Heavy-Traffic Limits for Queues with Periodic Arrival Processes

Ward Whitt

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

Abstract

We establish conventional heavy-traffic limits for the number of customers in a $G_t/GI/s$ queue with a periodic arrival process. We assume that the arrival counting process can be represented as the composition of a cumulative stochastic process that satisfies a FCLT and a deterministic cumulative rate function that is the integral of a periodic function. We establish three different heavy-traffic limits for three different scalings of the deterministic arrival rate function. The different scalings capture the three cases in which the predictable deterministic variability (i) dominates, (ii) is of the same order as, or (iii) is dominated by the stochastic variability in the arrival and service processes.

Keywords: heavy-traffic limits, queues with periodic arrival rates, heavy-traffic limits for periodic queues

1. Introduction

In this paper we establish heavy-traffic functional weak laws of large numbers (FWLLN's) and functional central limit theorems (FCLT's) for $G_t/GI/s$ queues with arrival processes having periodic arrival rate functions. The model has a fixed number of servers working in parallel, an unlimited waiting space and the first-come first-served service discipline. By "conventional heavy traffic," we mean that we allow the arrival rate to increase, but keep the maximum possible service rate fixed.

Conventional heavy-traffic approximations help understand the performance of complex queueing systems; see [1] for a review (especially Chapters 5 and 9). Highlights are Kingman [2] showing that the steady-state wait in

a $GI/GI/1$ queue can be approximately by an exponential random variable and Iglehart and Whitt [3, 4] showing that the entire waiting time (and queue length) process in a $G/G/s$ queue can be approximated by a reflected Brownian motion (RBM) with negative drift, which has an exponential steady-state distribution, where the mean of the exponential distribution and the drift and diffusion coefficients of the RBM depend on the basic rate and variability parameters of the arrival and service processes.

It is important to note that Mandelbaum and Massey [5] already developed conventional heavy-traffic approximations for queues with time-varying arrival processes. They analyzed the $M_t/M_t/1$ model and showed that the presence of time-varying arrival rate can introduce major complications; e.g., there is need for care in even properly defining a proper notion of traffic intensity; see §3 of [5]. In [5] an elaborate theory is developed, which demonstrates both complex performance, e.g., see Figures 3.1 and 4.1 in [5] and complex proof techniques, including the use of the Skorohod M_1 topology to treat the discontinuities evident in Figure 4.1 of [5].

In contrast, our goal is to expose the more elementary story that follows quite directly from [4] if we make additional simplifying assumptions. In particular, we only focus on the first-order performance. If there is an important story in the deterministic fluid approximation stemming from the FWLLN, then we focus on that FWLLN. We only consider the heavy-traffic FCLT when that provides the first-order description of performance, i.e., when the FWLLN is the same as for the model with constant arrival rate. We then can see if the time-varying arrival rate is insignificant in the heavy-traffic limits by seeing if it plays no role in either the FWLLN or the FCLT.

The theory in [4] implies that, under regularity conditions, a heavy-traffic FWLLN (FCLT) holds for the queue length process whenever FWLLN's (FCLT's) hold for the arrival and service processes. Thus, for periodic arrival processes, previous heavy-traffic FWLLN's and FCLT's can be applied if we have a FWLLN and a FCLT for the periodic arrival process. In particular, we can apply Theorem 1 of [4] and basic continuous mapping arguments to establish conventional heavy-traffic limits for the $G_t/GI/s$ model. This important consequence of [4] no doubt has been recognized, but evidently nothing has been published.

An important role in the conventional heavy-traffic limits is played by the scaling of both time and space. Roughly, the required scaling is the same as needed for a sequence of simple random walks to converge to a Brownian motion with drift: we need to scale time by some factor n and then scale space

by $1/\sqrt{n}$, with the mean step being $c/\sqrt{n}$. Since the mean step is related to the traffic intensity of the queue, n should be related to the traffic intensity ρ in the queueing system by $1 - \rho_n = 1/\sqrt{n}$. The important observation is that, in terms of the traffic intensity ρ the required time scaling is $(1-\rho)^{-2}$.

As we show here in Corollary 3.1, when time scaling is omitted from the deterministic arrival rate function in the standard heavy-traffic FCLT, the heavy-traffic limit with the time scaling is the same as if the periodic cycles in the periodic arrival rate function are getting shorter in the heavy-traffic limit as $\rho \uparrow 1$. As a consequence, there is still a heavy-traffic limit, but that limit is the same as if the periodic arrival rate were replaced by its long-run average. This phenomenon was first shown for the $M_t/GI/1$ model by Falin [6], but without mentioning any connection to time scaling. When the time scaling is included, the approximation stemming from the heavy-traffic FCLT is a reflection of the usual Brownian motion with drift plus a deterministic cumulative rate function associated with a periodic arrival rate function.

We are especially interested in the time scaling. In the main heavy-traffic FCLT, Theorem 3.2 here, the limit process is relatively complicated, so that it is not easy to compute the approximate performance measures. It thus may be necessary to exploit simulation in order to quantify performance. Nevertheless, the heavy-traffic limits can provide useful insight into the simulations. As illustrated by [7], heavy-traffic scaling can help understand numerical performance calculations, because greater regularity is revealed. Indeed, we were motivated to establish these heavy-traffic limits in our study of grey-box modeling of queueing systems in [8, 9], in which birth-and-death processes are fit to observations of a queue-length process. Specifically, the present paper arose in the study of that approach applied to periodic queues in [9].

Our approach has an important implication. By focusing on only the first-order performance, we determine when the predictable deterministic variability or the unpredictable stochastic variability dominates. We establish different heavy-traffic limits showing when the predictable deterministic variability (i) dominates, (ii) is of the same order as, or (iii) is dominated by the stochastic variability in the arrival and service processes. The more detailed analysis in [5] shows (i) that there may be different answers at different times and (ii) how to describe the refined second-order performance (diffusion approximation) for the first-order performance (deterministic fluid model) when the deterministic variability dominates.

We make two additional comments about [5]. First, since we tell only

part of the story in [5], it follows that the story can be deduced from the reasoning of [5], extended from the $M_t/M_t/1$ to $G_t/GI/s$ model, but that is a more difficult route. Second, the additional structure revealed in the more general analysis in [5] is also important for understanding the performance of queues with time-varying parameters.

In closing this introduction, we remark that the conventional heavy-traffic regime is quite different from the many-server heavy-traffic regime, which we also briefly discuss in §4 for comparison. Since there is no time scaling in the many-server heavy traffic regime, the many-server heavy-traffic approximations are more straightforward in engineering applications. There also is already a significant body of related literature on many-server heavy-traffic approximations for queues with time-varying arrival rates in [10, 11, 12, 13, 14]. In both settings, these results are facilitated by previous results concluding that heavy-traffic limits for the queue length depend on the arrival process through its FCLT. Thus in both cases it suffices to establish a FCLT for the arrival process with the appropriate scaling.

2. The Arrival Process Model

We will consider periodic stochastic arrival counting processes defined by

$$A(t) \equiv N(\Lambda(t)), \quad t \geq 0, \tag{1}$$

where N is a stochastic counting process satisfying a functional central limit theorem (FCLT), i.e.,

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

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 [1], and B_a is a standard (drift 0, variance 1) Brownian motion (BM), while Λ is a cumulative arrival rate function, satisfying

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0, \tag{3}$$

with λ being a periodic arrival rate function, which is assumed to be integrable over finite intervals with finite long-run average

$$\bar{\lambda} \equiv \lim_{t \to \infty} t^{-1} \Lambda(t). \tag{4}$$

Throughout this paper, we assume that the cumulative arrival rate function Λ in (3) 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 [15].

The construction in (1) is convenient for constructing non-Markov periodic arrival processes. It was suggested by [16] and also used by [17, 18] and no doubt others. 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 [1]), 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.

3. Conventional Heavy-Traffic Limits for the $G_t/GI/s$ Model

In this section we establish heavy-traffic limits for the queue-length process (number in system) in the $G_t/GI/s$ model, which has s homogeneous servers working in parallel, unlimited waiting room and customers entering service in order of arrival. We assume that the service times come from a sequence of independent and identically distributed (i.i.d.) random variables, which is independent of the arrival process. We let the mean service time be s and its scv be c_s^2. This choice of the mean makes the maximum total service rate be 1. We let the arrival process be periodic with the structure in (1)-(4).

We will construct a family of models indexed by the traffic intensity ρ and let ρ increase toward 1, its upper limit for stability. We will let the traffic intensity be determined by the deterministic arrival rate function λ, requiring that $\lambda_\rho = \rho\lambda$ for each ρ, where we let $\bar{\lambda} = 1$. As a consequence

the average arrival rate in model ρ is $\rho < 1$, so that we anticipate a proper steady-state distribution exists for each ρ.

Thus, we get an arrival process for each ρ by letting

$$A_\rho(t) \equiv N(\Lambda_\rho(t)), \quad t \geq 0. \tag{5}$$

We will establish heavy-traffic limits for the associated queue-length processes $Q_\rho \equiv \{Q_\rho(t) : t \geq 0\}$, where $Q_\rho(t)$ is the number of customers in the system ρ at time t. We will apply Theorem 1 (a) of [4]. From that source, we know that: (i) an important role is played by the scaling of time and space as $\rho \uparrow 1$ and (ii) there is a heavy-traffic FCLT for the properly scaled version of Q_ρ whenever there is an associated FCLT for the properly scaled version of the arrival processes A_ρ.

To apply the results of [4], in model ρ we need to scale (increase) time by a factor $(1-\rho)^{-2}$ and scale (decrease) space by a factor $1-\rho$. Hence, paralleling (2), we define the following scaled deterministic cumulative arrival rate functions

$$\hat{\Lambda}_\rho(t) \equiv (1-\rho)[\Lambda_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\rho\Lambda_f(t)], \quad t \geq 0, \tag{6}$$

for each ρ, $0 < \rho < 1$, and we assume that

$$\hat{\Lambda}_\rho(t) \to \Lambda_d(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \uparrow 1. \tag{7}$$

We assume that the two functions Λ_f and Λ_d appearing in (6) and (7) have the structure of cumulative arrival rate functions. By the way the scaling is chosen in (6), it is natural that the associated arrival rate functions λ_f and λ_d be periodic functions with averages $\bar{\lambda}_f = 1$ and $\bar{\lambda}_d = 0$.

We can obtain (6) and (7) if we let

$$\Lambda_\rho(t) \equiv (1-\rho)^{-2}\rho\Lambda_f((1-\rho)^2 t) + (1-\rho)^{-1}\Lambda_d((1-\rho)^2 t) + o((1-\rho)^{-1}) \tag{8}$$

for $t \geq 0$ with Λ_f and Λ_d as above, where $o(x)$ means that $o(x)/x \to 0$ as x approaches its limit. The notation for the limiting cumulative rate functions is chosen to indicate that Λ_f appears in a first-order fluid limit, while Λ_d appears in a second-order diffusion limit. The implications of (6)-(8) is that the period of the periodic cycles grows with ρ as ρ increases, specifically by a factor $(1-\rho)^{-2}$.

As usual, in applications, we think of our actual system being model ρ in a family of models. If we choose ρ so that the period is $(1-\rho)^{-2}t_0$, then it is

natural that the limiting cumulative rate functions Λ_f and Λ_d be integrals of periodic functions λ_f and λ_d that have period t_0. The time units have already been chosen so that the maximum service rate is 1. The traffic intensity ρ might be chosen so that, with these time units, the long-run drift rate is $-(1-\rho)$. We then can choose t_0 so that the period is $(1-\rho)^{-2}t_0$.

We start by establishing a first-order fluid limit, which describes the situation in which the predictable deterministic variability in $\Lambda_\rho(t)$ dominates the unpredictable stochastic variability in the process N. Let $\bar{A}_\rho$ and $\bar{Q}_\rho$ be fluid-scaled processes, defined as

$$\bar{A}_\rho(t) \equiv (1-\rho)^2 A_\rho((1-\rho)^{-2}t) \quad \text{and} \quad \bar{Q}_\rho(t) \equiv (1-\rho)^2 Q_\rho((1-\rho)^{-2}t), \quad t \geq 0. \tag{9}$$

Let $\Psi(x)$ be the one-dimensional one-sided reflection map, as in §§5.2 and 13.5 of [1], and let $e(t) \equiv t$, $t \geq 0$, be the identity map in $\mathcal{D}$.

Theorem 3.1. (*FWLLN*) *If, in addition to the conditions above,* $\bar{Q}_\rho(0) \Rightarrow q(0)$ *in* $\mathbb{R}$ *as* $\rho \uparrow 1$, *where* $q(0)$ *is a nonnegative real number, then*

$$\bar{A}_\rho \Rightarrow \Lambda_f \quad \textit{and} \quad \bar{Q}_\rho \Rightarrow \Psi(q(0) + \Lambda_f - e) \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad \rho \uparrow 1. \tag{10}$$

Proof. Apply the FWLLN for N in the form $(1-\rho)^2 N((1-\rho)^{-2}t) \Rightarrow t$ in $\mathcal{D}$ together with the limit $(1-\rho)^2\Lambda_\rho((1-\rho)^{-2}t) \to \Lambda_f(t)$ in $\mathcal{D}$ that follows from (6) plus the composition map, exploiting (5), and the continuous mapping theorem to get the stated limit for $\bar{A}_\rho$. Apply the FWLLN for the service process to a get a FWLLN for the net input process as in [1, 4]. Then apply the continuous mapping theorem again with the one-sided reflection map to get the stated FWLLN for $\bar{Q}_\rho$. ∎

We remark that we could strengthen the FWLLN conclusion to a FSLLN if we assumed that N_a obeyed a SLLN as well as a FCLT. We also remark that the fluid limit corresponds to an elementary ordinary differential equation (ODE) with boundary at 0. In particular, it evolves as $\dot{q}(t) = \lambda_f(t) - 1$ if $q(t) > 0$ and $\dot{q}(t) = (\lambda_f(t) - 1)^+$ if $q(t) = 0$. Interesting first-order deterministic behavior occurs if $\lambda_f(t) > 1$ over some intervals.

For the FCLT, we assume that Λ_f takes the trivial form, i.e., $\Lambda_f(t) \equiv t$, $t \geq 0$, because that is what is required for a direct application of [4]. That puts all the time-varying component in the diffusion scale. The more general case yields a diffusion refinement to the fluid limit in Theorem 3.1. It is substantially more complicated; it can be treated by applying the approach in [5].

With $\Lambda_f(t) \equiv t$, $t \geq 0$, we now define the appropriate diffusion-scaled processes.

$$\begin{aligned} \hat{A}_\rho(t) &\equiv (1-\rho)[A_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}t] \quad \text{and} \\ \hat{Q}_\rho(t) &\equiv (1-\rho)Q_\rho((1-\rho)^{-2}t), \quad t \geq 0. \end{aligned} \tag{11}$$

The FCLT describes the situation in which the predictable deterministic variability in $\Lambda_\rho(t)$ and the unpredictable stochastic variability in the process N are of the same order, so that both appear in the limit.

Theorem 3.2. (*FCLT*) *If, in addition to the conditions above,* $\Lambda_f(t) \equiv t$ *and* $\hat{Q}_\rho(0) \Rightarrow \hat{Q}(0)$ *in* $\mathbb{R}$ *as* $\rho \uparrow 1$, *where* $\hat{Q}(0)$ *is a nonnegative real-valued random variable independent of the arrival and service processes with* $P(\hat{Q}(0) < \infty) = 1$, *then*

$$\begin{aligned} \hat{A}_\rho &\Rightarrow c_a B_a + \Lambda_d - e \quad \textit{in} \quad \mathcal{D} \quad \textit{and} \\ \hat{Q}_\rho &\Rightarrow \Psi(\hat{Q}(0) + c_a B_a + \Lambda_d - e - c_s B_s) \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad \rho \uparrow 1, \end{aligned} \tag{12}$$

where B_s *is a BM with* B_s, B_a *and* $\hat{Q}(0)$ *being mutually independent, so that* $c_s B_s - c_a B_a \stackrel{\mathrm{d}}{=} \sqrt{c_s^2 + c_a^2}B$, *where* B *is a BM.*

Thus, we see that the limit process is a relection of a limiting net input process, which has constant drift -1, periodic deterministic component $\Lambda_d(t)$ and stochastic component $\sqrt{c_s^2 + c_a^2}B$, where B is a standard BM.

Proof. For the arrival process write

$$\begin{aligned} \hat{A}_\rho(t) &= (1-\rho)[N(\Lambda_\rho((1-\rho)^{-2}t)) - (1-\rho)^{-2}\rho t] \\ &= (1-\rho)[N(\Lambda_\rho((1-\rho)^{-2}t)) - \Lambda_\rho((1-\rho)^{-2}t) \\ &\quad +\Lambda_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\rho t] \\ &\quad +(1-\rho)[(1-\rho)^{-2}\rho t - (1-\rho)^{-2}t] \\ &\Rightarrow c_a B_a(t) + \Lambda_d(t) - t \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad \rho \uparrow 1, \end{aligned} \tag{13}$$

using $(1-\rho)^2\Lambda_\rho((1-\rho)^{-2}t) \Rightarrow t$ in $\mathcal{D}$ as $\rho \uparrow 1$ and the composition map with the time change for the first term on the second line. For the queue-length process, we apply Theorem 1 (a) of [4]. ∎

It remains to describe the situation in which the the unpredictable stochastic variability in the process N dominates the predictable deterministic variability in $\Lambda_\rho(t)$. However, that third setting is already described by the FCLT

in the case with $\Lambda_d(t) = 0$, $t \geq 0$. It is important to note that this case naturally arises if we do not scale time in the deterministic cumulative arrival rate function as we did in (6)-(8). This case was treated by [6], but without discussion of the scaling.

Corollary 3.1. (*unscaled arrival rate function*) *If we let* $\Lambda_\rho(t) = \rho\Lambda(t)$, *then the FCLT in Theorem* 3.2 *holds with* $\Lambda_d(t) \equiv 0$, $t \geq 0$.

Proof. The periodic structure implies that $\Lambda_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\rho t = O(1)$ as $\rho \uparrow 1$, because the difference is bounded by the integral of $\lambda(t)$ over one periodic cycle. Hence, $(1-\rho)[\Lambda_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\rho t] \to 0$ uniformly in t as $\rho \uparrow 1$. ∎

We remark that Corollary 3.1 is consistent with the limiting behavior of the $M_t/M/s$ model with a sinusoidal arrival rate function as the cycles become shorter in [19]. Contrary to initial intuition, as the rate of oscillation increases, the model behaves like the stationary model with constant arrival rate. Proper intuition is gained by focusing on the cumulative rate function, which is approaching a linear function.

4. Many-Server Heavy-Traffic Limits for Periodic Queues

Paralleling Theorem 1 of [4], for the many-server heavy-traffic regime it is also known for several models that the limit depends on the arrival process through its FCLT; e.g., for the $G_t/G/\infty$ model, see [11, 20]; for the $G_t/M/s + M$ model, see §7.5 of [21]. Moreover, the situation is much less complicated because no scaling of time is needed in the many-server heavy-traffic regime.

Now we consider a sequence of models indexed by n, where n is the overall arrival rate in model n, where the number of servers is allowed to grow with n as well. The relevant limit is

$$\hat{A}_n(t) \equiv n^{-1/2}[A_n(t) - nt] \Rightarrow \hat{A} \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{14}$$

We are thus interested in the form (14) takes when we use (1) and (2).

Theorem 4.1. (*many-server heavy-traffic FCLT*) *If, in addition to the conditions above,* (1) *and* (2) *hold with*

$$\begin{aligned} \hat{\Lambda}_n &\to \Lambda_d \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad n \to \infty, \quad \textit{where} \\ \hat{\Lambda}_n(t) &\equiv n^{-1/2}[\Lambda_n(t) - nt], \quad t \geq 0, \end{aligned} \tag{15}$$

then (14) *holds with* $\hat{A} = c_a B_a + \Lambda_d$*, where* B_a *is BM and* Λ_d *is the limit in* (15)*.*

Just as in §3, if a more general centering term Λ_f appears in (15), then the FCLT is much more complicated. In that case, the methods of [10] can be applied. As can be seen from [11, 20] and §7.5 of [21], the extra deterministic term Λ_d significantly complicates the associated approximating processes stemming from the associated FCLT for the queue-length process. However, in many applications it may be reasonable to assume that $\Lambda_d = 0$.

5. Conclusions

5.1. Summary

We have applied Theorem 1 of [4] to establish heavy-traffic limits for queues with periodic arrival processes in the conventional heavy-traffic regime in which $\rho \uparrow 1$ with a fixed number of servers. The case in which the predictable deterministic variability dominates is captured by Theorem 3.1; the case in which the unpredictable stochastic variability dominates is captured by Corollary 3.1; the most complicated case in which both forms of variability contribute is captured by Theorem 3.2. When both forms of variability contribute, the resulting limit process is relatively complicated, but the scaling can provide useful insight; e.g., it can be helpful for understanding simulations.

5.2. Extensions

The results easily can be generalized in several directions. First, the results are not limited to the scaling in (2). For example, we could have convergence to a stable process, which involves different scaling; see §§5.5, 6.3 and 6.4 and of [1]. Second, the results extend to non-periodic time-varying arrival processes, but there is simplicity in the present framework, which serves to communicate the main ideas. The main point here is to emphasize the importance of scaling. With a periodic arrival process and no time scaling, Corollary 3.1 shows that the limit is the same as if the arrival rate function is constant, equal to its long-run average. Third, the results also extend to stochastic arrival rate functions, if we can assume that there is joint convergence of the scaled processes in (2) and (7). That will yield the required limit for the scaled arrival process $\hat{A}_\rho$ in Theorem 3.2, but of course the limit process is even more complicated. Finally, under regularity

conditions, the results extend to open networks of queues if the assumptions hold for the external arrival processes to all the queues.

Acknowledgement. The author thanks NSF for research support (grants CMMI 1066372 and and 1265070).

References

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

[2] J. F. C. Kingman, The single server queue in heavy traffic, Proc. Camb. Phil. Soc. 77 (1961) 902–904.

[3] D. L. Iglehart, W. Whitt, Multiple channel queues in heavy traffic, I, Advances in Applied Probability 2 (1) (1970) 150–177.

[4] D. L. Iglehart, W. Whitt, Multiple channel queues in heavy traffic, II: Sequences, networks and batches, Advances in Applied Probability 2 (2) (1970) 355–369.

[5] A. Mandelbaum, W. A. Massey, Strong approximations for time-dependent queues, Mathematics of Operations Research 20 (1) (1995) 33–64.

[6] G. I. Falin, Periodic queues in heavy traffic, Advances in Applied Probability 21 (1989) 485–487.

[7] J. Abate, W. Whitt, Calculating transient characteristics of the Erlang loss model by numerical transform inversion, Stochastic Models 14 (3) (1998) 663–680.

[8] J. Dong, W. Whitt, Stochastic grey-box modeling of queueing systems: exploiting fitted birth-and-death processes, submitted to Queueing Systemss. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html (2014).

[9] J. Dong, W. Whitt, Stationary birth-and-death proceses fit to queues with periodic arrival rate functions, in preparation (2014).

[10] A. Mandelbaum, W. A. Massey, M. I. Reiman, Strong approximations for Markovian service networks, Queueing Systems 30 (1998) 149–201.

[11] G. Pang, W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues, Queueing Systems 65 (2010) 325–364.

[12] Y. Liu, W. Whitt, A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading, Oper. Res. Letters 40 (2012) 307–312.

[13] A. A. Puhalskii, On the $M_t/M_t/K_t + M_t$ queue in heavy traffic, Math. Methods Oper. Res. 78 (2013) 119–148.

[14] Y. Liu, W. Whitt, Many-server heavy-traffic limits for queues with time-varying parameters, Annals of Applied Probability 24 (1) (2014) 378–421.

[15] S. Kim, W. Whitt, Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes?, Manufacturing and Service Oper. Management 16 (3) (2014) 464–480.

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

[17] I. Gebhardt, B. L. Nelson, Transforming renewal processes for simulation of non-stationary arrival procsses, INFORMS Journal on Computing 21 (2009) 630–640.

[18] Y. Liu, W. Whitt, Stabilizing performance in networks of queues with time-varying arrival rates, Probability in the Engineering and Informational Sciences 24.

[19] W. Whitt, The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase, Management Science 37 (3) (1991) 307–314.

[20] G. Pang, W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues with dependent service times, Queueing Systems 73 (2) (2013) 119–146.

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