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}

July 29, 2014

Abstract

We consider a general $G_t/G_t/1$ single-server queue with unlimited waiting space and a time-varying arrival rate, where the the service rate at each time is subject to control. We first study the rate-matching control, where the 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 a square-root service-rate control, where the service rate exceeds the arrival rate by a constant multiple of the square root of the arrival rate. We show that this alternative service-rate control stabilizes the waiting time, but not the queue length, when the arrival rate changes very slowly relative to the average service time. This behavior is supported by a limit theorem supporting the pointwise-stationary approximation.

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 queueing system with a time-varying arrival rate function. 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 [3, 5, 10, 23] and references therein. 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.

Having a single-server queue where the the service rate is a continuous deterministic function subject to control is an idealization of what can occur 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 really 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 function is possible.

We start by considering 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

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

In considering this rate-matching control, we assume that the arrival rate function is known. In future work, we intend to consider the case in which the rate-matching control is used with an estimate of the arrival rate function obtained from data. 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 it stabilizes the queue length as $t \to \infty$ (to allow the effect of the initial condition to dissipate), but not the waiting time.

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 [7], extended for approximations of generalized Jackson networks in [17] and reviewed in §5.7 of [8], in §7 of [2] and elsewhere. Now, instead of allocating capacity to several queues in different locations, we allocate capacity to a single queue at different times. As an analog of Kleinrock's [7] square-root capacity allocation formula (appearing in (7.2) here), we also consider the *square-root service-rate control*

$$\mu(t) \equiv \lambda(t) + \beta\sqrt{\lambda(t)}, \quad t \geq 0. \tag{1.2}$$

We will identify a setting in which the square-root service-rate control in (1.2) is optimal in §7 by establishing a connection between the two problems.

Here is how this paper is organized: We start in §2 by defining the specific $G_t/G_t/1$ model, showing how to construct the service times, and showing that the queue length process in this model is a deterministic time transformation of the queue length process in an associated stationary $G/G/1$ model. In §3 we establish positive stabilization properties of the rate-matching control. In §4 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 theorem that yields a useful approximation for the time-varying waiting time distribution. In particular, Theorem 4.2 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 the heavy-traffic limit. Paralleling Theorem 2 and Corollary 1 of [10] for staffing multi-server queues, Theorem 4.3 shows that no control that asymptotically stabilizes the queue length in this heavy-traffic regime can simultaneously stabilize the virtual waiting time.

In §5 we consider the special case of a periodic arrival rate function in more detail. After Theorem 5.1 formalizes the notion of a periodic steady state, Theorem 5.2 establishes a periodic heavy-traffic limit for the waiting times of successive arrivals. As in [9] for multi-server queues, this illustrates a nearly periodic situation in which the limit depends on the order of the two iterated limits as $n \to \infty$ and $t \to \infty$.

Finally, in §7 we consider the square-root service-rate control in (1.2) that is an analog of the square-root capacity allocation formula in Kleinrock [7] and the square-root staffing formula in

[3, 10, 23]. We draw conclusions in §8.

2 The Model

For our results, we exploit a special composition construction of the arrival and service processes in order to obtain a general $G_t/G_t/1$ model. This is without loss of generality for the $M_t/M_t/1$ model, but is a restriction more generally.

In particular, we assume that the arrival process is defined by the composition

$$A(t) \equiv N_a(\Lambda(t)) = N_a(\int_0^t \lambda(s)\, ds), \quad t \geq 0, \tag{2.1}$$

where N_a is a rate-1 stochastic counting process satisfying a functional strong law of large numbers (FSLLN) and a functional central limit theorem (FCLT), i.e.,

$$\bar{N}_{a,n} \rightarrow e \quad \text{and} \quad \hat{N}_{a,n} \Rightarrow c_a B_a \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \rightarrow \infty, \tag{2.2}$$

with

$$\bar{N}_{a,n}(t) \equiv n^{-1} N_a(nt) \quad \text{and} \quad \hat{N}_{a,n}(t) \equiv n^{-1/2}[N_a(nt) - nt], \quad t \geq 0, \tag{2.3}$$

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 [21], while Λ is a deterministic cumulative arrival rate function, satisfying

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0, \tag{2.4}$$

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

$$\bar{\lambda} \equiv \lim_{t \rightarrow \infty} t^{-1} \Lambda(t). \tag{2.5}$$

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 (2.1) 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. More generally, the composition construction is convenient for constructing non-Markov counting processes with time-varying rates that satisfy FSLLN's and FCLT's; see §4.4 of [21] and [11]. This 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, 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.

As indicated in §1, we specify the random service requirements of successive customers separately from the service rate, which is deterministic and subject to control. For the first seven 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 (1.1). 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, independent of N_a, satisfying an FSLLN and an FCLT, i.e.,

$$\bar{N}_{s,n} \to e \quad \text{and} \quad \hat{N}_{s,n} \Rightarrow c_s B_s \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{2.6}$$

where

$$\bar{N}_{s,n}(t) \equiv n^{-1} N_s(nt) \quad \text{and} \quad \hat{N}_{s,n}(t) \equiv n^{-1/2}[N_s(nt) - nt], \quad t \geq 0, \tag{2.7}$$

with B_s being a standard BM, necessarily independent of B_a.

As usual, the queue length process can be defined as

$$Q(t) \equiv A(t) - D(t), \quad t \geq 0, \tag{2.8}$$

where $D(t)$ is the total number of departures in the interval $[0, t]$. We understand $D(t)$ to satisfy

$$D(t) \equiv N_s(\int_0^t \mu(s) 1_{\{Q(s)>0\}}\, ds) = N_s(\int_0^t (\lambda(s)/\rho) 1_{\{Q(s)>0\}}\, ds), \quad t \geq 0, \tag{2.9}$$

where 1_A is the indicator function, equal to 1 on A and 0 otherwise. Note that Q and D in (2.8) and (2.9) are defined in terms of each other. However, as in Lemma 2.1 of [14], there is a unique solution, as can be proved by induction on the successive events in the processes A and S.

2.1 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 (1.1) and the square-root service-rate control in (1.2), 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 (1.1) or (1.2). 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

$$B_k \quad = \quad D_{k-1} \vee A_k, \quad D_k = B_k + V_k \quad \text{and} \quad W_k = A_k - B_k, \quad k \geq 1, \tag{2.10}$$

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

$$S_k = \int_{B_k}^{B_k+V_k} \mu(s)\, ds, \quad k \geq 1. \tag{2.11}$$

If we let

$$M(t) \equiv \int_0^t \mu(s)\, ds, \quad t \geq 0, \tag{2.12}$$

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,

$$V_k = M^{-1}(S_k + M(B_k)) - B_k, \quad k \geq 1. \tag{2.13}$$

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.

Since the service-time formula (2.13) is somewhat complicated, it is helpful to have a useful practical approximation. That is achieved by employing local linear Taylor approximations

$$M(t + s) \approx M(t) + \mu(t)s \quad \text{and} \quad M^{-1}(t + s) = M^{-1}(t) + \frac{s}{\mu(M^{-1}(t))}, \tag{2.14}$$

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 (2.14). When we apply the Taylor approximation in (2.13), regarding S_k as a small perturbation about $M(B_k)$, we get

$$V_k = M^{-1}(M(B_k)) + \frac{S_k}{\mu(M^{-1}(M(B_k)))} - B_k = \frac{S_k}{\mu(B_k)}. \tag{2.15}$$

The approximation in (2.15) corresponds to assuming that each customer's service rate does not change during service. In particular, the service rate for each customer is the constant rate operating at the time the customer starts service. In the actual model, the service rate may keep changing, but this seems to be a reasonable approximation. Indeed, this could be the model assumption. Under heavy-traffic conditions, the difference will be negligible.

2.2 Time Transformation of Stationary Model

We now show that, with the rate-matching service-rate control in (1.1), we can circumvent the construction of the service times in (2.13) in order to deduce some important structure. (With this approach, we do *not* use the approximation in (2.15).) An important consequence of the composition construction in (2.1)-(2.9) 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

$$Q_1(t) \equiv A_1(t) - D_1(t), \quad t \geq 0, \tag{2.16}$$

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

$$D_1(t) \equiv N_s(\int_0^t \mu_1(s) 1_{\{Q_1(s)>0\}}\, ds) = N_s(\int_0^t \rho^{-1} 1_{\{Q_1(s)>0\}}\, ds), \quad t \geq 0. \tag{2.17}$$

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 2.1 (*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,*

$$(A(t), D(t), Q(t)) = (A_1(\Lambda(t)), D_1(\Lambda(t)), Q_1(\Lambda(t))), \quad t \geq 0. \tag{2.18}$$

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 (2.8) and (2.9) ((2.16) and (2.17)). 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 (2.9), we express D as

$$\begin{aligned} D(t) &= N_s(\int_0^t \rho^{-1}\lambda(s)1_{\{Q(s)>0\}}\, ds), \quad t \ge 0, \\ &= N_s(\int_0^{\Lambda(t)} \rho^{-1}1_{\{Q(\Lambda^{-1}(u))>0\}}\, du), \quad t \ge 0, \\ &= N_s(\int_0^{\Lambda(t)} \rho^{-1}1_{\{Q_1(u)>0\}}\, du) = D_1(\Lambda(t)), \quad t \ge 0, \end{aligned} \tag{2.19}$$

as claimed, where we have used $Q = Q_1 \circ \Lambda$ in the third step. As in the definitions (2.8) and (2.9), 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 (2.19). ■

3 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

$$\bar{d}_\rho(t) \equiv \frac{\int_0^t \lambda(s)1_{\{Q(s)>0\}}\, ds}{\Lambda(t)}. \tag{3.1}$$

In (3.1) 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

$$\bar{U}_{1,\rho}(t) \equiv t^{-1}\int_0^t 1_{\{W_1(s)>0\}}\, ds \equiv t^{-1}\int_0^t 1_{\{Q_1(s)>0\}}\, ds. \tag{3.2}$$

Theorem 3.1 (*stabilizing the average delay probability*) *Under the conditions above,*

$$\bar{U}_{1\rho}(t) \to \rho \quad \textit{and} \quad \bar{d}_\rho(t) \to \rho \quad \textit{in} \quad \mathbb{R} \quad \textit{as} \quad t \to \infty \tag{3.3}$$

for $\bar{U}_{1\rho}(t)$ in (3.2) *and $\bar{d}_\rho(t)$ in* (3.1).

To prove Theorem 3.1, 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 $\bar{S}_{1,n}(t) \equiv n^{-1}S_1(nt)$, $t \ge 0$.

Lemma 3.1 (*preliminary FSLLN's*) *Under the conditions above, the processes A and S_1 satisfy the FSLLN's*

$$\bar{A}_n \to e \quad \textit{and} \quad \bar{S}_{1,n} \to \rho^{-1} e \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad n \to \infty \quad \textit{w.p.1} \tag{3.4}$$

and the associated SLLN's

$$t^{-1} A(t) \to 1 \quad \textit{and} \quad t^{-1} S_1(t) \to \rho^{-1} \quad \textit{in} \quad \mathbb{R} \quad \textit{as} \quad t \to \infty \quad \textit{w.p.1} \tag{3.5}$$

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 [21]. Thus, the limit in (2.5) is equivalent to the stronger limit $\bar{\Lambda}_n \to e$ in $\mathcal{D}$ as $n \to \infty$, 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 $(x \circ y)(t) \equiv x(y(t))$: $\bar{A}_n = \bar{N}_{a,n} \circ \bar{\Lambda}_n \to e \circ e = e$, i.e., $\bar{A}_n(t) = \bar{N}_{a,n}(\bar{\Lambda}_n(t))$, $t \geq 0$; see §13.2 of [21]. Similarly, $\bar{S}_{1,n} = \bar{N}_{s,n} \circ \rho^{-1} e \to e \circ \rho^{-1} e = \rho^{-1} e$. 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 3.1. 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; see [16, 19]. The limit in (3.3) 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 3.1 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 (2.19) to obtain

$$\bar{d}_\rho(t) = \frac{\int_0^{\Lambda(t)} 1_{\{Q_1(u) > 0\}} \, du}{\Lambda(t)}. \tag{3.6}$$

Since $\Lambda(t) \to \infty$ as $t \to \infty$, 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 $t \to \infty$ whenever there is a proper steady-state distribution for $Q_1(t)$ as $t \to \infty$.

Theorem 3.2 (*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 $t \to \infty$, where*

$P(Q_1(\infty) < \infty) = 1$, *then also*

$$Q(t) \Rightarrow Q_1(\infty) \quad in \quad \mathbb{R} \quad as \quad t \to \infty, \tag{3.7}$$

and

$$P(W(t) > 0) = P(Q(t) \geq 1) \to \rho \quad as \quad t \to \infty. \tag{3.8}$$

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 $\Lambda^{-1}(t) \to \infty$ as $t \to \infty$, $Q(\Lambda^{-1}(t)) \Rightarrow Q_1(\infty)$ as $t \to \infty$, which directly implies that $Q(t) \Rightarrow Q_1(\infty)$ as $t \to \infty$ 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 (3.8). ■

4 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 2.1, we do *not* have $W(t) \stackrel{\mathrm{d}}{=} W_1(\Lambda(t))$, where $\stackrel{\mathrm{d}}{=}$ means equal in distribution. Unfortunately, the virtual waiting time is more complicated. We can write

$$P(W(t) > w) = \sum_{k=1}^{\infty} P(W(t) > w | Q(t) = k) P(Q(t) = k), \tag{4.1}$$

where

$$P(W(t) > w | Q(t) = k) = P(\inf \{u \geq 0 : D(t+u) - D(t) \geq k\} > w). \tag{4.2}$$

Theorem 3.2 shows that $Q(t)$ approaches a steady-state limit as $t \to \infty$ in considerable generality, but, because of the first passage time structure in (4.2), the conditional probability in (4.2) 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 (1.1). Afterwards, we establish a heavy-traffic limit theorem.

4.1 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 [21] 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. 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

$$\begin{aligned} A_{1,k} &\equiv U_{1,1} + \cdots + U_{1,k}, \\ W_{1,k} &\equiv [W_{1,k} + V_{1,k} - U_{1,k-1}]^+, \\ D_{1,k} &\equiv A_{1,k} + W_{1,k} + V_{1,k}, \quad k \geq 1, \end{aligned} \tag{4.3}$$

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.,

$$\begin{aligned} A_1(t) &\equiv \max\{k \geq 0 : A_{1,k} \leq t\}, \\ D_1(t) &\equiv \max\{k \geq 0 : D_{1,k} \leq t\}, \\ Q_1(t) &\equiv A_1(t) - D_1(t), \quad t \geq 0. \end{aligned} \tag{4.4}$$

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

$$W_1(t) \equiv W_{1,A_1(t)} + V_{1,A_1(t)} - (t - A_{1,A_1(t)}), \quad t \geq 0. \tag{4.5}$$

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 [21]. 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,

$$(A_k, D_k, C_k, Q_k) \equiv (\Lambda^{-1}(A_{1,k}), \Lambda^{-1}(D_{1,k}), \Lambda^{-1}(C_{1,k}), Q_{1,k}), \quad k \geq 1, \tag{4.6}$$

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

$$C(t) \equiv \max\{k \geq 0 : C_k \leq t\} \quad \text{and} \quad Q(t) \equiv Q_{C(t)}, \quad t \geq 0. \tag{4.7}$$

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

$$D(t) \equiv \max\{k \geq 0 : D_k \leq t\}, \quad t \geq 0. \tag{4.8}$$

Theorem 4.1 (*constructing the virtual waiting time*) *The virtual waiting time $W(t)$ can be represented as*

$$W(t) = \Lambda_t^{-1}(W_1(\Lambda(t)), \quad t \geq 0, \tag{4.9}$$

where Λ_t^{-1} is the inverse of

$$\Lambda_t(v) = \Lambda(t+v) - \Lambda(t), \quad v \geq 0 \quad \textit{and} \quad t \geq 0, \tag{4.10}$$

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

Proof. From (4.1) and (4.2),

$$\begin{aligned} W(t) &\equiv \inf\{u \geq 0 : D(t+u) - D(t) = Q(t)\} \\ &= \inf\{u \geq 0 : D_1(\Lambda(t+u)) - D_1(\Lambda(t)) = Q_1(\Lambda(t))\}, \quad t \geq 0, \end{aligned} \tag{4.11}$$

while

$$W_1(\Lambda(t)) = \inf\{v \geq 0 : D_1(\Lambda(t)+v) - D_1(\Lambda(t)) = Q_1(\Lambda(t))\}. \tag{4.12}$$

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

$$W_1(\Lambda(t)) = \Lambda(t+W(t)) - \Lambda(t) = \Lambda_t(W(t)), \quad t \geq 0, \tag{4.13}$$

for Λ_t defined in (4.10) above or, equivalently,

$$W(t) = \Lambda_t^{-1}(W_1(\Lambda(t))), \quad t \geq 0. \quad \blacksquare \tag{4.14}$$

We can use Theorem 4.1 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 4.1 (*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*

$$E[W(t)] = \int_0^\infty e^{-(1-\rho)\Lambda_t(x)/\rho}\, dx. \tag{4.15}$$

Proof. First the associated stationary $G/G/1$ model is $M/M/1$ with arrival rate 1 and service rate $1/\rho$, so that $P(W_1(t) > x) = e^{-(1-\rho)x/\rho}$ for large t. Next use the tail integral formula for the mean with (4.9) to write

$$\begin{aligned} E[W(t)] &= \int_0^\infty P(W(t) > x\, dx = \int_0^\infty P(\Lambda_t^{-1}(W_1(\Lambda(t)) > x\, dx \\ &= \int_0^\infty P((W_1(\Lambda(t)) > \Lambda_t(x)\, dx = \int_0^\infty e^{-(1-\rho)\Lambda_t(x)/\rho}\, dx. \end{aligned} \tag{4.16}$$

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 $E[W(t)] = \rho/(1-\rho)$. ■

4.2 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 increase toward 1; e.g., see Chapters 5 and 9 of [21]. We start by constructing a 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 $n = (1-\rho)^{-2}$ and we scale space by $n^{-1/2} = (1-\rho)$. 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 $\{\lambda_n : n \ge 1\}$ indexed by n, with each being continuous and strictly positive. Let associated scaled arrival rate functions and cumulative arrival rate functions be defined by

$$\bar{\lambda}_n(t) \equiv \lambda_n(nt) \quad \text{and} \quad \bar{\Lambda}_n(t) \equiv n^{-1}\Lambda_n(nt), \quad t \ge 0 \quad \text{and} \quad n \ge 1, \tag{4.17}$$

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

$$\tilde{\Lambda}_{n,t}(u) \equiv n^{-1/2}[\Lambda_n(nt + u\sqrt{n}) - \Lambda_n(nt)], \quad t \ge 0 \quad \text{and} \quad n \ge 1. \tag{4.18}$$

We assume that these scaled functions have the limits

$$\bar{\lambda}_n \to \lambda_f \quad \text{and} \quad \bar{\Lambda}_n \to \Lambda_f \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty \tag{4.19}$$

and

$$\tilde{\Lambda}_{n,t}(u) \to \lambda_f(t)u \quad \text{as} \quad n \to \infty \tag{4.20}$$

uniformly in t and u over bounded subintervals of $[0, \infty)$, where λ_f is a continuous and strictly positive. To be consistent with §2, 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

$$\hat{\Lambda}_n(t) \equiv n^{-1/2}[\Lambda_n(nt) - n\Lambda_f(t)], \quad t \geq 0 \quad \text{and} \quad n \geq 1, \tag{4.21}$$

and assume that

$$\hat{\Lambda}_n \to \Lambda_d \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{4.22}$$

where Λ_d is a continuous function, although the limit (4.22) will play no role in Theorem 4.2 below.

Since $\rho_n \to 1$ as $n \to \infty$, the service requirements remain $O(1)$ as $n \to \infty$. The time scaling makes the arrival rates and service rates be of order $O(n)$ as $n \to \infty$ means that we look over large time intervals, but the arrival rate and service rate remain $O(1)$ as $n \to \infty$, so that the service times are also $O(1)$ as $n \to \infty$. 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 $n \to \infty$, 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 $\dot{\lambda}_n(t) \approx \lambda_f(t/n)/n$. Thus, the arrival rate changes more slowly as n increases. That makes the model tends 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 (4.17)-(4.22) and the interpretation above.

Example 4.1 (*a sinusoidal example*) To illustrate, we start with the limit arrival rate function $\lambda_f(t)$ and proceed backwards to construct the sequence of arrival rate functions with this limit, using the usual scaling. Let $\lambda_f(t) \equiv 1 + \beta \sin(\gamma t)$ for $0 < \beta < 1$ and $\gamma > 0$ and 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 $\lambda_n(t) \equiv \lambda_f(t/n)$ 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

$$\tilde{\Lambda}_{n,t}(u) = \sqrt{n}[\Lambda_f(t + u/\sqrt{n}) - \Lambda_f(t)] \to \lambda_f(t)u \tag{4.23}$$

as $n \to \infty$ uniformly in t and u, by the definition of a derivative, consistent with the assumptions in (4.19) and (4.20).

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

$$\Lambda_n(t) \equiv n\Lambda_f(t/n) + \sqrt{n}\Lambda_d(t/n). \tag{4.24}$$

With (4.24), we have

$$\bar{\Lambda}_n(t) = \Lambda_f(t) + n^{-1/2}\Lambda_d(t) \quad \text{and} \quad \hat{\Lambda}_n(t) = \Lambda_d(t) \tag{4.25}$$

so that again $\bar{\Lambda}_n \to \Lambda_f$ and $\hat{\Lambda}_n \to \Lambda_d$ in $\mathcal{D}$. Instead of (4.23), we now have

$$\tilde{\Lambda}_{n,t}(u) = \sqrt{n}[\Lambda_f(t + u/\sqrt{n}) - \Lambda_f(t)] + [\Lambda_d(t + u/\sqrt{n}) - \Lambda_d(t)] \tag{4.26}$$

so that, just as before, $\tilde{\Lambda}_{n,t}(u) \to \lambda_f(t)u$ as $n \to \infty$ 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 $n \to \infty$ is that the cycles in the periodic arrival rate function are of length $O(n)$, where $n = (1-\rho)^{-2}$. ■

As a consequence of (4.17)-(4.22), we have associated limits for the scaled arrival process. To state them, let

$$\begin{aligned} \bar{A}_n(t) &\equiv n^{-1}N_a(\Lambda_n(nt), \quad \hat{A}_n(t) \equiv n^{-1/2}[A_n(nt) - n\Lambda_f(t)] \quad \text{and} \\ & \qquad \tilde{A}_{n,t}(u) \equiv n^{-1/2}[A_n(nt + u\sqrt{n}) - A_n(nt)], \quad t \ge 0 \quad \text{and} \quad n \ge 1. \end{aligned} \tag{4.27}$$

Lemma 4.1 (*limits for the scaled arrival process*) *Under the scaling above, we have the FSLLN*

$$\bar{A}_n = \bar{N}_{a,n} \circ \bar{\Lambda}_n \to \Lambda_f \quad in \quad \mathcal{D} \quad as \quad n \to \infty \quad w.p.1, \tag{4.28}$$

the associated FCLT

$$\hat{A}_n = \hat{N}_{a,n} \circ \bar{\Lambda}_n + \hat{\Lambda}_n \Rightarrow B \circ \Lambda_f + \Lambda_d \quad in \quad \mathcal{D} \quad as \quad n \to \infty. \tag{4.29}$$

and

$$\tilde{A}_{n,t}(u) \to \lambda_f(t)u \quad as \quad n \to \infty \tag{4.30}$$

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 [21]. For (4.30), we use the fact that tightness associated with the weak convergence of $\hat{N}_{a,n}$ in (2.2) implies that

$$n^{-1/2}[N_a(nt + u\sqrt{n}) - N_a(nt)] \Rightarrow u \quad \text{as} \quad n \to \infty \tag{4.31}$$

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

$$\begin{aligned} n^{-1/2}[A_n(nt + \sqrt{n}u) - A_n(nt)] &= n^{-1/2}[N_a(\Lambda_n(nt + u\sqrt{n})) - N_a(\Lambda_n(nt))] \\ = n^{-1/2}[N_a(\Lambda_n(nt) &+ \lambda_f(t)u\sqrt{n} + o(\sqrt{n})) - N_a(\Lambda_n(nt))] \Rightarrow \lambda_f(t)u \end{aligned} \tag{4.32}$$

uniformly in t and u within finite time intervals. We use the convergence $\bar{\Lambda}_n \to \Lambda_f$ 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

$$\hat{Q}_{1,n}(t) \equiv n^{-1/2} Q_{1,n}(nt), \quad t \geq 0, \tag{4.33}$$

so that $\hat{Q}_n(t) = \hat{Q}_{1,n}(\bar{\Lambda}_n(nt))$, $t \geq 0$ by Theorem 2.1. Let $W_n(t)$ be the virtual waiting time at time t in model n and define the associated scaled processes

$$\hat{W}_n(t) \equiv n^{-1/2} W_n(nt), \quad t \geq 0. \tag{4.34}$$

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 4.2 (*heavy-traffic limit for the time-varying waiting time*) *Let the system start empty. Under the scaling assumptions above, including* (4.17)-(4.20),

$$(\hat{Q}_n, \hat{W}_n) \Rightarrow (\hat{Q}, \hat{W}) \quad in \quad \mathcal{D}^2 \quad as \quad n \to \infty, \tag{4.35}$$

where

$$\hat{W}(t) \equiv \hat{Q}(t)/\lambda_f(t) \quad \textit{and} \quad \hat{Q}(t) \equiv R(\Lambda_f(t); -1, c_a^2 + c_s^2), \quad t \ge 0. \tag{4.36}$$

As a consequence, for each $T > 0$,

$$\sup_{0 \le t \le T} \{\hat{W}_n(t) - (\hat{Q}_n(t)/\lambda_f(t))|\} \Rightarrow 0 \quad \textit{as} \quad n \to \infty \tag{4.37}$$

and, for each $x \ge 0$,

$$P(\hat{Q}_n(t) > x) \to e^{-2x/(c_a^2+c_s^2)} \quad \textit{and} \quad P(\lambda_f(t)\hat{W}_n(t) > x) \to e^{-2x/(c_a^2+c_s^2)} \tag{4.38}$$

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

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 [21] and the continuity of the inverse function used in the first passage-time, as discussed in §§5.7, 13. 6 and 13.7 of [21]. The essential argument follows §5.4 of the Internet Supplement of [21], drawing on Theorem 13.7.4 of [21], but we will give a direct proof.

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

$$\hat{A}_{1,n}(t) \equiv n^{-1/2}[N_a(nt) - nt], \quad t \ge 0 \quad \text{and} \quad n \ge 1 \tag{4.39}$$

and

$$\hat{S}_{1,n}(t) \equiv n^{-1/2}[N_s(nt/\rho_n) - nt] = n^{-1/2}[N_s(nt/\rho_n) - nt/\rho_n] + t \quad t \ge 0 \quad \text{and} \quad n \ge 1. \tag{4.40}$$

As a consequence,

$$(\hat{A}_{1,n}\hat{S}_{1,n}) \Rightarrow (B_a, B_s + e) \quad \text{in} \quad \mathcal{D}^2, \tag{4.41}$$

where B_a and B_b 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 [21] to obtain

$$\hat{Q}_{1,n} \Rightarrow R(\cdot) \equiv R(\cdot; -1, c_a^2 + c_s^2) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{4.42}$$

so that $\hat{Q}_n = \hat{Q}_{1,n} \circ \bar{\Lambda}_n \Rightarrow R(\Lambda_f(\cdot))$ in $\mathcal{D}$ as $n \to \infty$, by applying the continuous mapping theorem with the composition map without centering, as in §13.2 of [21].

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

$$\hat{W}_n(t) \quad \equiv \quad n^{-1/2}W_n(nt) = \inf\{u \ge 0 : D_n(nt + u\sqrt{n}) - D_n(nt) \ge Q_n(nt)\}$$

$$
\begin{aligned}
&= \inf\{u \geq 0 : n^{-1/2}[D_n(nt + u\sqrt{n}) - D_n(nt)] \geq n^{-1/2}Q_n(nt)\} \\
&= \inf\{u \geq 0 : \hat{D}_{n,t}(u) \geq \hat{Q}_n(t)\}, \qquad (4.43)
\end{aligned}
$$

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 $n \to \infty$; we will now show that $\hat{D}_{n,t}(u) \to u\lambda_f(t)$ 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

$$
n^{-1/2}[B_{1,n}(\Lambda_n(nt + u\sqrt{n}) - B_{1,n}(\Lambda_n(nt))] \to u\lambda_f(t) \qquad (4.44)
$$

uniformly in t and u in $[0,T]$ by applying condition (4.20) and the FCLT for $\hat{B}_{1,n}$ contained in Theorem 9.3.4 of [21]. From the assumed FCLT for N_s in (2.6), we obtain the desired convergence $\hat{D}_{n,t}(u) \to u\lambda_f(t)$ 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 (4.37), where we already have established that $\hat{Q}_n \Rightarrow \hat{Q}$ with the specified distribution in (4.36). The joint limit in (4.35) then follows by the convergence-together theorem, as in Theorem 11.4.7 of [21].

The last two limits in (4.38) follow immediately from (4.35) 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 $t \to \infty$. ■

Remark 4.1 (*the resulting approximation*) The limit in (4.38) 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 function λ and the traffic intensity ρ, using $\lambda(t) \equiv \lambda_\rho(t) \approx \lambda_f((1-\rho)^2 t)$, then we get

$$
E[Q_\rho(t)] \approx \frac{c_a^2 + c_s^2}{2(1-\rho)} \approx \frac{\rho(c_a^2 + c_s^2)}{2(1-\rho)}, \qquad (4.45)
$$

and

$$
E[W_\rho(t)] \approx \frac{c_a^2 + c_s^2}{2(1-\rho)\lambda(t)} = \frac{(c_a^2 + c_s^2)}{2(1-\rho)\rho\bar{\mu}(\lambda(t)/\bar{\lambda})} \approx \frac{\rho(c_a^2 + c_s^2)}{2(1-\rho)\bar{\mu}(\lambda(t)/\bar{\lambda})}, \qquad (4.46)
$$

where $\bar{\mu} = \bar{\lambda}/\rho$ is the limiting average of $\mu(t)$, which exists by (1.1) and (2.5). 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 [18] for a discussion of such refinements to direct heavy-traffic approximations. That final formula in (4.46) 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

$$W_\rho(t) \approx \frac{Q_\rho(t)}{\lambda(t)}, \quad t \geq 0. \quad \blacksquare \tag{4.47}$$

Remark 4.2 (*Application of Corollary 4.1*) 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 4.1 to obtain a limit for the mean waiting time consistent with Theorem 4.2 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 $\rho_n/(1-\rho_n) \sim \sqrt{n}$ as $n \to \infty$, so that $E[\hat{W}_{1,n}^*] \equiv E[W_{1,n}^*]/\sqrt{n} \to 1$ as $n \to \infty$. Then Corollary 4.1 implies that

$$E[\hat{W}_n(t)] \equiv E[W_n(nt)/\sqrt{n}] \to \frac{1}{\lambda_f(t)} \quad \text{as} \quad n \to \infty. \tag{4.48}$$

Paralleling (4.16), the reasoning is

$$\begin{aligned} E[\hat{W}_n(t)] &= \int_0^\infty P(W_n(nt) > x\sqrt{n})\, dx = \int_0^\infty P(\Lambda_{n,t}^{-1}(W_{1,n}^*) > x\sqrt{n})\, dx \\ &= \int_0^\infty P(\hat{W}_{1,n}^* > \tilde{\Lambda}_{n,t}(x))\, dx = \int_0^\infty e^{-\tilde{\Lambda}_{n,t}(x)/(1-(1/\sqrt{n}))}\, dx \\ &\quad \to \int_0^\infty e^{-\lambda_f(t)x}\, dx = \frac{1}{\lambda_f(t)} \quad \text{as} \quad n \to \infty. \quad \blacksquare \end{aligned} \tag{4.49}$$

We formalize the qualitative conclusion about the implications of time variability to be drawn from formula (4.46) in the following corollary.

Corollary 4.2 *In the heavy-traffic limit of Theorem* 4.2*, 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,*

$$\frac{\lambda^\downarrow}{\lambda^\uparrow} \leq \frac{E[W(t_1)]}{E[W(t_2)]} \leq \frac{\lambda^\uparrow}{\lambda^\downarrow} \quad \textit{for all} \quad t_1, t_2 \quad \textit{such that} \quad t_1 > t \quad \textit{and} \quad t_2 > t. \tag{4.50}$$

In applications we have a single model with a fixed traffic intensity ρ. The applied relevance of the heavy-traffic limit in Theorem 4.2 will depend on the limiting cumulative rate function Λ_f in (4.19). 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 4.1, so we return to that example.

Example 4.2 (*the sinusoidal example without time scaling*) We now return to Example 4.1 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 $\bar{\Lambda}_n^{no}(t) = n^{-1}\lambda_f(nt) \to t$ in $\mathcal{D}$ as $n \to \infty$ and $\tilde{\Lambda}_{n,t}^{no}(u) = n^{-1/2}[\Lambda_f(nt + u\sqrt{n}) - \Lambda_f(nt)] \to u\bar{\lambda} = u$ as $n \to \infty$ 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 so not 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

$$\Lambda_n^{\#}(t) \equiv n\Lambda_f(t), \quad t \geq 0, \tag{4.51}$$

without including the time scaling in Λ_n above. Then we still get the limits in (4.19) and (4.20), 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 (4.19) and (4.20), and thus also in Theorem 4.2, reveal no impact of the time variability. ■

Paralleling Theorem 2 and Corollary 1 of [10] 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 4.3 (*stabilizing both in heavy traffic*) *Let the system start empty. Let the scaling assumptions above apply, including* (4.17)-(4.20), *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* $n \to \infty$, *where* $\hat{Q}(t) \Rightarrow \hat{Q}(\infty)$ *as* $t \to \infty$ *with* $0 < E[\hat{Q}(\infty)] < \infty$. *Then*

$$(\hat{Q}_n, \hat{W}_n) \Rightarrow (\hat{Q}, \hat{W}) \quad \text{in} \quad \mathcal{D}^2 \quad \text{as} \quad n \to \infty, \tag{4.52}$$

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* $n \to \infty$ *and then* $t \to \infty$ *unless* $\lambda_f(t) \to \lambda_f(\infty)$ *as* $t \to \infty$.

Proof. We can apply the second half of the proof of Theorem 4.2. Given the assumed convergence $\hat{Q}_n \Rightarrow \hat{Q}$ in $\mathcal{D}$ as $n \to \infty$, we can apply the tightness that follows from this convergence to deduce that

$$n^{-1/2}[Q_n(nt + u\sqrt{n}) - Q_n(nt)] \Rightarrow 0 \quad \text{as} \quad n \to \infty \tag{4.53}$$

uniformly in t and u over finite intervals. Combined with the limit for $\tilde{A}_{n,t}$ in (4.30), (4.53) implies that

$$\tilde{D}_{n,t}(u) \equiv n^{-1/2}[D_n(nt + u\sqrt{n}) - D_n(nt)] \Rightarrow \lambda_f(t)u \quad \text{as} \quad n \to \infty \tag{4.54}$$

uniformly in t and u over finite intervals. Thus the limit(4.52) and the subsequent results hold by the proof of Theorem 4.2. ▪

5 A Periodic Arrival Rate Function

Let us now consider the special case of a periodic arrival rate function λ with period c; see [6, 15] for background. 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

$$\{(A_1(t+s) - A_1(s), D_1(t+s) - D_1(s), Q_1(t+s)) : t \ge 0\} \Rightarrow \{(A_1^*(t), D_1^*(t), Q_1^*(t) : t \ge 0\} \tag{5.1}$$

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

5.1 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 5.1 (*periodic steady state*) *If λ is periodic with period c and* (5.1) *holds, then*

$$\begin{aligned} &\{(A(t+kc) - A(kc), D(t+kc) - D(kc), Q(t+kc), W(t+kc)) : t \ge 0\} \\ &\quad \Rightarrow \{(A^*(t), D^*(t), Q^*(t), W^*(t) : t \ge 0\} \quad \textit{in} \quad \mathcal{D}^4 \quad \textit{as} \quad k \to \infty, \end{aligned} \tag{5.2}$$

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 $\{(A^*(t+kc) - A^*(kc), D^*(t+kc) - D^*(kc), Q^*(t+kc)), W^*(t+kc) : t \ge 0\}$ in $\mathcal{D}^4$ is independent of k.

Proof. With the assumptions, Theorem 2.1 implies that (5.2) holds for the triple (A, D, Q). Then (4.1) and (4.2) imply that the same is true for W. (Theorem 4.2 and and (4.46) 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 $k \to \infty$ 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.,

$$P(W_{n,\infty} > w) = \frac{\int_0^{nc} \lambda_n(s) P(W_n^*(s) > w)\, ds}{nc\bar{\lambda}}; \tag{5.3}$$

See Proposition A1 in the Appendix of [12].

However, Theorem 4.2 provides a heavy-traffic limit as $n \to \infty$ for the integrand in (5.3), which is independent of the time argument s. Hence, we see that the limit in (4.38) should apply to $\hat{W}_{n,\infty}$ as well as $\lambda_f(t)\hat{W}_n(t)$; i.e., paralleling (4.46), we have the associated heavy-traffic approximation

$$E[W_{\rho,\infty}] \approx \frac{\rho(c_a^2 + c_s^2)}{2(1-\rho)\bar{\mu}}. \tag{5.4}$$

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 (5.3) should hold. If we consider possible arrival times, then we should focus on $E[W(t)]$, which is periodic.

5.2 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 §5 above, which has a periodic limit. This shows that the order of the two limits as $t \to \infty$ and as $n \to \infty$ cannot be interchanged, just as for the multi-server queues with deterministic service times in [9]. In the heavy-traffic limit, the arrival times occur at fixed places within the cycle.

To state the limit, let

$$\hat{Z}_n(t) \equiv n^{-1/2} W_{n,\lfloor nt \rfloor}, \quad t \geq 0 \quad \text{and} \quad n \geq 1, \tag{5.5}$$

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

Theorem 5.2 (*heavy-traffic limit for the waiting times of successive arrivals*) *Let the system start empty. Under the scaling assumptions above, including* (4.17)-(4.20),

$$\hat{Z}_n \Rightarrow \hat{Z} = \hat{W} \circ \Lambda_f^{-1} \quad in \quad \mathcal{D} \quad as \quad n \to \infty, \tag{5.6}$$

where $\hat{Z}_n$ *is defined in* (5.5) *and*

$$\hat{Z}(t) \stackrel{\mathrm{d}}{=} \frac{R(t; -1, c_a^2 + c_s^2)}{\lambda_f(\Lambda_f^{-1}(t))}, \quad t \geq 0, \tag{5.7}$$

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 $n \to \infty$ 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, $\bar{A}_n^{-1} \to \Lambda_f^{-1}$ in $\mathcal{D}$ as $n \to \infty$; see §13.6 of [21]. Hence, by the continuous mapping theorem with composition, we have the claimed (5.6). We then obtain (5.7) from (4.36).

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 (5.4) can be obtained from (5.7) if we consider t sufficiently large that we replace the RBM with its exponential steady-state distribution and we replace $\lambda_f(\Lambda_f^{-1})$ in the denominator by its long-run average $\bar{\lambda}_f = 1$. As in [9], the periodic heavy-traffic limit shows the possibility of nearly periodic behavior for systems in practice.

6 Simulation Experiments for the Rate-Matching Control

Theorems 2.1 and 5.1 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 2.1 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 5.1 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 $(Q^*(t+c), W^*(t+c)) \stackrel{\mathrm{d}}{=} (Q^*(t), W^*(t))$ 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 (4.6) requires that we be able to compute Λ^{-1}, while Theorem 4.1 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

$$\Lambda^{-1}(kc) = \Lambda(kc) = kc \quad \text{for all} \quad k \geq 1. \tag{6.1}$$

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

$$\Lambda^{-1}(kc + t) = kc + \Lambda^{-1}(t), \quad 0 \leq t \leq c. \tag{6.2}$$

Hence, we could compute, table 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 4.1 revisited*) *Suppose that*

$$\lambda(t) = 1 + \beta \sin{(\gamma t)}, t \geq 0, \tag{6.3}$$

so that

$$\Lambda(t) = t - (\beta/\gamma)(\cos{(\gamma t)} - 1), t \geq 0, \tag{6.4}$$

and

$$\Lambda_t(u) = u - (\beta/\gamma)(\cos{(\gamma(t+u))} - \cos{(\gamma t)}), t \geq 0, \tag{6.5}$$

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

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

But how do we calculate $\Lambda^{-1}(t)$ and $\Lambda_t^{-1}(u)$? ■

We now observe that the heavy-traffic scaling of time and space in §4 makes the approximate simulation method in (2.15) more appropriate as n increases. As observed just prior to Example 4.1, the service requirements and service times are of order $O(1)$ as $n \to \infty$. 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 $n \to \infty$. This provides strong support for approximation (2.15), showing that it is asymptotically correct as $n \to \infty$.

7 A Control to Stabilize the Expected Waiting Time

In this section we examine the square-root service-rate control in (1.2) as a way to stabilize the waiting time instead of the queue length. 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.,

$$s(t) = m(t) + \beta\sqrt{m(t)}, \tag{7.1}$$

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 β in (7.1) in order to focus on a particular performance measure; see [3, 10, 23] and references therein.

An analog in our setting is the square-root service-rate control in (1.2). From Theorem 4.2, we see that if we are interested in stabilizing the expected virtual waiting time $E[W(t)]$, the rate matching control in (1.1) overstaffs when the arrival rate $\lambda(t)$ is relatively large and understaffs when it is relatively low. Formula (1.2) acts to correct that bias. We now show that the squar-root service-rate control in (1.2) is asymptotically optimal with respect to an appropriate criterion with an appropriate time scaling.

To establish this positive asymptotic result, we exploit connections to the earlier work on optimal capacity allocation in [2, 7, 8, 22] 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 $\sum_{i=1}^{n} r_i\mu_i \le B$, 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 is the product-form steady-state distribution for open queueing networks, under which the n queues are mutually independent in steady state, so that 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 (7.1). 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 (scv, variance divided by the square of the mean) of an interarrival times and a service time (which is exact for $M/M/1$), and the product form is approximately valid, the optimal allocations are

$$\mu_i = \lambda_i + \frac{(B-\Lambda)\sqrt{\lambda_i r_i (c_{a,i}^2 + c_{s,i}^2)}}{\sum_{j=1}^n \sqrt{\lambda_j r_j (c_{a,j}^2 + c_{s,j}^2)}}. \tag{7.2}$$

We make three initial observations: First, if $r_i(c_{a,i}^2 + c_{s,i}^2)$ is independent of i, then (7.2) looks more like (7.1). Second, we note that the theoretical bases for (7.1) and (7.2) are quite different. Formula (7.1) 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 (7.2) 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 (7.2) 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 (1.1) would be optimal.

The nice analysis leading to (7.2) 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

$$\frac{1}{m}\sum_{k=1}^{m} E[W(kT/m)] \tag{7.3}$$

and we have the service rate constraint

$$\int_0^T \mu(t)\,dt > \rho^{-1}\int_0^T \lambda(t)\,dt \quad \text{for} \quad 0 < \rho < 1. \tag{7.4}$$

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, 4, 13, 20]. Assuming that the PSA is valid as an approximation, then (7.1) is optimal. We state the asymptotic result for Markovian systems that follows from [20].

Theorem 7.1 (*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 control subject to the constraint that $\mu(t) > \lambda(t)$ for all t, $0 \le t \le T$. 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* (7.3) *subject to the constraint in* (7.4)*, then the PSA control in* (1.2) *is asymptotically optimal as $n \to \infty$.*

Proof. We combine the asymptotic result in [20], 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 [7]. ∎

7.1 Simulation Experiments for the Square-Root Control

Here we simulate only the $M_t/M_t/1$ model. We generate the arrival counting process A as an NHPP with arrival rate $\lambda(t)$ at time t. We generate a potential service process $S(t)$ as another NHPP with service rate function $\mu(t) = \lambda(t) + \beta\sqrt{\lambda(t)}$. W let a net input process bee defined by

$$X(t) = A(t) - S(t), \quad t \ge 0. \tag{7.5}$$

We then define the qqueue length process by applying the one-dimensional reflection map, as in §14.5 of [21],

$$Q(t) = X(t) - \inf_{0 \le s \le t} \{X(s)\}, \quad t \ge 0. \tag{7.6}$$

After simulating this system, we see the arrival times and the waiting times of successive customers. We see the waiting times W_k that go with the arrival times A_k. We average the waiting times for all arrival times in a bin of the periodic cycle.

8 Conclusion

We have studied a general $G_t/G_t/1$ single-server queue with time-varying arrival rate function $\lambda(t)$ where the model of the random customer service requirements are specified but the deterministic service rate $\mu(t)$ is subject to control. The specific model specified in §2 involves a composition construction of the arrival and service processes in (2.1) and (2.9), starting from fixed stochastic processes N_a and N_s that satisfy a FSLLN and a FCLT.

In §§2-5 we studied the rate-matching service rate control in with $\mu(t) \equiv \lambda(t)/\rho$ for selected traffic intensity ρ, $0 < \rho < 1$. Theorem 2.1 shows that the composition construction makes the performance triple $(A(t), D(t), Q(t))$ a deterministic time transformation of the triple in a stationary model. As a consequence, Theorem 3.1 and Theorem 3.2 showed that the average delay probability as defined in (3.1) and the virtual delay probability $P(W(t) > 0)$ both converge to ρ as $t \to \infty$. In fact, Theorem 3.2 shows that the entire queue-length process typically converges to a proper steady-state distribution.

Nevertheless, the tail probability of delay remains time-dependent. Theorem 4.2 establishes a heavy-traffic limit showing that, for sufficiently large t, the virtual waiting time $W(t)$ can be approximated by an exponential random variable with a time-varying mean, which is inversely proportional to the relative arrival rate $\lambda(t)/\bar{\lambda}$, as in (4.46). Crucial for that limit is time scaling within the arrival rate function, as specified in (4.17)-(4.22). For a periodic arrival rate function, the periodic cycle should grow as $(1-\rho)^{-2}$ as $\rho \uparrow 1$. Theorem 4.3 shows that no control that asymptotically stabilizes the queue length in this heavy-traffic regime can simultaneously stabilize the virtual waiting time. Theorem 5.2 establishes a periodic heavy-traffic limit for the waiting times of successive arrivals, providing another example in which the limit depends on the order of the two iterated limits as $n \to \infty$ and $t \to \infty$.

Finally, in §7 we consider the square-root service-rate control in (1.2) that is an analog of the square-root staffing formula (7.1) for multi-server queues from [3, 10, 23] and the square-root capacity allocation formula (7.2) for Markovian open Jackson networks from Kleinrock [7]. Theorem 7.1 shows that it is optimal for $M_t/M_t/1$ models in the limit from [20] supporting the pointwise stationary approximation.

Acknowledgement

The author gratefully acknowledges the helpful comments of Columbia doctoral student Ni Ma and the support from NSF grants CMMI grant 1265070.

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] G. R. Bitran and S. Dasu. A review of open queueing network models of manufacturing systems. *Queueing Systems*, 12:95–134, 1992.

[3] M. Defraeye and I. Van Niewenhuyse. 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.

[4] L. V. Green and P. J. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.*, 37:84–97, 1991.

[5] B. He, Y. Liu, and W. Whitt. Stabilizing performance in nonstationary queues with non-Poisson arrivals. Columbia University, working paper, 2014.

[6] D. P. Heyman and W. Whitt. The asymptoic behavior of queues with time-varying arrival. *Journal of Applied Probability*, 21(1):143–156, 1984.

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

[8] L. Kleinrock. *Queueing Systems*, volume 2. Wiley, New York, 1976.

[9] Y. Liu and W. Whitt. Nearly periodic behavior in the the overloaded $G/D/S+GI$ queue. *Stochastic Systems*, 1(2):340–410, 2011.

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

[11] W. A. Massey and W. Whitt. Unstable asymptotics for nonstationary queues. *Math. Oper. Res.*, 19 (2):267–291, 1994.

[12] 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.

[13] 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.

[14] G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, 4:193–267, 2007.

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

[16] S. Stidham. A last word on $L = \lambda W$. *Oper. Res.*, 22:417–421, 1974.

[17] L. M. Wein. Capacity allocation in generalized Jackson networks. *Oper. Res. Letters*, 8:143–146, 1989.

[18] W. Whitt. Refining diffusion approximations for queues. *Oper. Res. Letters*, 1(5):165–169, 1982.

[19] W. Whitt. A review of $L = \lambda W$. *Queueing Systems*, 9:235–268, 1991.

[20] 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.

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

[22] R. W. Wolfe. *Stochastic Modeling and the Theory of Queues.* Prentice-Hall, Englewood Cliffs, NJ, 1989.

[23] 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.