Many-Server Limits for Periodic Infinite-Server Queues

Ward Whitt

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

December 17, 2015

Abstract

To better understand (i) the long-run average performance of periodic many-server queues and (ii) the consequence of fitting a stationary birth-and-death (BD) process to a segment of the sample path of the queue-length process from a many-server system, we establish many-server heavy-traffic fluid limits for the periodic $M_t/GI/\infty$ model. These yield relatively simple fluid approximations for both the steady-state distribution and the fitted rates. For the special case of sinusoidal arrival rates, (i) the limiting steady-state distribution has an arcsine law, while (ii) the fluid birth and death rates are linear over a bounded interval and equal, with the interval end points and the slope depending on the service distribution. The fluid rates are usually increasing functions of the state, but we show that they can be decreasing functions of the state for long deterministic service times (exceeding a half sine cycle).

Keywords: periodic queues, periodic arrival rates, periodic steady state, birth-and-death processes, fitting birth-and-death processes to data, many-server heavy-traffic limits.

1 Introduction

This paper is a sequel to [6, 7] in which we began to investigate what can be learned from fitting a stationary state-dependent birth-and-death (BD) process to a stochastic process $\{Q(t) : t \geq 0\}$ that takes values on the nonnegative integers and makes all its transitions in unit steps, but may not itself be a BD process. In doing so, we are primarily thinking of the queue-length (number in system) in a complex queueing system; e.g., it may be the evolving content of one queue in a complex network, such as a healthcare system. The observed sample path may be from system data or computer simulations. Fitting BD rates to data can be useful to estimate the steady-state distribution and diagnose what model is appropriate.

1.1 Estimating Steady-State Distributions via BD Rates

Suppose that $Q(t)$ has a proper limiting steady-state probability mass function (pmf) α as $t \to \infty$; i.e., $P(Q(t) = k) \to \alpha_k$ as $t \to \infty$ for each $k \geq 0$, where $\sum_{k=0}^{\infty} \alpha_k = 1$. Given system data, i.e., a segment $\{Q(s) : 0 \leq s \leq t\}$ of the sample path, a standard way to estimate the pmf α is to calculate the proportion of time spent in each state; i.e., if $T_k(t)$ is the total time spent in state k during $[0, t]$, then we estimate α_k by

$$\bar{\alpha}_k(t) \equiv \frac{T_k(t)}{t}, \quad k \geq 0, \tag{1}$$

where $\equiv$ denotes equality by definition

Given that all transitions of the stochastic process $\{Q(t) : t \geq 0\}$ are unit transitions, then there is an alternative way to estimate α that may be attractive. Let $A_k(t)$ and $D_k(t)$ be the number of arrivals and departures, respectively, observed in state k over $[0, t]$. State-dependent birth and death rates can be estimated by

$$\bar{\lambda}_k(t) \equiv \frac{A_k(t)}{T_k(t)} \quad \text{and} \quad \bar{\mu}_k(t) \equiv \frac{D_k(t)}{T_k(t)}. \tag{2}$$

We then estimate the steady-state distribution by solving the local-balance equations for a BD process; i.e., we let $\bar{\alpha}^e(t) \equiv \{\bar{\alpha}^e_k(t) : k \geq 0\}$ be the solution to the equation

$$\bar{\alpha}^e_k(t)\bar{\lambda}_k(t) = \bar{\alpha}^e_{k+1}(t)\bar{\mu}_k(t), \quad k \geq 0, \tag{3}$$

with the additional property that $\sum_{k=0}^{\infty} \bar{\alpha}^e_k(t) = 1$ We use the superscript e to denote that the vector $\bar{\alpha}^e_k(t)$ is obtained from the estimated BD rates in (2) via (3).

If the stochastic process $\{Q(t) : t \geq 0\}$ is actually a BD process, then (2) is the natural model-fitting procedure [2, 15, 30]. In fact, these are the maximum likelihood estimators for these rates. Then (3) is the standard way to find the steady-state pmf α. However, we are not restricting attention to BD processes.

We have found that this procedure can be effective for the non-Markov stationary many-server $M/GI/s + GI$ queueing model with customer abandonment (the $+GI$), which has i.i.d. service times and patience times with general distributions. In [25] we found that this model could be well approximated by the associated $M/M/s + M_{(k)}$ model, where $M_{(k)}$ denotes a state-dependent abandonment rate. The queue-length process in the approximating BD model was constructed directly (analytically) in [25], but we also have observed that the statistical fitting approach tends to produce a very similar BD model.

However, here we have in mind more general applications, in which we may not be sure what model is appropriate. It is remarkable that, without making any stochastic assumptions, the two empirical steady-state probability vectors $\bar{\alpha}(t)$ in (1) and $\bar{\alpha}^e(t)$ in (3) are intimately related: If $Q(0) = Q(t)$, then the two probability vectors $\bar{\alpha}(t)$ and $\bar{\alpha}^e(t)$ constructed from the sample path over $[0,t]$ are *identical*. More generally, they are stochastically ordered; see Theorem 1 of [27]. Moreover, under minor regularity conditions, $\bar{\alpha}^e_k(t)$ and $\bar{\alpha}_k(t)$ are both consistent estimators of α_k, i.e.,

$$\alpha_k \equiv \lim_{t\to\infty} \bar{\alpha}_k(t) = \lim_{t\to\infty} P(Q(t) = k) = \lim_{t\to\infty} \bar{\alpha}^e_k(t); \tag{4}$$

see Chapter 4 of [10] and Corollary 4.1 of [27]. (These properties can be regarded as consequences of rate-conservation, like Little's law, i.e., the fundamental queueing relation $L = \lambda W$, e.g., see [10, 16] and references therein.) In this paper we will be concerned with the limiting rates, as is relevant if we have large samples. When we refer to estimates such as $\bar{\lambda}_k$ without a time argument, we understand that it represents the limit as $t \to \infty$. The BD indirect estimation of the steady-state pmf α via (2) and (3) is an attractive alternative to the direct method in (1) because the BD rates tend to have more structure than the steady-state pmf itself, typically being nondecreasing functions, even linear or piecewise-linear, as can be seen from the simulation examples in [6, 7].

Moreover, in [6] we found for the $GI/GI/s$ model that the transient behavior (as revealed by appropriate first passage times) could be effectively approximated by a one-parameter time transformation of the fitted BD process.

1.2 The Indirect Estimation Procedure for Periodic $M_t/GI/s$ Queues

The indirect estimation procedure was investigated for periodic $M_t/GI/s$ models in [7]. Of course, $P(Q(t) = k$ does not converge as $t \to \infty$ in the periodic $M_t/GI/s$ model, so we need to explain.

Suppose that the arrival-rate function is periodic with period c. The stochastic process $\{Q(t) : t \ge 0\}$ has a dynamic steady-state pmf $\alpha(t)$, $0 \le t < c$ (a family of pmf's indexed by t), and an overall steady-state pmf α^c if the following limits are well defined probability vectors:

$$\begin{aligned} \alpha_k(t) &\equiv \lim_{n\to\infty} P(Q(nc+t) = k) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} 1_{\{Q(jc+t)=k\}}, \quad 0 \le t < c, \quad \text{and} \\ \alpha^c_k &\equiv \frac{1}{c} \int_0^c \alpha_k(t)\, dt = \lim_{t\to\infty} \frac{1}{t} \int_0^t 1_{\{Q(s)=k\}}\, ds, \quad k \ge 0. \end{aligned} \tag{5}$$

Moreover, α^c can be directly regarded as a limiting steady-state pmf (a special case of the pmf α defined above) if we randomize the initial time uniformly over the interval $[0,c]$. It is the steady-state pmf α^c that we consider for periodic queues. The overall steady-state pmf α^c is of practical interest, e.g., for a hospital emergency room, because it reveals the long-run frequency of different occupancy levels and thus the average congestion and the average use of resources.

The numerical examples in [7] were for the sinusoidal arrival rate function

$$\lambda(t) \equiv \bar{\lambda}\lambda_1(t), \quad \text{where} \quad \lambda_1(t) \equiv 1 + \beta \sin(\gamma t), \tag{6}$$

which has cycle length $c = 2\pi/\gamma$. There are three parameters: (i) the average arrival rate $\bar{\lambda}$, (ii) the relative amplitude β and (iii) the time scaling factor γ or, equivalently, the cycle length $c = 2\pi/\gamma$.

In §5 of [7] it was shown for that it was possible to fit the rather complex stationary state-dependent birth-rate function associated with the $M_t/GI/s$ model having the arrival-rate function in (6) using a parametric function with only two parameters, in particular,

$$\lambda^p_k \equiv a \arctan b(k-c) + d, \tag{7}$$

which is nondecreasing in k with finite limits as k increases and decreases, and has the parameter four-tuple (a, b, c, d). For the arrival-rate function in (6), we found that it suffices to let $c = d = \bar{\lambda}$, so that leaves only the two parameters a and b. Consistent with expectations, the indirect estimation procedure was especially effective in obtaining a reasonable estimate of α with relatively small sample sizes, especially in the tails where there are only a few data points.

1.3 A Diagnostic Tool for Model Fitting

In addition to providing an alternative way to estimate the steady-state distribution, the fitted BD rates also provides a diagnostic tool to help determine what stochastic models may be appropriate for a queueing process arising in a complex setting. As discussed in [6], the fitted BD process may serve as a grey-box model as in [3, 17].

Our study of the $GI/GI/s$ models in [6] and $M_t/GI/s$ models in [7] show that the fitted rates in these models tend to have consistent structure that helps determine if such a model is appropriate. From these studies of the $GI/GI/s$ and $M_t/GI/s$ models, we find that these models have signatures. If these features are not seen in data analysis, then the candidate model is likely to be inappropriate. For example, when s is not too small, for a wide class of these models (i.e., for different arrival and service processes) the fitted death rates tend to have the approximate form

$$\bar{\mu}_k \approx (k \wedge s)\mu, \quad k \geq 0, \tag{8}$$

when the mean service time is $E[S] = 1/\mu$, where $a \wedge b \equiv \min\{a, b\}$ and $\equiv$ denotes equality by definition. (This holds exactly for exponential service distributions.) The fitting can detect systematic deviations from these s-server models, e.g., caused by time-varying staffing or agent absenteeism, as discussed in [13, 26].

Figure 1 (a variant of Figure 23 of [7]) illustrates by showing the fitted BD rates obtained from 25 weeks of data from an Israeli emergency department studied in [1, 29].

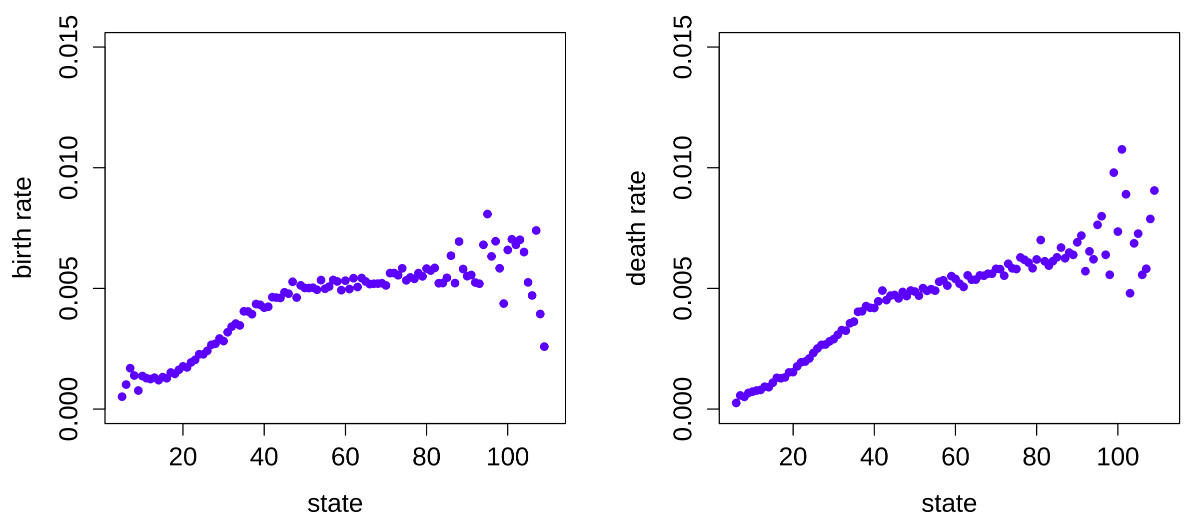

Figure 1: The fitted state-dependent birth rate $\bar{\lambda}_k$ (left) and death rate $\bar{\mu}_k$ (right) obtained from arrival and departure data in an Israeli emergency department over 25 weeks, taken from [29]

It is well known that the arrivals to an ED vary strongly over time, just as in most service systems. Thus, a natural candidate rough aggregate model for an ED is the $M_t/GI/\infty$ queue, which has a nonhomogeneous Poisson process (NHPP) as its arrival process, i.i.d. service (length-of-stay, LoS) times with some general (non-exponential, perhaps lognormal) distribution, s servers, unlimited waiting space and service in order of arrival. The assumption that the length-of-stay random variables are i.i.d. might be postulated under the assumption that the length of stay should only depend on the patient's medical condition.

Given an understanding of what the fitted BD rates look like in $M_t/GI/\infty$ models, the fitted rates for the ED in Figure 1 are very revealing. As anticipated, the fitted birth rates are roughly consistent with an NHPP (M_t) arrival process having a periodic arrival rate function, but the fitted death rates are inconsistent with the IS model having i.i.d. service times; e.g. see Figures 1, 4, 5 and 8 in [7]. The fitted death rates are approximately piecewise-linear with the slope changing at about $k = 40$, but the death-rate function does not look like (8) for either $s = 40$ or $s = \infty$. These tentative conclusions about the ED based on the analysis of $M_t/GI/s$ queues are strongly supported by further data analysis in [29]. The data analysis in [29] supports an $M_t/G_t/\infty$, where there is strong time-dependence in the LoS distribution (the G_t) as well as the arrival rate function. That conclusion in turn is consistent with other observations, e.g., see [1, 24] and references there. The fitted BD is convenient because it quickly exposes the difficulty with a model using i.i.d. service times.

1.4 New Asymptotic Formulas

The studies [6, 7] have exposed important properties of the fitted BD rates, but they have not generated explicit formulas for the fitted BD rates, except in the special cases in which the fitted BD rates are obviously simple because of Markovian structure. For example, the limiting fitted death-rate function in $G_t/M/s$ models (where the service times are independent of the arrival process) is always given by (8).

The present paper contributes by establishing many-server heavy-traffic (MSHT) limits that yield explicit first-order fluid approximations for the fitted BD rates in the time-varying $M_t/GI/\infty$ infinite-server (IS) model. These limiting rates show what the fitted rates should look like in large-scale IS models. Thus the results here facilitate model diagnosis.

We also establish new limits for the steady-state distribution of the periodic $M_t/GI/\infty$, clarifying and extending results for the $M_t/M/\infty$ model with a sinusoidal arrival-rate function in [28]. We show that for all models with sinusoidal arrival rate, the limiting steady-state distribution is the arcsine law on an interval with end points depending on the model parameters. Theorem 3.3 provides a simple explanation through a new representation for the mean function.

Here is how the paper is organized: We start in §2 by providing background on the time-varying $M_t/GI/\infty$ IS model. In §3 we propose explicit approximations for the steady-state distribution and the BD rates. Then in §4 we establish MSHT limits that show the approximations are asymptotically correct. In §4.2 we also establish limits for state occupation times in more general time-varying IS models.

In §5 we establish additional results for the case of sinusoidal arrival rates, which are often used in studies of queues with time-varying arrival rates, e.g., [14]. In §6 we carefully examine the case of deterministic service times. We show that the fitted rates can be decreasing functions of the state with long deterministic service times. We give detailed formulas for exponential and hyperexponential service times in §7. In §8 we show that the local-balance method for getting the steady-state distribution from the fitted BD rates described in §1.1 does not extend directly to the fluid limit. Finally, we draw conclusions in §9.

2 Background

2.1 The $M_t/GI/\infty$ Queueing Model

The stochastic queueing models considered in this paper are $M_t/GI/\infty$ infinite-server (IS) queueing models, having a nonhomogeneous Poisson process (NHPP, the M_t) as an arrival process with a time-varying arrival-rate function $\lambda(t)$. For simplicity, we assume that the arrival-rate function is differentiable and bounded above and below, i.e., we assume that there are constants λ_L and λ_U such that

$$0 < \lambda_L \leq \lambda(t) \leq \lambda_U < \infty \quad \text{for all} \quad t. \tag{9}$$

The service times of successive customers are assumed to be independent and identically distributed (i.i.d.) random variables, each distributed as a random variable S with a general cumulative distribution function (cdf) G having finite mean $E[S] = 1/\mu$. There are infinitely many servers, so that each customer enters service immediately upon arrival. We assume that the arrival process is independent of the service times. We assume that the system started empty in the distant past, but that condition can be relaxed by letting $\lambda(t) = 0$ before some starting time t_0.

By Theorem 1 of [9], the number in system at time t, denoted by $Q(t)$, has a Poisson distribution for each t with mean function

$$m(t) \equiv E[Q(t)] \quad = \quad E[\lambda(t - S_e)]E[S] = E[S] \int_0^\infty \lambda(t - s) dG_e(s), \quad t \geq 0, \tag{10}$$

where S_e is a random variable with the stationary-excess cdf G_e associated with the service-time cdf G, i.e.,

$$G_e(t) \equiv P(S_e \leq t) \equiv \frac{1}{E[S]} \int_0^t (1 - G(s))\, ds, \quad t \geq 0. \tag{11}$$

Moreover, the departure process in the $M_t/GI/\infty$ model is an NHPP with departure rate function

$$\delta(t) = E[\lambda(t - S)] = \int_0^\infty \lambda(t - s) dG(s), \quad t \geq 0. \tag{12}$$

In the special case of exponential (M) service, we have $S_e \stackrel{\mathrm{d}}{=} S$, where $\stackrel{\mathrm{d}}{=}$ means "equal in distribution," so that, by combining (10) and (12), we obtain the familiar conclusion

$$\delta(t) = \frac{m(t)}{E[S]} = m(t)\mu. \tag{13}$$

2.2 The $M_t/GI/\infty$ Queueing Models with Periodic Arrival Rates

We will focus on the special case of the $M_t/GI/\infty$ IS model in which the arrival-rate function λ is periodic with periodic c and average arrival rate $\bar{\lambda}$. The mean $m(t)$ in (10) and the departure rate $\delta(t)$ in (12) become periodic functions when the arrival-rate function is periodic. If the system starts empty at time 0, then $m(t)$ and $\delta(t)$ converge (in a nondecreasing way) to those periodic expressions, as shown in §2.2 of [7].

For the IS model we have the following stronger result.

Theorem 2.1 (*regenerative structure*) *For the $M_t/GI/\infty$ model with periodic arrival-rate function having period c that starts empty at time 0, the stochastic process $\{Q(nc+t) : n \geq 1\}$ is a regenerative process for each t, $0 \leq t < c$, with the epochs n at which $Q(nc+t) = 0$ serving as regeneration epochs. Moreover, the mean time between successive regenerations is $1/P(X(m(t)) = 0) < \infty$,* *where $X(m)$ is a random variable with the Poisson distribution having mean m and $m(t)$ is the periodic mean function. Thus, the limits in (5) are valid. Moreover, α^c can be regarded as a special case of the direct steady-state pmf α in §1.1 if we randomize the initial time uniformly over the interval $[0, c]$.*

Proof. From the convergence of $m(t)$ to the periodic limit, we have $Q(nc + t) \Rightarrow X(m(t))$ as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution. Thus, $P(Q(nc+t) = 0) \to P(X(m(t)) > 0)$ as $n \to \infty$, where $m(t) > 0$ by virtue of (9). We then note that the sum on the right in the expression for $\alpha_k(t)$ in (5) when $k = 0$ is the averaged discrete-time renewal counting function, so that the stated limit follows from the law of large numbers (LLN) for renewal processes, which implies that the limit must coincide with the mean time between successive regenerations. The regenerative structure then makes all the processes cumulative processes, e.g., as in [11]. The associated expressions for α^c in (5) are elementary consequences. ∎

For the special case of a sinusoidal arrival rate function, many structural properties were established in [8]. In particular, for the arrival function in (6) (and starting empty in the distant past),

$$m(t) = \bar{\lambda} E[S] m_1(t), \quad \text{where} \quad m_1(t) = 1 + \beta \left(\mathcal{C} \sin(\gamma t) - \mathcal{S} \cos(\gamma t) \right), \tag{14}$$

with S_e distributed according to G_e as in (11) and

$$\mathcal{C} \equiv E[\cos(\gamma S_e)] \quad \text{and} \quad \mathcal{S} \equiv E[\sin(\gamma S_e)]. \tag{15}$$

For M service, $\mathcal{C} = 1/(1 + \gamma^2)$ and $\mathcal{S} = \gamma/(1 + \gamma^2)$. From (14), (10) and (12), we see that the corresponding formula for the departure rate is

$$\delta(t) = \bar{\lambda} \delta_1(t), \quad \text{where} \quad \delta_1(t) = 1 + \beta \left(\mathcal{C}' \sin(\gamma t) - \mathcal{S}' \cos(\gamma t) \right) \tag{16}$$

and

$$\mathcal{C}' \equiv E[\cos(\gamma S)] \quad \text{and} \quad \mathcal{S}' \equiv E[\sin(\gamma S)]. \tag{17}$$

2.3 The MSHT Fluid Approximation for Large Scale

We will be interested in the fluid approximation of the periodic $M_t/GI/\infty$ IS model, obtained by letting the scale parameter (the average arrival rate) $\bar{\lambda}$ get large. Following common practice, we will consider a sequence of $M_t/GI/\infty$ models indexed by n, where $\bar{\lambda}_n = n$, $n \ge 1$. We let the service-time cdf G be held fixed independent of n and let the arrival rate function be $\lambda_n(t) = n\lambda_1(t)$.

Let $Q_n(t)$ be the number in system as a function of n. Let $A_n(t)$ and $D_n(t)$ count the number of arrivals and departures over the interval $[0, t]$, $t \ge 0$, respectively, again as a function of n. Let $(m_n(t), \lambda_n(t), \delta_n(t))$ be the triple $(m(t), \lambda(t), \delta(t))$ in §2.1 as a function of the scale parameter n. Because of the linearity of the model, we have the scaling property

$$n^{-1}(m_n(t), \lambda_n(t), \delta_n(t)) = (m_1(t), \lambda_1(t), \delta_1(t)) \quad \text{for all} \quad n \ge 1, \tag{18}$$

where $(m_1(t), \lambda_1(t), \delta_1(t))$ is understood to be $(m_n(t), \lambda_n(t), \delta_n(t))$ when $n = 1$; see §4 of [18]. As a regularity condition, we assume that (9) holds for $\lambda_1(t)$, which implies that $0 < m_1(t) < \infty$ for all t as well.

Moreover, by the weak LLN (WLLN) for the Poisson distribution,

$$n^{-1} \left(Q_n(t), A_n(t), D_n(t) \right) \Rightarrow (m_1(t), \lambda_1(t), \delta_1(t)) \quad \text{as} \quad n \to \infty \quad \text{for each} \quad t > 0. \tag{19}$$

The WLLN provides support for the simple fluid approximation

$$(Q_n(t), A_n(t), D_n(t)) \approx (nm_1(t), n\lambda_1(t), n\delta_1(t))\,. \tag{20}$$

See [19], Theorem 3.1 of [23] and references therein for more general functional LLN (FWLLN), yielding unofrm convergence over bounded intervals.

3 Simple Fluid Approximations

In this section we propose simple fluid approximations for the steady-state distribution and the fitted BD rates in the periodic $M_t/GI/\infty$ IS model with large scale based on the asymptotics in §2.3.

3.1 The Overall Steady-State Distribution

We now develop an approximation for the steady-state distribution α^c in (5). For simplicity, we focus on the associated cumulative distribution function (cdf). Let Z_n be a random variable with the steady-state cdf of the scaled queue length $\bar{Q}_n(t) \equiv n^{-1}Q_n(t)$, using the scaling in §2.3. Assuming that the system started empty in the distant past, so that the process $\{\bar{Q}_n(t) : t \geq 0\}$ is in dynamic steady state, with overall average value 1 and period c. Then

$$P(Z_n \leq 1 + x) \equiv \frac{1}{c}\int_0^c P(\bar{Q}_n(t) \leq 1 + x)\, dt \tag{21}$$

Moreover, from the FWLLN version of (19) in Theorem 3.1 of [23] and the continuous mapping theorem, it follows that $Z_n \Rightarrow Z$ as $n \to \infty$, where Z gives the fluid model steady-state cdf, i.e.,

$$P(Z \leq 1 + x) \equiv \frac{1}{c}\int_0^c 1_{\{m_1(t) \leq 1+x\}}\, dt, \tag{22}$$

provided that $m_1(t) = 1 + x$ for only finitely many t. We propose (22) as a relatively simple approximation for (21).

We illustrate with an explicit formula for the cdf of Z with the sinusoidal arrival rate function in (6) that we will prove in §5.2.

Theorem 3.1 (*fluid steady-state cdf for sinusoidal arrival rates*) *Consider the fluid model associated with the $M_t/GI/\infty$ model with the sinusoidal arrival-rate function in* (6) *having parameter triple $(\bar{\lambda}, \beta, \gamma)$ with $\bar{\lambda} = 1$. Then*

$$P(Z \geq 1 - x) = P(Z \leq 1 + x) = \frac{1}{2} + \frac{1}{\pi}\arcsin\left(x/\beta S_U\right), \quad 0 \leq x \leq \beta s_U, \tag{23}$$

where s_U is given in (26)*; i.e., $Z - 1$ has the arcsine cdf on $[-\beta s_U, \beta s_U]$. Thus, the variance of Z is $\beta^2 s_U^2/2$. For the special case of an exponential service distribution, $s_U = 1/\sqrt{1+\gamma^2}$, as indicated in Corollary 7.1 in §7.*

Theorem 3.1 improves Theorem 3.1 of [28] for the $M_t/M/\infty$ special case and extends it to general service-time distributions. (The notation is slightly different here; the results above agree with [28].) We also establish a local version of Theorem 3.1 in §5.2.

Theorem 3.2 (*local limit for sinusoidal arrival rates*) *In the setting of Theorem 3.1,*

$$\frac{1}{t}\int_0^t 1_{\{Q_n(s)=\lfloor n(1+x)\rfloor\}}\,ds \to \frac{\gamma}{2\pi}\int_0^{2\pi/\gamma} P(Q_n(s)=\lfloor n(1+x)\rfloor)\,ds \tag{24}$$

as $t\to\infty$ and

$$\frac{\gamma}{2\pi}\int_0^{2\pi/\gamma} P(Q_n(s)=\lfloor n(1+x)\rfloor)\,ds \to \frac{2}{\beta s_U\sqrt{1-(x/\beta s_U)^2}} \quad as \quad n\to\infty. \tag{25}$$

The limit in (25) is the arcsine pdf on the interval $[-\beta s_U, \beta s_U]$.

3.2 Simplified Formulas for Sinusoidal Arrival Rates

The relatively clean mathematical results in Theorems 3.1 and 3.2 primarily follow from a simple representation of the mean function $m(t)$ in the $M_t/GI/\infty$ model with sinusoidal arrival rate, which is primarily based on [8]. The representation in [8] is as the linear function of two trigonometric functions in (14). The alternative representation is in terms of a single trigonometric function centered at the times t^* where the extreme values occur. In particular, in §5 we prove the following extension of [8]. A previous simplification was given in Theorem 6.3 of [21]. The following is essentially equivalent to Corollary 7.1 of [20].

Theorem 3.3 (*the mean function with GI service*) *Consider the $M_t/GI/\infty$ model with mean service time $E[S]=1$ and sinusoidal arrival rate function in (6), starting empty in the distant past. Assume that neither $\mathcal{S}$ nor $\mathcal{C}$ in (15) is 0. Then all extreme points of $s(t)\equiv(m_1(t)-1)/\beta$ and thus $m_1(t)$ and $m(t)$ in (14) are attained at the points $t^*+(k\pi/\gamma)$ for integer k, where*

$$t^*\equiv t^*(\gamma)=\frac{\pi}{2\gamma}+\frac{1}{\gamma}\tan^{-1}(\mathcal{S}/\mathcal{C}) \quad and \quad s_U\equiv\sup_{t\ge 0}(s(t))=[\mathcal{S}^2+\mathcal{C}^2]^{1/2}. \tag{26}$$

In addition,

$$s(t^*+t)=s(t^*)\cos(\gamma t)=s(t^*-t) \quad for\ all \quad t, \tag{27}$$

where

$$s(t^*)=\frac{\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))(\mathcal{S}^2+\mathcal{C}^2)}{\mathcal{C}}=\pm s_U, \tag{28}$$

If $\mathcal{S}>0$ and $\mathcal{C}>0$, then $s(t^*)=+s_U$.

Comparing (16) to (14), we see that the departure rate from an $M_t/GI/\infty$ queue with a sinusoidal arrival rate function also can be represented in terms of a single trigonometric function, which facilitates analysis of networks of IS queues with sinusoidal arrival rates, as in [21].

Corollary 3.1 (*the departure-rate function with GI service*) *Consider the $M_t/GI/\infty$ model with sinusoidal arrival rate function in (6), starting empty in the distant past. Assume that neither $\mathcal{S}'$ nor $\mathcal{C}'$ in (17) is 0. All extreme points of $\delta(t)$ in (16) are attained at the points $t^{**}+(k\pi/\gamma)$ for integer k, where*

$$t^{**}\equiv t^{**}(\gamma)=\frac{\pi}{2\gamma}+\frac{1}{\gamma}\tan^{-1}(\mathcal{S}'/\mathcal{C}') \quad and \quad \delta_U\equiv\sup_{t\ge 0}(\delta(t))=[(\mathcal{S}')^2+(\mathcal{C}')^2]^{1/2}. \tag{29}$$

Then

$$\delta(t^{**}+t)=\delta(t^{**})\cos(\gamma t)=\delta(t^{**}-t) \quad for\ all \quad t, \tag{30}$$

where

$$\delta(t^{**})=\frac{\cos(\tan^{-1}(\mathcal{S}'/\mathcal{C}'))((\mathcal{S}')^2+(\mathcal{C}')^2)}{\mathcal{C}'}=\pm\delta_U, \tag{31}$$

If $\mathcal{S}'>0$ and $\mathcal{C}'>0$, then $\delta(t^{**})=+\delta_U$.

3.3 The Proposed Fluid Approximation for the Fitted BD Rates

We now define a state-dependent fitted fluid input rate function λ^f and a state-dependent fitted fluid output rate function μ^f, as functions of the basic fluid model triple $(m_1, \lambda_1, \delta_1)$ in (18)-(20), analogous to the fitted birth and death rates for stochastic models introduced and studied in [6, 7]. The general idea is that the fitted input rate $\lambda^f(x)$ in state x should be the average time-dependent input rate $\lambda_1(t)$ over all times t at which $m_1(t) = x$. Similarly, the fitted output rate $\mu^f(x)$ in state x should be the average time-dependent output rate $\delta_1(t)$ over all times t at which $m_1(t) = x$.

The present fluid model setting is more elementary because there is no randomness; the functions are all deterministic. However, there is some question about how the average should be computed. We make some assumptions on the states x that we consider. First, we consider x for which the inverse $m_1^{-1}(x) \equiv \{t \in [0, c] : m_1(t) = x\}$ is a finite set. That might hold for all x, but we only require it for the x we consider. We contend that the average should be a weighted average, depending on the derivative $m_1(t)$, denoted by $\dot{m}_1(t)$. We shall want to apply the inverse function theorem, so we assume that the derivative $\dot{m}_1(t)$ is well defined and positive at all times t such that $m_1(t) = x$ for each state x we consider. Let $n_m(x)$ be the number of points in the set $m_1^{-1}(x)$ and let $t_j^m(x)$ be the j^{th} such point ordered within $[0, c)$. Then we let

$$\lambda^f(x) = \sum_{j=1}^{n_m(x)} p_j^m(x) \lambda_1(t_j^m(x)) \quad \text{and} \quad \mu^f(x) = \sum_{j=1}^{n_m(x)} p_j^m(x) \delta_1(t_j^m(x))), \tag{32}$$

where

$$p_j^m(x) \equiv \frac{a_j^m(x)}{\sum_{j=1}^{n_m(x)} a_j^m(x)} \quad \text{and} \quad a_j^m(x) \equiv \frac{1}{|\dot{m}_1(t_j^m(x))|}. \tag{33}$$

Our second assumption implies that $0 < a_j^m(x) < \infty$ for the x we consider. Paralleling (20), we propose the following fluid approximations for the fitted BD rates in model n:

$$\bar{\lambda}_{n,\lfloor nx \rfloor}(\infty) \approx n\lambda^f(x) \quad \text{and} \quad \bar{\mu}_{n,\lfloor nx \rfloor}(\infty) \approx n\mu^f(x) \tag{34}$$

for all states x such that $m_1(x) > 0$.

We next provide support for our formulation with the weights $p_j^m(x)$ in (32) and (33) by establishing a heavy-traffic limit for the fitted birth rates in the $M_t/GI/\infty$ IS model.

4 The Many-Server Heavy-Traffic Limit for the Fitted Rates

4.1 MSHT Limit for the Fitted BD Rates

We consider the sequence of periodic $M_t/GI/\infty$ IS queueing models indexed by the average arrival rate, $\bar{\lambda}_n = n$, introduced in §2.3, but now we focus on a periodic arrival-rate function, assuming that the system started empty in the distant past.

We now state our main result. For that purpose, let $\lfloor x \rfloor$ be the greatest integer less than or equal to x.

Theorem 4.1 (*heavy-traffic limit of the fitted rates*) *Consider the sequence of $M_t/GI/\infty$ models indexed by n starting empty in the distant past, where $\lambda_1(t)$ is a bounded differentiable periodic function. Consider a state x such that $m_1^{-1}(x)$ is a finite set and the derivative $\dot{m}_1(t)$ is well* *defined and positive at all times t such that $m_1(t) = x$. Then, the fitted birth and death rates satisfy*

$$\bar{\lambda}_{n,k}(t) \equiv \frac{\int_0^t 1_{\{Q_n(s)=k\}} \lambda_n(s)\, ds}{\int_0^t 1_{\{Q_n(s)=k\}}\, ds} \to \frac{\int_0^c P(Q_n(s)=k)\lambda_n(s)\, ds}{\int_0^c P(Q_n(s)=k)\, ds} \equiv \bar{\lambda}_{n,k}(\infty) \quad \text{and} \tag{35}$$

$$\bar{\mu}_{n,k}(t) \equiv \frac{\int_0^t 1_{\{Q_n(s)=k\}} k(\delta_1(s)/m_1(s))\, ds}{\int_0^t 1_{\{Q_n(s)=k\}}\, ds} \to \frac{\int_0^c P(Q_n(s)=k) k(\delta_1(s)/m_1(s))\, ds}{\int_0^c P(Q_n(s)=k)\, ds} \equiv \bar{\mu}_{n,k}(\infty)$$

as $t \to \infty$. Moreover,

$$n^{-1}\bar{\lambda}_{n,\lfloor nx \rfloor}(\infty) \to \lambda^f(x) \quad \text{and} \quad n^{-1}\bar{\mu}_{n,\lfloor nx \rfloor}(\infty) \to \mu^f(x) \quad \text{as} \quad n \to \infty, \tag{36}$$

for all x such that $m_1(x) > 0$, where $\lambda^f(x)$ and $\mu^f(x)$ are defined in (32) with $p_j^m(x)$ defined in (33). In addition,

$$\mu^f(x) = x \sum_{j=1}^{n_m(x)} p_j^m(x)(\delta_1(t_j^m(x))/m_1(t_j^m(x))). \tag{37}$$

Proof. The first expression for $\bar{\lambda}_{n,k}(t)$ in (35) is elementary because the arrival rate depends on time, but not on the state. The first expression $\bar{\mu}_{n,k}(t)$ in (35) follows from Theorem 2.6 of [7], which states that, conditional on $Q(t) = k$, the departure rate at time t is

$$\delta_k(t) = k g_{k,t}(0) = \frac{k\mu E[\lambda(t-S)]}{E[\lambda(t-S_e)]} = \frac{k\delta(t)}{m(t)}. \tag{38}$$

The limits in (35) follow from Theorem 2.1, treating the numerators and denominators separately. For the scaling limits in (36), the main new part of Theorem 4.1, we exploit the fact the $Q_n(t)$ has a Poisson distribution with mean $nm_1(t)$ for each t and $n \geq 1$, which is asymptotically Gaussian with that mean and variance. We prove the remaining limit in (36) by establishing limits for the numerators and denominators on the right in (35) separately. The factor n^{-1} on the left in (36) is removed by writing $\lambda_n(t) = n\lambda_1(t)$ and $k = n(k/n)$ for $k = \lfloor nx \rfloor$ in the expression for $\bar{\mu}_{n,k}(t)$. The denominators are interesting in their own right because they are directly limits for occupation times. Thus, we establish a MSHT limit for occupation times in a general sequence of $M_t/GI/\infty$ models scaled as above in §4.2 below. Given the continuity of the rate functions $\lambda_1(t)$ and $\delta_1(t)/m_1(t)$, it suffices to apply the same arguments to the numerator and denominator, so the full proof is completed in §4.2 below. This logic for the fitted death rates directly leads to the expression (37), but the two formulas for $\mu^f(x)$ in (32) and (37) are actually equivalent, because at the times $t_j^m(x)$, we have $m_1(t_j^m(x)) = x$, which makes $x\delta_1(t_j^m(x))/m_1(t_j^m(x)) = \delta_1(t_j^m(x))$. ■

4.2 A MSHT Limit for Occupation Times

We now complete the proof of Theorem 4.1 by establishing MSHT limits for the denominators in (35). The denominators are directly limits for occupation times, for which there is a substantial literature, e.g., [4], but evidently nothing previously for time-varying models. Thus, we establish a MSHT limit for occupation times in a general sequence of $M_t/GI/\infty$ models scaled as above, without requiring that the arrival-rate function be periodic.

Theorem 4.2 (*MSHT occupation time limits*) *Consider a general sequence of $M_t/GI/\infty$ IS models with scaling as in §2.3. If m_1 is differentiable at t_0 with $m_1(t_0) = x$ and $\dot{m}_1(t_0) \neq 0$, then for each u, $-\infty < u < +\infty$,*

$$\sqrt{n} P(Q_n(t_0 + (u/\sqrt{n})) = \lfloor nx \rfloor) \to \frac{\phi(u\dot{m}_1(t_0)/\sqrt{x})}{\sqrt{x}} \quad \text{as} \quad n \to \infty. \tag{39}$$

Moreover, there are times t_1 and t_2 such that $t_1 < t_0 < t_2$ with $m_1(t) = x$ for $t \in [t_1, t_2]$ only at t_0. For any such times,

$$n \int_{t_1}^{t_2} P(Q_n(s) = \lfloor nx \rfloor)\, ds \to \frac{1}{\dot{m}_1(t_0)} \quad \textit{as} \quad n \to \infty. \tag{40}$$

We first establish three lemmas used in the proof of Theorem 4.2.

Lemma 4.1 *For all real x,*

$$e^{\sqrt{n}x}(1 + (x/\sqrt{n}))^{-n} \to e^{x^2} \quad \textit{as} \quad n \to \infty. \tag{41}$$

Proof. Use natural logarithms and the expansion $\log(1+x) = x - (x^2/2) + O(x^3)$ as $x \to 0$ to get

$$\begin{aligned} \log(e^{\sqrt{n}x}(1 + (x/\sqrt{n})^{-n}) &= \sqrt{n}x - n\log(1 + (x/\sqrt{n})) \\ &= \sqrt{n}x - \sqrt{n}x + \frac{x^2}{2} - O(x^3/3\sqrt{n}) \to \frac{x^2}{2} \quad \text{as} \quad n \to \infty. \quad \blacksquare \end{aligned} \tag{42}$$

We use Lemma 4.1 to establish two limits for the Poisson distribution. Let ϕ be the pdf of a standard $N(0,1)$ Gaussian random variable. Let $f(n) \sim g(n)$ as $n \to \infty$ mean that $f(n)/g(n) \to 1$ as $n \to \infty$. Let $X(n)$ have a Poisson distribution with mean n.

Lemma 4.2 (*first Poisson limit*) *Let $X(n)$ have a Poisson distribution with mean n. For all real x,*

$$\sqrt{n}P(X(n) = \lfloor n + x\sqrt{n} \rfloor) \to \phi(x) \quad \textit{as} \quad n \to \infty. \tag{43}$$

Proof. Using Stirling's formula in line 2 and Lemma 4.1 in line 5, we have as $n \to \infty$

$$\begin{aligned} \sqrt{2\pi n}P(X(n) = \lfloor n + x\sqrt{n} \rfloor) &= \frac{\sqrt{2\pi n}e^{-n}n^{\lfloor n + x\sqrt{n} \rfloor}}{(\lfloor n + x\sqrt{n} \rfloor)!} \\ &\sim \frac{\sqrt{2\pi n}e^{-n}n^{n + x\sqrt{n}}}{(n + x\sqrt{n})^{n + x\sqrt{n}}e^{-(n + x\sqrt{n})}\sqrt{2\pi(n + x\sqrt{n})}} \\ &\sim e^{x\sqrt{n}}(1 + (x/\sqrt{n}))^{-(n + x\sqrt{n})} \\ &\sim e^{x\sqrt{n}}(1 + (x/\sqrt{n}))^{-n}(1 + (x/\sqrt{n}))^{-x\sqrt{n}} \\ &\sim e^{x^2/2}e^{-x^2} = e^{-x^2/2} \quad \text{as} \quad n \to \infty, \end{aligned} \tag{44}$$

which is equivalent to what is claimed. $\blacksquare$

Lemma 4.3 (*second Poisson limit*) *Let $X(n)$ have a Poisson distribution with mean n. For all real x,*

$$\sqrt{n}P(X(n + x\sqrt{n}) = n) \to \phi(x) \quad \textit{as} \quad n \to \infty. \tag{45}$$

Proof. Using Stirling's formula in line 2 and Lemma 4.1 in line 4, we have as $n \to \infty$

$$\begin{aligned}
\sqrt{2\pi n}P(X(n+x\sqrt{n})=n) &= \frac{\sqrt{2\pi n}e^{-(n+x\sqrt{n})}(n+x\sqrt{n})^n}{n!} \\
&\sim \frac{\sqrt{2\pi n}e^{-n}e^{-x\sqrt{n}}(n+x\sqrt{n})^n}{e^{-n}n^n\sqrt{2\pi n}} \\
&\sim e^{-x\sqrt{n}}(1+(1/x\sqrt{n}))^n \sim e^{-x^2/2} \quad \text{as} \quad n \to \infty, \qquad (46)
\end{aligned}$$

which is equivalent to what is claimed. ∎

Remark 4.1 (*local limit theorem*) Lemma 4.2 can also be regarded as a consequence of the local central limit theorem, because the Poisson random variable $X(n)$ can be represented as the sum of n i.i.d. random variables distributed as $X(1)$; e.g., see [5, 12, 22].

We now complete the proof of Theorem 4.2.

Proof of Theorem 4.2. Let $m \equiv nx$. We apply a Taylor expansion of m_1 about t_0, writing $m_1(t_0+(u/\sqrt{n})) = m_1(t_0)+u\dot{m}_1(t_0)/\sqrt{n}+o(1/\sqrt{n})$ as $n \to \infty$. Then, applying Lemma 4.3 in the last step, we have

$$\begin{aligned}
\sqrt{n}P(Q_n(t_0+(u/\sqrt{n}))=\lfloor nx \rfloor) &= \sqrt{n}P(X(n(m_1(t_0)+\frac{u\dot{m}_1(t_0)}{\sqrt{n}}+o(1/\sqrt{n}))=\lfloor nx \rfloor) \\
&= \frac{\sqrt{nx}}{\sqrt{x}}P(X(nx+\frac{\sqrt{nx}u\dot{m}_1(t_0)}{\sqrt{x}}+o(\sqrt{nx}))=\lfloor nx \rfloor) \\
&= \frac{\sqrt{m}}{\sqrt{x}}P(X(m+\frac{\sqrt{m}u\dot{m}_1(t_0)}{\sqrt{x}}+o(\sqrt{m}))=\lfloor m \rfloor) \\
&\to \frac{\phi(u\dot{m}_1(t_0)/\sqrt{x})}{\sqrt{x}} \quad \text{as} \quad m \to \infty \qquad (47)
\end{aligned}$$

and so also as $n \to \infty$, justifying (39). For (40), we apply (39). First, we observe that

$$n\int_{t_1}^{t_2} P(Q_n(s)=\lfloor nx \rfloor)\,ds = \lim_{a\to\infty} n\int_{t_0-a/\sqrt{n}}^{t_0+a/\sqrt{n}} P(Q_n(s)=\lfloor nx \rfloor)\,ds. \qquad (48)$$

Then, by successively making the change of variables $s \equiv t_0+u/\sqrt{n}$ so that $ds = du/\sqrt{n}$, $v \equiv u\dot{m}_1(t_0)/\sqrt{x}$ and $m \equiv nx$, we obtain

$$\begin{aligned}
n\int_{t_0-a/\sqrt{n}}^{t_0+a/\sqrt{n}} P(Q_n(s)=\lfloor nx \rfloor)\,ds &= \sqrt{n}\int_{-a}^{+a} P(X(nx+\sqrt{nx}u\dot{m}_1(t_0)/\sqrt{x}+o(\sqrt{n}))=\lfloor nx \rfloor)\,du \\
&= \frac{\sqrt{nx}}{\dot{m}_1(t_0)}\int_{-a}^{+a} P(X(nx+\sqrt{nx}v+o(\sqrt{nx}))=\lfloor nx \rfloor)\,dv \\
&= \frac{\sqrt{m}}{\dot{m}_1(t_0)}\int_{-a}^{+a} P(X(m+\sqrt{m}v+o(\sqrt{m}))=\lfloor m \rfloor)\,dv \\
&\to \frac{1}{\dot{m}_1(t_0)}\int_{-a}^{+a} \phi(v)\,dv \quad \text{as} \quad m \to \infty \quad \text{or} \quad n \to \infty \\
&\to \frac{1}{\dot{m}_1(t_0)} \quad \text{as} \quad a \to \infty. \quad \blacksquare \qquad (49)
\end{aligned}$$

4.3 Aggregate Estimator for Large Scale

Good estimates of the fitted rates $\bar{\lambda}_{n,k}$ and $\bar{\mu}_{n,k}$ from data tends to require a very large amount of data as the scale n increases. Notice that the probability $P(Q_n(t) = k)$ appearing in (35) is at most of order $O(1/\sqrt{n})$ because $Q_n(t)$ has a Poisson distribution with mean $nm_1(t) = O(n)$; see Theorem 4.2. Thus large samples may be needed to obtain accurate estimations, even for the most relevant states k, when n is very large.

To address this problem, we propose aggregate estimators of the birth and death rates, and show that they are asymptotically equivalent to the direct estimators. We do so by showing that they too converge to the fluid rate functions λ^f and μ^f after scaling. In particular, for large n, we propose estimating the birth rate by $\bar{\lambda}_{n,k} \approx \lambda_k^{ag}(mc; n, \epsilon)$ for suitably large m and suitably small ϵ, where

$$\bar{\lambda}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon) \quad \equiv \quad \frac{\int_0^{mc} 1_{\{\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon \rfloor\}} \lambda_n(t)\, dt}{\int_0^{mc} 1_{\{\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon)\rfloor\}}\, dt} \tag{50}$$

for each $n \ge 1$, $m \ge 1$ and $\epsilon > 0$.

The corresponding aggregate estimator for the death rate is somewhat more complicated. We would first write

$$\bar{\mu}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon) \quad \equiv \quad \frac{\int_0^{mc} \sum_{k=\lfloor nx \rfloor+1}^{\lfloor n(x+\epsilon)\rfloor} 1_{\{Q_n(t)=k\}} k(\delta_1(t)/m_1(t))\, dt}{\int_0^{mc} 1_{\{\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon \rfloor\}}\, dt}, \tag{51}$$

which is not so convenient. However, we can bound $\bar{\mu}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon)$ above and below by more convenient estimators, i.e.,

$$\bar{\mu}^{ag,L}_{\lfloor nx \rfloor}(mc; n, \epsilon) \le \bar{\mu}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon) \le \bar{\mu}^{ag,U}_{\lfloor nx \rfloor}(mc; n, \epsilon), \tag{52}$$

where

$$\begin{aligned} \bar{\mu}^{ag,L}_{\lfloor nx \rfloor}(mc; n, \epsilon) &\equiv \frac{\int_0^{c} 1_{\{\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon)\rfloor\}} \lfloor nx \rfloor (\delta_1(t)/m_1(t))\, dt}{\int_0^{c} 1_{\{\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon)\rfloor\}}\, dt} \quad \text{and} \\ \bar{\mu}^{ag,U}_{\lfloor nx \rfloor}(mc; n, \epsilon) &\equiv \frac{\int_0^{c} 1_{\{\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon)\rfloor\}} \lfloor n(x+\epsilon) \rfloor (\delta_1(t)/m_1(t))\, dt}{\int_0^{c} 1_{\{\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon)\rfloor\}}\, dt}, \end{aligned} \tag{53}$$

because $\lfloor nx \rfloor \le k \le \lfloor n(x+\epsilon) \rfloor$ in the numerator of (50)

If we let $n \uparrow \infty$ and then $\epsilon \downarrow 0$, then we get the same limits for $\bar{\mu}^{ag,L}_{\lfloor nx \rfloor}(mc; n, \epsilon)$ and $\bar{\mu}^{ag,U}_{\lfloor nx \rfloor}(mc; n, \epsilon)$ and so for all three. It would be natural to use the two bounds as direct estimators, knowing that the more precise formulation falls in between.

As before, from Theorem 2.1, we have

$$\bar{\lambda}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon) \to \bar{\lambda}^{ag}_{\lfloor nx \rfloor}(\infty; n, \epsilon) \quad \text{and} \quad \bar{\mu}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon) \to \bar{\mu}^{ag}_{\lfloor nx \rfloor}(\infty; n, \epsilon) \tag{54}$$

as $m \to \infty$, where

$$\bar{\lambda}^{ag}_{\lfloor nx \rfloor}(\infty; n, \epsilon) \quad \equiv \quad \frac{\int_0^{c} P(\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon)\rfloor \lambda_n(t)\, dt}{\int_0^{c} P(\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon)\rfloor)\, dt} \tag{55}$$

and

$$\bar{\mu}^{ag}_{\lfloor nx \rfloor}(\infty; n, \epsilon) \quad \equiv \quad \frac{\int_0^{c} \sum_{k=\lfloor nx \rfloor+1}^{\lfloor n(x+\epsilon)\rfloor} P(Q_n(t)=k) \rfloor k(\delta_1(t)/m_1(t))\, dt}{\int_0^{c} P(\lfloor nx \rfloor \le Q_n(t) \le \lfloor n(x+\epsilon)\rfloor)\, dt}, \tag{56}$$

and similarly for the two bounds in (53).

Theorem 4.3 (*MSHT asymptotics for the aggregate estimators*) *Consider the sequence of periodic $M_t/GI/\infty$ models starting empty in the distant past. Consider a state x such that $m_1^{-1}(x)$ is a finite set and the derivative $\dot{m}_1(t)$ is well defined and positive at all times t such that $m_1(t) = x$. If first $n \to \infty$ and then $\epsilon \downarrow 0$, then*

$$n^{-1}\bar{\lambda}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon) \to \lambda^f(x) \quad \textit{and} \quad n^{-1}\bar{\mu}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon) \to \mu^f(x) \tag{57}$$

for each $m \geq 1$, where $\lambda^f(x)$ and $\mu^f(x)$ are defined in (32) with $p_j^m(x)$ defined in (33). The same limit holds for the two bounding death rates in (53).

Proof. We apply Theorem 3.1 of [23] to get the functional weak LLN (FWLLN)

$$\sup_{0 \leq t \leq T} \{|n^{-1}Q_n(t) - m_1(t)|\} \Rightarrow 0 \quad \text{as} \quad n \to \infty \quad \text{for each} \quad T, \quad 0 < T < \infty. \tag{58}$$

We observe that the floor function is asymptotically negligible as $n \to \infty$, so that it can be ignored in the limit. We then apply the continuous mapping theorem with the integral functional to get the first limit on n. We use the assumption that the inverse of m_1, $m_1^{-1}(x)$, is a finite set for each x to have the set of discontinuities of the indicator function be a finite set. For each $t_j^m(x) \in m^{-1}(x)$, the limit of $[m^{-1}(x+\epsilon) - m^{-1}(x)]/\epsilon$ as $\epsilon \to 0$ is the absolute value of the derivative of $m^{-1}(x)$, which is $1/|\dot{m}_1(t_j^m(x))|$, by virtue of the inverse function theorem. The argument for the two founding fitted death rates is essentially the same. The limit for $\bar{\mu}^{ag}_{\lfloor nx \rfloor}(mc; n, \epsilon)$ then follows by a sandwiching argument. ■

As usual, let $f(n) = o(n)$ as $n \to \infty$ mean that $f(n)/n \to 0$ as $n \to \infty$.

Corollary 4.1 *Under the conditions of Theorems 4.1 and 4.3,*

$$\bar{\lambda}_{n,\lfloor nx \rfloor}(\infty) - \bar{\lambda}^{ag}(\lfloor nx \rfloor; n, \epsilon) = o(n) \quad \textit{as} \quad n \to \infty \quad \textit{and} \quad \epsilon \downarrow 0. \tag{59}$$

Remark 4.2 (*differentiability*) By (10) and (12), a natural sufficient condition to have the functions λ_1, m_1 and δ_1 all be bounded and differentiable is simply to have that condition imposed on λ_1.

Remark 4.3 (*conjectured joint limit*) The iterated limit in which first $n \to \infty$ and then $\epsilon \to 0$ in Theorem 4.1 provides theoretical support for the approximation

$$\lambda^e_{n,k_n} \approx n\lambda^f(x) \quad \text{and} \quad \mu^e_{n,k_n} \approx n\mu^f(x) \quad \text{when} \quad x = k_n/n \tag{60}$$

and n is not too small. We conjecture that there is direct convergence as well when $n \to \infty$ with $k_n/n \to x$ under regularity conditions, but that remains to be shown.

Remark 4.4 (*exponential service*) If the service-time distributon is exponential, then we have $\mu_f(x) = \mu x$, as we should because $\delta_1(t)/m_1(t) = \mu$ for all t.

5 The $M_t/GI/\infty$ Fluid Model with Sinusoidal Arrival Rate

We start by proving the results in §3.1 and §3.2, in reverse order. We then discuss the relation of these results to previous results for the $M_t/M/\infty$ model in [28]. Afterward, we turn to the fitted fluid input and output rate functions.

5.1 Proof of the Theorem in §3.2

We use a basic trigonometric identity.

Lemma 5.1 (*trigonometric identity*) *For strictly positive real numbers x and y,*

$$\tan^{-1}(x/y) = \cos^{-1}(y/[x^2+y^2]^{-1/2}) \tag{61}$$

and

$$\cos(\tan^{-1}(x/y)) = \frac{y}{\sqrt{x^2+y^2}}. \tag{62}$$

For more general real numbers x and y that are not both 0, (61) and (62) hold, except for an ambiguity about the sign.

Proof of Theorem 3.3. First, the formulas in (26) are Theorem 4.1 and Corollary 4.2 of [8], which are conveniently proved using the complex variables, starting with $\sin(\theta) = [e^{i\theta} - e^{-i\theta}]/2i$ and $\cos(\theta) = [e^{i\theta} + e^{-i\theta}]/2$. Then (27) is an elaboration of Theorem 4.3 of [8], which is proved using the sine and cosine addition formulas. In particular, from (26), using the addition formulas in lines three and four, along with Lemma 5.1 in line six,

$$\begin{aligned}
s(t^*+t) &= \mathcal{C}\sin(\gamma[t^*+t]) - \mathcal{S}\cos(\gamma[t^*+t)] \\
&= \mathcal{C}\sin\left(\frac{\pi}{2} + \tan^{-1}(\mathcal{S}/\mathcal{C}) + \gamma t\right) - \mathcal{S}\cos\left(\frac{\pi}{2} + \tan^{-1}(\mathcal{S}/\mathcal{C}) + \gamma t\right) \\
&= \mathcal{C}\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}) + \gamma t) + \mathcal{S}\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}) + \gamma t) \\
&= \mathcal{C}\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))\cos(\gamma t) - \mathcal{C}\sin(\tan^{-1}(\mathcal{S}/\mathcal{C}))\sin(\gamma t) \\
&\quad + \mathcal{S}\sin(\tan^{-1}(\mathcal{S}/\mathcal{C}))\cos(\gamma t) + \mathcal{S}\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))\sin(\gamma t) \\
&= \frac{[\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))(\mathcal{S}^2+\mathcal{C}^2)}{\mathcal{C}}\cos(\gamma t) \\
&= s(t^*)\cos(\gamma t) = \pm s_U \cos(\gamma t),
\end{aligned} \tag{63}$$

where $s(t^*) = +s_U$ if $\mathcal{S} > 0$ and $\mathcal{C} > 0$, as claimed in (27) and (28). ∎

Remark 5.1 (*the sign of $\mathcal{C}$ and the location of the first extreme point*) *If $\mathcal{C} > 0$, then $s(t^*) > 0$ and $\pi/2\gamma < t^* < \pi/\gamma$, i.e., the first extreme point occurs in the second quarter cycle. However, if $\mathcal{C} < 0$, the then the sign of $s(t^*)$ is ambiguous, which can be attributed to the location of the first extreme point after $\pi/2\gamma$. For many examples, such as exponential and hyperexponential service distributions, this anomaly cannot occur, but it can occur, as we illustrate with the deterministic service distribution in §6.*

5.2 Proofs of the Theorems in §3.1

Proof of Theorem 3.1. From Theorem 3.3, $m_1(t) = 1 + \beta s(t)$ for $s(t)$ in (**??**). From (27), for $x \ge 0$,

$$\begin{aligned}
P(Z \le 1 + x) &\equiv \frac{1}{c}\int_0^c 1_{\{s(t) \le (x/\beta)\}}\, dt \\
&= \frac{\gamma}{2\pi}\int_0^{2\pi/\gamma} 1_{\{cos(\gamma t) \le (x/\beta s_U)\}}\, dt \\
&= \frac{1}{2} + \frac{\gamma}{\pi}\int_0^{2\pi/\gamma} 1_{\{cos(\gamma t) \le (x/\beta s_U)\}}\, dt
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2} + \frac{\gamma}{\pi}\left[\frac{\pi}{2\gamma} - \frac{\arccos\left(x/\beta s_U\right)}{\gamma}\right] \\
&= \frac{1}{2} + \frac{1}{\pi}\left[\frac{\pi}{2} - \arccos\left(x/\beta s_U\right)\right] \\
&= \frac{1}{2} + \frac{1}{\pi}\arcsin\left(x/\beta s_U\right),
\end{aligned} \tag{64}
$$

as claimed.

Proof of Theorem 3.2. The limit in (24) follows from Theorem 2.1. The limit in (25) follows from the limit (40) in Theorem 4.2. That yields

$$(2\pi n/\gamma)\alpha^c_{n,\lfloor n(1+x)\rfloor} \to \frac{2}{|\dot{m}_1(t^m_j(1+x))|} \quad \text{as} \quad n \to \infty, \tag{65}$$

where $\dot{m}_1(t) = \beta\dot{s}(t)$, so that

$$|\dot{m}_1(t^* + t)| = |\beta\dot{s}(t^* + t)| = \beta\gamma s_U|\sin(\gamma t)| \tag{66}$$

for t^* and s_U in (26). because $m_1(t^* + t) = \beta s_U \cos(\gamma t)$,

$$t^{m*}_1(1+x) = t^m_1(1+x) - t^* = \arccos\left(x/\beta s_U\right)\gamma. \tag{67}$$

Because $\sin\left(\arccos\left(x\right)\right) = \sqrt{1-x^2}$ for $0 \le x \le 1$ (as can be seen by letting $y = \arccos\left(x\right)$ and applying the identity $\sin^2(y) + \cos^2(y) = 1$),

$$\frac{2}{|\dot{m}_1(t^m_j(1+x))|} = \frac{2}{\beta s_U\sqrt{1-(x/\beta s_U)^2}}, \quad |x| \le \beta s_U, \tag{68}$$

which is the arcsine pdf on the interval $[-\beta s_U, \beta s_U]$. ∎

5.3 The $M_t/M/\infty$ Steady-State Distribution with Large Scale

Figures 1 and 2 of [28] show the steady-state pmf and cdf in the periodic $M_t/M/\infty$ model with the sinusoidal arrival rate function in (6) for $\beta = 10/35 = 0.286$, $\gamma = 1$ and four values of $n = \bar{\lambda} = 10$, 35, 100 and 1000. We are especially interested in the large-scale case with $n = \bar{\lambda} = 1000$.

Figure 2 repeats part of Figure 1 of [28] showing the pmf for $n = \bar{\lambda} = 1000$ and three values of $\gamma = 1/8$, 1 and 8. For $\gamma = 8$, we see the approximately Gaussian form characteristic of small scale, but as γ decreases we see the arcsine form. The high values of the density at the extremes can be understood by observing that they occur where the derivative of $m(t)$ is 0, and so changes relatively slowly.

In contrast, Figure 2 of [28] shows that the cdf is actually quite close to piecewise-linear, if not exactly piecewise-linear. That apparent inconsistency can be resolved by increasing the horizontal scale, as we do below in Figure 3. Figure 3 shows that the limiting cdf is not actually linear for $\bar{\lambda} = 1000$. Nevertheless, we see that the cdf view in Figure 3 shows that it is reasonable to regard the steady-state distribution of $m_1(t)$ as being approximately uniform over the interval $[1 - \beta s_U, 1 + \beta s_U]$.

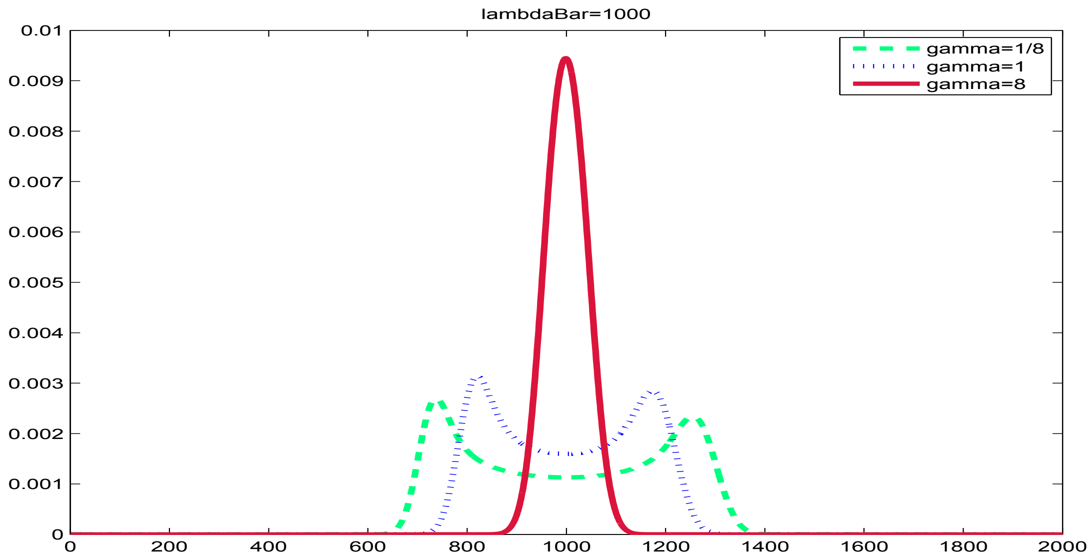

Figure 2: The steady-state pmf in the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (6) for $\bar{\lambda} = 1000$ $\beta = 10/35 = 0.286$ and three values of γ: 1/8, 1 and 8.

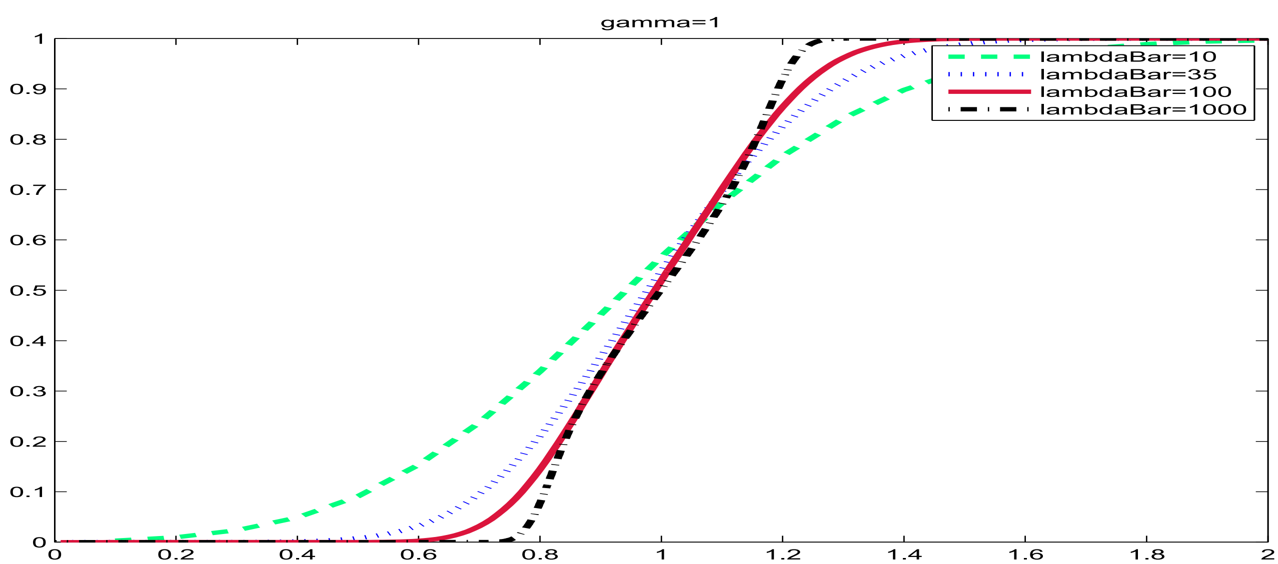

Figure 3: The cdf of the scaled steady-state random variable $\bar{Z}_n$ in the $M_t/M/\infty$ model with the sinusoidal arrival rate function in (6) for $\beta = 10/35 = 0.286$, $\gamma = 1$ and four values of $n = \bar{\lambda} = 10$, 35, 100 and 1000.

5.4 The Fitted Fluid Input and Output Rate Functions

Let λ^f and μ^f be the associated fitted fluid input and output rate functions. Paralleling our use of $s(t)$ in §3.2 to study the mean function $m(t)$, we introduce associated scaled fluid input rate function λ^s and μ^s, defined by

$$\lambda^s(y) \equiv [\lambda^f(1+\beta y) - 1]/\beta \quad \text{for} \quad -s_U \le y \le s_U. \tag{69}$$

By (27), the fluid rate function satisfies

$$\lambda^f(m_1(t^*+t)) = [\lambda_1(t^*+t) + \lambda_1(t^*-t)]/2 \tag{70}$$

Hence, we have

$$\lambda^s(s(t^*+t)) = [\sin(\gamma(t^*+t)) + \sin(\gamma(t^*-t))]/2 \tag{71}$$

for $s(t)$ in §3.2 and

$$\lambda^f(m_1(t^*+t)) = 1 + \beta\lambda^s(s(t^*+t)) \quad \text{for all} \quad t. \tag{72}$$

We define the associated scaled fluid output rate μ^s based on μ^f, defined as in (69), where μ^f is defined just as λ^f in (70) using $\delta_1(t) = E[\lambda_1(t-S)]$ in (12) instead of λ_1; i.e.,

$$\mu^s(s(t^*+t)) = E[\sin(\gamma(t^*-S+t)) + \sin(\gamma(t^*-S-t))]/2. \tag{73}$$

The extra expectation in (73) makes the algebra more complicated.

To work with the death rates, we need to relate the trigonometric integrals $E[\sin(\gamma S)]$ and $E[\cos(\gamma S)]$ to their counterparts $\mathcal{S}$ and $\mathcal{C}$ with S replaced by S_e. We use an elementary trigonometric identity connecting S and S_e.

Lemma 5.2 (*trigonometric identity for S and S_e*) *For any nonnegative random variable S with mean $E[S] = 1$ and any real number $\gamma > 0$,*

$$E[\sin(\gamma S)] = \gamma E[\cos(\gamma S_e)] = \gamma\mathcal{C} \quad \textit{and} \quad E[\cos(\gamma S)] = 1 - \gamma\mathcal{S}. \tag{74}$$

Hence, if $E[\cos(\gamma S)] \neq 1$, then $\mathcal{S} > 0$.

Proof. Apply integration by parts, e.g.,

$$\begin{aligned} E[\sin(\gamma S)] &= \int_0^\infty \sin(\gamma x) g(x)\, dx \\ &= -\sin(\gamma x) G^c(x)|_0^\infty - \int_0^\infty G^c(x)(-\gamma)\cos(\gamma x)\, dx = \gamma\mathcal{C}. \quad \blacksquare \end{aligned} \tag{75}$$

Theorem 5.1 (*the scaled fitted fluid rates*) *Consider the $M_t/GI/\infty$ fluid model with $E[S] = 1$ and the sinusoidal arrival rate function λ_1 in (6), where we start empty in the distant past. As in Theorem 3.3, assume that neither $\mathcal{S}$ nor $\mathcal{C}$ is 0. That implies that $s(t)$ is not identically 0 and that the mean function $m(t)$ is not a constant function. For t^* in (26), the associated scaled fitted fluid input and output rates satisfy*

$$\lambda^s(s(t^*+t)) = \mu^s(s(t^*+t)) = \cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))\cos(\gamma t) = \left(\frac{\mathcal{C}}{\sqrt{\mathcal{S}^2+\mathcal{C}^2}}\right)\cos(\gamma t), \tag{76}$$

so that

$$\lambda^s(s(t^*+t)) = \mu^s(s(t^*+t)) = \mu^s(s(t^*-t)) = \lambda^s(s(t^*-t)) \quad \textit{for all} \quad t. \tag{77}$$

As a consequence, $\lambda^f(x) = \mu^f(x)$ is a linear function of x, i.e.,

$$\lambda^f(x) = \mu^f(x) = 1 + C(\gamma)(x-1) \quad \textit{for} \quad 1 - \beta s_U \le x \le 1 + \beta s_U \tag{78}$$

for $s_U = \sqrt{\mathcal{S}^2 + \mathcal{C}^2}$ *as in* (26)*, where*

$$C(\gamma) \equiv \frac{\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))}{s_U} = \frac{\mathcal{C}}{\mathcal{S}^2 + \mathcal{C}^2}. \tag{79}$$

In general, the signs of $s(t^*) = \cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))$ *in* (76) *and* $C(\gamma)$ *in* (79) *are ambiguous, but if* $\mathcal{S} > 0$ *and* $\mathcal{C} > 0$*, then both are positive, so that* $\lambda^s(s(t^* + t))$ *and* $\lambda^s(s(t^* - t))$ *are decreasing functions of* t *in the interval* $[0, \pi/\gamma]$*, and* $\lambda^f(x)$ *is an increasing function of* x *over* $[1-\beta s_U, 1+\beta s_U]$*.*

Proof. We apply the symmetry of the function $s(t)$ about t^* as shown in (27). For (76), we apply (71), using the trigonometric sum and difference formulas in the third and fourth lines, to get

$$\begin{aligned}
\lambda^s(s(t^* + t)) &= [\sin(\gamma(t^* + t)) + \sin(\gamma(t^* - t))]/2 \\
&= [\sin((\pi/2) + \tan^{-1}(\mathcal{S}/\mathcal{C}) + \gamma t) + \sin((\pi/2) + \tan^{-1}(\mathcal{S}/\mathcal{C}) - \gamma t)]/2 \\
&= [\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}) + \gamma t) + \cos(\tan^{-1}(\mathcal{S}/\mathcal{C}) - \gamma t)]/2 \\
&= \cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))\cos(\gamma t),
\end{aligned} \tag{80}$$

where $\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))$ is given in (62). Similarly, using (12),

$$\begin{aligned}
\mu^s(s(t^* + t)) &= E[\sin(\gamma(t^* - S + t)) + \sin(\gamma(t^* - S - t))]/2 \\
&= E[\sin((\pi/2) + \tan^{-1}(\mathcal{S}/\mathcal{C}) - \gamma S + \gamma t) + \sin((\pi/2) + \tan^{-1}(\mathcal{S}/\mathcal{C}) - \gamma S - \gamma t)]/2 \\
&= E[\cos(\tan^{-1}(\mathcal{S}/\mathcal{C}) - \gamma S + \gamma t) + \cos(\tan^{-1}(\mathcal{S}/\mathcal{C}) - \gamma S - \gamma t)]/2 \\
&= E[\cos(\tan^{-1}(\mathcal{S}/\mathcal{C})) - \gamma S]\cos(\gamma t),
\end{aligned} \tag{81}$$

where, by applying Lemma 5.2 in the last step,

$$\begin{aligned}
\mu^s(s(t^*)) &= E[\cos(\tan^{-1}(\mathcal{S}/\mathcal{C})) - \gamma S] \\
&= \cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))E[cos(\gamma S)] + \sin(\tan^{-1}(\mathcal{S}/\mathcal{C}))E[\sin(\gamma S)] \\
&= \cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))E[cos(\gamma S)] + (\mathcal{S}/\mathcal{C})E[sin(\gamma S)] \\
&= \cos(\tan^{-1}(\mathcal{S}/\mathcal{C}))\left(E[cos(\gamma S)] + \left(\frac{\mathcal{S}}{\mathcal{C}}\right)E[sin(\gamma S)]\right) \\
&= \cos(\tan^{-1}(\mathcal{S}/\mathcal{C})). \quad \blacksquare
\end{aligned} \tag{82}$$

6 Fluid Model for the $M_t/D/\infty$ Model

To illustrate the ambiguities about the sign of $\mathcal{S}$ and $\mathcal{C}$ in (**??**) and in the trigonometric function in Lemma 5.1, we carefully treat the special case of a deterministic (D) service-time distribution, amplifying §6 of [8]. In the process, we show that the bounds in Corollary 4.1 of [7] are attained asymptotically. We also show that the fitted fluid input and output rate functions can be decreasing functions of the state. In addition, we show that this property occurs in simulations of $M_t/D/\infty$ queues.

Let the cumulative arrival rate function associated with a general arrival rate function be $\lambda(t)$, starting from time 0, be

$$\Lambda(t) \equiv \int_0^t \lambda(s)\,ds. \tag{83}$$

With D service times having value 1,

$$m(t) = \Lambda(t) - \Lambda(t-1) = \int_{t-1}^t \lambda(s)\,ds. \tag{84}$$

Let $D(t)$ be the cumulative output by time t, starting from time 0, and let $\delta(t)$ be the departure rate function. Clearly,

$$D(t) = \Lambda(t-1) \quad \text{and} \quad \delta(t) = \lambda(t-1). \tag{85}$$

In the special case of sinusoidal arrival rate function in (6), starting out empty in the distant past, we can carry out the integration in (83) and (84) to get

$$\Lambda(t) = \bar{\lambda}\,(t + (\beta/\gamma)(1 - \cos(\gamma t))) \tag{86}$$

and

$$m(t) = \bar{\lambda}\,(1 + (\beta/\gamma)(\cos(\gamma(t-1)) - \cos(\gamma t)))\,. \tag{87}$$

Let

$$s(t) = (m(t) - \bar{\lambda})/\beta. \tag{88}$$

Theorem 6.1 (*the mean function with D service*) *Consider the $M_t/D/\infty$ model with sinusoidal arrival rate function in* (6)*, starting empty in the distant past, where the service times are of length* 1*. Let*

$$t^* \equiv t^*(\gamma) \equiv (\pi/2\gamma) + 1/2 \quad \textit{and} \quad s_U \equiv s_U(\gamma) \equiv (2/\gamma)(\sin(\gamma/2)). \tag{89}$$

Then

$$s(t^* \pm t) = s_U \cos(\gamma t). \tag{90}$$

Hence, all extreme points of $m(t)$ are attained at the points $t^* + (k\pi/\gamma)$ for integers k and the values of these extreme points are $m(t^*) = \bar{\lambda}\,(1 \pm \beta s_U)$.

Proof. It suffices to apply (87) and (88) to directly calculate (90). We use the standard trigonometric sum and difference formulas in the second and third lines to write

$$\begin{aligned} s(t^*+t) &= (1/\gamma)[\cos(\gamma((\pi/2\gamma) - (1/2) + t)) - \cos(\gamma((\pi/2\gamma) + (1/2) + t))] \\ &= (1/\gamma)[\sin((\gamma/2) - \gamma t)) + \sin((\gamma/2) + \gamma t))] \\ &= (2/\gamma)\sin(\gamma/2)\cos(\gamma t) = s_U \cos(\gamma t). \end{aligned} \tag{91}$$

Since cos is an even function, bounded above by 1, with the maximum attained at time 0, the stated properties of $m(t)$ as a function of t follow.

We remark that it is also possible to apply Corollary 4.2 of [8]. To do so, we first observe that $E[\cos(\gamma S_e)] = \sin(\gamma)/\gamma$ and $E[\sin(\gamma S_e)] = (1 - \cos(\gamma))/\gamma$. Then we observe that $\sin(\gamma) = 2\sin(\gamma/2)\cos(\gamma/2)$ and $1 - \cos(\gamma) = 2(1 - \cos(\gamma/2)^2) = 2\sin(\gamma/2)^2$. Combining these, we see that $E[\sin(\gamma S_e)]/E[\cos(\gamma S_e)] = \tan(\gamma/2)$, so that $t^* = t_\lambda + 1/2$, as can be calculated directly in this case with D service. We can also apply the same trigonometric relations with Corollary 4.2 of [8] to show that $(E[\cos(\gamma S_e)]^2 + E[\sin(\gamma S_e)]^2)^{1/2} = 2\sin(\gamma/2)/\gamma$. ∎

Corollary 6.1 (*more for the mean function with D service*) *Consider the setting of Theorem 6.1. For $0 < \gamma < 2\pi$, $m(t^* \pm t)$ is a strictly decreasing function of t over the interval $[0, \pi/\gamma]$, so that $m(t)$ attains its maximum*

$$m_U = \bar{\lambda}\,(1 + \beta s_U) \tag{92}$$

at time t^*, where t^* and s_U are defined in (89)*, while $m(t^* \pm t)$ is a strictly increasing function of t over the interval $[\pi/\gamma, 2\pi/\gamma]$, so that $m(t)$ attains its minimum*

$$m_L = \bar{\lambda}\,(1 - \beta s_U) \tag{93}$$

at time $t^* + (\pi/\gamma)$. *In addition,*

$$m_U(\gamma) \to \bar{\lambda}(1+\beta) \quad \text{and} \quad m_L(\gamma) \to \bar{\lambda}(1-\beta) \quad \text{as} \quad \gamma \to 0, \tag{94}$$

while

$$m_U(\gamma) \to \bar{\lambda} \quad \text{and} \quad m_L(\gamma) \to \bar{\lambda} \quad \text{as} \quad \gamma \to \infty. \tag{95}$$

However, for $2\pi < \gamma < 4\pi$, $\sin(\gamma/2) \le 0$, *so that* $m(t^* \pm t)$ *is a strictly increasing function of* t *over the interval* $[0, \pi/\gamma]$, *so that* $m(t)$ *attains its maximum* m_U *in* (92) *at time* $t^* + (\pi/\gamma) = (3\pi/\gamma) + 1/2$, *while* $m(t^* \pm t)$ *is a strictly decreasing function of* t *over the interval* $[\pi/\gamma, 2\pi/\gamma]$, *so that* $m(t)$ *attains its minimum* m_L *in* (93) *at time* $t^* + (2\pi/\gamma)$ *and at time* t^*.

Proof. These subsequent results depend on the behavior of $\sin(\gamma/2)$ and $\cos(\gamma t)$ in (92) and (89). First, $\sin(\gamma/2)$ is nonnegative for $0 \le \gamma \le 2\pi$. Then $\cos(\gamma t)$ is a strictly decreasing function of t on $[0, \pi/\gamma]$. For the limits in (94), we use the asymptotic expression $\sin(x)/x \to 1$ as $x \to 0$. ∎

We see anomalous behavior in the mean when the D service time is an integer multiple of a full cycle.

Corollary 6.2 (*full-cycle service times*) *Consider the setting of Theorem 6.1, where the mean service time is exactly one full cycle, i.e.,* $1 = 2\pi/\gamma$, *or an integer multiple. Then*

$$m(t) = m_U = \bar{\lambda} \quad \text{for all} \quad t. \tag{96}$$

Proof. Apply (87). ∎

However, if we exclude the case in Corollary 6.2, then m has a well defined inverse on a subinterval of the real line.

Corollary 6.3 (*inverse of the mean function*) *Consider the setting of Theorem 6.1, where the mean service time is not an integer multiple of a full cycle, i.e., we do not have* $1 = 2k\pi/\gamma$ *for some integer* k. *Then* m *is a strictly monotone function on the interval* $[t^* - (\pi/\gamma), t^*]$ *and so has a well defined inverse* m^{-1} *on the interval* $[\bar{\lambda}(1 - s_U), \bar{\lambda}(1 + s_U)]$, *where* t^* *and* s_U *are defined in* (89). *In particular,*

$$m^{-1}(x) = t^* + \frac{1}{\gamma} \arccos\left((x/\bar{\lambda}) - 1)/s_U\right) \quad \text{for} \quad \bar{\lambda}(1 - \beta s_U) \le x \le \bar{\lambda}(1 + \beta s_U). \tag{97}$$

Proof. Apply (90) to construct the inverse directly. To verify, observe that $m^{-1}(m(t^* + t) = t^{(*)} + t$. ∎

We now turn to the fluid limit. Let λ^f and μ^f be the associated fitted fluid input and output rate functions. Let λ^f_U and λ^f_L be the corresponding maximum and minimum of the fitted fluid input rate function λ^f and similarly for μ^f.

Theorem 6.2 (*the fitted fluid rates*) *Consider the* $M_t/D/\infty$ *model in the setting of Theorem 6.1, where the mean service time is not an integer multiple of a full cycle* $2\pi/\gamma$, *so that* $m(t)$ *is not a constant function. For* t^* *in* (89), *the associated fitted fluid input and output rates satisfy*

$$\lambda^f(m_1(t^* + t)) = \mu^f(m_1(t^* + t)) = \frac{\lambda(t^* + t) + \lambda(t^* - t)}{2\bar{\lambda}} = 1 + \beta(\cos(\gamma/2)\cos(\gamma t)) \tag{98}$$

As a consequence,

$$\lambda^f(x) = \mu^f(x) = 1 + C(\gamma)(x - 1) \quad \text{for} \quad 1 - \beta s_U \le x \le 1 + \beta s_U \tag{99}$$

for s_U *in* (89), *where*

$$C(\gamma) \equiv \beta(\gamma/2)\cot(\gamma/2), \tag{100}$$

which is a strictly decreasing function of γ *on* $[0,\pi]$ *satisfying* $C(0)=\beta$ *and* $C(\pi)=0$. *Hence, for* $0<\gamma<\pi$, *we have* $0<C(\gamma)<1$, *so that* λ^f *is not the identity function.*

Proof. We apply the symmetry of the function $m(t)$ about t^*, as shown in Theorem 4.3 of [8] or directly above in (90). For (98), we apply (27), using the trigonometric sum and difference formulas in the fourth and fifth lines, to get

$$\begin{aligned}
\lambda^f(m_1(t^*+t)) &= [\lambda_1(t^*+t)+\lambda_1(t^*-t)]/2\\
&= 1+(\beta/2)[\sin(\gamma((\pi/2\gamma)+(1/2)+t))+\sin(\gamma((\pi/2\gamma)+(1/2)-t))]\\
&= 1+(\beta/2)[\sin((\pi/2)+(\gamma/2)+\gamma t))+\sin((\pi/2)+(\gamma/2)-\gamma t))]\\
&= 1+(\beta/2)[\cos((\gamma/2)+\gamma t)+\cos((\gamma/2)-\gamma t)]\\
&= 1+\beta[\cos(\gamma/2)\cos(\gamma t)]
\end{aligned} \tag{101}$$

and

$$\begin{aligned}
\mu^f(m_1(t^*+t)) &= [\delta_1(t^*+t)+\delta_1(t^*-t)]/2 = [\lambda_1(t^*-1+t)+\lambda_1(t^*-1-t)]/2\\
&= 1+(\beta/2)[\sin(\gamma((\pi/2\gamma)-(1/2)+t))+\sin(\gamma((\pi/2\gamma)-(1/2)-t))]\\
&= 1+(\beta/2)[\sin((\pi/2)-(\gamma/2)+\gamma t))+\sin((\pi/2)-(\gamma/2)-\gamma t))]\\
&= 1+(\beta/2)[\cos((-\gamma/2)+\gamma t)+\cos((-\gamma/2)-\gamma t)] \qquad (102)\\
&= 1+\beta[\cos(-\gamma/2)\cos(\gamma t)] = 1+\beta[\cos(\gamma/2)\cos(\gamma t)] = \lambda^f(m_1(t_m+t). \quad \blacksquare
\end{aligned}$$

Then we apply (98) to observe that $\lambda^f(x)=\lambda^f(m_1(t^*+t))$ when $m_1(t^*+t)=x$ or, equivalently, when $m^{-1}(x)=t^*+t$, so that $t=m^{-1}(x)-t^*$. Hence, we can apply (97) to get

$$\begin{aligned}
\lambda^f(x) &= 1+\beta[\cos(\gamma/2)\cos(\gamma(m^{-1}(x)-t^*)))\\
&= 1+\beta\cos(\gamma/2)(x-1)/s_U,
\end{aligned} \tag{103}$$

which reduces to (99). $\blacksquare$

Corollary 6.4 (*more for the fitted fluid rate functions*) *In the setting of Theorem 6.2, there are four different cases for* $0<\gamma<4\pi$, *depending on whether* γ *belongs to one of the four subintervals:* (*i*) $(0,\pi)$, (*ii*) $(\pi,2\pi)$, (*iii*) $(2\pi,3\pi)$ *or* (*iv*) $(3\pi,4\pi)$, *leading to* $\lambda^f=\mu^f$ *being a strictly increasing* (*decreasing*) *function in cases* (*i*) *and* (*iii*) ((*ii*) *and* (*iv*)). *In particular,*

(*i*) *If* $0<\gamma<\pi$, *then* $\lambda^f(m_1(t^*+t))$ *is a strictly decreasing function of* t *on the interval* $[0,\pi/\gamma]$ *attaining its maximum of* $1+\beta(\cos(\gamma/2)$ *at time* $t=0$, *and its minimum of* $1-\beta(\cos(\gamma/2)$ *at time* π/γ. *Since* $m_1(t^*+t)$ *is also a strictly decreasing function of* t *on the interval* $[0,\pi/\gamma]$ *in this case,* $\lambda^f=\mu^f$ *is strictly increasing on* $[m_L/\bar{\lambda}, m_U/\bar{\lambda}]$ *and*

$$\lambda^f_U(\gamma)=\mu^f_U(\gamma)=\lambda^f(t^*)=1+\beta\cos(\gamma/2) \quad \textit{and} \quad \lambda^f_L(\gamma)=1-\beta\cos(\gamma/2). \tag{104}$$

Moreover,

$$\lambda^f_U(\gamma)=\mu^f_U(\gamma)\to 1+\beta \quad \textit{and} \quad \lambda^f_L(\gamma)=\mu^f_U(\gamma)\to 1-\beta \quad \textit{as} \quad \gamma\to 0. \tag{105}$$

(*ii*) *However, if* $\pi<\gamma<2\pi$, *then* $\lambda^f(m_1(t^*+t))$ *is a strictly increasing function of* t *on the interval* $[0,\pi/\gamma]$ *attaining its maximum of* $1-\beta(\cos(\gamma/2)>1$ *at time* $t=\pi\gamma$, *and its minimum of*

$1+\beta(\cos(\gamma/2) < 1$ *at time* $t=0$. *Since* $m_1(t^*+t)$ *is still a strictly decreasing function of* t *on the interval* $[0,\pi/\gamma]$, *in this case* $\lambda^f = \mu^f$ *is strictly decreasing on* $[m_L/\bar{\lambda}, m_U/\bar{\lambda}]$ *and*

$$\lambda_U^f(\gamma) = \mu_U^f(\gamma) = \lambda^f(t_m) = 1 - \beta\cos(\gamma/2) \quad \text{and} \quad \lambda_L^f(\gamma) = 1 + \beta\cos(\gamma/2). \tag{106}$$

(*iii*) *If* $2\pi < \gamma < 3\pi$, *then* $\lambda^f(m_1(t^*+t))$ *is a strictly increasing function of* t *on the interval* $[0,\pi/\gamma]$, *but now* $m_1(t^*+t)$ *is also a strictly increasing function of* t *on the interval* $[0,\pi/\gamma]$. *Hence, just as in case* (*i*), $\lambda^f = \mu^f$ *is strictly increasing on* $[m_L/\bar{\lambda}, m_U/\bar{\lambda}]$.

(*iv*) *If* $3\pi < \gamma < 4\pi$, *then* $\lambda^f(m_1(t^*+t))$ *is a strictly decreasing function of* t *on the interval* $[0,\pi/\gamma]$, *just as in case* (*i*), *but now* $m_1(t^*+t)$ *is a strictly decreasing function of* t *on the interval* $[0,\pi/\gamma]$. *Hence, just as in case* (*ii*), $\lambda^f = \mu^f$ *is strictly decreasing on* $[m_L/\bar{\lambda}, m_U/\bar{\lambda}]$.

Proof. We observe that $\cos(\gamma/2) > 0$ in cases (i) and (ii), whereas $\cos(\gamma/2) < 0$ in cases (ii) and (iii). The overall complexity occurs because the cases for $m_1(t^*+t)$ and $\lambda^f(m_1(t^*+t))$ are different. ∎

Corollary 6.5 (*half-cycle service times*) *Consider the* $M_t/D/\infty$ *model with sinusoidal arrival rate function in* (6), *starting empty in the distant past, where the mean service time is exactly one half cycle, i.e.,* $1 = \pi/\gamma$. *Then*

$$\lambda^f(x) = \mu^f(x) = 1 \quad \text{for all} \quad x. \tag{107}$$

Proof. Apply (98). ∎

7 Exponential Service and Relatives

Most results for the case of exponential service are given in §5 of [8]. We see that the fitted fluid input and output rate functions are especially nice in this case.

Corollary 7.1 (*M service*) *In addition to the assumptions of* §5, *suppose that the service-time distribution is exponential* (M) *with* $E[S]=1$. *Then* S_e *is distributed as* S, $\mathcal{C} = 1/(1+\gamma^2) > 0$ *and* $\mathcal{S} = \gamma/(1+\gamma^2) > 0$ *for all* $\gamma > 0$. *Thus,* $s_U = 1/\sqrt{1+\gamma^2}$ *and* $C(\gamma) = 1$, *so that* $\lambda^f = \mu^f$ *is the identity function, i.e.,*

$$\lambda^f(x) = \mu^f(x) = x \quad \text{for all} \quad x, \quad 1 - \beta s_U \le x \le 1 + \beta s_U. \tag{108}$$

Some of the nice structure for M service carries over to H_k service, i.e., a mixture of k exponential cdf's. Let the mean-1 service time S have an H_k cdf, so that

$$G^c(x) \equiv 1 - G(x) \equiv \sum_{i=1}^k p_i e^{-\mu_i x}, \quad x \ge 0, \tag{109}$$

where

$$\sum_{i=1}^k p_i = 1 \quad \text{and} \quad \sum_{i=1}^k p_i/\mu_i = E[S] = 1. \tag{110}$$

Corollary 7.2 (H_k *service*) *In addition to the assumptions of* §5*, suppose that the service-time distribution is hypexponential* (H_k) *with* $E[S] = 1$ *as in* (109) *and* (110)*. Then* S_e *has an* H_k *distribution with*

$$P(S_e > x) = \sum_{i=1}^{k} q_i e^{-\mu_i x}, \quad x \geq 0, \quad \textit{where} \quad q_i = pi/\mu_i \quad \textit{and} \quad \sum_{i=1}^{k} q_i = 1. \tag{111}$$

Consequently, for all $\gamma > 0$,

$$0 < \mathcal{C} = \sum_{i=1}^{k} q_i \frac{1}{1 + (\gamma/\mu_i)^2} < \frac{1}{1+\gamma^2} \quad \textit{and} \quad 0 < \mathcal{S} = \sum_{i=1}^{k} q_i \frac{(\gamma/\mu_i}{1 + (\gamma/\mu_i)^2} < \frac{\gamma}{1+\gamma^2}. \tag{112}$$

Hence, letting $s_{U,GI}$ *denoting* s_U *as a function of the mean-*1 *service-time distribution,*

$$s_{U,H_k} < s_{U,M}. \tag{113}$$

For all non-M H_k *distributions,* λ^f *is an increasing function, but not equal to the identity function.*

Proof. First, (111) follows immediately from (109) and the definition of S_e.

8 Local Balance with Large Scale

From our results above, we see that we cannot apply the local balance equation (3) to obtain the limiting steady-state pmf from the large scale approximations of the fitted BD rates. For example, the natural approximating fitted birth and death rates with large scale for the $M_t/M/\infty$ model based on Theorems 4.1 and 5.1 plus Corollary 7.1 are

$$\bar{\lambda}_{n,k} = \bar{\lambda}_{n,k} = k, \quad n(1 - \beta s_U) \leq k \leq n(1 + \beta s_U)$$

This BD process on the interval $[n(1 - \beta s_U), n(1 + \beta s_U)]$ has steady-state pmf

$$\bar{\alpha}_{n,n(1-\beta s_U)+k} = r_{n,k} / \sum_{j=n(1-\beta s_U)}^{n(1+\beta s_U)} r_{n,j}, \quad 0 \leq k \leq 2n\beta s_U,$$

where, by telescoping products,

$$r_{n,k} = \frac{\prod_{j=n(1-\beta s_U)}^{n(1-\beta s_U)+k-1} \lambda_{n,j}}{\prod_{j=n(1-\beta s_U)+1}^{n(1-\beta s_U)+k} \mu_{n,j}} = \frac{n(1-\beta s_U)}{n(1-\beta s_U)+k}.$$

We thus, see that $r_{n,k}$ decreases from just below 1 at $k = n(1 - \beta s_U) + 1$ to $(1 - \beta s_U)/(1 + \beta s_U)$ at $k = n(1 + \beta s_U) - 1$. Thus $\alpha_{n,k}$ is decreasing, approximately linearly. Clearly, this distribution is not the arcsine pdf appearing in Theorem 3.2. Thus, we conclude that this fluid scaling is too crude to retain the local-balance characterization of the steady-state distribution discussed in §1.1.

9 Conclusions

In this paper we established many-server heavy-traffic fluid limits for the steady-state distribution and the fitted birth and death rates in periodic $M_t/GI/\infty$ models. The simple fluid approximation for the steady-state cdf in §3 should serve as a useful reference. Theorems 3.1 and 3.2 expose the simple arcsine limit for sinusoidal arrival rates.

Theorem 4.1 establishes many-server heavy-traffic (MSHT) limits for the fitted birth and death rates in general periodic $M_t/GI/\infty$ queueing models. Since the estimation tends to require a great amount of data when the scale increases, we also proposed alternative aggregate estimators. Theorem 4.3 shows that these aggregate estimators also converge to the same limits after scaling.

We also obtained stronger explicit results for the special case of sinusoidal arrival-rate functions. Theorem 6.2 shows that the limiting fitted birth and death rates are equal and linear over a finite interval, with these being the restriction of the identity function if and only if the service-time distribution is exponential. Formula (78) shows that the linear functions for different service-time distributions coincide at the overall average arrival rate. An explicit expression for the slope is given in (79).

In §7 we gave explicit formulas for the fitted rates with exponential and hyperexponential service-time distributions. In these cases, the limiting fitted rates are always strictly increasing, but that is not the case for all service-time distributions. In §6 we carefully analyzed the case of deterministic service times, exposing unusual behavior; e.g., we showed that the limiting fitted rate functions are actually decreasing for some cases of long service times.

In §8 we observed that the steady-state distribution of the BD process using the asymptotic rates does not necessarily coincide with the many-server heavy-traffic limit of the steady-state distributions.

There are many directions for future research. For this same model, it remains to consider more refined scaling, which may retain the local-balance characterization of the steady-state distribution in §1.1. It remains to establish corresponding results for the finite-capacity $M_t/GI/s$ model and other models.

Acknowledgement
The author thanks James Dong for previous related work, including extensive simulations, that improved the author's understanding of this subject. The author acknowledges support from NSF grant CMMI 1265070.

References

[1] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y. and Yom-Tov, G. (2015). Patient flow in hospitals: a data-based queueing-science perspective. Stochastic Systems, published online, DOI-10.1214/14-SSY153.

[2] Billingsley, P. (1961). *Staistical Inference for Markov Processes*. Chicago: University of Chicago press.

[3] Bohlin, T. (2006). *Practical Grey-Box Process Identification*. London: Springer.

[4] Darling, D. A. and Kac, M. (1957). On occupatioon times for markoff processes. *Trans Amer Math Soc* 84(2):444–458.

[5] Davis, B. and McDonald, D. (1995). An elementary proof of the local central limit theorem. *Journal of Theoretical Probability* 8(5):693–701.

[6] Dong, J. and Whitt, W. (2015). Stochastic grey-box modeling of queueing systems: fitting birth-and-death processes to data. *Queueing Systems* 79:391–426.

[7] Dong, J. and Whitt, W. (2015). Using birth-and-death processes to estimate the steady-state distribution of a periodic queue. To appear in *Naval Research Logistics*.

[8] Eick, S. G., Massey, W. A. and Whitt, W. (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci* 39:241–252.

[9] Eick, S. G., Massey, W. A. and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Oper Res* 41:731–742.

[10] El-Taha, M. and Stidham, S. (1999). *Sample-Path Analysis of Queueing Systems*. Boston: Kluwer.

[11] Glynn, P. W. and Whitt, W. (1993). Limit theorems for cumulative processes. *Stochastic Processes and Their Applications* 47:299–314.

[12] Gnedenko, B. V. (1948). On the local limit theorem in the theory of probability. *Uspekhi Mat Nauk* 3:187–194.

[13] Green, L. V., Kolesar, P. J. and Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper Management* 16:13–29.

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

[15] Keiding, N. (1975). Maximum likelhood estimation in the birth-and-death process. *Ann Statist* 3:363–372.

[16] Kim, S.-H. and Whitt, W. (2013). Statistical analysis with Little's law. *Operations Research,* 61(4):1030–1045.

[17] Kristensen, N. R., Madsen, H. and Jorgensen, S. B. (2004). Parameter estimation in stochastic grey-box models. *Automaticaa* 40(2):225–237.

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

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

[20] Liu, Y. and Whitt, W. (2014). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences* 28(4):414–449.

[21] Massey, W. A. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13(1):183–250.

[22] McDonald, D. (2005). The local limit theorem: a historical perspective. *Journal of Iranian Statistical Society* 4(2):73–86.

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

[24] Shi, P., Chou, M. C., Dai, J. G. and Sim, J. (2015). Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science* articles in advance:doi10.1287/mnsc.2014.2112.

[25] Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Sci* 51:221–235.

[26] Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* 15(1):88–102.

[27] Whitt, W. (2012). Fitting birth-and-death queueing models to data. *Statistics and Probability Letters* 82:998–1004.

[28] Whitt, W. (2014). The steady-state distribution of the $M_t/M/\infty$ queue with a sinusoidal arrival rate function. *Operations Research Letters* 42:311–318.

[29] Whitt, W. and Zhang, X. (2015). A data-generated queueing model of an emergency department. In preparation, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[30] Wolff, R. W. (1965). Problems for statistical inference for birth and death queueing models. *Operations Research* 13:343–357.