Operations Research Letters 44 (2016) 796–800

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Heavy-traffic limits for a single-server queue leading up to a critical point

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

ARTICLE INFO

Article history:
Received 10 August 2016
Received in revised form
21 September 2016
Accepted 6 October 2016
Available online 14 October 2016

Keywords:
Heavy-traffic limits
Nonstationary queues
Queues with time-varying arrival rates
Onset of critical loading
Rush hour
Transition through saturation

ABSTRACT

We establish heavy-traffic limits for the arrival and workload processes in a single-server queue with a time-varying arrival-rate function. We establish limits at and before a critical point, the onset of critical loading, where the arrival-rate function approaches its critical value from below. We extend results by Newell (1968) and Mandelbaum and Massey (1995) and present alternative views of the interesting scaling constants that arise in these limits.

1. Introduction

The purpose of this paper is to extend, and contribute to a better understanding of, diffusion approximations and heavy-traffic limits for a single-server queue with time-varying arrival-rate function that were established by Newell [15–18] and Mandelbaum and Massey [11]; also see [12,8,13,4]. As in our recent paper [22] for queues with periodic arrival-rate functions, we develop these limits in the standard framework for heavy-traffic limits, as in [6,7,19,21].

In this paper we focus on one special case: the limiting behavior at and before an isolated critical point, at what is called the onset of critical loading in [11], where the arrival rate approaches the critical value from below. Thus, this paper relates to only Theorem 3.4 in [11]. For that result, we generalize the setting from $M_t/M/1$ to $G_t/G/1$ and give alternative developments, leading to alternative scaling and alternative interpretations of it.

In particular, we consider a single-server queue with unlimited waiting room having service times with mean 1, which fixes the time scale. We consider a sequence of models with an associated sequence of arrival processes having time-varying arrival-rate functions. We will establish heavy-traffic limits, which involve scaling time, so that we are looking at intervals over which many customers arrive and are served. We assume that there is an isolated critical point, which we take to be time 0. In particular, we assume that the arrival-rate function satisfies

$$\lambda(0) = 1 \quad \text{and} \quad \lambda(t) < 1 \quad \text{for } t < 0. \tag{1}$$

Moreover, for simplicity, we assume that λ is nondecreasing in t before time 0. (There is no mass of workload from the distant past contributing to the buildup of congestion at the critical point.) As observed in [15], the congestion at times $t < 0$ is less, often much less, than the steady-state distribution with the instantaneous traffic intensity $\rho(t) = \lambda(t)$, because the traffic intensity was previously at lower values.

The approaches in [15–18,11] are quite different from [6,7,19,21], even though they can be related. First, Newell [15–18] makes a direct diffusion process approximation and then analyzes the Fokker–Planck partial differential equation for the time-varying cumulative distribution function. The papers [15–17] are landmark contributions to queueing theory, but they are challenging to understand, because they both develop the diffusion approximation and analyze it. It turns out that the diffusion process is the limiting diffusion in the heavy-traffic limits, with the key asymptotic properties captured by the parameters of that diffusion process. In contrast, as in [6,7,19,21], our approach emphasizes scaling, so it avoids looking directly at the detailed evolution of the diffusion process, but that remains to be done to calculate explicit approximations. In [23] we develop a robust queueing approach to calculate such explicit approximations.

The more modern [11] exploits strong approximations. For scaling, it starts with an initial arrival-rate function and expands

E-mail address: ww2040@columbia.edu.

http://dx.doi.org/10.1016/j.orl.2016.10.005
0167-6377/© 2016 Elsevier B.V. All rights reserved.

time about each fixed point in that arrival function. As a consequence of that expansion, the relevant long-time behavior of the arrival-rate function is determined by the local behavior of the initial function, as exposed by a Taylor-series expansion, which requires extra regularity assumptions. This approach is helpful for analyzing highly structural mathematical models, as we illustrate with a sinusoidal example in Example 4.1.

Here is how this paper is organized. In Section 2 we formulate our arrival process model. In Section 3 we establish a heavy-traffic functional central limit theorem (FCLT) for the arrival process. In Section 4 we show how that arrival process FCLT can be recast in a setting in which we expand time within a fixed initial arrival-rate function, as in [11].

Motivated by the desire to develop an approach that is helpful for applications, in Section 5 we show how the heavy-traffic FCLT can be expressed in yet a different way using drift scaling, which is in the spirit of §4.3 of [20]. We think that it is natural to first fit the drift function and then afterwards choose the appropriate time and space scaling that goes with that drift function. That approach directly yields the appropriate space scaling and the diffusion process approximation, which the FCLT's imply should perform well when the drift constant is suitably small. Given a FCLT for the arrival process, corresponding heavy-traffic FCLT's follow for the standard queueing processes using the continuous mapping approach in [21]. In Section 6 we illustrate by establishing the heavy-traffic FCLT for the workload process. Finally, we briefly draw conclusions in Section 7.

2. The arrival process model

As in [22], we construct the arrival process A by composing a process assumed to satisfy a FCLT and a deterministic cumulative arrival-rate function. In particular, we let the stochastic arrival counting processes defined by

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

where N is a stochastic counting process satisfying a FCLT, i.e.,

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt] \Rightarrow c_a B_a(t) \quad \text{in } \mathcal{D} \text{ as } n \to \infty, \tag{3}$$

where $\Rightarrow$ denotes convergence in distribution in the function space $\mathcal{D}$ of right-continuous real-valued functions on the interval $[0, \infty)$ with left limits, as in [21], and B_a is a standard (drift 0, variance 1) Brownian motion (BM), while Λ is a cumulative arrival-rate function, satisfying $\Lambda(t) \equiv \int_0^t \lambda(s)\, ds$, $t \geq 0$, with λ being the arrival-rate function, which is assumed to be integrable over finite intervals.

The construction in (2) is convenient for constructing non-Markov nonstationary arrival processes. It was suggested in [14] and also used in [3,5,22,23]. However, it is important to recognize that, even though it allows very general stochastic processes N, including renewal processes and much more (see §4.4 of [21]), this model is highly structured, having all unpredictable stochastic variability associated with the process N, with its FCLT behavior captured by the single variability parameter c_a, while all the predictable deterministic variability associated with the deterministic arrival-rate function λ and its associated cumulative rate function Λ. More generally, we might contemplate a time-varying variability parameter. In the present context, if the process N is a renewal counting process, then c_a is the square root of c_a^2, the squared coefficient of variation (scv, variance divided by the square of the mean) of an interarrival time. From an engineering perspective, the tractability produced by reducing the impact of the stochastic variability to the single parameter c_a^2 may be essential for drawing useful conclusions about system performance.

Throughout this paper, we assume that the cumulative arrival-rate function Λ is deterministic, but it is significant that the results here can be extended to cover the case in which arrival-rate function is a stochastic process, which can be important in applications. For example, service system arrival process data often indicate overdispersion caused by day-to-day variation, as discussed in [9].

3. A conventional heavy-traffic FCLT for the arrival process

Given the composition representation of the arrival process in (2) and the assumed FCLT in (3), we can obtain a conventional FCLT for the arrival process A defined in (2), which involves scaling time by n and space by $1/\sqrt{n}$, and then letting $n \to \infty$, if we write

$$\hat{A}_n(t) \equiv n^{-1/2}[A_n(nt) - nt] \quad \text{and}$$
$$\hat{\Lambda}_n(t) \equiv n^{-1/2}[\Lambda_n(nt) - nt], \quad t \geq 0, \tag{4}$$

where

$$A_n(t) \equiv N(\Lambda_n(t)) \quad \text{and} \quad \Lambda_n(t) \equiv \int_0^t \lambda_n(s)\, ds, \quad t \geq 0, \tag{5}$$

and we make appropriate assumptions about the deterministic arrival-rate functions $\lambda_n(t)$, which requires that $\lambda_n(t)$ remain close to 1 for large time intervals about $t = 0$. As in [22], it is important that we scale time and space in the deterministic cumulative arrival rate functions Λ_n in (4).

We find it convenient to work in reverse time, because the workload process then can be represented as a simple supremum of the net input process; see Section 6. The reverse-time construction is discussed in [23], which develops a time-varying robust queueing approximation based on the supremum representation. Hence, we measure time backwards from time 0, so that $A(t)$ counts the number of arrivals in $[-t, 0]$. This section is devoted to establishing a FCLT for the process A. We remark that a FCLT also holds in forward time or in intervals $[t_1, t_2]$ with $t_1 < 0 < t_2$ by the same argument.

As a main example in our reverse-time framework, we focus on arrival-rate functions that decay in a power away from the critical point at time 0; i.e.,

$$\lambda_n(t) = 1 - c_n t^p, \quad t \geq 0, \tag{6}$$

for some real number $p \geq 0$. In comparison to [11] and Section 4, note that p is not restricted to being an integer, but $p = 1$ and $p = 2$ are especially interesting.

We emphasize that we are concerned with large time, i.e., time scaled by n as $n \to \infty$, so that we need to consider the scaled version

$$\lambda_n(nt) = 1 - c_n(nt)^p, \quad t \geq 0, \tag{7}$$

where we allow $n \to \infty$. Note that $p = 0$ in (6) and (7) corresponds to a constant arrival rate of $1 - c_n$, which produces a constant negative drift of $1 - c_n$. The following result covers this stationary model as a special case.

Theorem 3.1 (*Conventional FCLT for the Arrival Process*). *If, in addition to the FCLT for $\hat{N}_n$ in (3),*

$$\hat{\Lambda}_n \to \hat{\Lambda} \quad \text{in } \mathcal{D}, \tag{8}$$

for $\hat{\Lambda}_n$ defined in (4) and (5), then

$$\hat{A}_n \Rightarrow c_a B_a + \hat{\Lambda} \quad \text{in } \mathcal{D} \text{ as } n \to \infty. \tag{9}$$

Under assumption (6), the limit in (8) holds if and only if

$$c_n n^{(2p+1)/2} \to c \quad \text{as } n \to \infty,\ 0 < c < \infty, \tag{10}$$

in which case

$$\hat{\Lambda}(t) = \frac{-ct^{p+1}}{p+1}, \quad t \geq 0. \tag{11}$$

Proof. To obtain (9), we apply the standard argument for preservation of convergence with composition and centering, as in §13.3 of [21]. In particular, we write

$$\begin{aligned} \hat{A}_n &= n^{-1/2}[N(\Lambda_n(nt)) - nt] \\ &= n^{-1/2}[N(\Lambda_n(nt)) - \Lambda_n(nt)] + n^{-1/2}[\Lambda_n(nt) - nt] \\ &\Rightarrow c_a B_a + \hat{\Lambda}. \end{aligned}$$

To establish the limit of the first term in the second line, we use the limit $\bar{\Lambda}_n \Rightarrow e$ in $\mathcal{D}$, where $\bar{\Lambda}_n(t) \equiv n^{-1}\Lambda_n(nt)$ and $e(t) \equiv t, t \geq 0$, which is a consequence of (4) and (8). We can then apply Theorem 11.4.5 of [21] to obtain the joint convergence $(\hat{A}_n, \bar{\Lambda}_n, \hat{\Lambda}_n) \Rightarrow (\hat{A}, e, \hat{\Lambda})$ in $\mathcal{D}^3$. Then we can apply §13.3 of [21].

Turning to (10), we integrate (6) to get

$$\Lambda_n(t) = t - \frac{c_n t^{p+1}}{p+1}, \quad t \geq 0,$$

which implies that

$$\hat{\Lambda}_n(t) = \frac{-c_n n^{(2p+1)/2} t^{p+1}}{p+1}, \quad t \geq 0,$$

from which (10) and (11) follow directly. ■

Observe that, in contrast to [15–17], we obtain the insightful scaling constants without directly working with Brownian motion or the limiting diffusion process. (Here we obtain the required constants c_n in (6) and (7) for given n; we show how we can reverse this step in Section 5.) We exploit the scaling in the FCLT for N in (3). That logic provides a basis for generalizations, as we now indicate. This shows that there are many more possibilities than suggested by previous work.

Remark 3.1 (*Alternative Scaling and Limit Processes*). Analogs of Theorem 3.1 follow immediately if we replace condition (3) with alternative limit processes and scaling. For example, if the interarrival times are i.i.d. with a heavy-tailed distribution, having finite mean but an infinite variance, then under regularity conditions, the limit process would be a stable process and the spatial scaling for $\hat{N}_n$ in (3) would be by $n^{-1/\alpha}$ for $1 < \alpha < 2$, as in §6.3 and §4.5 of [21]. Because of the discontinuities in the limit process, the mode of convergence should involve the M_1 topology, as in Theorem 6.3.1 of [21]. The spatial scaling in $\hat{\Lambda}_n$ would need to be changed to $n^{-1/\alpha}$ to match the spatial scaling of $\hat{N}_n$ in (3), which in turn changes (10) in the presence of (6) and (7). Similar modifications cover strong dependence in the arrival process, as in §4.6 and §4.7 of [21].

Remark 3.2 (*Simple and Complex Models*). The power arrival-rate functions in (6) are appealing for their simplicity, but Theorem 3.1 applies much more generally. For the second half of Theorem 3.1 to yield useful approximations, we require that the arrival-rate function approximately satisfy (6) for large t. To see that the assumed convergence in (8) can hold much more generally, we give one example: suppose that, instead of (6),

$$\lambda_n(t) = 1 - c_{1,n}t^{p_1} - c_{2,n}t^{p_2}, \quad t \geq 0.$$

Then the limit in (8) holds if and only if

$$c_{j,n}n^{(2p_j+1)/2} \to c_j \quad \text{as } n \to \infty, \ 0 < c < \infty, \ \text{for } j = 1, 2,$$

in which case

$$\hat{\Lambda}(t) = -\frac{c_1 t^{p_1+1}}{p_1+1} - \frac{c_2 t^{p_2+1}}{p_2+1}, \quad t \geq 0.$$

Clearly, such complications in the limit function $\hat{\Lambda}$ make the limit process $\hat{A}$ harder to work with.

4. Expanding time within a smooth function

Motivated by [11], we now consider an alternative approach to the arrival-rate function, where we expand time in a fixed smooth function. By smooth, we mean that it is infinitely differentiable, so that we can construct Taylor series approximations. (That excludes non-integer p in (6).) In particular, given a smooth arrival-rate function λ satisfying (1), and maintaining our reverse-time perspective, we assume that

$$\lambda_n(t) \equiv \lambda(t/b_n), \quad t \geq 0, \tag{12}$$

where $b_n \to \infty$ as $n \to \infty$, so that we can expand in a Taylor series approximation

$$\lambda_n(t) = 1 + \frac{\lambda^{(k)}(0)(t/b_n)^k}{k!} + o((1/b_n)^k) \quad \text{as } n \to \infty \tag{13}$$

for some integer $k \geq 1$, where $-\infty < \lambda^{(k)}(0) < 0$ by (1) and the reverse-time view. We keep the scaling constant b_n in (12) general so that we can adjust to fit in the framework of Theorem 3.1.

Theorem 4.1 (*Expanding Time within a Fixed Arrival-Rate Function*). *If, in addition to the FCLT for $\hat{N}_n$ in (3), (12) and (13) hold for an initial smooth function λ and a nonnegative integer k, then*

$$\Lambda_n(t) = t + \frac{\lambda^{(k)}(0)t^{k+1}}{b_n^k(k+1)!} + o((1/b_n)^k) \quad \textit{as } n \to \infty, \tag{14}$$

so that (8)–(11) hold for $p = k$ if and only if

$$b_n/n^{(2k+1)/2k} \to b, \quad 0 < b < \infty, \tag{15}$$

in which case $\hat{\Lambda}$ is as in (11) with $c = 1/b^k$.

Proof. From (14) and (4),

$$\hat{\Lambda}_n(t) = \frac{\lambda^{(k)}(0)(nt)^{k+1-(1/2)}}{b_n^k(k+1)!} + o((n/b_n)^k),$$

where $o((n/b_n)^k)$ is $o(n^{-1/2})$ by (15). ■

We think that this approach is appealing for mathematical simplicity and elegance, but it seems not so well suited for developing a suitable approximation to match data. This approach is appealing to treat structured mathematical models, as we illustrate with the next example.

Example 4.1 (*Sinusoidal Arrival-Rate Function*). Given the literature on queues with time-varying arrival-rate functions, e.g., [1,10,22], it is natural to focus on the sinusoidal case. To expand the time around the peak, we consider the cosine function over a half cycle, which is an even function, and thus invariant under time reversal. In particular, let

$$\lambda(t) = \beta\cos(t), \quad -\pi < t < \pi, \tag{16}$$

which makes $\lambda(t)$ decreasing away from its peak of β at time 0. Let $\Lambda(t) = 0$ before $-\pi$ in forward time.

Hence, recalling that $\cos(t) = 1 - t^2/2 + o(t)$ as $t \to 0$, from (12) and (13), we obtain

$$\lambda_n(t) = 1 - \frac{\beta(t/b_n)^2}{4} + o(1/b_n)^2 \quad \text{as } n \to \infty \tag{17}$$

for $-\pi < t < \pi$. Hence, (8)–(11) hold for $p = 2$; i.e., $c_n n^{5/2} \to c$ and $\hat{\Lambda}(t) = -ct^3/3, t \ge 0$. Thus the spatial scaling is by $O(n^{-1/5})$, which is consistent with [17,11].

5. Drift scaling

It is often insightful for applications of heavy-traffic limits to scale by the drift. For the stationary model, we can express the scaled queue length as

$$\hat{Q}_\delta(t) = \delta Q(\delta^{-2}t), \quad t \ge 0, \text{ for } \delta \equiv 1 - \rho, \tag{18}$$

e.g., see §4.3 of [20]. In our setting, we have mean service time 1 and thus a time-varying drift $1 - \lambda(t)$, $t \ge 0$. If we consider structured arrival-rate functions, then we can represent the drift using a constant. Thus, paralleling (6), we can write the drift as a power function of t determined by a constant $\delta > 0$ via

$$\lambda_\delta(t) = 1 - \delta t^p, \quad \text{so that } \Lambda_\delta(t) = t - \frac{\delta t^{p+1}}{p+1}, \ t \ge 0. \tag{19}$$

For applications, we think that it is natural to start by fitting the two parameters δ and p in the function in (19) and then consider the diffusion approximation determined by those parameters. The heavy-traffic limit theorems indicate that the quality of the direct approximation should improve as δ decreases.

To state the appropriate FCLT as $\delta \downarrow 0$ for fixed p, we use the associated δ-scaled processes.

$$\begin{aligned} \hat{\Lambda}_\delta(t) &\equiv \delta^{1/(2p+1)}[\Lambda_\delta(\delta^{-2/(2p+1)}t) - \delta^{-2/(2p+1)}t] \quad \text{and} \\ \hat{A}_\delta(t) &\equiv \delta^{1/(2p+1)}[A_\delta(\delta^{-2/(2p+1)}t) - \delta^{-2/(2p+1)}t] \\ &\text{for } A_\delta(t) \equiv N(\Lambda_\delta(t)), \end{aligned} \tag{20}$$

and $t \ge 0$.

Theorem 5.1 (*FCLT for the Arrival Process with Drift Scaling*). *If, in addition to the FCLT for $\hat{N}_n$ in (3), $\hat{\Lambda}_\delta \to \hat{\Lambda}$ in $\mathcal{D}$ as $\delta \to 0$ for $\hat{\Lambda}_\delta$ defined in (19) and (20), then*

$$\hat{A}_\delta \Rightarrow c_a B_a + \hat{\Lambda} \quad \text{in } \mathcal{D} \text{ as } \delta \to 0, \tag{21}$$

where $\hat{A}_\delta$ is defined in (20). Under assumption (19), $\hat{\Lambda}_\delta = \hat{\Lambda}$ for all δ, where

$$\hat{\Lambda}(t) = \frac{-t^{p+1}}{p+1}.$$

Proof. The first assertion in (21) is just (8) in Theorem 3.1 with $n \to \infty$ replaced by $\delta^{-2/(2p+1)} \to \infty$ as $\delta \to 0$. The last assertion follows by direct calculation. ■

6. Heavy-traffic limits for the workload process

We can obtain associated heavy-traffic limits for queueing processes of interest in the $G_t/GI/1$ model by applying one of these established arrival process FCLT's with established results in [6,7,19,21]. We illustrate by applying Theorem 3.1 to the workload process, representing the remaining work in service time in the system at time t. The simple approach for the workload follows [19], which is possible because it is simple for single-server queues, unlike for the multi-server queues considered in [6,7].

We assume that the sequence of service times $\{V_k\}$, also indexed in reverse time, is independent of the arrival process and satisfies a FCLT; in particular,

$$\hat{S}_n \Rightarrow c_s B_s \quad \text{in } \mathcal{D} \text{ as } n \to \infty, \tag{22}$$

where B_s is a (standard) BM independent of the BM B_a in Theorem 3.1 and

$$\hat{S}_n(t) \equiv n^{-1/2} \sum_{k=1}^{\lfloor nt \rfloor} (V_k - 1), \quad t \ge 0,$$

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

The workload at time $-t_1$ starting empty at time $-t_2$, where $-\infty < -t_2 < -t_1 \le 0$ in forward time, can be represented in the reverse-time framework as

$$\begin{aligned} &W_{t_1}(t_2) \equiv \Psi_{t_1}(X)(t_2) \equiv \sup_{\{t_1 \le s \le t_2\}} \{X(s) - X(t_1)\}, \\ &0 \le t_1 < t_2 < \infty, \end{aligned} \tag{23}$$

where $X(s) = Y(s) - s$ is the reverse-time net-input process over the interval $[0, s]$ or, equivalently, the forward-time net input over $[-s, 0]$, with Y being the random sum

$$Y(s) \equiv \sum_{k=1}^{A(s)} V_k, \quad s \ge 0.$$

Now construct the scaled processes

$$\begin{aligned} \hat{Y}_n(s) &\equiv n^{-1/2}[Y_n(ns) - ns], \quad s \ge 0, \\ \hat{X}_n(s) &\equiv n^{-1/2} X_n(ns), \quad s \ge 0, \\ \hat{W}_{t_1,n}(s) &\equiv n^{-1/2} W_{t_1,n}(ns), \quad s \ge 0. \end{aligned}$$

Let $\overset{\mathrm{d}}{=}$ denote equal in distribution, applied here for stochastic processes, and let $N(m, \sigma^2)$ denote a normal or Gaussian random variable with mean m and variance σ^2.

Theorem 6.1 (*Heavy-traffic FCLT for the Workload Process*). *If the conditions of Theorem 3.1 hold in addition to the assumptions in this section above, then*

$$\begin{aligned} &(\hat{A}_n, \hat{S}_n, \hat{Y}_n, \hat{X}_n, \hat{W}_{t_1,n}) \Rightarrow (\hat{A}, \hat{S}, \hat{Y}, \hat{X}, \hat{W}_{t_1}) \\ &\text{in } \mathcal{D}^5 \text{ as } n \to \infty, \end{aligned} \tag{24}$$

where

$$(\hat{A}, \hat{S}, \hat{Y}, \hat{X}, \hat{W}_{t_1}) = (c_a B_a + \hat{\Lambda}, c_s B_s, \hat{A} + \hat{S}, \hat{Y}, \Psi_{t_1}(\hat{X})), \tag{25}$$

for Ψ_{t_1} defined in (23) above, with B_a and B_s being two independent BM's. As a consequence, $\hat{X}$ is the Gaussian process

$$\hat{X} \overset{\mathrm{d}}{=} c_x B + \hat{\Lambda}, \tag{26}$$

where $c_x^2 \equiv c_a^2 + c_s^2$ and B is BM, so that $\hat{X}(t) \overset{\mathrm{d}}{=} N(\hat{\Lambda}(t), (c_a^2 + c_s^2)t)$ for all t with $0 \le t \le t_1$.

Proof. First, by Theorem 11.4.4 of [21], under the conditions of Theorem 3.1 and the assumptions above,

$$(\hat{A}_n, \hat{S}_n) \Rightarrow (c_a B_a + \hat{\Lambda}, c_s B_s) \quad \text{in } \mathcal{D}^2 \text{ as } n \to \infty, \tag{27}$$

where $\hat{\Lambda}$ is given in Theorem 3.1, while B_a and B_s are two independent standard Brownian motion processes, as stated for the first two terms in (24). Convergence of the third and fourth terms in (24) follow from preservation of convergence under composition and centering, just as in the proof of Theorem 3.1. Finally, convergence of the scaled workload process follows from the continuous mapping theorem for the supremum over bounded intervals. The usual reflection map is replaced by an ordinary supremum in reverse time. ■

We suggest applying Theorem 6.1 to approximate the workload $W_n(nt_1)$ at any time $nt_1 \leq 0$, starting empty in the infinite past with arrival-rate functions satisfying (1) and the assumptions above, by

$$W_n(nt_1) \approx \sqrt{n} \lim_{t_2 \to \infty} \{\hat{W}_{t_1}(t_2)\}$$
$$= \sqrt{n} \sup_{\{s \geq t_1\}} \{c_x B(s - t_1) + \hat{\Lambda}(s) - \hat{\Lambda}(t_1)\}.$$

Because of condition (1) and the conditions in Theorem 6.1, we anticipate that the supremum above will be attained, and at relatively small s. The approximation is relatively direct at the critical point when $t_1 = nt_1 = 0$.

Remark 6.1 (*Limit Interchange and Initial Conditions*)**.** The approximation above is not directly justified by a limit, because it involves an interchange of limits. We want to first let $t_2 \to \infty$ and then let $n \to \infty$, but we develop our approximation by changing the order of these two limits. That parallels what is done to approximate the steady-state distribution with a stationary model, which has been rigorously addressed in some cases, e.g., see [2].

We emphasize that Theorems 3.1–6.1 are closely related to Theorem 3.4 of [11]. The formulation of Theorem 3.4 in [11] also has a problem with the initial conditions. In particular, the formulation of Theorem 3.4 in [11] is tantamount to starting empty at some finite time in the past.

Remark 6.2 (*Before and After the Critical Time*)**.** All the FCLT's can be extended to time intervals $[t_1, t_2]$ with $t_1 < 0 < t_2$ under essentially the same conditions. That will leave the scaling unchanged. This extension can help to analyze the impact after time 0. However, the required notation gets more complicated; e.g., see Theorem 3.4 in [11].

7. Conclusions

We have established heavy-traffic limits for the arrival and workload processes in the $G_t/GI/1$ queue with a time-varying arrival-rate function. We have focused on the limiting behavior at and before a critical point, where the arrival-rate function approaches its critical value from below, which has been called the onset of critical loading in [11]. Our results extend and complement previous results in [15–18,8,13,11,4,22].

In Section 3 we established a heavy-traffic FCLT for the arrival process in the standard heavy-traffic framework in [21], which follows [6,7]. Given the usual time scaling by n and space scaling by $1/\sqrt{n}$, this requires conditions on the associated sequence of arrival-rate functions. For arrival-rate functions of the form $\lambda_n(t) = 1 - c_n t^p$, viewed in reverse time back from the critical point at time 0, the FCLT requires the scaling $c_n n^{(2p+1)/2} \to c$ as $n \to \infty$, $0 < c < \infty$, in (10) and leads to the deterministic component $\hat{\Lambda}(t) = -ct^{p+1}/(p+1)$ in the diffusion process limit. Note that the exponent of t in $\hat{\Lambda}(t)$ is necessarily one higher than the exponent of t in $\lambda_n(t)$, which occurs naturally because the cumulative arrival-rate function is the integral of the arrival-rate function.

The most interesting cases in that limit theorem are the linear and quadratic cases, which arise for $p = 1$ and $p = 2$. These lead to spatial scaling by $n^{-1/3}$ and $n^{-1/5}$. These correspond, respectively, to the transition through saturation in [15] and the mild rush hour in [17]. While these two cases yield important insights, there are many other possibilities, as illustrated in Remarks 3.1 and 3.2. An important conclusion of this paper is that we should not assume that the spatial scaling in heavy-traffic limits for queues with time-varying arrival rates should be restricted to $n^{-1/3}$ and $n^{-1/5}$.

In Section 4 we showed how that arrival process FCLT can be recast in a setting in which we expand time within a fixed initial arrival-rate function, as in [11]. That framework uses a Taylor series expansion, which restricts the possible cases. To obtain limits of the same form as in Section 3, we impose conditions on the time scaling of the arrival-rate functions. This case arises naturally when we consider highly structured mathematical models like the sinusoidal arrival rate function in Example 4.1.

Motivated by the desire to develop an approach that is useful for applications, in Section 5 we show how the heavy-traffic FCLT can be expressed in yet a different way using drift scaling. The idea is that we would fit the model by first observing the essential large-time-scale form of the drift. In particular, given that we directly observe (estimate) that the arrival-rate function has the form $\lambda_\delta(t) = 1 - \delta t^p, t \geq 0$, as in (19), we develop appropriate time and space scaling in terms of the parameter δ in order to obtain a heavy-traffic limit. The required time scaling is by $\delta^{-2/(2pk+1)}$, while the associated space scaling is by $\delta^{1/(2pk+1)}$. As before, the limit process $\hat{\Lambda}$ is a power with one higher exponent than $1 - \lambda_\delta(t)$.

Associated heavy-traffic FCLT's for the usual queueing processes follow quickly given the limits for the arrival process, as in [21]. In Section 6 we illustrate by establishing the heavy-traffic FCLT for the workload process.

Acknowledgment

The author thanks NSF for research support (grants CMMI 1265070 and 1634133).

References

[1] S.G. Eick, W.A. Massey, W. Whitt, $M_t/G/\infty$ queues with sinusoidal arrival rates, Manage. Sci. 39 (1993) 241–252.
[2] D. Gamarnik, A. Zeevi, Validity of heavy traffic steady-state approximations in generalized Jackson networks, Adv. Appl. Probab. 16 (1) (2006) 56–90.
[3] I. Gebhardt, B.L. Nelson, Transforming renewal processes for simulation of non-stationary arrival procsses, INFORMS J. Comput. 21 (2009) 630–640.
[4] H. Hannappa, R. Jain, A. Ward, A queueing model with independent arrivals, and its fluid and diffusion limits, Queueing Syst. 80 (1–2) (2014) 71–103.
[5] B. He, Y. Liu, W. Whitt, Staffing a service system with non-Poisson nonstationary arrivals, Probab. Engrg. Inform. Sci. 30 (3) (2016) published online on June 13.
[6] D.L. Iglehart, W. Whitt, Multiple channel queues in heavy traffic, I, Adv. Appl. Probab. 2 (1) (1970) 150–177.
[7] D.L. Iglehart, W. Whitt, Multiple channel queues in heavy traffic, II: Sequences, networks and batches, Adv. Appl. Probab. 2 (2) (1970) 355–369.
[8] J. Keller, Time-dependent queues, SIAM Rev. 24 (1982) 401–412.
[9] S. Kim, W. Whitt, Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? Manuf. Serv. Oper. Manage. 16 (3) (2014) 464–480.
[10] Y. Liu, W. Whitt, Many-server heavy-traffic limits for queues with time-varying parameters, Ann. Appl. Probab. 24 (1) (2014) 378–421.
[11] A. Mandelbaum, W.A. Massey, Strong approximations for time-dependent queues, Math. Oper. Res. 20 (1) (1995) 33–64.
[12] W.A. Massey, Nonstationary queues (thesis), Stanford University, 1981.
[13] W.A. Massey, Asymptotic analysis of the time-varying $M/M/1$ queue, Math. Oper. Res. 10 (1985) 305–327.
[14] W.A. Massey, W. Whitt, Unstable asymptotics for nonstationary queues, Math. Oper. Res. 19 (2) (1994) 267–291.
[15] G.F. Newell, Queues with time dependent arrival rates, I: the transition through saturation, J. Appl. Probab. 5 (1968) 436–451.
[16] G.F. Newell, Queues with time dependent arrival rates, II: the maximum queue and the return to equilibrium, J. Appl. Probab. 5 (1968) 579–590.
[17] G.F. Newell, Queues with time dependent arrival rates, III: a mild rush hour, J. Appl. Probab. 5 (1968) 591–606.
[18] G.F. Newell, Applications of Queueing Theory, second ed., Chapman and Hall, London, 1982.
[19] W. Whitt, Weak convergence theorems for priority queues: Preemptive-resume discipline, J. Appl. Probab. 8 (1971) 74–94.
[20] W. Whitt, Planning queueing simulations, Manage. Sci. 35 (11) (1989) 1341–1366.
[21] W. Whitt, Stochastic-Process Limits, Springer, New York, 2002.
[22] W. Whitt, Heavy-traffic limits for queues with periodic arrival processes, Oper. Res. Lett. 42 (2014) 458–461.
[23] W. Whitt, W. You, Time-Varying Robust Queueing, Columbia University, 2016, http://www.columbia.edu/~ww2040/allpapers.html.