Q_1(t)$ with constant arrival rate 1 and constant service rate $1/\rho$ by a simple time transformation. In particular, let the arrival process of Q_1 be $A_1 \equiv N_a$ and let the queue length and departure process be defined as

$$Q_1(t) \equiv A_1(t) - D_1(t), \quad t \geq 0, \quad (3.17)$$

where $A_1 \equiv N_a$ and $D_1(t)$ is the total number of departures in the interval $[0, t]$. We understand $D_1(t)$ to satisfy

$$D_1(t) \equiv N_s \left(\int_0^t \mu_1(s) 1_{\{Q_1(s) > 0\}} ds \right) = N_s \left(\int_0^t \rho^{-1} 1_{\{Q_1(s) > 0\}} ds \right), \quad t \geq 0. \quad (3.18)$$

Let Λ^{-1} be the inverse of the continuous strictly increasing function Λ , so that $\Lambda(\Lambda^{-1}(t)) = \Lambda^{-1}(\Lambda(t)) = t$, $t \geq 0$.

Theorem 3.1 (*time transformation of a stationary model*) For (A, D, Q) with the rate-matching service-rate control and the stationary single-server model (A_1, D_1, Q_1) defined above,

$$(A(t), D(t), Q(t)) = (A_1(\Lambda(t)), D_1(\Lambda(t)), Q_1(\Lambda(t))), \quad t \geq 0. \quad (3.19)$$

Proof. The relation between A and A_1 holds by definition. We will establish the relation between the pair (Q, D) and the pair (Q_1, D_1) together, paralleling their definitions via (3.7) and (3.8) ((3.17) and (3.18)). We will exploit the change of variables $s = \Lambda^{-1}(u)$ or $u = \Lambda(s)$ and the associated differential relation $du = \lambda(s)ds$. Starting with (3.8), we express D as

$$\begin{aligned} D(t) &= N_s \left(\int_0^t \rho^{-1} \lambda(s) 1_{\{Q(s) > 0\}} ds \right), \quad t \geq 0, \\ &= N_s \left(\int_0^{\Lambda(t)} \rho^{-1} 1_{\{Q(\Lambda^{-1}(u)) > 0\}} du \right), \quad t \geq 0, \\ &= N_s \left(\int_0^{\Lambda(t)} \rho^{-1} 1_{\{Q_1(u) > 0\}} du \right) = D_1(\Lambda(t)), \quad t \geq 0, \end{aligned} \quad (3.20)$$

as claimed, where we have used $Q = Q_1 \circ \Lambda$ in the third step. As in the definitions (3.7) and (3.8), we can use induction on the transition epochs of the processes N_a and N_s to verify that there is a unique solution for (D, Q) and for (D_1, Q_1) that must be related by (3.20). ■

4 Basic Stabilization of the Rate-Matching Service-Rate Control

We first show that the rate-matching service-rate control always stabilizes (as time evolves) the proportion of arrivals that are delayed, which we define (as a function of the traffic intensity ρ) by

$$\bar{d}_\rho(t) \equiv \frac{\int_0^t \lambda(s) 1_{\{Q(s) > 0\}} ds}{\Lambda(t)}. \quad (4.1)$$

In (4.1) we weight the server busy event at s , which is $1_{\{Q(s) > 0\}}$, by the relative likelihood of an arrival at time s during the interval $[0, t]$, which is $\lambda(s)/\Lambda(t)$. In the case of constant arrival rate, $\bar{d}_\rho(t)$ reduces to the utilization over $[0, t]$, defined by

$$\bar{U}_{1,\rho}(t) \equiv t^{-1} \int_0^t 1_{\{W_1(s) > 0\}} ds \equiv t^{-1} \int_0^t 1_{\{Q_1(s) > 0\}} ds. \quad (4.2)$$

Theorem 4.1 (*stabilizing the average delay probability*) *Under the conditions above,*

$$\bar{U}_{1\rho}(t) \rightarrow \rho \quad \text{and} \quad \bar{d}_\rho(t) \rightarrow \rho \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty \quad (4.3)$$

for $\bar{U}_{1\rho}(t)$ in (4.2) and $\bar{d}_\rho(t)$ in (4.1).

To prove Theorem 4.1, we use FSLLN's and SLLN's for the arrival and service processes. Let $S_1(t) \equiv N_s(t/\rho)$ be the counting process associated with the successive partial sums of the service times in the system with constant rates, paralleling $A_1 = N_a$ for the arrival process. By direct assumption, A_1 satisfies a FSLLN and thus also an ordinary SLLN. We now show that is also true of A and S_1 . Let $\bar{A}_n(t) \equiv n^{-1}A(nt)$ and $\bar{S}_{1,n}(t) \equiv n^{-1}S_1(nt)$, $t \geq 0$.

Lemma 4.1 (*preliminary FSLLN's*) *Under the conditions above, the processes A and S_1 satisfy the FSLLN's*

$$\bar{A}_n \rightarrow e \quad \text{and} \quad \bar{S}_{1,n} \rightarrow \rho^{-1}e \quad \text{in } \mathcal{D} \quad \text{as } n \rightarrow \infty \quad \text{w.p.1} \quad (4.4)$$

and the associated SLLN's

$$t^{-1}A(t) \rightarrow 1 \quad \text{and} \quad t^{-1}S_1(t) \rightarrow \rho^{-1} \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty \quad \text{w.p.1} \quad (4.5)$$

Proof. First the FSLLN's and SLLN's are actually equivalent in this setting of a single process; see Ch. 1 of the internet supplement to [36]. Thus, the limit in (3.4) is equivalent to the stronger limit $\bar{\Lambda}_n \rightarrow e$ in \mathcal{D} as $n \rightarrow \infty$, where $\bar{\Lambda}_n(t) = \Lambda(nt)/n$, $t \geq 0$. We can obtain the FSLLN's by the continuity of the composition map, which is defined by $(x \circ y)(t) \equiv x(y(t))$: $\bar{A}_n = \bar{N}_{a,n} \circ \bar{\Lambda}_n \rightarrow e \circ e = e$, i.e., $\bar{A}_n(t) = \bar{N}_{a,n}(\bar{\Lambda}_n(t))$, $t \geq 0$; see §13.2 of [36]. Similarly, $\bar{S}_{1,n} = \bar{N}_{s,n} \circ \rho^{-1}e \rightarrow e \circ \rho^{-1}e = \rho^{-1}e$. Then the ordinary SLLN's are obtained by applying the projection map from \mathcal{D} to \mathbb{R} taking x to $x(t)$ at $t = 1$, which is also continuous at all t that are continuity points of x . ■

Proof of Theorem 4.1. We first deduce the conclusion for the system with queue-length process Q_1 , having constant arrival and service rates. For that system, we can apply the sample-path version of Little's law to the service facility, using the notation $L^* = \lambda^*W^*$; see [30, 34]. The limit in (4.3) for $\bar{U}_{1,\rho}$ to be established is then L^* . The LLN for N_a with limit 1 is λ^* . Since the service rate is constant, $N_s(\rho^{-1}t)$ counts the number of partial sums of the service times that are less than or equal to t . Since the SLLN of S_1 established in Lemma 4.1 is equivalent to the SLLN of the service times, we see that the average of the service times approaches ρ , which is W^* . Since the limits for λ^* and W^* hold, the limit for $\bar{U}_{1,\rho}$ holds as well with $L^* = \lambda^*W^* = 1 \times \rho = \rho$.

For the second limit, perform a change of variables as in (3.20) to obtain

$$\bar{d}_\rho(t) = \frac{\int_0^{\Lambda(t)} 1_{\{Q_1(u) > 0\}} du}{\Lambda(t)}. \quad (4.6)$$

Since $\Lambda(t) \rightarrow \infty$ as $t \rightarrow \infty$, we can apply the first result. ■

Finally, we conclude this section by observing that there is a proper limiting steady-state distribution for $Q(t)$ as $t \rightarrow \infty$ whenever there is a proper steady-state distribution for $Q_1(t)$ as $t \rightarrow \infty$.

Theorem 4.2 (*stabilizing the queue-length distribution and the steady-state delay probability*) *Let $Q_1(t)$ be the queue length process when $\lambda(t) = 1$, $t \geq 0$. If $Q_1(t) \Rightarrow Q_1(\infty)$ as $t \rightarrow \infty$, where $P(Q_1(\infty) < \infty) = 1$, then also*

$$Q(t) \Rightarrow Q_1(\infty) \quad \text{in } \mathbb{R} \quad \text{as } t \rightarrow \infty, \quad (4.7)$$

and

$$P(W(t) > 0) = P(Q(t) \geq 1) \rightarrow \rho \quad \text{as } t \rightarrow \infty. \quad (4.8)$$

Proof. Let Λ^{-1} be the inverse of the continuous strictly increasing function Λ . It follows that $\{Q(\Lambda^{-1}(t)) : t \geq 0\}$ is distributed as $\{Q_1(t) : t \geq 0\}$. Since Λ^{-1} is deterministic with $\Lambda^{-1}(t) \rightarrow \infty$ as $t \rightarrow \infty$, $Q(\Lambda^{-1}(t)) \Rightarrow Q_1(\infty)$ as $t \rightarrow \infty$, which directly implies that $Q(t) \Rightarrow Q_1(\infty)$ as $t \rightarrow \infty$ as well, which in turn immediately implies the associated limit. Given Little's law for the system with Q_1 , we have $P(Q_1(\infty) > 0) = \rho$ in (4.8). ■

5 The Virtual Waiting Time with the Rate-Matching Control

Often we are interested in the distribution or the moments of the virtual waiting time $W(t)$. Unlike Theorem 3.1, we do *not* have $W(t) \stackrel{d}{=} W_1(\Lambda(t))$, where $\stackrel{d}{=}$ means equal in distribution. Unfortunately, the virtual waiting time is more complicated. We can write

$$P(W(t) > w) = \sum_{k=1}^{\infty} P(W(t) > w | Q(t) = k) P(Q(t) = k), \quad (5.1)$$

where

$$P(W(t) > w | Q(t) = k) = P(\inf \{u \geq 0 : D(t+u) - D(t) \geq k\} > w). \quad (5.2)$$

Theorem 4.2 shows that $Q(t)$ approaches a steady-state limit as $t \rightarrow \infty$ in considerable generality, but, because of the first passage time structure in (5.2), the conditional probability in (5.2) is in general time varying.

In this section we first develop an explicit expression for the virtual waiting time $W(t)$ with the rate-matching service-rate control in (2.1). Afterwards, we establish a heavy-traffic limit theorem.

5.1 An Explicit Expression

To develop an explicit expression for the virtual waiting time for the rate-matching service-rate control, we exploit the connection to the stationary $G/G/1$ model. For the base $G/G/1$ model we assume that the interarrival times $U_{1,k}$ of the counting process $A_1 \equiv N_a$ and the service times $V_{1,k}$ of the counting process $S_1 \equiv N_s \circ \rho^{-1}e$ have been specified.

Given the interarrival times and service times, we use the classical Lindley recursion as on p. 207 of [36] that maps the interarrival times $U_{1,k}$ and the service times $V_{1,k}$ into the waiting times $W_{1,k}$ in the stationary $G/G/1$ model. (Specifically, $W_{1,k}$ is the waiting time before starting service of the k^{th} arrival, which occurs at time A_k , assuming that the system starts empty at time 0.) The formulas for the arrival times $A_{1,k}$ and departure times $D_{1,k}$ as well as the waiting times $W_{1,k}$ are through the equations

$$\begin{aligned} A_{1,k} &\equiv U_{1,1} + \cdots + U_{1,k}, \\ W_{1,k+1} &\equiv [W_{1,k} + V_{1,k} - U_{1,k+1}]^+, \\ D_{1,k} &\equiv A_{1,k} + W_{1,k} + V_{1,k}, \quad k \geq 1, \end{aligned} \tag{5.3}$$

where $[x]^+ \equiv \max\{0, x\}$ and $W_{1,1} \equiv 0$. The associated arrival counting process $A_1(t)$ and departure counting process $D_1(t)$ are constructed as inverse processes, while the queue length process $Q_1(t)$ is their difference, i.e.,

$$\begin{aligned} A_1(t) &\equiv \max\{k \geq 0 : A_{1,k} \leq t\}, \\ D_1(t) &\equiv \max\{k \geq 0 : D_{1,k} \leq t\}, \\ Q_1(t) &\equiv A_1(t) - D_1(t), \quad t \geq 0. \end{aligned} \tag{5.4}$$

We then can construct the virtual waiting time at time t in terms of the waiting time of the last arrival before time t , $W_{1,A_1(t)}$, by

$$W_1(t) \equiv [W_{1,A_1(t)} + V_{1,A_1(t)} - (t - A_{1,A_1(t)})]^+, \quad t \geq 0. \tag{5.5}$$

A short S program to convert the sequence $\{(U_{1,k}, V_{1,k}, W_{1,k}) : k \geq 1\}$ into the associated sequence $\{(A_{1,k}, D_{1,k}, C_{1,k}, Q_{1,k}) : k \geq 1\}$, where $C_{1,k}$ is the time of the k^{th} change in the queue length process (caused by an arrival or a departure) and $Q_{1,k} = Q_1(C_{1,k})$ is the queue length at

time $C_{1,k}$, is given on p. 210 of [36]. Similarly, the associated virtual waiting time in the $G/G/1$ model at change time $C_{1,k}$ is then $W_1(C_{1,k})$.

We then obtain a relatively simple construction of the associated sequence $\{(A_k, D_k, C_k, Q_k) : k \geq 1\}$ for our $G_t/G_t/1$ model with time-varying arrival rate function λ , in particular,

$$(A_k, D_k, C_k, Q_k) \equiv (\Lambda^{-1}(A_{1,k}), \Lambda^{-1}(D_{1,k}), \Lambda^{-1}(C_{1,k}), Q_{1,k}), \quad k \geq 1, \quad (5.6)$$

where Λ^{-1} is the inverse of Λ , which is well defined because Λ is strictly increasing and continuous.

Then for any $t \geq 0$, we can construct the queue length at time t by setting

$$C(t) \equiv \max \{k \geq 0 : C_k \leq t\} \quad \text{and} \quad Q(t) \equiv Q_{C(t)}, \quad t \geq 0. \quad (5.7)$$

Similarly, for any $t \geq 0$, we can construct the departure counting process at time t by setting

$$D(t) \equiv \max \{k \geq 0 : D_k \leq t\}, \quad t \geq 0. \quad (5.8)$$

Theorem 5.1 (*constructing the virtual waiting time*) *The virtual waiting time $W(t)$ can be represented as*

$$W(t) = \Lambda_t^{-1}(W_1(\Lambda(t))), \quad t \geq 0, \quad (5.9)$$

where $W_1(t)$ is the waiting time at time t in the associated stationary $G/G/1$ model and Λ_t^{-1} is the inverse of

$$\Lambda_t(v) = \Lambda(t+v) - \Lambda(t), \quad v \geq 0 \quad \text{and} \quad t \geq 0, \quad (5.10)$$

which is strictly increasing and continuous. If $W_1(t)$ has its stationary distribution W_1^* , then $W(t) \stackrel{d}{=} \Lambda_t^{-1}(W_1^*)$.

Proof. From (5.1) and (5.2),

$$\begin{aligned} W(t) &\equiv \inf \{u \geq 0 : D(t+u) - D(t) = Q(t)\} \\ &= \inf \{u \geq 0 : D_1(\Lambda(t+u)) - D_1(\Lambda(t)) = Q_1(\Lambda(t))\}, \quad t \geq 0, \end{aligned} \quad (5.11)$$

while

$$W_1(\Lambda(t)) = \inf \{v \geq 0 : D_1(\Lambda(t)+v) - D_1(\Lambda(t)) = Q_1(\Lambda(t))\}. \quad (5.12)$$

Thus we have $\Lambda(t) + W_1(\Lambda(t)) = \Lambda(t + W(t))$ or

$$W_1(\Lambda(t)) = \Lambda(t + W(t)) - \Lambda(t) = \Lambda_t(W(t)), \quad t \geq 0, \quad (5.13)$$

for Λ_t defined in (5.10) above or, equivalently, the claimed formula (5.9). ■

We can use Theorem 5.1 to give an explicit integral formula for the mean $E[W(t)]$ in the $M_t/M_t/1$ model. Hence we can numerically compute the mean in this case.

Corollary 5.1 (*mean wait in the $M_t/M_t/1$ model*) For the $M_t/M_t/1$ model with the rate-matching service-rate control, if t is large so that $W_1(t)$ can be regarded as being in steady state, then

$$E[W(t)] = \rho \int_0^\infty e^{-(1-\rho)\Lambda_t(x)/\rho} dx. \quad (5.14)$$

Proof. First the associated stationary $G/G/1$ model is $M/M/1$ with arrival rate 1 and service rate $1/\rho$, so that $P(W_1(t) > x) = \rho e^{-(1-\rho)x/\rho}$ for large t . Next use the tail integral formula for the mean with (5.9) to write

$$\begin{aligned} E[W(t)] &= \int_0^\infty P(W(t) > x) dx = \int_0^\infty P(\Lambda_t^{-1}(W_1(\Lambda(t))) > x) dx \\ &= \int_0^\infty P(W_1(\Lambda(t)) > \Lambda_t(x)) dx = \int_0^\infty \rho e^{-(1-\rho)\Lambda_t(x)/\rho} dx. \end{aligned} \quad (5.15)$$

As a sanity check, note that if λ is constant, then the model is $M/M/1$ with arrival rate 1 and service rate ρ , so that $E[W(t)] = \rho^2/(1-\rho)$. ■

To illustrate Theorems 5.1 and Corollary 5.1, we consider a sinusoidal arrival rate function.

Example 5.1 (*a sinusoidal example*) Consider the sinusoidal arrival rate function

$$\lambda(t) \equiv 1 + \beta \sin(\gamma t) \quad \text{for } 0 < \beta < 1 \quad \text{and} \quad \gamma > 0 \quad (5.16)$$

with parameters $\beta = 0.2$ and $\gamma = 1.0$. Since Theorem 5.1 and Corollary 5.1 are for the Markovian $M_t/M_t/1$ model, we consider that model. We plot that arrival rate function over one cycle together with the periodic steady-state time-varying mean wait computed numerically using matlab from Corollary 5.1 in the case $\rho = 0.9$ in Figure 4. Figure 4 shows that the waiting time is not stabilized. Consistent with previous results about time-varying queues, e.g., as in [7, 8], we see that the peak and trough of the mean waiting time lag behind the corresponding peak and trough of the arrival rate function. In this example, the peak mean waiting time occurs slightly before the trough of the arrival rate.

5.2 A Heavy-Traffic Limit for the Virtual Waiting Time

We now obtain a heavy-traffic limit for $W(t)$ that provides helpful insight. As usual with heavy-traffic limits of single-server queues, we scale time and space as we allow the traffic intensity to

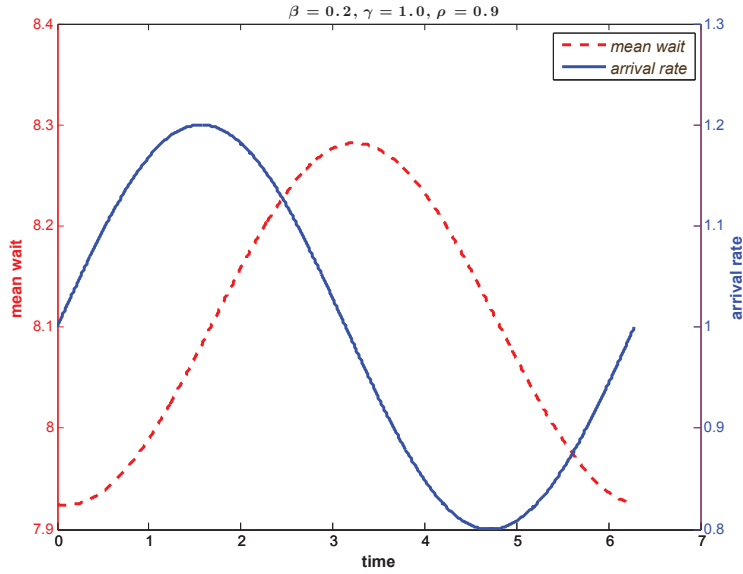


Figure 4: The periodic steady-state time-varying mean wait $E[W(t)]$ (with values shown on the left vertical axis) in the $M_t/M_t/1$ model with the rate-matching service-rate control in (2.1) and the sinusoidal arrival rate $\lambda(t) \equiv 1 + \beta \sin(\gamma t)$ with $\beta = 0.2$ and $\gamma = 1.0$ (with values shown on the right vertical axis) for traffic intensity $\rho = 0.9$ and no scaling over one cycle $[0, 2\pi]$, computed from Corollary 5.1.

increase toward 1; e.g., see Chapters 5 and 9 of [36]. We start by constructing a sequence of the models with constant arrival and service rates, corresponding to the triple (A_1, D_1, Q_1) indexed by n . As usual, we let the traffic intensity in model n be $\rho_n = 1 - (1/\sqrt{n})$, we scale time by $n = (1 - \rho_n)^{-2}$ and we scale space by $n^{-1/2} = (1 - \rho_n)$. We achieve these traffic intensities by scaling the service requirements, i.e., we let $S_{1,n}(t) \equiv N_s(t/\rho_n)$ for ρ_n just specified.

To obtain interesting limits that capture the time-varying arrival rate, we consider a sequence of arrival rate functions $\{\lambda_n : n \geq 1\}$ indexed by n , with each being continuous and strictly positive. Let associated scaled arrival rate functions and cumulative arrival rate functions be defined by

$$\bar{\lambda}_n(t) \equiv \lambda_n(nt) \quad \text{and} \quad \bar{\Lambda}_n(t) \equiv n^{-1}\Lambda_n(nt), \quad t \geq 0 \quad \text{and} \quad n \geq 1, \quad (5.17)$$

so that $\bar{\Lambda}_n(t) = \int_0^t \bar{\lambda}_n(s) ds$. We also introduce a refined scaling involving increments of order \sqrt{n} in Λ_n about time nt . For that purpose, let

$$\tilde{\Lambda}_{n,t}(u) \equiv n^{-1/2}[\Lambda_n(nt + u\sqrt{n}) - \Lambda_n(nt)], \quad t \geq 0 \quad \text{and} \quad n \geq 1. \quad (5.18)$$

We assume that these scaled functions have the limits

$$\bar{\lambda}_n \rightarrow \lambda_f \quad \text{and} \quad \bar{\Lambda}_n \rightarrow \Lambda_f \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \rightarrow \infty \quad (5.19)$$

and

$$\tilde{\Lambda}_{n,t}(u) \rightarrow \lambda_f(t)u \quad \text{as } n \rightarrow \infty \quad (5.20)$$

uniformly in t and u over bounded subintervals of $[0, \infty)$, where λ_f is a continuous and strictly positive. To be consistent with §3, we assume that λ_f has a long-run average $\bar{\lambda}_f = 1$. As a further regularity condition, we assume that $\lambda_n(t)$ is uniformly bounded for all n and t .

We also specify a refined “diffusion scale” scaling with

$$\hat{\Lambda}_n(t) \equiv n^{-1/2}[\Lambda_n(nt) - n\Lambda_f(t)], \quad t \geq 0 \quad \text{and} \quad n \geq 1, \quad (5.21)$$

and assume that

$$\hat{\Lambda}_n \rightarrow \Lambda_d \quad \text{in } \mathcal{D} \quad \text{as } n \rightarrow \infty, \quad (5.22)$$

where Λ_d is a continuous function, although the limit (5.22) will play no role in Theorem 5.2 below.

Extending (2.2) in a natural way, we let the arrival process in model n be

$$A_n(t) \equiv N_a(\Lambda_n(t)) = N_a\left(\int_0^t \lambda_n(s) ds\right), \quad t \geq 0, \quad (5.23)$$

where N_a is the fixed base process introduced before in (2.2). Since N_a is a rate-1 process and $\bar{\lambda}_f = 1$, the long-run arrival rate is 1.

Since $\rho_n \rightarrow 1$ as $n \rightarrow \infty$, the service requirements as specified by $S_{1,n}(t) \equiv N_s(t/\rho_n)$ remain $O(1)$ as $n \rightarrow \infty$. The time scaling in (5.17)-(5.20) makes the arrival rates and service rates of the scaled arrival and service processes become of order $O(n)$ as $n \rightarrow \infty$, meaning that we look over large time intervals. As usual, the heavy-traffic scaling of space and time will make the queue lengths and waiting times be of order $O(\sqrt{n})$. Hence, the service times are asymptotically negligible compared to the waiting times, but both are asymptotically negligible compared to the time scale n .

Even though the arrival rate at time t remains $O(1)$ as $n \rightarrow \infty$, the arrival rate function is affected significantly by the scaling, because it is changing more slowly as n increases. In particular, the arrival rate at time t is $\lambda_n(t) \approx \lambda_f(t/n)$, so it has derivative $\dot{\lambda}_n(t) \approx \dot{\lambda}_f(t/n)/n$. Thus, the arrival rate changes more slowly as n increases. That makes the model tend to be in steady-state at each time t with arrival rate $\lambda_n(t)$, service rate $\lambda_n(t)/\rho_n$ and constant traffic intensity $\rho_n = 1 - (1/\sqrt{n})$. It is significant that the steady-state behavior at time t itself depends on t , because the operative arrival rate itself is a function of time.

The following example may help to understand the scaling in (5.17)-(5.22) and the interpretation above.

Example 5.2 (*scaling in the sinusoidal example*) To illustrate, we return to Example 5.1, but now modified to be in the asymptotic framework just introduced. In particular, we start with the limit arrival rate function $\lambda_f(t)$ defined as in (5.16) (with the subscript notation added now) and proceed backwards to construct the sequence of arrival rate functions with this limit, using the usual scaling. Let $\Lambda_f(t) \equiv \int_0^t \lambda_f(s) ds$ for $t \geq 0$. Let $\Lambda_n(t) \equiv n\Lambda_f(t/n)$, so that

$$\lambda_n(t) \equiv \lambda_f(t/n) = 1 + \beta \sin(\gamma t/n) \quad \text{for } 0 < \beta < 1 \quad \text{and} \quad \gamma > 0 \quad (5.24)$$

and $\dot{\lambda}_n(t) = \dot{\lambda}_f(t/n)/n$. From the perspective of the arrival rate function in model n , we see that the scaling corresponds to slowing time down by a factor of n , making the periodic cycles get longer as the scale n gets larger.

Then, by construction, $\bar{\lambda}_n(t) \equiv \lambda_n(nt) = \lambda_f(t)$, $\bar{\Lambda}_n(t) \equiv n^{-1}\Lambda_n(nt) = \Lambda_f(t)$ and $\hat{\Lambda}_n(t) = 0 \equiv \Lambda_d(t)$ for all n and t , while

$$\tilde{\Lambda}_{n,t}(u) = \sqrt{n}[\Lambda_f(t + u/\sqrt{n}) - \Lambda_f(t)] \rightarrow \lambda_f(t)u \quad (5.25)$$

as $n \rightarrow \infty$ uniformly in t and u , by the definition of a derivative, consistent with the assumptions in (5.19) and (5.20).

In order to have Λ_d play a role, we can define a more general family of arrival rate functions,

$$\Lambda_n(t) \equiv n\Lambda_f(t/n) + \sqrt{n}\Lambda_d(t/n). \quad (5.26)$$

With (5.26), we have

$$\bar{\Lambda}_n(t) = \Lambda_f(t) + n^{-1/2}\Lambda_d(t) \quad \text{and} \quad \hat{\Lambda}_n(t) = \Lambda_d(t) \quad (5.27)$$

so that again $\bar{\Lambda}_n \rightarrow \Lambda_f$ and $\hat{\Lambda}_n \rightarrow \Lambda_d$ in \mathcal{D} . Instead of (5.25), we now have

$$\tilde{\Lambda}_{n,t}(u) = \sqrt{n}[\Lambda_f(t + u/\sqrt{n}) - \Lambda_f(t)] + [\Lambda_d(t + u/\sqrt{n}) - \Lambda_d(t)] \quad (5.28)$$

so that, just as before, $\tilde{\Lambda}_{n,t}(u) \rightarrow \lambda_f(t)u$ as $n \rightarrow \infty$ uniformly in t and u over any bounded interval, now exploiting the assumed continuity of Λ_d . We use the bounded interval to obtain uniform continuity.

In applications, we would want our system to be system n for some n . For any n to be appropriate, the long-run average arrival rate should be unchanged at 1, but since the length of the sinusoidal cycle in λ_f is $2\pi/\gamma$, the length of the sinusoidal cycle in λ_n should be $2\pi n/\gamma$. The key relationship assumed as $n \rightarrow \infty$ is that the cycles in the periodic arrival rate function are of length $O(n)$, where $n = (1 - \rho)^{-2}$. ■

As a consequence of (5.17)-(5.22), we have associated limits for the scaled arrival process. To state them, let

$$\begin{aligned}\bar{A}_n(t) &\equiv n^{-1}N_a(\Lambda_n(nt), & \hat{A}_n(t) &\equiv n^{-1/2}[A_n(nt) - n\Lambda_f(t)] \quad \text{and} \\ \tilde{A}_{n,t}(u) &\equiv n^{-1/2}[A_n(nt + u\sqrt{n}) - A_n(nt)], \quad t \geq 0 \quad \text{and} \quad n \geq 1.\end{aligned}\quad (5.29)$$

Lemma 5.1 (*limits for the scaled arrival process*) *Under the scaling above, we have the FSLLN*

$$\bar{A}_n = \bar{N}_{a,n} \circ \bar{\Lambda}_n \rightarrow \Lambda_f \quad \text{in } \mathcal{D} \quad \text{as } n \rightarrow \infty \quad \text{w.p.1}, \quad (5.30)$$

the associated FCLT

$$\hat{A}_n = \hat{N}_{a,n} \circ \bar{\Lambda}_n + \hat{\Lambda}_n \Rightarrow B \circ \Lambda_f + \Lambda_d \quad \text{in } \mathcal{D} \quad \text{as } n \rightarrow \infty. \quad (5.31)$$

and

$$\tilde{A}_{n,t}(u) \rightarrow \lambda_f(t)u \quad \text{as } n \rightarrow \infty \quad (5.32)$$

uniformly in t and u within finite intervals.

Proof. Apply the continuous mapping theorem with the composition map, with and without centering; see §§13.2 and 1.3.3 of [36]. For (5.32), we use the fact that tightness associated with the weak convergence of $\hat{N}_{a,n}$ in (3.1) implies that

$$n^{-1/2}[N_a(nt + u\sqrt{n}) - N_a(nt)] \Rightarrow u \quad \text{as } n \rightarrow \infty \quad (5.33)$$

uniformly in t and u within bounded time intervals. In particular,

$$\begin{aligned}n^{-1/2}[A_n(nt + \sqrt{n}u) - A_n(nt)] &= n^{-1/2}[N_a(\Lambda_n(nt + u\sqrt{n})) - N_a(\Lambda_n(nt))] \\ &= n^{-1/2}[N_a(\Lambda_n(nt) + \lambda_f(t)u\sqrt{n} + o(\sqrt{n})) - N_a(\Lambda_n(nt))] \Rightarrow \lambda_f(t)u\end{aligned}\quad (5.34)$$

uniformly in t and u within finite time intervals. We use the convergence $\bar{\Lambda}_n \rightarrow \Lambda_f$ to deduce that $\Lambda_n(nt) < cnt$ for some constant c for all suitably large n . ■

We now introduce the scaled queueing processes, using the usual heavy-traffic scaling. Let

$$\hat{Q}_{1,n}(t) \equiv n^{-1/2}Q_{1,n}(nt), \quad t \geq 0, \quad (5.35)$$

so that $\hat{Q}_n(t) = \hat{Q}_{1,n}(\bar{\Lambda}_n(nt))$, $t \geq 0$ by Theorem 3.1. Let $W_n(t)$ be the virtual waiting time at time t in model n and define the associated scaled processes

$$\hat{W}_n(t) \equiv n^{-1/2}W_n(nt), \quad t \geq 0. \quad (5.36)$$

Let \mathcal{D}^k be the k -fold product space of \mathcal{D} with itself with the usual product topology. Let $R(t; a, b)$ be reflected Brownian motion (RBM) with drift $-a$ and diffusion coefficient b .

Theorem 5.2 (*heavy-traffic limit for the time-varying waiting time*) *Let the system start empty. Under the scaling assumptions above, including (5.17)-(5.20),*

$$(\hat{Q}_n, \hat{W}_n) \Rightarrow (\hat{Q}, \hat{W}) \quad \text{in } \mathcal{D}^2 \quad \text{as } n \rightarrow \infty, \quad (5.37)$$

where

$$\hat{W}(t) \equiv \hat{Q}(t)/\lambda_f(t) \quad \text{and} \quad \hat{Q}(t) \equiv R(\Lambda_f(t); -1, c_a^2 + c_s^2), \quad t \geq 0. \quad (5.38)$$

with λ_f in (5.16). As a consequence, for each $T > 0$,

$$\sup_{0 \leq t \leq T} \{\hat{W}_n(t) - (\hat{Q}_n(t)/\lambda_f(t))\} \Rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (5.39)$$

and, for each $x \geq 0$,

$$P(\hat{Q}_n(t) > x) \rightarrow e^{-2x/(c_a^2 + c_s^2)} \quad \text{and} \quad P(\lambda_f(t)\hat{W}_n(t) > x) \rightarrow e^{-2x/(c_a^2 + c_s^2)} \quad (5.40)$$

as first $n \rightarrow \infty$ and then $t \rightarrow \infty$.

Proof. We rely on the basic heavy-traffic FCLT for the standard $G/G/1$ queue covering the triple $(A_{1,n}, Q_{1,n}, D_{1,n})$ and related processes, as given in Theorem 9.3.4 of [36] and the continuity of the inverse function used in the first passage-time, as discussed in §§5.7, 13.6 and 13.7 of [36]. The essential argument follows §5.4 of the Internet Supplement of [36], drawing on Theorem 13.7.4 of [36], but we will give a direct proof.

First, we define the sequence of scaled processes associated with the arrival and service processes

$$\hat{A}_{1,n}(t) \equiv n^{-1/2}[N_a(nt) - nt], \quad t \geq 0 \quad \text{and} \quad n \geq 1 \quad (5.41)$$

and

$$\hat{S}_{1,n}(t) \equiv n^{-1/2}[N_s(nt/\rho_n) - nt] = n^{-1/2}[N_s(nt/\rho_n) - nt/\rho_n] + t \quad t \geq 0 \quad \text{and} \quad n \geq 1. \quad (5.42)$$

As a consequence,

$$(\hat{A}_{1,n}, \hat{S}_{1,n}) \Rightarrow (B_a, B_s + e) \quad \text{in } \mathcal{D}^2, \quad (5.43)$$

where B_a and B_s are independent BM's. Thus, $\hat{A}_{1,n} - \hat{S}_{1,n} \Rightarrow B_a - B_s - e$ in \mathcal{D} and we can apply Theorem 9.3.4 of [36] to obtain

$$\hat{Q}_{1,n} \Rightarrow R(\cdot) \equiv R(\cdot; -1, c_a^2 + c_s^2) \quad \text{in } \mathcal{D} \quad \text{as } n \rightarrow \infty, \quad (5.44)$$

so that $\hat{Q}_n = \hat{Q}_{1,n} \circ \bar{\Lambda}_n \Rightarrow R(\Lambda_f(\cdot))$ in \mathcal{D} as $n \rightarrow \infty$, by applying the continuous mapping theorem with the composition map without centering, as in §13.2 of [36].

We now come to the more difficult part of the argument. Let $D_n(t)$ and $D_{1,n}$ be the departure processes associated with system n . Since

$$\begin{aligned} \hat{W}_n(t) &\equiv n^{-1/2}W_n(nt) = \inf \{u \geq 0 : D_n(nt + u\sqrt{n}) - D_n(nt) \geq Q_n(nt)\} \\ &= \inf \{u \geq 0 : n^{-1/2}[D_n(nt + u\sqrt{n}) - D_n(nt)] \geq n^{-1/2}Q_n(nt)\} \\ &= \inf \{u \geq 0 : \hat{D}_{n,t}(u) \geq \hat{Q}_n(t)\}, \end{aligned} \tag{5.45}$$

where $\hat{D}_{n,t}(u) \equiv n^{-1/2}[D_n(nt + u\sqrt{n}) - D_n(nt)]$. We have observed that $\hat{Q}_n \Rightarrow R(\Lambda_f(\cdot))$ in \mathcal{D} as $n \rightarrow \infty$; we will now show that $\hat{D}_{n,t}(u) \rightarrow u\lambda_f(t)$ uniformly in t and u over the time intervals $[0, T]$ for $0 < T < \infty$.

For that purpose, let $B_{1,n}(t)$ be the amount of time that the server has been busy in the interval $[0, t]$ in the system with queue length $Q_{1,n}$. Since $D_n(nt) = D_{1,n}(\Lambda_n(nt))$, we have $D_n(nt) = N_s(\rho_n^{-1}B_{1,n}(\Lambda_n(nt)))$, $t \geq 0$. We obtain

$$n^{-1/2}[B_{1,n}(\Lambda_n(nt + u\sqrt{n})) - B_{1,n}(\Lambda_n(nt))] \rightarrow u\lambda_f(t) \tag{5.46}$$

uniformly in t and u in $[0, T]$ by applying condition (5.20) and the FCLT for $\hat{B}_{1,n}$ contained in Theorem 9.3.4 of [36]. From the assumed FCLT for N_s in (3.5), we obtain the desired convergence $\hat{D}_{n,t}(u) \rightarrow u\lambda_f(t)$ uniformly in t and u over the time intervals $[0, T]$ for $0 < T < \infty$. From there we can apply the continuity of the inverse function used in the first passage time. This argument directly implies (5.39), where we already have established that $\hat{Q}_n \Rightarrow \hat{Q}$ with the specified distribution in (5.38). The joint limit in (5.37) then follows by the convergence-together theorem, as in Theorem 11.4.7 of [36].

The last two limits in (5.40) follow immediately from (5.37) by applying the continuous mapping theorem with the projection at t because the direct limit $R(t; -1, (c_a^2 + c_s^2))$ converges in distribution to an exponential random variable with mean $(c_a^2 + c_s^2)/2$ as $t \rightarrow \infty$. ■

Paralleling Theorem 2 and Corollary 1 of [22] for many-server queues, we now show that *any* service rate control that stabilizes the queue length in heavy-traffic cannot also stabilize the virtual waiting time at the same time.

Theorem 5.3 (*stabilizing both in heavy traffic*) *Let the system start empty. Let the scaling assumptions above apply, including (5.17)-(5.20), but consider any service-rate control that stabilizes*

the queue length in the sense that $\hat{Q}_n \Rightarrow \hat{Q}$ in \mathcal{D} as $n \rightarrow \infty$, where $\hat{Q}(t) \Rightarrow \hat{Q}(\infty)$ as $t \rightarrow \infty$ with $0 < E[\hat{Q}(\infty)] < \infty$. Then

$$(\hat{Q}_n, \hat{W}_n) \Rightarrow (\hat{Q}, \hat{W}) \quad \text{in } \mathcal{D}^2 \quad \text{as } n \rightarrow \infty, \quad (5.47)$$

where $\hat{W}(t) \equiv \hat{Q}(t)/\lambda_f(t)$, $t \geq 0$. As a consequence, $\hat{W}_n(t)$ is not stabilized asymptotically as first $n \rightarrow \infty$ and then $t \rightarrow \infty$ unless $\lambda_f(t) \rightarrow \lambda_f(\infty)$ as $t \rightarrow \infty$.

Proof. We can apply the second half of the proof of Theorem 5.2. Given the assumed convergence $\hat{Q}_n \Rightarrow \hat{Q}$ in \mathcal{D} as $n \rightarrow \infty$, we can apply the tightness that follows from this convergence to deduce that

$$n^{-1/2}[Q_n(nt + u\sqrt{n}) - Q_n(nt)] \Rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (5.48)$$

uniformly in t and u over finite intervals. Combined with the limit for $\tilde{A}_{n,t}$ in (5.32), (5.48) implies that

$$\tilde{D}_{n,t}(u) \equiv n^{-1/2}[D_n(nt + u\sqrt{n}) - D_n(nt)] \Rightarrow \lambda_f(t)u \quad \text{as } n \rightarrow \infty \quad (5.49)$$

uniformly in t and u over finite intervals. Thus the limit(5.47) and the subsequent results hold by the proof of Theorem 5.2. ■

Remark 5.1 (*the resulting approximation*) The limit in (5.40) leads to approximating $Q_\rho(t)$ and $W_\rho(t)$ by exponential random variables if t is not too small. It also leads to a time-varying approximation for the time-varying mean. In particular, if we express the limiting arrival rate function λ_f in terms of the original arrival rate λ and the traffic intensity ρ , using $\lambda(t) \equiv \lambda_\rho(t) \approx \lambda_f((1 - \rho)^2 t)$, then we get

$$E[Q_\rho(t)] \approx \frac{c_a^2 + c_s^2}{2(1 - \rho)} \approx \frac{\rho(c_a^2 + c_s^2)}{2(1 - \rho)}, \quad (5.50)$$

and

$$E[W_\rho(t)] \approx \frac{c_a^2 + c_s^2}{2(1 - \rho)\lambda(t)} = \frac{(c_a^2 + c_s^2)}{2(1 - \rho)\rho\bar{\mu}(\lambda(t)/\bar{\lambda})} \approx \frac{\rho(c_a^2 + c_s^2)}{2(1 - \rho)\bar{\mu}(\lambda(t)/\bar{\lambda})}, \quad (5.51)$$

where $\bar{\mu} = \bar{\lambda}/\rho$ is the limiting average of $\mu(t)$, which exists by (2.1) and (3.4). The last approximation in each case is obtained to make the approximation consistent with the exact result for the $M/M/1$ model, and is justified by using $\rho \approx 1$; see [32] for a discussion of such refinements to direct heavy-traffic approximations. That final formula in (5.51) differs from the familiar heavy-traffic approximation for the steady-state wait in a $G/G/1$ queue, $E[W] \approx \rho(c_a^2 + c_s^2)/2(1 - \rho)\mu$, by simply inserting the relative arrival rate $\lambda(t)/\bar{\lambda}$ in the denominator. (We assume that t is sufficiently

large, or we have different initial conditions, so that a steady-state formula would be appropriate otherwise.) The joint limit also leads to the pathwise approximation

$$W_\rho(t) \approx \frac{Q_\rho(t)}{\bar{\mu}(\lambda(t)/\lambda)}, \quad t \geq 0. \quad \blacksquare \quad (5.52)$$

Remark 5.2 (*Application of Corollary 5.1*) For the sequence of $M_t/M_t/1$ models with long-run average arrival rates $\bar{\lambda}_n = 1$ and average service rate $1/\rho_n = 1/(1 - (1/\sqrt{n}))$, we can apply Corollary 5.1 to obtain a limit for the mean waiting time consistent with Theorem 5.2 under the assumed scaling. Again assume that $W_{1,n}(t)$ can be regarded as being in steady state distributed as $W_{1,n}^*$ with mean $\rho_n^2/(1 - \rho_n) \sim \sqrt{n}$ as $n \rightarrow \infty$, so that $E[\hat{W}_{1,n}^*] \equiv E[W_{1,n}^*]/\sqrt{n} \rightarrow 1$ as $n \rightarrow \infty$. Then Corollary 5.1 implies that

$$E[\hat{W}_n(t)] \equiv E[W_n(nt)/\sqrt{n}] \rightarrow \frac{1}{\lambda_f(t)} \quad \text{as } n \rightarrow \infty. \quad (5.53)$$

Paralleling (5.15), the reasoning is

$$\begin{aligned} E[\hat{W}_n(t)] &= \int_0^\infty P(W_n(nt) > x\sqrt{n}) dx = \int_0^\infty P(\Lambda_{n,t}^{-1}(W_{1,n}^*) > x\sqrt{n}) dx \\ &= \int_0^\infty P(\hat{W}_{1,n}^* > \tilde{\Lambda}_{n,t}(x)) dx = (1 - (1/\sqrt{n})) \int_0^\infty e^{-\tilde{\Lambda}_{n,t}(x)/(1 - (1/\sqrt{n}))} dx \\ &\rightarrow \int_0^\infty e^{-\lambda_f(t)x} dx = \frac{1}{\lambda_f(t)} \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (5.54)$$

However, for the approximation, we should not simply replace $1 - \rho_n = 1 - (1/\sqrt{n})$ by its limit 1. Instead, the approximation should be

$$E[\hat{W}_n(t)] \equiv E[W_n(nt)/\sqrt{n}] \approx \frac{(1 - (1/\sqrt{n}))^2}{\lambda_f(t)}, \quad n \geq 2. \quad (5.55)$$

To compare these to the limit, we would write instead

$$\frac{E[\hat{W}_n(t)]}{(1 - (1/\sqrt{n}))^2} \approx \frac{1}{\lambda_f(t)}. \quad \blacksquare \quad (5.56)$$

Example 5.3 (*more with the sinusoidal example*) To illustrate Theorem 5.2 and Remark 5.2, we again consider the sinusoidal example in Example 5.1 and 5.2 with base arrival rate function $\lambda_f(t) \equiv 1 + \beta \sin(\gamma t)$ with parameters $\beta = 0.2$ and $\gamma = 1.0$. We now consider a sequence of $M_t/M_t/1$ models indexed by n with the same base arrival rate function $\lambda_f(t)$ and the scaling in this section with $1 - \rho_n = 1/\sqrt{n}$. As in Example 5.2, we let $\lambda_n(t) \equiv \lambda_f(t/n)$, $t \geq 0$ and $n \geq 1$. We now plot the scaled periodic steady-state mean waiting time function, using (5.56), over one (scaled) cycle $[0, 2\pi]$ in Figure 5. Figure 5 strongly supports the heavy-traffic limit established in Theorem 5.2, confirming that the scaled mean waiting time is not stabilized by the rate-matching control and has the indicated form. \blacksquare

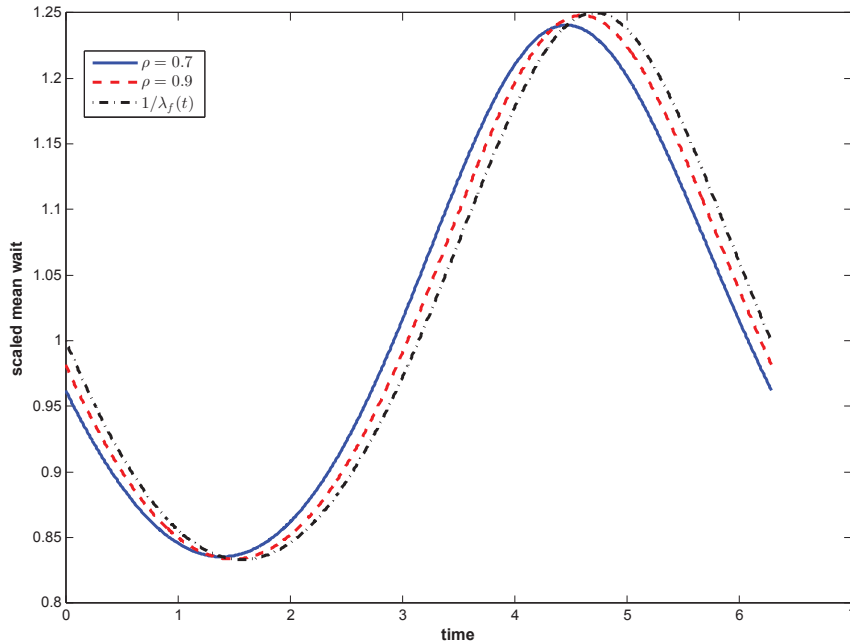


Figure 5: The scaled periodic steady-state time-varying mean wait $E[\hat{W}_n(t)]/\rho_n^2 \equiv E[W_n(nt)]/\rho_n^2\sqrt{n}$ as in (5.56) with the scaled sinusoidal arrival rate function having $\beta = 0.2$ and $\gamma = 1.0$ for $\rho_n = 0.7$ and 0.9 , compared to the limit $1/\lambda_f(t)$, over one cycle $[0, 2\pi]$.

We formalize the qualitative conclusion about the implications of time variability to be drawn from formula (5.51) in the following corollary.

Corollary 5.2 *In the heavy-traffic limit of Theorem 5.2, the approximating time-varying mean wait at time t is decreasing in the relative arrival rate $\lambda(t)/\bar{\lambda}$, being largest when $\lambda(t)/\bar{\lambda}$ is smallest. If $\lambda^\downarrow \leq \lambda(t) \leq \lambda^\uparrow$ for all $t \geq 0$, then provided that t sufficiently large,*

$$\frac{\lambda^\downarrow}{\lambda^\uparrow} \leq \frac{E[W(t_1)]}{E[W(t_2)]} \leq \frac{\lambda^\uparrow}{\lambda^\downarrow} \quad \text{for all } t_1, t_2 \text{ such that } t_1 > t \text{ and } t_2 > t. \quad (5.57)$$

In applications we have a single model with a fixed traffic intensity ρ . The applied relevance of the heavy-traffic limit in Theorem 5.2 will depend on the limiting cumulative rate function Λ_f in (5.19). To usefully approximate an observed time-varying arrival rate, it is important that Λ_f have time variability seen in the application. We now want to see the consequence of omitting the time scaling of the arrival rate functions in Example 5.2, so we return to that example.

Example 5.4 (*the sinusoidal example without time scaling*) We now return to Example 5.2 and suppose instead that we do not include the time scaling as n increases. It is natural to approach this through the arrival rate function. If we do so, then we would have $\lambda_n^{no}(t) = \lambda_f(t)$ and thus $\Lambda_n^{no}(t) = \Lambda_f(t)$ for all n . Having done this, we see that $\bar{\Lambda}_n^{no}(t) = n^{-1}\lambda_f(nt) \rightarrow t$ in \mathcal{D} as $n \rightarrow \infty$

and $\tilde{\Lambda}_{n,t}^{no}(u) = n^{-1/2}[\Lambda_f(nt + u\sqrt{n}) - \Lambda_f(nt)] \rightarrow u\bar{\lambda} = u$ as $n \rightarrow \infty$ uniformly in t and u , because we are looking at the average of λ_f over an interval of length $u\sqrt{n}$ multiplied by u . Hence, we do not see the impact of the periodicity in the limit.

We might instead omit the time scaling in the cumulative arrival rate function. Then we would have the cumulative arrival rate function

$$\Lambda_n^\#(t) \equiv n\Lambda_f(t), \quad t \geq 0, \quad (5.58)$$

without including the time scaling in Λ_n above. Then we still get the limits in (5.19) and (5.20), but now $\Lambda_f^\#(t) = t$ and $\lambda_f^\#(t) = 1$ for all $t \geq 0$. Thus, if we do not scale time, the limits in (5.19) and (5.20), and thus also in Theorem 5.2, reveal no impact of the time variability. This example is consistent with [9] and Corollary 3.1 in [37]. ■

6 A Periodic Arrival Rate Function

Let us now consider the special case of a periodic arrival rate function λ with period c ; see [17, 29] for background. (The sinusoidal function in Examples 5.1-5.4 is a special case.) In addition, we assume that the stationary model (A_1, D_1, Q_1) has a limiting steady-state version, by which we mean the following process limit

$$\{(A_1(t+s) - A_1(s), D_1(t+s) - D_1(s), Q_1(t+s)) : t \geq 0\} \Rightarrow \{(A_1^*(t), D_1^*(t), Q_1^*(t)) : t \geq 0\} \quad (6.1)$$

in \mathcal{D}^3 as $s \rightarrow \infty$, where Q_1^* is a stationary process, while (A_1^*, D_1^*) has stationary increments.

6.1 A Periodic Steady State

With these assumptions, we can deduce that our model has a periodic steady state. The following expresses a process version of that periodic steady state. It is significant that the one-dimensional marginals $Q(t)$ have a simple limiting steady-state distribution, independent of the periodic structure, but the 2-dimensional (and higher) marginals $(Q(t_1), Q(t_2))$ only have a limiting periodic steady-state distribution, with the periodic structure.

Theorem 6.1 (*periodic steady state*) *If λ is periodic with period c and (6.1) holds, then*

$$\begin{aligned} & \{(A(t+kc) - A(kc), D(t+kc) - D(kc), Q(t+kc), W(t+kc)) : t \geq 0\} \\ & \Rightarrow \{(A^*(t), D^*(t), Q^*(t), W^*(t)) : t \geq 0\} \quad \text{in } \mathcal{D}^4 \quad \text{as } k \rightarrow \infty, \end{aligned} \quad (6.2)$$

where (Q^*, W^*) is a periodic process with the marginal distribution of $Q^*(t)$ in \mathbb{R} independent of t , while (A^*, D^*) has periodic increments, i.e., the distribution of $\{(A^*(t+kc) - A^*(kc), D^*(t+kc) - D^*(kc), Q^*(t+kc), W^*(t+kc)) : t \geq 0\}$ in \mathcal{D}^4 is independent of k .

Proof. With the assumptions, Theorem 3.1 implies that (6.2) holds for the triple (A, D, Q) . Then (5.1) and (5.2) imply that the same is true for W . (Theorem 5.2 and (5.51) yield an approximation for that periodic steady-state variable $W^*(t)$.) ■

In this context of a periodic steady-state distribution, under regularity conditions, the waiting times of successive arrivals will directly have a steady-state distribution. For example, if the arrival process N_a is a renewal process with a non-lattice interarrival-time distribution, then the waiting time of the k^{th} arrival $W_{n,k}$ should converge to a proper steady-state limit $W_{n,\infty}$ as $k \rightarrow \infty$ for each n , because the arrivals do not occur at fixed places within a cycle. The periodic arrival rate implies that the steady-state wait $W_{n,\infty}$ should be a continuous mixture of $W_n^*(s)$ over a cycle, i.e.,

$$P(W_{n,\infty} > w) = \frac{\int_0^{nc} \lambda_n(s) P(W_n^*(s) > w) ds}{nc\bar{\lambda}}; \quad (6.3)$$

see Proposition A1 in the Appendix of [26].

However, Theorem 5.2 provides a heavy-traffic limit as $n \rightarrow \infty$ for the integrand in (6.3), which is independent of the time argument s . Hence, we see that the limit in (5.40) should apply to $\hat{W}_{n,\infty}$ as well as $\lambda_f(t)\hat{W}_n(t)$; i.e., paralleling (5.51), we have the associated heavy-traffic approximation

$$E[W_{\rho,\infty}] \approx \frac{\rho(c_a^2 + c_s^2)}{2(1-\rho)\bar{\mu}}. \quad (6.4)$$

As a consequence, the expected waiting time of successive arrivals is also stabilized by the rate-matching service rate control. However, this occurs, not because the expected waiting time is independent of the time of arrival, but because successive arrivals might occur anywhere in the periodic cycle. That is, we focus on $W_{n,k}$, the waiting time of the k^{th} arrival as k gets large, which has no fixed arrival time within a cycle. We can only conclude that (6.3) should hold. If we consider possible arrival times, then we should focus on $E[W(t)]$, which is periodic.

6.2 A Heavy-Traffic Limit for the Waiting Times of Successive Arrivals

We now show that a heavy-traffic limit can be obtained for the waiting time sequence $\{W_{n,k} : k \geq 0\}$ in the periodic setting of §6 above, which has a periodic limit. This shows that the order of the two limits as $t \rightarrow \infty$ and as $n \rightarrow \infty$ cannot be interchanged, just as for the multi-server queues

with deterministic service times in [21]. In the heavy-traffic limit, the arrival times occur at fixed places within the cycle.

To state the limit, let

$$\hat{Z}_n(t) \equiv n^{-1/2} W_{n, \lfloor nt \rfloor}, \quad t \geq 0 \quad \text{and} \quad n \geq 1, \quad (6.5)$$

where $\lfloor x \rfloor$ is the floor function denoting the greatest integer less than or equal to x .

Theorem 6.2 (*heavy-traffic limit for the waiting times of successive arrivals*) *Let the system start empty. Under the scaling assumptions above, including (5.17)-(5.20),*

$$\hat{Z}_n \Rightarrow \hat{Z} = \hat{W} \circ \Lambda_f^{-1} \quad \text{in } \mathcal{D} \quad \text{as } n \rightarrow \infty, \quad (6.6)$$

where \hat{Z}_n is defined in (6.5) and

$$\hat{Z}(t) \stackrel{d}{=} \frac{R(t; -1, c_a^2 + c_s^2)}{\lambda_f(\Lambda_f^{-1}(t))}, \quad t \geq 0, \quad (6.7)$$

with $\lambda_f(\Lambda_f^{-1}(t))$ being a periodic function with period $c\bar{\lambda}$.

Proof. Note that $\|\hat{Z}_n - \hat{W}_n \circ \bar{A}_n^{-1}\|_T \Rightarrow 0$ as $n \rightarrow \infty$ for any $T > 0$. Any difference is due to multiple arrivals at the same time, which is $o(\sqrt{n})$ uniformly in t over bounded intervals by the tightness of \hat{A}_n . By the continuous mapping theorem with the inverse map, $\bar{A}_n^{-1} \rightarrow \Lambda_f^{-1}$ in \mathcal{D} as $n \rightarrow \infty$; see §13.6 of [36]. Hence, by the continuous mapping theorem with composition, we have the claimed (6.6). We then obtain (6.7) from (5.38).

For the final statement, since λ_f is periodic with period c , we have $\Lambda_f(nc + t) = nc\bar{\lambda} + \Lambda_f(t)$, $0 \leq t \leq c$. As a consequence, $\Lambda_f^{-1}(nc\bar{\lambda} + t) = nc + \Lambda_f^{-1}(t)$, $0 \leq t \leq c\bar{\lambda}$. Since λ_f is periodic with period c , $\lambda_f(nc + \Lambda_f^{-1}(t)) = \lambda_f(\Lambda_f^{-1}(t))$, $0 \leq t \leq c\bar{\lambda}$ and $\lambda_f(\Lambda_f^{-1}(nc\bar{\lambda} + t)) = \lambda_f(\Lambda_f^{-1}(t))$, $0 \leq t \leq c\bar{\lambda}$, showing that indeed $\lambda_f(\Lambda_f^{-1}(t))$ is a periodic function with period $c\bar{\lambda}$. ■

We remark that the steady-state approximation in (6.4) can be obtained from (6.7) if we consider t sufficiently large that we replace the RBM with its exponential steady-state distribution and we replace $\lambda_f(\Lambda_f^{-1}(t))$ in the denominator by its long-run average $\bar{\lambda}_f = 1$. As in [21], the periodic heavy-traffic limit shows the possibility of nearly periodic behavior for systems in practice.

Theorems 3.1 and 6.1 are useful for conducting simulation experiments in order to evaluate the time-varying behavior of the queue length $Q(t)$ and the virtual waiting time $W(t)$ with the rate-matching service-rate control. First, Theorem 3.1 implies that $Q(t)$ approaches the steady-state limiting distribution of $Q_1(t)$ in the associated stationary $G/G/1$ model (assuming that it has a

proper limiting steady-state distribution). Hence, it suffices to start by simulating the stationary $G/G/1$ model in a conventional way.

Second, Theorem 6.1 implies that, if the arrival rate function is periodic with period c , then the stochastic process $\{(Q(t), W(t)) : t \geq 0\}$ has a periodic steady-state distribution $\{(Q^*(t), W^*(t)) : t \geq 0\}$, where $(Q^*(t+c), W^*(t+c)) \stackrel{d}{=} (Q^*(t), W^*(t))$ for all $t \geq 0$. Hence, if we consider examples with periodic arrival processes, then we can observe when the periodic steady-state is reached, and thus know when the impact of the initial conditions will have dissipated.

Formula (5.6) requires that we be able to compute Λ^{-1} , while Theorem 5.1 requires that we be able to compute Λ_t^{-1} . That task is simplified if we have a periodic function. In particular, if λ is periodic with periodic cycle c and with long-run average $\bar{\lambda} = 1$, then

$$\Lambda^{-1}(kc) = \Lambda(kc) = kc \quad \text{for all } k \geq 1. \quad (6.8)$$

As a consequence, it suffices to know the inverse over just one cycle, because

$$\Lambda^{-1}(kc+t) = kc + \Lambda^{-1}(t), \quad 0 \leq t \leq c. \quad (6.9)$$

Hence, we could compute, tabulate and apply the values of $\Lambda^{-1}(ck/n)$ for $1 \leq k \leq n$ to compute relevant inverse function values.

Example 6.1 (*Example 5.2 revisited*) Consider the sinusoidal arrival rate function $\lambda = \lambda_f$ in (5.16), so that

$$\Lambda(t) = t - (\beta/\gamma)(\cos(\gamma t) - 1), t \geq 0, \quad (6.10)$$

and

$$\Lambda_t(u) = u - (\beta/\gamma)(\cos(\gamma(t+u)) - \cos(\gamma t)), t \geq 0, \quad (6.11)$$

Also note that, since the periodic cycles are of length $2\pi/\gamma$, we have

$$\Lambda(2k\pi/\gamma) = 2k\pi/\gamma = \Lambda^{-1}(2k\pi/\gamma) \quad \text{for all } k \geq 1.$$

We conclude this section by observing that the heavy-traffic scaling of time and space in §5 makes the approximate simulation method in (3.16) more appropriate as n increases. As observed just prior to Example 5.2, the service requirements and service times are of order $O(1)$ as $n \rightarrow \infty$. However, the arrival rate and service rate change more slowly as n increases. Indeed, the derivative of the service rate is $O(1/n)$ as $n \rightarrow \infty$. This provides strong support for approximation (3.16), showing that it is asymptotically correct as $n \rightarrow \infty$ with the scaling.

7 Controls to Stabilize the Expected Waiting Time

In this section we consider controls aimed to stabilize the time-varying waiting time $W(t)$.

7.1 An Impossibility Result

Paralleling Theorem 5.3, we now show that any control that stabilizes the waiting time distribution $P(W(t) > x)$, $x \geq 0$, cannot also stabilize the mean number waiting in queue $E[(Q(t) - 1)^+]$ unless the arrival rate function is constant.

Theorem 7.1 (*impossibility of stabilizing both*) *Consider a $G_t/G_t/1$ system starting empty in the distant past. Suppose that a service-rate control makes the distribution of $W(t)$ independent of t with finite positive mean EW and a convex cdf $P(W > x)$, as occurs when W has a monotone pdf. Then the only arrival rate functions for which the mean number waiting in queue $E[(Q(t) - 1)^+]$ is constant, independent of t , are the constant arrival rate functions.*

Proof. First, we can express the mean number waiting in queue in terms of the time-varying waiting time distribution using the time-varying version of Little's law as in [2], and then perform a change of variables to get

$$E[(Q(t) - 1)^+] = \int_{-\infty}^t \lambda(s)P(W(s) > t - s) ds = \int_0^\infty \lambda(t - s)P(W(t - s) > s) ds, \quad t \geq 0, \quad (7.1)$$

If $W(t) \stackrel{d}{=} W$ for all t with $0 < EW < \infty$, then $p(x) \equiv P(W > x)/EW$ with $p(x) \equiv 0$ for $x < 0$ is a pdf on the entire real line. Suppose that $E[(Q(t) - 1)^+] = m$ for all t . Then

$$\int_{-\infty}^\infty \lambda_m(t - s)p(s) ds = 0, \quad (7.2)$$

where $\lambda_m(t) \equiv \lambda(t) - m$, which is bounded and continuous by the assumption on λ . We now deduce, under the assumed structure on p , that equation (7.2) implies that we must have $\lambda_m(t) = 0$ for all t , which implies that $\lambda(t)$ is constant.

Since the left side of (7.2) is a convolution integral as in §V.4 of [11], we can apply characteristic functions (cf's, which correspond to Fourier transforms) as in §XV of [11] to get $\hat{\lambda}_m(t)\hat{p}(t) = 0$ for all t for the cf's. We then can apply symmetrization as in §V.5 and §XV of [11] to get

$$|\hat{\lambda}_m(t)\hat{p}(t)|^2 = |\hat{\lambda}_m(t)|^2 \cdot |\hat{p}(t)|^2 = 0 \quad \text{for all } t. \quad (7.3)$$

The advantage of this approach is that the characteristic function $|\hat{p}|^2$ is a real-valued functions of t .

Since the pdf f corresponding to the cf $|\hat{p}|^2$ can be represented as

$$f(x) = \int_0^\infty p(x+s)p(s) ds,$$

we see that the monotonicity and convexity of p for $x \geq 0$ is inherited by f . We can then apply the assumed structure of the cdf $P(W > x)$ and Theorem 8.7 of [5] to deduce that $|\hat{p}(t)|^2$ is nonnegative for all t . Finally, without loss of generality, by applying a limiting argument as in Exercise 13 on p. 528 in §XV of [11], we can assume that $|\hat{p}(t)|^2 > 0$ for all t . We thus can divide by the positive real number to deduce that

$$|\hat{\lambda}_m(t)|^2 = 0 \quad \text{for all } t, \quad \text{so that } \hat{\lambda}_m(t) = 0 \quad \text{for all } t, \quad (7.4)$$

which in turn implies that λ is a constant function. ■

The conditions on the distribution of W in Theorem 7.1 are reasonable, but presumably they can be relaxed. Since

$$E[Q(t)] = P(Q(t) \geq 1)E[(Q(t) - 1)^+ | Q(t) \geq 1], \quad t \geq 0, \quad (7.5)$$

we see that in heavy traffic, under regularity conditions, $P(Q(t) \geq 1) \approx 1$ and $E[(Q(t) - 1)^+]/E[Q(t)] \approx 1$, so that the result applies to $E[Q(t)]$ in heavy traffic, consistent with Theorem 5.3.

7.2 The First Square-Root Service-Rate Control

We now examine the two square-root service-rate controls in (2.3) and (2.4) as ways to stabilize the waiting time. For multi-server models with time-varying arrival rates, the various approaches to server staffing (choosing a time-varying number of servers) in order to stabilize the performance of a queueing system with a time-varying arrival rate function lead to a *square-root staffing formula*, i.e.,

$$s(t) = m(t) + \xi \sqrt{m(t)}, \quad (7.6)$$

where $m(t)$ is an appropriate offered load, corresponding to an expected number of busy servers in an associated infinite-server model, with different methods to find the quality-of-service parameter ξ in (7.6) in order to focus on a particular performance measure; see [6, 22, 39] and references therein.

To apply (7.6) here, we need to define an appropriate offered load in the present setting, which we do not take up here. A direct analog in our setting is the square-root service-rate control in (2.3). From Theorem 5.2, we see that, if we are interested in stabilizing the expected virtual waiting time

$E[W(t)]$, then the rate matching control in (2.1) overstaffs when the arrival rate $\lambda(t)$ is relatively large and understaffs when it is relatively low. Formula (2.3) acts to correct that bias. We now show that the square-root service-rate control in (2.3) is asymptotically optimal with respect to an appropriate criterion (Which is not stabilizing) with an appropriate time scaling.

To establish this positive asymptotic result, we exploit connections to the earlier work on optimal capacity allocation in [3, 19, 20, 38] mentioned in §1. The goal in that work is to allocate service rates μ_i to each of n single-server queues with specified arrival rates λ_i . The object is to minimize the total expected *steady-state* waiting time at all queues, $\sum_{i=1}^n E[W_i]$ subject to a budget constraint $\sum_{i=1}^n r_i \mu_i \leq B$, where r_i is the cost of allocating rate μ_i at queue i and $B > \Lambda \equiv \sum_{i=1}^n r_i \lambda_i$. (The waiting time is the elapsed time from customer arrival to starting service.)

The key to a simple solution of the steady-state capacity-allocation problem is the product-form steady-state distribution for open queueing networks. Since the n queues are mutually independent in steady state, the allocation of μ_i affects queue i but no other queue. The product form is exact for a Markovian Jackson network, where in steady state each queue behaves as an $M/M/1$ model, and can be a reasonable approximation for a generalized Jackson network, where each queue behaves as an $GI/GI/1$ model. Interestingly, this problem is also solved by a square-root formula much like (7.6). Assuming that $E[W_i] \approx \lambda_i(c_{a,i}^2 + c_{s,i}^2)/2(\mu_i - \lambda_i)$, where $c_{a,i}^2$ and $c_{s,i}^2$ are the squared coefficients of variation (sev, variance divided by the square of the mean) of an interarrival time and a service time (which is exact for $M/M/1$), and the product form is approximately valid, the optimal allocations are

$$\mu_i = \lambda_i + \frac{(B - \Lambda) \sqrt{\lambda_i r_i (c_{a,i}^2 + c_{s,i}^2)}}{\sum_{j=1}^n \sqrt{\lambda_j r_j (c_{a,j}^2 + c_{s,j}^2)}}. \quad (7.7)$$

We make three initial observations: First, if $r_i(c_{a,i}^2 + c_{s,i}^2)$ is independent of i , then (7.7) looks more like (7.6). Second, we note that the theoretical bases for (7.6) and (7.7) are quite different. Formula (7.6) can be explained by the central limit theorem (e.g., the number of busy servers in the $M_t/GI/\infty$ infinite-server model is Poisson, and thus approximately Gaussian, with mean and variance equal to $m(t)$), whereas formula (7.7) follows from basic optimization theory (the form of the convex objective function with $\mu_i - \lambda_i$ in the denominator of each term and the independence of the queues). Third, the form of the solution in (7.7) depends critically on the form of the objective function. If we want to balance the ratio of the mean waiting time to the mean service time or minimize the sum of these ratios, then the rate-matching service rate control in (2.1) would be optimal.

The nice analysis leading to (7.7) would apply to our time-varying arrival-rate setting under two-conditions: (i) if we had a similar objective function involving the sum of the mean waiting times at different times, and (ii) if we could assume that the performance of the queue at one time is approximately independent of its performance at another time, with the allocation of capacity at one time not affecting the performance at any other time.

To consider condition (i), we first need to replace the steady-state waiting time by the time-varying virtual waiting time, $W(t)$, i.e., the time an arrival at time t would have to wait if there were an arrival at time t . Condition (i) should be approximately satisfied if we elect to minimize the average mean time-varying expected waiting time, i.e., if for some $T > 0$ and $m > 1$, the objective function is

$$\frac{1}{m} \sum_{k=1}^m E[W(kT/m)] \quad (7.8)$$

and we have the service rate constraint

$$\int_0^T \mu(t) dt > \rho^{-1} \int_0^T \lambda(t) dt \quad \text{for } 0 < \rho < 1. \quad (7.9)$$

However, condition (ii) is more problematic. Clearly, condition (ii) cannot hold exactly, because the performance at any time depends on the history prior to that time. Nevertheless, it might hold approximately. Indeed, for queues with time-varying arrival rates, condition (ii) is captured by the *pointwise-stationary approximation* (PSA), discussed in [1, 14, 27, 35]. Assuming that the PSA is valid as an approximation, then (7.6) is optimal. We state the asymptotic result for Markovian systems that follows from [35], which has been established.

Theorem 7.2 (*asymptotic optimality in the PSA scaling*) *Consider the Markovian $M_t/M_t/1$ model with the time-varying arrival rate $\lambda(t)$ and service rate $\mu(t)$, where $\mu(t)$ is subject to the constraint that $\mu(t) > \lambda(t)$ for all t , $0 \leq t \leq T$. Consider a sequence of models indexed by n in which both the arrival rate function and the service rate function in model n are multiplied by n . If the goal is to choose a service rate function $\mu(t)$ to minimize the objective function (7.8) subject to the constraint in (7.9), then the PSA control in (2.3) is asymptotically optimal as $n \rightarrow \infty$.*

Proof. We combine the asymptotic result in [35], which shows that the system asymptotically has the steady state distribution of an $M/M/1$ queue at each time with the traffic intensity at that time, independent of other times, and the optimization in [19]. ■

We conclude by (i) observing that Theorem 7.2 evidently extends to $G_t/G_t/1$ models, but producing a control analogous to (7.7) and (ii) emphasizing that goal expressed by the objective

function in Theorem 7.2 is not stable performance and, indeed, Figure 2 shows that $E[W(t)]$ is not fully stabilized with the square root control in (2.3) even though PSA is appropriate in that example.

7.3 The Second Square-Root Service-Rate Control

We now consider the second square-root service-rate control. It too is based on assuming that we have time scaling so that the PSA is appropriate. Thus, we directly assume that the expected time-varying waiting time at each time t can be approximated by

$$E[W(t)] \approx \frac{\rho(t)V}{\mu(t)(1-\rho(t))}, \quad t \geq 0, \quad (7.10)$$

as would be appropriate if the $G_t/G_t/1$ model were in steady state at each time t with arrival rate $\lambda(t)$, service rate $\mu(t)$ and traffic intensity $\rho(t) \equiv \lambda(t)/\mu(t) < 1$, where V is a variability formula, which we take to be $V \equiv (c_a^2 + c_s^2)/2$.

We then assume that the goal is to choose $\mu(t)$ to stabilize $E[W(t)]$ at the target w for all t . Thus, from (7.10), we have the equation

$$w = \frac{\lambda(t)V}{\mu(t)^2 - \mu(t)\lambda(t)}, \quad (7.11)$$

which leads to the quadratic equation in $x = \mu(t)$

$$wx^2 - \lambda(t)wx - \lambda(t)V = 0, \quad (7.12)$$

which has the solution given in (2.4) for $\zeta \equiv 4V/w$.

Since $\sqrt{1+\epsilon} = 1 + \epsilon/2 + o(\epsilon^2)$ as $\epsilon \downarrow 0$, we see that if we let the target $w \rightarrow \infty$, then the solution to (7.12) approaches

$$x \equiv \mu(t, w) = \lambda(t) + \frac{V}{w}, \quad (7.13)$$

That can be formalized by considering a sequence of models indexed by n , where we increase the arrival and service rates with n even more than in Theorem 7.2. As for Theorem 7.2, we consider $M_t/M_t/1$ models.

Theorem 7.3 (*heavy-traffic behavior of the service rate control (7.13) with PSA*) *Consider a sequence of $M_t/M_t/1$ models indexed by n with $\lambda_n(t) \equiv n^p \lambda(t)$ for $p > 1$, where $\lambda(t)$ is continuous and strictly positive. Suppose that the service-rate control in (7.13) is used as a function of the target w in the form $\mu_n(t) = \lambda_n(t) + n^{p-1/2}/w$ for $t \geq 0$. Then, for each $t > 0$, $\sqrt{n}(1 - \rho_n(t)) \rightarrow 1/w\lambda(t)$,*

$$n^{-1/2}Q_n(t) \Rightarrow w\lambda(t)X \quad \text{and} \quad n^{p-(1/2)}W_n(t) \Rightarrow wX \quad \text{as} \quad n \rightarrow \infty, \quad (7.14)$$

where X is a mean-1 exponential random variable, so that the scaled waiting time is stabilized in the limit.

Proof. The proof can be a variant of the proof of the PSA limit in [35] used in the proof of Theorem 7.2. For any $T > 0$, there exists λ^\downarrow and λ^\uparrow such that $0 < \lambda^\downarrow \leq \lambda(t) \leq \lambda^\uparrow < \infty$, $0 \leq t \leq T$. We can thus construct bounding stationary $M/M/1$ models with the maximum arrival rate and minimum service rate over the interval $[0, T]$ for each n . These are ordered by sample path stochastic order. As a consequence, we deduce that the set of random variables $\{n^{-1/2}Q_n(t) : 0 \leq t \leq T, n \geq 1\}$ is stochastically bounded.

We next observe that we can approximate the $M_t/M_t/1$ model in an interval $[t_0 - \delta, t]$ by a stationary $M/M/1$ model with constant arrival rate $\lambda_n(t)$ and associated service rates. By the continuity of λ , for any $\epsilon > 0$, there is a $\delta > 0$ such that $\lambda(t) - \epsilon \leq \lambda(s) \leq \lambda(t) + \epsilon$, $t - \delta \leq s \leq t$. Hence, bounds can be constructed above and below that are arbitrarily close; i.e., we can consider $\lambda(s)$ as essentially constant for $s \in [t - \delta, t]$. Because $\{n^{-1/2}Q_n(t - \delta) : n \geq 1\}$ is stochastically bounded and the scaling of the arrival rate function, we can apply standard heavy-traffic arguments to deduce that $n^{-1/2}Q_n(t) \Rightarrow w\lambda(t)X$ as $n \rightarrow \infty$; i.e., we get the steady-state of the RBM limit at time t . We then can apply (5.1) and (5.2) to deduce that $n^{p-(1/2)}W_n(t) \Rightarrow wX$, completing the proof. ■

Evidently Theorem 7.3 extends to a large class of $G_t/G_t/1$ queues, provided that approximation (7.10) is appropriate. Exact formulas apply for $M_t/G_t/1$ queues, while a large class of heavy-traffic limits are consistent with (7.10) as well.

Remark 7.1 (*applying the reasoning to the first square-root service-rate control in (2.3)*) We can also apply the reasoning in (7.10) and (7.11) to see what happens if we aim to stabilize the mean waiting time at w . Combining (7.11) with (2.3), we see that the solution is

$$w = \frac{(V/\xi)}{1 + \sqrt{\lambda(t)}} \quad (7.15)$$

which shows that we should not succeed even with long cycles where PSA is a good approximation.

8 Simulation Examples

We have evaluated the performance of the staffing algorithms developed in this paper by conducting simulation experiments. For all experiments, we estimate the mean values by performing 10,000 i.i.d. replications of individual runs. For $\gamma = 0.001$, we use the time interval $[0, 2 \times 10^4]$, which

is about three cycles. We first consider the rate-matching service-rate control in (2.1) with traffic intensity $\rho = 0.8$. We generate the exact service times using (3.12), but for such long cycles we find that the approximations in (3.13) and (3.14) yield visually identical plots; that is not true for larger γ such as 0.1.

We illustrate the results in this paper by displaying simulation results for the $M_t/M_t/1$ model with the sinusoidal arrival rate function $\lambda(t) \equiv 1 + \beta \sin \gamma t$, as in (5.16), with parameter $\beta = 0.2$ and several values of γ . We primarily focus on $\gamma = 0.001$; see Figures 1-3. Since a sine cycle is $2\pi/\gamma$, such a small γ makes for long cycles, one cycle being of length 6.28×10^3 . Consistent with Figures 1-3 and Theorems 3.1 and 4.2, corresponding estimates of the time-varying delay probability show that, just like $E[Q(t)]$, it is stabilized by the rate-matching control (2.1), but not by the two square-root controls (2.3) and (2.4).

Figure 3 shows the performance estimates for the second square-root service-rate control in (2.4) with parameter $\zeta = 1.0$, again with the same arrival process as before. Since $V = 1$ in this case and $\zeta = 4V/w$ for equation (7.12), we see that $\zeta = 1$ should stabilize the mean waiting time at 4.0, provided that PSA is appropriate. Figure 3 shows that this second square-root service-rate control in (2.4) indeed stabilizes the mean waiting time at the targeted value 4.0 after an initial transient, but does not stabilize the mean queue length.

The stabilization of $E[Q(t)]$ and $P(W(t) > 0) = P(Q(t) > 0)$, but not of $E[W(t)]$, by the rate-matching control, as illustrated by Figure 1, remains true for all γ , as illustrated by the corresponding plots for $\gamma = 0.1$ in Figure 6. The simulation run is now over the time interval $[0, 2000]$, but only the last three cycles are shown, where each is of length $2\pi \approx 6.28$. Figure 6 shows that the mean $E[W(t)]$ fluctuates less over each cycle with larger γ . There is now more averaging, because the average waiting time is about one half a cycle.

However, this good property of the rate-matching control (holding for all γ) does not apply to the second square-root control, because it depends on the validity of the PSA. The waiting times cease to be stabilized for larger values of γ , where PSA no longer holds, as illustrated by Figure 7, which shows the case of $\gamma = 0.1$.

As $\gamma \rightarrow \infty$, the cumulative arrival rate function approaches a linear function and the arrival process approaches a stationary Poisson process, so that the performance of the $M_t/M_t/1$ model with the rate matching control approaches that of the stationary $M/M/1$ model (proved as in [33]). That phenomenon can already be seen from $\gamma = 10$, where $E[W(t)]$, $E[Q(t)]$ and $P(W(t) > 0)$ are all stabilized by *all* the controls, as shown in Figure 8 for the rate-matching control. Similar plots

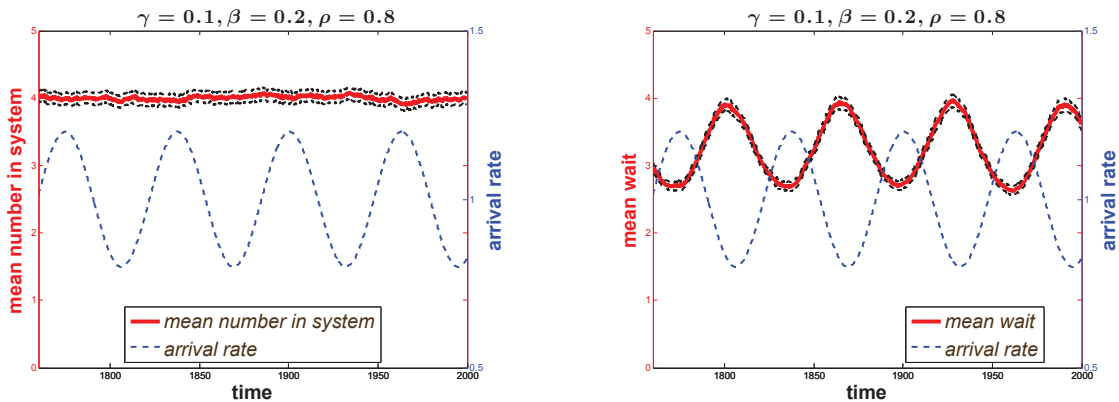


Figure 6: Simulation estimates of the time-varying mean number in the system, $E[Q(t)]$ (left) and the mean waiting time, $E[W(t)]$ (right) for the $M_t/M_t/1$ model with the rate-matching service-rate control in (2.1) for the same sinusoidal arrival rate function except shorter cycles based on $\gamma = 0.1$.

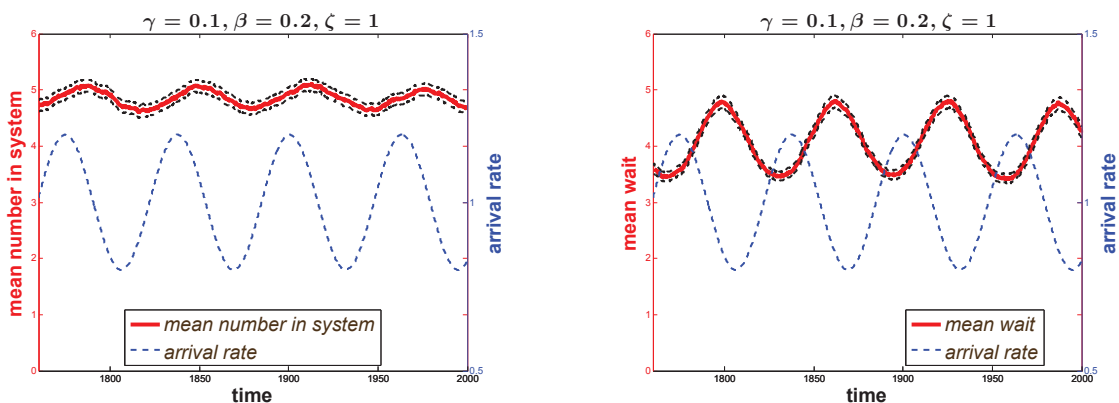


Figure 7: Simulation estimates of the time-varying mean number in the system, $E[Q(t)]$ (left) and the mean waiting time, $E[W(t)]$ (right) for the $M_t/M_t/1$ model, having the same sinusoidal arrival rate with shorter cycles ($\gamma = 0.1$), with the second square-root service-rate control in (2.4).

hold for the two square-root service-rate controls too for $\gamma = 10$. As for $\gamma = 1$, each simulation run is over the interval $[0, 2000]$, but only the last three cycles are shown. In this case each cycle is of length $2\pi/10 \approx 0.628$. Significant averaging is to be expected since the average waiting time is now several cycle lengths.

Acknowledgement

The author gratefully acknowledges support from NSF grant CMMI grant 1265070 and the help of Columbia doctoral student Ni Ma, including all reported numerical and simulation results.

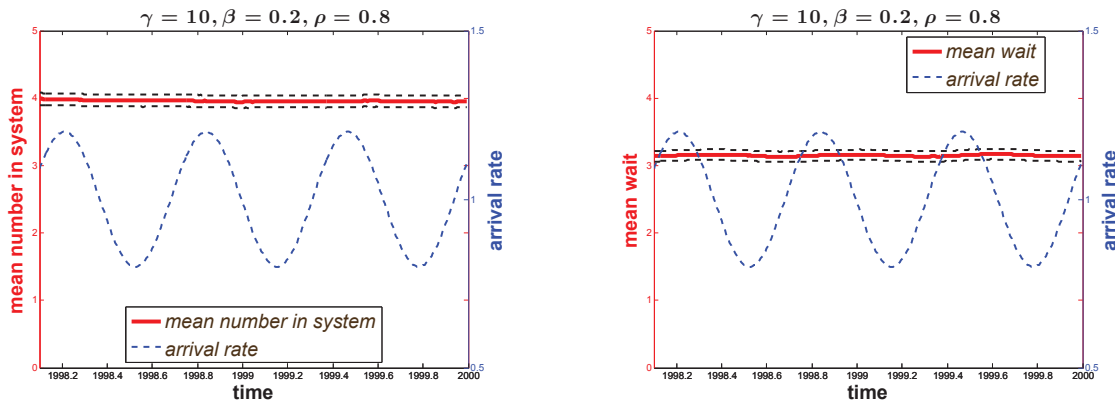


Figure 8: Simulation estimates of the time-varying mean number in the system, $E[Q(t)]$ (left) and the mean waiting time, $E[W(t)]$ (right) for the $M_t/M_t/1$ model with the rate-matching service-rate control in (2.1) for the same sinusoidal arrival rate function with very short cycles based on $\gamma = 10$.

References

- [1] A. Bassamboo, J. M. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.*, 54(3):419–435, 2006.
- [2] D. Bertsimas and G. Mourtzinou. Transient laws of nonstationary queueing systems and their applications. *Queueing Systems*, 25:315–359, 1997.
- [3] G. R. Bitran and S. Dasu. A review of open queueing network models of manufacturing systems. *Queueing Systems*, 12:95–134, 1992.
- [4] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Stat. Assoc.*, 100:36–50, 2005.
- [5] D. C. Champeney. *A Handbook of Fourier Theorems*. Cambridge University Press, Cambridge, UK, 1987.
- [6] M. Defraeye and I. Van Nieuwenhuysse. 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.
- [7] S. G. Eick, W. A. Massey, and W. Whitt. The physics of the $M_t/G/\infty$ queue. *Oper. Res.*, 41:731–742, 1993.
- [8] S. G. Eick, W. A. Massey, and W. Whitt. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.*, 39:241–252, 1993.
- [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] W. Feller. *An Introduction to Probability Theory and its Applications*. John Wiley, New York, second edition edition, 1971.
- [12] W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.
- [13] I. Gebhardt and B. L. Nelson. Transforming renewal processes for simulation of non-stationary arrival processes. *INFORMS Journal on Computing*, 21:630–640, 2009.
- [14] L. V. Green and P. J. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.*, 37:84–97, 1991.
- [15] L. V. Green, P. J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management*, 16:13–29, 2007.

- [16] B. He, Y. Liu, and W. Whitt. Staffing a service system with non-Poisson nonstationary arrivals. Columbia University, working paper, 2015.
- [17] D. P. Heyman and W. Whitt. The asymptotic behavior of queues with time-varying arrival. *Journal of Applied Probability*, 21(1):143–156, 1984.
- [18] O. B. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Sci.*, 42:1383–1394, 1996.
- [19] L. Kleinrock. *Communication Nets: Stochastic Message Flow and Delay*. Dover, New York, 1964.
- [20] L. Kleinrock. *Queueing Systems*, volume 2. Wiley, New York, 1976.
- [21] Y. Liu and W. Whitt. Nearly periodic behavior in the the overloaded $G/D/S + GI$ queue. *Stochastic Systems*, 1(2):340–410, 2011.
- [22] Y. Liu and W. Whitt. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.*, 60(6):1551–1564, 2012.
- [23] Y. Liu and W. Whitt. Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 24, 2014.
- [24] N. Ma and W. Whitt. Stabilizing the expected waiting time in a single-server queue with time-varying arrival rate. Columbia University, in preparation, 2014.
- [25] W. A. Massey and W. Whitt. Unstable asymptotics for nonstationary queues. *Math. Oper. Res.*, 19(2):267–291, 1994.
- [26] W. A. Massey and W. Whitt. A stochastic model to capture space and time dynamics in wireless communication systems. *Prob. in the Engineering and Informational Sciences*, 8:541–569, 1994.
- [27] W. A. Massey and W. Whitt. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability*, 9(4):1130–1155, 1997.
- [28] G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, 4:193–267, 2007.
- [29] T. Rolski. Queues with nonstationary inputs. *Queueing Systems*, 5:113–130, 1989.
- [30] S. Stidham. A last word on $L = \lambda W$. *Oper. Res.*, 22:417–421, 1974.
- [31] L. M. Wein. Capacity allocation in generalized Jackson networks. *Oper. Res. Letters*, 8:143–146, 1989.
- [32] W. Whitt. Refining diffusion approximations for queues. *Oper. Res. Letters*, 1(5):165–169, 1982.
- [33] W. Whitt. Departures from a queue with many busy servers. *Mathematics of Operations Research*, 9(4):534–544, 1984.
- [34] W. Whitt. A review of $L = \lambda W$. *Queueing Systems*, 9:235–268, 1991.
- [35] W. Whitt. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science*, 37(3):307–314, 1991.
- [36] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.
- [37] W. Whitt. Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters*, 42:458–461, 2014.
- [38] R. W. Wolfe. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [39] G. Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing and Service Oper. Management*, 16(2):283–299, 2014.