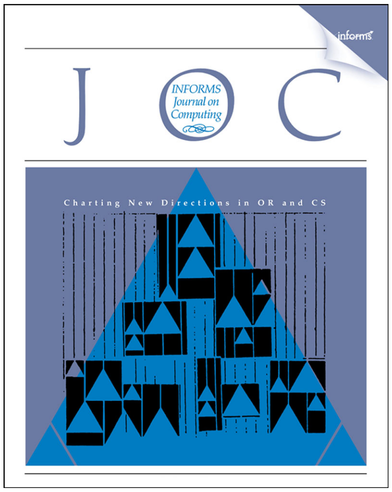


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Approximations for the Queue Length Distributions of Time-Varying Many-Server Queues

Jamol Pender, Young Myoung Ko

To cite this article:
Jamol Pender, Young Myoung Ko (2017) Approximations for the Queue Length Distributions of Time-Varying Many-Server Queues. INFORMS Journal on Computing 29(4):688-704. https://doi.org/10.1287/ijoc.2017.0760

informs
http://pubsonline.informs.org/journal/ijoc/

INFORMS JOURNAL ON COMPUTING
Vol. 29, No. 4, Fall 2017, pp. 688–704
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

# Approximations for the Queue Length Distributions of Time-Varying Many-Server Queues

**Jamol Pender,**[a] **Young Myoung Ko**[b]

[a] School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853; [b] Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Gyeongbuk, 37673, Korea

**Contact:** jjp274@cornell.edu (JP); youngko@postech.ac.kr, http://orcid.org/0000-0003-0659-6688 (YMK)

**Received:** May 17, 2016
**Revised:** December 11, 2016
**Accepted:** March 6, 2017
**Published Online:** August 23, 2017

**https://doi.org/10.1287/ijoc.2017.0760**

**Copyright:** © 2017 INFORMS

**Abstract.** This paper presents a novel and computationally efficient methodology for approximating the queue length (the number of customers in the system) distributions of time-varying non-Markovian many-server queues (e.g., $G_t/G_t/n_t$ queues), where the number of servers ($n_t$) is large. Our methodology consists of two steps. The first step uses phase-type distributions to approximate the general interarrival and service times, thus generating an approximating $Ph_t/Ph_t/n_t$ queue. The second step develops strong approximation theory to approximate the $Ph_t/Ph_t/n_t$ queue with fluid and diffusion limits whose mean and variance can be computed using ordinary differential equations. However, by naively representing the $Ph_t/Ph_t/n_t$ queue as a Markov process by expanding the state space, we encounter the lingering phenomenon*even* when the queue is *overloaded*. Lingering typically occurs when the mean queue length is equal or near the number of servers, however, in this case it also happens when the queue is overloaded and this time is not of zero measure. As a result, we develop an alternative representation for the queue length process that avoids the lingering problem in the overloaded case, thus allowing for the derivation of a Gaussian diffusion limit. Finally, we compare the effectiveness of our proposed method with discrete event simulation in a variety parameter settings and show that our approximations are very accurate.

**History:** Accepted by Winfried K. Grassmann, Area Editor for Computational Probability & Analysis.
**Funding:** This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning [NRF-2014R1A1A1005019].
**Supplemental Material:** The online supplement is available at https://doi.org/10.1287/ijoc.2017.0760.

**Keywords:** time-varying non-Markovian queues • many-server queues • fluid and diffusion limits • strong approximations

## 1. Introduction

Real-world applications of large-scale queueing systems such as data centers, call centers, and healthcare centers have time-varying and dynamic behavior. Furthermore, the arrival and service processes are not necessarily Markovian in general (Brown et al. 2005, Arfeen et al. 2013, Nelson and Taaffe 2004a). Many of the recent studies on large-scale non-Markovian queues rely on the asymptotic approaches that utilize fluid and diffusion limits as described in Billingsley (1999) and Whitt (2002). Research on non-Markovian systems has progressed to the point of analyzing underloaded systems (a.k.a. the offered-load model, infinite-server queues) as a result of their analytical or numerical tractability (Whitt 1982; Glynn 1982; Eick et al. 1993; Nelson and Taaffe 2004a, b). Studies on the delay model, e.g., $M_t/G_t/n_t$, $G_t/M_t/n_t$, $G_t/G_t/n_t$ queues, have been conducted from the context of fluid queues or heavy traffic diffusion models in the Halfin-Whitt regime (Halfin and Whitt 1981; Puhalskii and Reiman 2000; Pang and Whitt 2009; Reed 2009; Whitt 2006; Liu and Whitt 2012, 2014a, b).

This paper uses the *uniform acceleration* method coupled with strong approximations and accelerates parameters while keeping the traffic intensity constant, see for example (Kurtz 1978, Mandelbaum et al. 1998, Hampshire et al. 2006). Kurtz (1978) establishes strong approximation theorems for state-dependent continuous time Markov chains (CTMCs) having differentiable rate functions. Extending Kurtz (1978), Mandelbaum et al. (1998) consider time-varying parameters and non-differentiable rate functions such as min$(\cdot,\cdot)$ that commonly occur in the analysis of queues. Mandelbaum et al. (2002) prove that the strong approximation results developed in Kurtz (1978) can also be applied when the fluid limit stays at the nondifferentiable points of rate functions for a measure-zero amount of time. However, in some queueing processes, it is hard to avoid the measure-zero assumption. See for instance Niyirora and Pender (2017) and Hampshire and Massey (2005, 2010), Hampshire et al. (2009b, a) where optimal staffing methods force staffing at the nondifferentiable points.

To address the issue of when the fluid limit is near the nondifferentiable points of the rate functions for more than a measure-zero amount of time Ko and Gautam (2013) propose a Gaussian-based approximation method that achieves better approximation quality. Massey and Pender (2011, 2013) improve the result of Ko and Gautam (2013) by incorporating the skewness of the queueing process and by expanding the queue length process in terms of Hermite polynomials, which are orthogonal with respect to the Gaussian distribution. In the same spirit, the work of Pender (2014, 2015b, a, c) extends the results of Massey and Pender (2013) and explores the impact of the kurtosis through a Gram-Charlier expansion and using other distributions as closure approximations. More work by Engblom and Pender (2014) also proves that spectral expansions as closure approximations for the functional Kolmogorov forward equations of the queue length process are provably optimal in an $L^2$ sense for approximating the moments of nonstationary birth-death processes. Although the spectral approach offers great insight especially for higher moments of the queue length processes and provable error bounds on the approximation error on the moments, fluid and diffusion limits also offer complementary insight for the sample path behavior of the queueing process.

In the spirit of fluid and diffusion limits, Liu and Whitt (2012) prove a weak law of large numbers limit for the $G_t/GI/n_t{+}GI$ queue and extend the work of Mandelbaum et al. (1998) in the sense that they consider non-Markovian interarrival, service and abandonment times. However, the service times are not time-varying and the limit does not converge almost surely as the limit in this work. In a follow-up paper, Liu and Whitt (2014b) provide a heavy-traffic diffusion limit for $G_t/M/s_t{+}GI$ queues. The methodology used by Liu and Whitt (2014b) is to paste together the overloaded and underloaded intervals of the nonstationary queueing process. Thus, they explicitly avoid the case where the number of servers is equal to the fluid limit. As shown in Mandelbaum et al. (2002), Ko and Gautam (2013), Liu and Whitt (2014b), it appears reasonable to approximate the queue length process with a Gaussian process. However, estimating the parameters of a Gaussian process depends on both fluid and diffusion limits. Lastly, Reed (2009) and Dai et al. (2010) uses the continuous mapping approach to prove diffusion limits for queues with general and phase type service respectively. Although this work was a significant advance in the many server literature, Reed (2009), Dai et al. (2010) do not explore the impact of nonstationary arrival and service times and this work generalizes their work in this regard. Lastly, since our approximations are for nonstationary processes, the approximations are universally useful and apply in any regime.

Using phase-type distributions for approximating general distributions in queueing analysis is not new, see for example Barbour (1976). The matrix-geometric method (MGM) described in Neuts (1981) is a well-known approach for the analysis of non-Markovian queues. MGM, however, can only handle phase-type distributions with a small number of phases due to state space explosion. Nelson and Taaffe (2004a) develop a method based on the partial-moment differential equations for the analysis of $Ph_t/Ph_t/\infty$ queues that accurately estimates the moments of the number of entities in the system. The number of differential equations to evaluate the first two moments is $m_A + m_S - 1 + m_A m_S(m_S + 1)$, where $m_A$ and $m_S$ are the number of phases in the interarrival and service time distributions, respectively. The result, however, is not applicable to the delay models, such as $Ph_t/Ph_t/n_t$ queues studied in our paper. Creemers et al. (2014) devise a phase-type approximation algorithm for small-to-medium-sized queues (2–10 servers) using two-moment matching procedures, however, the downfall is that the method does not scale well with the number servers and it has a high computational cost when the number of servers is large. Our goal is to remove this dependence on the number of servers since it is very limiting in a computational sense, especially in large-scale service systems.

### 1.1. Main Contributions of Paper

The contributions of this work can be summarized as follows. First, we consider the dynamics of a $G_t/G_t/n_t$ queue. The $G_t/G_t/n_t$ queueing model is relatively intractable since we are unable to derive the exact distribution of the queue length as a function of time. Thus, we first approximate the general and non-Markovian arrival and service distributions with phase-type distributions with an appropriate number of phases. This reduces our problem to analyzing the $Ph_t/Ph_t/n_t$ queue, which is more tractable than its general counterpart. Second, we derive fluid and diffusion limits for a $Ph_t/Ph_t/n_t$ queue using *uniform acceleration* coupled with strong approximations of time changed Poisson processes. Unfortunately, when we naively keep track of the number of customers being served in each phase and the number of customers in the system separately, we encounter the *lingering* issue; the fluid limit stays at nondifferentiable points during some intervals having positive measure. This prevents us from deriving a Gaussian or continuous diffusion limit. Thus, another important contribution of our work is our proposal of an alternative Markovian formulation of the queueing process that enables us to successfully obtain the diffusion limit. One attractive feature of our method is its computational efficiency. The number of ordinary differential equations to obtain the fluid and diffusion limits is $O([m_A + m_S]^2)$ and it does *not* depend on the number of servers, $n_t$ like other numerical methods by Creemers et al. (2014). The number of phases used for approximating interarrival and service time distributions is 8–10 and the numerical solution is reached in less than a minute using a commercial solver (e.g., MATLAB). Most previous studies only use two phases for matching first two moments because the increase in the dimension of the state space makes the analysis extremely difficult otherwise. Lastly, we prove the fluid and diffusion limits in a different manner than what is given in Mandelbaum et al. (1998). This is because the proof of Lemma 9.3 of Mandelbaum et al. (1998) depends on the untrue assertion that if a sequence of non-negative random variables defined on the same probability space is tight, then it has a finite limit superior almost surely. Moreover, in Remark 1 of Puhalskii (2013) it is also shown that there are issues with establishing the martingale property in the proof of Lemma 9.1.

### 1.2. Organization of Paper

The remainder of this paper is organized as follows. Section 2 describes the $G_t/G_t/n_t$ queueing model and the problem settings. Section 3 builds a mathematical model for describing the dynamics of the system for the $Ph_t/Ph_t/n_t$ queue. We explain the impact of the lingering problem and introduce an alternative sample path representation for analyzing it. Section 4 constructs the fluid and diffusion limit theorems as approximations for the sample path dynamics of the queueing process in the finite server setting. Section 5 discusses the infinite server setting and provides the fluid and diffusion limits for the infinite server queueing model. Section 6 discusses the numerical examples used to validate the effectiveness of our proposed approach. Section 7 concludes and offers suggestions for future research.

## 2. Problem Description

We consider a $G_t/G_t/n_t$ queue, a time-varying version of a $G/G/n$ queue, with a general time-varying arrival process, a general time-varying service time distribution, and a time-varying number of servers. The system has an infinite capacity of waiting space and customers in the waiting space are served under the first-come, first-served discipline. Let $X(t)$ denote the number of customers in the system at time $t$ and $\bar{x}(t)$ denote the corresponding fluid limit. We assume that the fluid limit $(\bar{x}(t))$ alternates between the underloaded (i.e., $\bar{x}(t) < n_t$) and overloaded (i.e., $\bar{x}(t) > n_t$) regimes and hits the critically loaded regime (i.e., $\bar{x}(t) = n_t$) at most a countable number of times. The performance measures of interest are $\mathrm{E}[X(t)]$, $\mathrm{Var}[X(t)]$ and, if possible, the distribution of $X(t)$ for all time $0 \le t \le T$ and $T < \infty$.

More specifically, we analyze a $Ph_t/Ph_t/n_t$ queue as an approximation of the $G_t/G_t/n_t$ queue since phase-type distributions are dense in all positive-support distributions and the use of phase-type distribution in queueing analysis does not lose generality significantly (Barbour 1976, Whitt 1982, and Asmussen et al. 1996).

A phase-type distribution with $m$ phases represents the time taken from an initial state to an absorbing state of a continuous time Markov chain with the following infinitesimal generator matrix:

$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{s} & \mathbf{S} \end{pmatrix},$$

where $\mathbf{0}$ is a $1 \times m$ zero vector, $\mathbf{s}$ is an $m \times 1$ vector, and $\mathbf{S}$ is an $m \times m$ matrix. Note $\mathbf{s} = -\mathbf{S}\mathbf{e}$ where $\mathbf{e}$ is an $m \times 1$ vector of ones. The matrix $\mathbf{S}$ and the initial distribution $\boldsymbol{\alpha}$ which is a $1 \times m$ vector identify the phase-type distributions. Finding the best phase-type distribution for approximating a general distribution is beyond the scope of this paper, and we refer to the reader to a large number of references (Bobbio et al. 2005, Johnson and Taaffe 1991, Yu et al. 2012, Botta and Harris 1986, Feldmann and Whitt 1998, Ou et al. 1997, Asmussen et al. 1996, Osogami and Harchol-Balter 2006). To give the reader a better understanding of our methodology, we describe the fitting algorithm that we use in Section 6.

We assume that our phase-type distributions have initial distributions, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and infinitesimal generator matrices, $\mathbf{Q}_A$ and $\mathbf{Q}_S$, for the arrival process and service times respectively. The number of phases in $\mathbf{S}_A$ and $\mathbf{S}_S$ is $m_A$ and $m_S$, respectively. The matrices $\mathbf{S}_A$ and $\mathbf{S}_S$, and the vectors $\mathbf{s}_A$ and $\mathbf{s}_S$ can be expressed as

$$\mathbf{S}_A = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1m_A} \\ \vdots & \vdots & \vdots \\ \lambda_{m_A 1} & \cdots & \lambda_{m_A m_A} \end{pmatrix}, \quad \mathbf{s}_A = (\lambda_{10}, \ldots, \lambda_{m_A 0})' \quad (1)$$

$$\mathbf{S}_S = \begin{pmatrix} \mu_{11} & \cdots & \mu_{1m_S} \\ \vdots & \vdots & \vdots \\ \mu_{m_S 1} & \cdots & \mu_{m_S m_S} \end{pmatrix}, \quad \mathbf{s}_S = (\mu_{10}, \ldots, \mu_{m_S 0})', \quad (2)$$

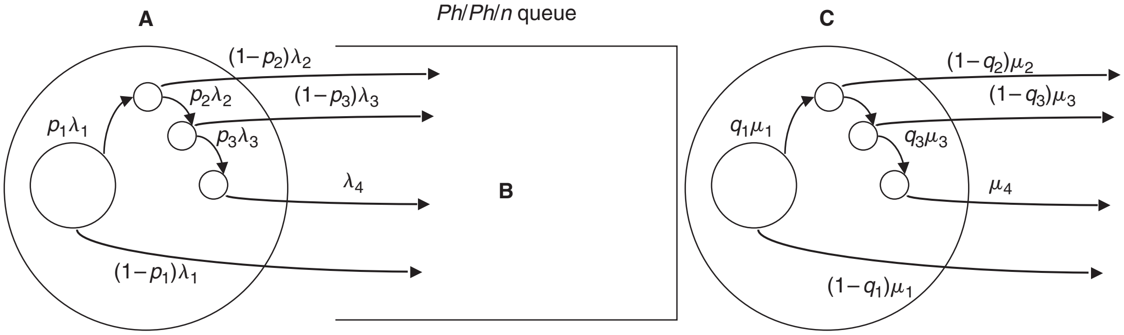
where $\lambda_{jk}$'s and $\mu_{il}$'s agree with the definition of the infinitesimal generator matrices, $\mathbf{Q}_A$ and $\mathbf{Q}_S$. Note that the time-varying extension can be achieved by replacing $\lambda_{jk}$ and $\mu_{il}$ with $\lambda_{jk}(t)$ and $\mu_{il}(t)$ and making sure that their integrals are locally bounded away from infinity.

## 3. The Queueing Model

With the phase-type distributions described in Section 2, we build a mathematical queueing model to describe the dynamics of the $Ph_t/Ph_t/n_t$ queue. We assume that the system starts with no customers.

Figure 1 illustrates an example of $Ph/Ph/n$ queue with Coxian inter-arrival and service times. To model the $Ph_t/Ph_t/n_t$ queue, we need to keep track of the phase in which the arriving customer is (area **A** in Figure 1), the number of customers being served in each phase (area **C**), and the number of customers

**Figure 1.** $Ph/Ph/n$ Queue with Coxian Distributions



in the waiting space (area **B**). We let $U_i(t)$ be the number of customers in phase $i$ of the arrival process at time $t$, $X_j(t)$ be the number of customers being served in phase $j$ of the service process, and $Z(t)$ be the total number of customers in the system. Note that the number of customers in the waiting space is $Z(t) - \sum_{i=1}^{m_S} X_i(t) \geq 0$ and $\sum_{i=1}^{m_A} U_i(t) = 1$ for all $t > 0$. Then, the state of the system $\mathbf{V}(t) = (U_1(t), \ldots, U_{m_A}, X_1(t), \ldots, X_{m_S}, Z(t))'$ is the solution to the following integral equations:

$$
\begin{aligned}
U_j(t) = U_j(0) &+ \sum_{k\neq j}^{m_A} Y_{kj}^{A}\left(\int_0^t \lambda_{kj} U_k(s)\,ds\right)\\
&- \sum_{k\neq j}^{m_A} Y_{jk}^{A}\left(\int_0^t \lambda_{jk} U_j(s)\,ds\right)\\
&- \sum_{k\neq j}^{m_A}\sum_{l=1}^{m_S} Y_{jkl}^{I}\left(\int_0^t \lambda_{j0}\alpha_k\beta_l U_j(s)\mathbf{1}_{\{Z(s)\leq n\}}\,ds\right)\\
&- \sum_{k\neq j}^{m_A} Y_{jk}^{Q}\left(\int_0^t \lambda_{j0}\alpha_k U_j(s)\mathbf{1}_{\{Z(s)> n\}}\,ds\right)\\
&+ \sum_{k\neq j}^{m_A}\sum_{l=1}^{m_S} Y_{kjl}^{I}\left(\int_0^t \lambda_{k0}\alpha_j\beta_l U_k(s)\mathbf{1}_{\{Z(s)\leq n\}}\,ds\right)\\
&+ \sum_{k\neq j}^{m_A} Y_{kj}^{Q}\left(\int_0^t \lambda_{k0}\alpha_j U_k(s)\mathbf{1}_{\{Z(s)> n\}}\,ds\right)\\
&\qquad\text{for } 1 \leq j \leq m_A, \qquad (3)
\end{aligned}
$$

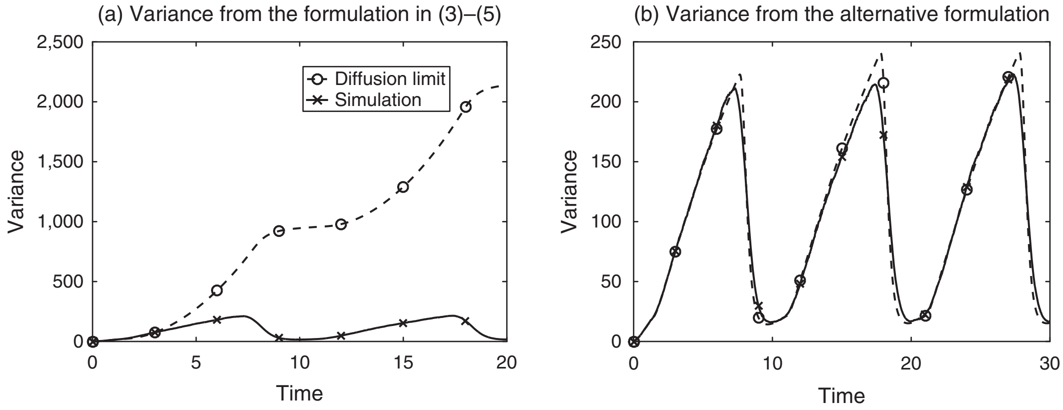
$$
\begin{aligned}
X_i(t) = &\sum_{j=1}^{m_A}\sum_{k=1}^{m_A} Y_{jki}^{I}\left(\int_0^t \lambda_{j0}\alpha_k\beta_i U_j(s)\mathbf{1}_{\{Z(s)\leq n\}}\,ds\right)\\
&+ \sum_{l\neq i}^{m_S} Y_{li}^{S}\left(\int_0^t \mu_{li} X_l(s)\,ds\right)\\
&- \sum_{l\neq i}^{m_S} Y_{il}^{S}\left(\int_0^t \mu_{il} X_i(s)\,ds\right)\\
&- Y_{i0}^{D}\left(\int_0^t \mu_{i0} X_i(s)\mathbf{1}_{\{Z(s)\leq n\}}\,ds\right)\\
&- \sum_{l\neq i}^{m_S} Y_{il}^{D}\left(\int_0^t \mu_{i0} X_i(s)\mathbf{1}_{\{Z(s)> n\}}\beta_l\,ds\right)\\
&+ \sum_{l\neq i}^{m_S} Y_{li}^{D}\left(\int_0^t \mu_{l0} X_l(s)\mathbf{1}_{\{Z(s)> n\}}\beta_i\,ds\right)\\
&\qquad\text{for } 1 \leq i \leq m_S, \qquad (4)
\end{aligned}
$$

$$
\begin{aligned}
Z(t) = &\sum_{j=1}^{m_A}\sum_{k=1}^{m_A}\sum_{l=1}^{m_S} Y_{jkl}^{I}\left(\int_0^t \lambda_{j0}\alpha_k\beta_l U_j(s)\mathbf{1}_{\{Z(s)\leq n\}}\,ds\right)\\
&+ \sum_{j=1}^{m_A}\sum_{k=1}^{m_A} Y_{jk}^{Q}\left(\int_0^t \lambda_{j0}\alpha_k U_j(s)\mathbf{1}_{\{Z(s)> n\}}\,ds\right)\\
&- \sum_{i=1}^{m_S} Y_{i0}^{D}\left(\int_0^t \mu_{i0} X_i(s)\mathbf{1}_{\{Z(s)\leq n\}}\,ds\right)\\
&- \sum_{i=1}^{m_S}\sum_{l=1}^{m_S} Y_{il}^{D}\left(\int_0^t \mu_{i0} X_i(s)\mathbf{1}_{\{Z(s)> n\}}\beta_l\,ds\right). \qquad (5)
\end{aligned}
$$

For notational convenience, Equations (3)–(5) represent the dynamics of a $Ph/Ph/n$ queue. As mentioned in Section 2, we can obtain the time-varying extension by replacing $\lambda_{jk}$, $\mu_{il}$ and $n$ with $\lambda_{jk}(t)$, $\mu_{il}(t)$, and $n(t)$ respectively under mild conditions given in Mandelbaum et al. (1998). Poisson processes, $Y_{kj}^{A}(\cdot)$'s count the number of transitions from phase $k$ to phase $j$ of the arrival process. When the waiting space is empty ($Z(t) \leq n$), Poisson processes, $Y_{jkl}^{I}(\cdot)$'s, count the number of departures from phase $j$ of the arrival process to phase $l$ of the service process according to the initial distribution $\boldsymbol{\beta}$ and the arrival process restarts from phase $k$ according to the initial distribution $\boldsymbol{\alpha}$. When the waiting space is not empty ($Z(t) > n$), Poisson processes, $Y_{jk}^{Q}(\cdot)$'s, count the number of departures from phase $j$ of the arrival process to the waiting space and a new arrival process begins in phase $k$. Poisson processes, $Y_{li}^{S}(\cdot)$'s, count the internal transitions from phase $l$ to phase $j$ of the service process. When the waiting space is empty, Poisson processes, $Y_{i0}^{D}(\cdot)$'s, count the number of departures from phase $i$ of the service process. When the waiting space is not empty, Poisson processes, $Y_{il}^{D}(\cdot)$'s, count the number of departures from phase $i$ and a new customer enters phase $l$ from the waiting space. Note that the Poisson processes explained previously have rate 1 (with random time changes) and are mutually independent.

We can easily figure out that the rate functions in Equations (3)–(5) (the integrands in Poisson processes) are not differentiable with respect to the elements of the state space vector, $\mathbf{V}(t)$. Thus, before applying the uniform acceleration, we conduct a quick check to find

**Figure 2.** Variance Estimation of Exp. 7



(a) Variance from the formulation in (3)–(5)

(b) Variance from the alternative formulation

whether the time during which the fluid limit stays at the nondifferentiable points has measure zero.

Let $\mathbf{v}(t) = (\bar{u}_1(t), \ldots, \bar{u}_{m_A}(t), \bar{x}_1(t), \ldots, \bar{x}_{m_S}(t), \bar{z}(t))'$ be the fluid limit of $\mathbf{V}(t)$. We check the Poisson process, $Y_{il}^D(\cdot)$ in Equation (4). The fluid limit for $Y_{il}^D(\cdot)$ is $\mu_{i0}\bar{x}_i(t)\mathbf{1}_{\{\bar{z}(t)>n\}}$. When $\bar{z}(t)$ hits $n$, the non-differentiable point, $\sum_{i=1}^{m_S} \bar{x}(t) = n$. However, during the overloaded time $\{t\colon\ \bar{z}(t) > n\}$ which can have strictly positive measure in our setting, $\sum_{i=1}^{m_S} \bar{x}(t)$ remains unchanged (i.e., $\sum_{i=1}^{m_S} \bar{x}(t) = n$). This implies that the subvector $(\bar{x}_1(t), \ldots, \bar{x}_{m_S}(t))'$ moves on the hyperplane during the overloaded period and we cannot obtain the diffusion limit from the result of Kurtz (1978) and Mandelbaum et al. (2002). When we try to apply fluid and diffusion limits with Equations (3)–(5) just ignoring the issue, we observe a huge gap between simulation and the numerical solution. For example (Exp. 7 in Section 6), Figure 2(a) shows the gap between the simulated variance and the variance from the diffusion limit. We devise an alternative formulation which can significantly improve the approximation accuracy (see Figure 2(b)).

The issue occurs because $\sum_{i=1}^{m_S} \bar{x}(t) = n$ during the overloaded period. The alternative formulation avoids this situation but requires an additional assumption that the phase-type distribution for service times has a unique initial state. Such distributions include the Erlang distribution and the Coxian distribution. According to Asmussen et al. (1996), the Coxian distribution provides almost the same quality of fit as the general phase-type distribution with the same number of phases. One reason is that the Coxian and generalized hyperexponential distribution, which are specific classes of phase-type distributions, are also dense in the class of positive-support distributions, see for example Sasaki et al. (2004). Thus, the additional assumption of restricting to the Coxian class, therefore, may not be quite restrictive. Without loss of generality, we assume the unique initial state is phase 1. The main idea is to maintain the waiting space inside phase 1 and control transition rates from phase 1 so that the system serves at most $n$ customers. We have the same state space except for $Z(t)$ because $X_1(t)$ accounts for customers in the waiting space. Using this representation, we can now write our new formulation of the queueing process as follows:

$$\begin{aligned}
U_j(t) = U_j(0) &+ \sum_{k\neq j}^{m_A} Y_{kj}^A\left(\int_0^t \lambda_{kj} U_k(s)\,ds\right)\\
&- \sum_{k\neq j}^{m_A} Y_{jk}^A\left(\int_0^t \lambda_{jk} U_j(s)\,ds\right)\\
&- \sum_{k\neq j}^{m_A} Y_{jk}^I\left(\int_0^t \lambda_{j0}\alpha_k U_j(s)\,ds\right)\\
&\qquad + \sum_{k\neq j}^{m_A} Y_{kj}^I\left(\int_0^t \lambda_{k0}\alpha_j U_k(s)\,ds\right)\\
&\qquad\qquad \text{for } 1 \le j \le m_A, \qquad (6)
\end{aligned}$$

$$\begin{aligned}
X_1(t) = &\sum_{j=1}^{m_A}\sum_{k=1}^{m_A} Y_{jk}^I\left(\int_0^t \lambda_{j0}\alpha_k U_j(s)\,ds\right)\\
&+ \sum_{l\neq 1}^{m_S} Y_{l1}^S\left(\int_0^t \mu_{l1} X_l(s)\,ds\right)\\
&- \sum_{l\neq 1}^{m_S} Y_{1l}^S\left(\int_0^t \mu_{1l}\left[\mathbf{1}_{\{\sum_{r=1}^{m_S} X_r(s)\le n\}} X_1(s)\right.\right.\\
&\qquad \left.\left. + \mathbf{1}_{\{\sum_{r=1}^{m_S} X_r(s)>n\}}\left(n - \sum_{r=2}^{m_S} X_r(s)\right)^+\right] ds\right)\\
&- Y_1^D\left(\int_0^t \mu_{10}\left[\mathbf{1}_{\{\sum_{r=1}^{m_S} X_r(s)\le n\}} X_1(s)\right.\right.\\
&\qquad \left.\left. + \mathbf{1}_{\{\sum_{r=1}^{m_S} X_r(s)>n\}}\left(n - \sum_{r=2}^{m_S} X_r(s)\right)^+\right] ds\right). \qquad (7)
\end{aligned}$$

$$\begin{aligned}
X_i(t) = &\,Y_{1i}^S\left(\int_0^t \mu_{1i}\left[\mathbf{1}_{\{\sum_{r=1}^{m_S} X_r(s)\le n\}} X_1(s)\right.\right.\\
&\qquad \left.\left. + \mathbf{1}_{\{\sum_{r=1}^{m_S} X_r(s)>n\}}\left(n - \sum_{r=2}^{m_S} X_r(s)\right)^+\right] ds\right)\\
&+ \sum_{l=2,\, l\neq i}^{m_S} Y_{li}^S\left(\int_0^t \mu_{li} X_l(s)\,ds\right)
\end{aligned}$$

$$- \sum_{l \neq i}^{m_S} Y_{il}^S\left(\int_0^t \mu_{il} X_i(s)\, ds\right) - Y_i^D\left(\int_0^t \mu_{i0} X_i(s)\, ds\right) \quad \text{for } 2 \le i \le m_S. \qquad (8)$$

Poisson processes, $Y_{kj}^A(\cdot)$'s and $Y_{li}^S(\cdot)$'s, are the same as those in Equations (3) and (4). Poisson processes, $Y_{jkl}^I(\cdot)$'s in Equation (3) are now replaced by $Y_{jk}^I(\cdot)$'s because the initial state of the service process is phase 1, that is, we do not need the index of the starting phase in the service process. Then, Poisson processes, $Y_{jk}^I(\cdot)$'s count the number of departures from phase $j$ that restart from phase $k$ of the arrival process according to the initial distribution $\boldsymbol{\alpha}$. Note that we do not have to count the number of departures that restart from the same phase, i.e., we do not count the case of $j = k$. Poisson processes, $Y_i^D(\cdot)$'s count departures from phase $i$ of the service process. Note that the Poisson processes explained previously have rate 1 (with random time changes) and are mutually independent. We can verify that the issue is not incurred in Equations (6)–(8). In the following section we describe the fluid and diffusion approximations.

### 3.1. Lipschitz Representation

It turns out that we can write our new formulation in terms of Lipschitz rate functions. This representation will aid us tremendously when proving the fluid and diffusion limit theorems for the queueing model.

$$\begin{aligned}
U_j(t) = U_j(0) &+ \sum_{k \neq j}^{m_A} Y_{kj}^A\left(\int_0^t \lambda_{kj} U_k(s)\, ds\right) \\
&- \sum_{k \neq j}^{m_A} Y_{jk}^A\left(\int_0^t \lambda_{jk} U_j(s)\, ds\right) \\
&- \sum_{k \neq j}^{m_A} Y_{jk}^I\left(\int_0^t \lambda_{j0} \alpha_k U_j(s)\, ds\right) \\
&+ \sum_{k \neq j}^{m_A} Y_{kj}^I\left(\int_0^t \lambda_{k0} \alpha_j U_k(s)\, ds\right) \quad \text{for } 1 \le j \le m_A,
\end{aligned}$$

$$\begin{aligned}
X_1(t) = \sum_{j=1}^{m_A} \sum_{k=1}^{m_A} &Y_{jk}^I\left(\int_0^t \lambda_{j0} \alpha_k U_j(s)\, ds\right) \\
&+ \sum_{l \neq 1}^{m_S} Y_{l1}^S\left(\int_0^t \mu_{l1} X_l(s)\, ds\right) \\
&- \sum_{l \neq 1}^{m_S} Y_{1l}^S\left(\int_0^t \mu_{1l}\left[\left(X_1(s) \wedge \left(n - \sum_{r=2}^{m_S} X_r(s)\right)^+\right)\right] ds\right) \\
&- Y_1^D\left(\int_0^t \mu_{10}\left[\left(X_1(s) \wedge \left(n - \sum_{r=2}^{m_S} X_r(s)\right)^+\right)\right] ds\right).
\end{aligned}$$

$$\begin{aligned}
X_i(t) = Y_{1i}^S&\left(\int_0^t \mu_{1i}\left[\left(X_1(s) \wedge \left(n - \sum_{r=2}^{m_S} X_r(s)\right)^+\right)\right] ds\right) \\
&+ \sum_{l=2, l \neq i}^{m_S} Y_{li}^S\left(\int_0^t \mu_{li} X_l(s)\, ds\right) \\
&- \sum_{l \neq i}^{m_S} Y_{il}^S\left(\int_0^t \mu_{il} X_i(s)\, ds\right) \\
&- Y_i^D\left(\int_0^t \mu_{i0} X_i(s)\, ds\right) \quad \text{for } 2 \le i \le m_S.
\end{aligned}$$

## 4. Fluid and Diffusion Approximations

In this section, we now provide our second main contribution of the paper, fluid and diffusion limit theorems for the queue length process. However, we first provide some definitions for notational convenience of the reader that will be used throughout the rest of the paper.

$$\mathbf{V}(t) = (U_1(t), \ldots, U_{m_A}(t), X_1(t), \ldots, X_{m_S}(t))'.$$
$$\mathbf{v} = (u_1, \ldots, u_{m_A}, x_1, \ldots, x_{m_S})'.$$

$\mathbf{d}_{jk}^A$: $(m_A + m_S) \times 1$ vector, $j$th element is $-1$, $k$th element is 1, and other elements are 0.
$\mathbf{d}_{jk}^I$: $(m_A + m_S) \times 1$ vector, $j$th element is $-1$, $k$th element is 1, and other elements are 0.
$\mathbf{d}_{il}^S$: $(m_A + m_S) \times 1$ vector, $(m_A + i)$th element is $-1$, $(m_A + l)$th element is 1, and other elements are 0.
$\mathbf{d}_i^D$: $(m_A + m_S) \times 1$ vector, $(m_A + i)$th element is $-1$, and other elements are 0.

$f_{jk}^A(t, \mathbf{v})$: rate function (integrand) in $Y_{jk}^A(\cdot)$.
$f_{jk}^I(t, \mathbf{v})$: rate function (integrand) in $Y_{jk}^I(\cdot)$.
$f_{il}^S(t, \mathbf{v})$: rate function (integrand) in $Y_{il}^S(\cdot)$.
$f_i^D(t, \mathbf{v})$: rate function (integrand) in $Y_i^D(\cdot)$.
$W_{jk}^A(t), W_{jk}^I(t), W_{il}^S(t), W_i^D(t)$: mutually independent standard Brownian motions.

$$\begin{aligned}
\mathbf{F}(t, \mathbf{v}) = &\sum_{j=1}^{m_A} \sum_{k=1, k \neq j}^{m_A} \mathbf{d}_{jk}^A f_{jk}^A(t, \mathbf{v}) + \sum_{j=1}^{m_A} \sum_{k=1}^{m_A} \mathbf{d}_{jk}^I f_{jk}^I(t, \mathbf{v}) \\
&+ \sum_{i=1}^{m_S} \sum_{l=1, l \neq i}^{m_S} \mathbf{d}_{il}^S f_{il}^S(t, \mathbf{v}) + \sum_{i=1}^{m_S} \mathbf{d}_i^D f_i^D(t, \mathbf{v}).
\end{aligned}$$

$$\begin{aligned}
d\mathbf{H}(t, \mathbf{v}) = &\sum_{j=1}^{m_A} \sum_{k=1, k \neq j}^{m_A} \mathbf{d}_{jk}^A \sqrt{f_{jk}^A(t, \mathbf{v})}\, dW_{jk}^A(t) \\
&+ \sum_{j=1}^{m_A} \sum_{k=1}^{m_A} \mathbf{d}_{jk}^I \sqrt{f_{jk}^I(t, \mathbf{v})}\, dW_{jk}^I(t) \\
&+ \sum_{i=1}^{m_S} \sum_{l=1, l \neq i}^{m_S} \mathbf{d}_{il}^S \sqrt{f_{il}^S(t, \mathbf{v})}\, dW_{il}^S(t) \\
&+ \sum_{i=1}^{m_S} \mathbf{d}_i^D \sqrt{f_i^D(t, \mathbf{v})}\, dW_i^D(t).
\end{aligned}$$

$$\begin{aligned}
\mathbf{G}(t, \mathbf{v}) = &\sum_{j=1}^{m_A} \sum_{k=1, k \neq j}^{m_A} \mathbf{d}_{jk}^A \mathbf{d}'^A_{jk} f_{jk}^A(t, \mathbf{v}) + \sum_{j=1}^{m_A} \sum_{k=1}^{m_A} \mathbf{d}_{jk}^I \mathbf{d}'^I_{jk} f_{jk}^I(t, \mathbf{v}) \\
&+ \sum_{i=1}^{m_S} \sum_{l=1, l \neq i}^{m_S} \mathbf{d}_{il}^S \mathbf{d}'^S_{il} f_{il}^S(t, \mathbf{v}) + \sum_{i=1}^{m_S} \mathbf{d}_i^D \mathbf{d}'^D_i f_i^D(t, \mathbf{v}).
\end{aligned}$$

With the aforementioned definitions, we rewrite Equations (6)–(8) in a vector form as follows:

$$\begin{aligned}\mathbf{V}(t) ={}& \mathbf{V}(0) + \sum_{j=1}^{m_A}\sum_{k=1,k\neq j}^{m_A} \mathbf{d}_{jk}^{A} Y_{jk}^{A}\left(\int_0^t f_{jk}^{A}(s,\mathbf{V}(s))\,ds\right)\\ &+ \sum_{j=1}^{m_A}\sum_{k=1}^{m_A} \mathbf{d}_{jk}^{I} Y_{jk}^{I}\left(\int_0^t f_{jk}^{I}(s,\mathbf{V}(s))\,ds\right)\\ &+ \sum_{i=1}^{m_S}\sum_{l=1,l\neq i}^{m_S} \mathbf{d}_{il}^{S} Y_{il}^{S}\left(\int_0^t f_{il}^{S}(s,\mathbf{V}(s))\,ds\right)\\ &+ \sum_{i=1}^{m_S} \mathbf{d}_{i}^{D} Y_{i}^{D}\left(\int_0^t f_{i}^{D}(s,\mathbf{V}(s))\,ds\right).\end{aligned}$$

Following the procedure of the uniform acceleration in Mandelbaum et al. (1998) and Kurtz (1978), we define a sequence of processes $\{\mathbf{V}^{\eta}(t), \eta \geq 1, t \geq 0\}$, where

$$\begin{aligned}\mathbf{V}^{\eta}(t) ={}& \mathbf{V}^{\eta}(0) + \sum_{j=1}^{m_A}\sum_{k=1,k\neq j}^{m_A} \mathbf{d}_{jk}^{A} Y_{jk}^{A}\left(\eta\int_0^t f_{jk}^{A}(s,\bar{\mathbf{V}}^{\eta}(s))\,ds\right)\\ &+ \sum_{j=1}^{m_A}\sum_{k=1}^{m_A} \mathbf{d}_{jk}^{I} Y_{jk}^{I}\left(\eta\int_0^t f_{jk}^{I}(s,\bar{\mathbf{V}}^{\eta}(s))\,ds\right)\\ &+ \sum_{i=1}^{m_S}\sum_{l\neq i}^{m_S} \mathbf{d}_{il}^{S} Y_{il}^{S}\left(\eta\int_0^t f_{il}^{S}(s,\bar{\mathbf{V}}^{\eta}(s))\,ds\right)\\ &+ \sum_{i=1}^{m_S} \mathbf{d}_{i}^{D} Y_{i}^{D}\left(\eta\int_0^t f_{i}^{D}(s,\bar{\mathbf{V}}^{\eta}(s))\,ds\right),\end{aligned}$$

where $\bar{\mathbf{V}}^{\eta}(t) = \mathbf{V}^{\eta}(t)/\eta$.

Note that we accelerate the arrival rate by accelerating the sum of $U_j^{\eta}(t)$ for $t \geq 0$, that is,

$$\sum_{j=1}^{m_A} U_j^{\eta}(t) = \eta, \quad \text{for } t \geq 0.$$

### 4.1. Fluid Limit Theorem

Then, we have the following proposition for the fluid limit:

**Theorem 1.** *Suppose* $\mathbf{V}^{\eta}(0)/\eta \to \mathbf{v}(0)$ *as* $\eta \to \infty$, *then*

$$\lim_{\eta\to\infty} \frac{\mathbf{V}^{\eta}(t)}{\eta} = \mathbf{v}(t) \text{ almost surely,}$$

*where* $\mathbf{v}(t) = (u_1(t), \ldots, u_{m_A}(t), x_1(t), \ldots, x_{m_S}(t))'$ *is the solution to the following system of ordinary differential equations:*

$$\begin{aligned}\frac{d}{dt}\mathbf{v}(t) ={}& \sum_{j=1}^{m_A}\sum_{k\neq j}^{m_A} \mathbf{d}_{jk}^{A} f_{jk}^{A}(t,\mathbf{v}(t)) + \sum_{j=1}^{m_A}\sum_{k\neq j}^{m_A} \mathbf{d}_{jk}^{I} f_{jk}^{I}(t,\mathbf{v}(t))\\ &+ \sum_{i=1}^{m_S}\sum_{l\neq i}^{m_S} \mathbf{d}_{il}^{S} f_{il}^{S}(t,\mathbf{v}(t)) + \sum_{i=1}^{m_S} \mathbf{d}_{i}^{D} f_{i}^{D}(t,\mathbf{v}(t)). \quad (9)\end{aligned}$$

**Proof.** See the online supplement.

### 4.2. Diffusion Limit Theorem

Now that we have the fluid limit, $\mathbf{v}(t)$, we can derive the diffusion limit as follows:

**Theorem 2.** *Let* $\mathbf{D}^{\eta}(t) = \sqrt{\eta}(\mathbf{V}^{\eta}(t)/\eta - \mathbf{v}(t))$, *then we have that*

$$\lim_{\eta\to\infty} \mathbf{D}^{\eta}(t) = \mathbf{D}(t) \text{ in distribution,}$$

*where* $\mathbf{D}(t)$ *is the solution to the following stochastic differential equation*

$$d\mathbf{D}(t) = d\mathbf{H}(t,\mathbf{v}(t)) + \partial\mathbf{F}(t,\mathbf{v}(t))\mathbf{D}(t)\,dt,$$

*and* $\partial\mathbf{F}(t,\mathbf{v})$ *is the gradient matrix of* $\mathbf{F}(t,\mathbf{v})$ *with respect to* $\mathbf{v}$. *If* $\mathbf{D}(0)$ *is a constant or normally distributed, then* $\{\mathbf{D}(t), t \geq 0\}$ *is a Gaussian process* (Arnold 1992).

**Proof.** See the online supplement.

Now that we have fluid and diffusion limits for the queue length process, we can therefore, for a large $\eta$, give an approximation for the original model as

$$\mathbf{V}^{\eta}(t) \approx \eta\mathbf{v}(t) + \sqrt{\eta}\,\mathbf{D}(t).$$

One should note that by increasing $\eta$ also implies that we are effectively increasing the number of servers along with other parameters (Mandelbaum et al. 2002). Therefore, if the number of servers is sufficiently large in the original setting (i.e., $\eta = 1$), we can approximate $\mathbf{V}(t)$ as follows:

$$\mathbf{V}(t) \approx \mathbf{v}(t) + \mathbf{D}(t).$$

Since $\{\mathbf{D}(t), t \geq 0\}$ is a Gaussian process, $\{\mathbf{V}(t), t \geq 0\}$ is approximately a Gaussian process. If we have the mean vector and the covariance matrix of $\mathbf{D}(t)$, we can approximately identify the queue length distributions as follows:

**Proposition 1** (Mean and Covariance Matrix of $\mathbf{D}(t)$, Arnold 1992). *Let* $\mathbf{M}(t) = \mathrm{E}[\mathbf{D}(t)]$ *and* $\boldsymbol{\Sigma}(t) = \mathrm{Cov}[\mathbf{D}(t), \mathbf{D}(t)]$. *Then,* $\mathbf{M}(t)$ *and* $\boldsymbol{\Sigma}(t)$ *are the unique solution to the following ordinary equations:*

$$\frac{d}{dt}\mathbf{M}(t) = \partial\mathbf{F}(t,\mathbf{v}(t))\mathbf{M}(t), \quad (10)$$

$$\begin{aligned}\frac{d}{dt}\boldsymbol{\Sigma}(t) ={}& \partial\mathbf{F}(t,\mathbf{v}(t))\boldsymbol{\Sigma}(t)\\ &+ \boldsymbol{\Sigma}(t)\partial\mathbf{F}(t,\mathbf{v}(t))' + \mathbf{G}(t,\mathbf{v}(t)). \quad (11)\end{aligned}$$

*If* $\mathbf{M}(0) = \mathbf{0}$, $\mathbf{M}(t) = \mathbf{0}$ *for all* $t \geq 0$.

Recall that we start with an empty queue, which implies that we do not have to solve Equation (10), i.e., $\mathbf{M}(t) = \mathbf{0}$ for all $t \geq 0$.

By solving differential Equations (9) and (11), we can approximate $\mathrm{E}[\mathbf{V}(t)]$ and $\mathrm{Cov}[\mathbf{V}(t), \mathbf{V}(t)]$ as follows:

$$\mathrm{E}[\mathbf{V}(t)] \approx \mathbf{v}(t),$$
$$\mathrm{Cov}[\mathbf{V}(t), \mathbf{V}(t)] \approx \boldsymbol{\Sigma}(t).$$

Let $X(t)$ be the number of customers in the system at time $t$. Then,

$$X(t) = \sum_{i=1}^{m_S} X_i(t).$$

Note that $\{X(t), t \geq 0\}$ is approximately a Gaussian process and we can obtain the mean and variance of $X(t)$ as follows:

$$\mathrm{E}[X(t)] = \sum_{i=1}^{m_S} \mathrm{E}[X_i(t)],$$

$$\mathrm{Var}[X(t)] = \sum_{i=1}^{m_S} \mathrm{Var}[X_i(t)] + 2 \sum_{i=1}^{m_S-1} \sum_{l=i+1}^{m_S} \mathrm{Cov}[X_i(t), X_l(t)].$$

### 4.3. Probability of Delay and Excessive Delay

Now armed with our fluid and diffusion approximations, we can also approximate other performance measures other than the mean and variance of the queue length process. One of the most important performance measures is the probability of delay or the probability that a customer must wait for service when they arrive to the queue. However, we derive an approximation for a more general quantity called the probability of excessive delay, i.e.,

$$\mathbb{P}(\text{Delay} > z) = \mathbb{P}(W(t) > z),$$

where $W(t)$ is the waiting time of a customer that joins the queue at time $t$. Thus, given our fluid and diffusion approximations for the mean and variance of the queue length we can derive a Gaussian approximation for the probability of excessive delay as

$$\begin{aligned}\mathbb{P}(W(t) > z) &\approx \mathbb{P}\left(\frac{(X(t) - n(t))^+}{\mu \cdot n(t)} > z\right) \\ &\approx \mathbb{P}\left(\frac{(x(t) + \sigma(t) \cdot \tilde{Z} - n(t))^+}{\mu \cdot n(t)} > z\right) \\ &= \mathbb{P}\left(\tilde{Z} > \frac{n(t) - x(t) + z \cdot \mu \cdot n(t)}{\sigma(t)}\right) \\ &= \bar{\Phi}\left(\frac{n(t) - x(t) + z \cdot \mu \cdot n(t)}{\sigma(t)}\right),\end{aligned}$$

where $x(t) = \sum_{i=1}^{m_S} x_i(t)$, $\tilde{Z}$ is a standard Gaussian random variable, and $\sigma(t)$ is the standard deviation of the diffusion limit corresponding to $X(t)$. Moreover, when $z = 0$, our expression reduces to the probability of delay, i.e.,

$$\begin{aligned}\mathbb{P}(\mathit{Delay}) &= \mathbb{P}(W(t) > 0) \\ &\approx \mathbb{P}(X(t) \geq n(t)) \\ &\approx \mathbb{P}\left(\tilde{Z} \geq \frac{n(t) - x(t)}{\sigma(t)}\right) \\ &\approx \bar{\Phi}\left(\frac{n(t) - x(t)}{\sigma(t)}\right).\end{aligned}$$

## 5. The Infinite Server Case

In this section, we demonstrate that we can also apply our fluid and diffusion limits in the infinite server setting as well. This provides first and second order approximations for the queue length process that was first studied by Nelson and Taaffe (2004a). However, we rigorously justify our approximations by limit theorems.

### 5.1. Infinite Server Representation

In the infinite server setting we have the following representation for the queue length process,

$$\begin{aligned}U_j(t) = U_j(0) &+ \sum_{k \neq j}^{m_A} Y_{kj}^A\left(\int_0^t \lambda_{kj} U_k(s)\, ds\right) \\ &- \sum_{k \neq j}^{m_A} Y_{jk}^A\left(\int_0^t \lambda_{jk} U_j(s)\, ds\right) \\ &- \sum_{k \neq j}^{m_A} Y_{jk}^I\left(\int_0^t \lambda_{j0}\alpha_k U_j(s)\, ds\right) \\ &+ \sum_{k \neq j}^{m_A} Y_{kj}^I\left(\int_0^t \lambda_{k0}\alpha_j U_k(s)\, ds\right) \quad \text{for } 1 \leq j \leq m_A, \\ X_1(t) = \sum_{j=1}^{m_A}\sum_{k=1}^{m_A} &Y_{jk}^I\left(\int_0^t \lambda_{j0}\alpha_k U_j(s)\, ds\right) \\ &+ \sum_{l \neq 1}^{m_S} Y_{l1}^S\left(\int_0^t \mu_{l1} X_l(s)\, ds\right) \\ &- \sum_{l \neq 1}^{m_S} Y_{1l}^S\left(\int_0^t \mu_{1l} X_1(s)\, ds\right) \\ &- Y_1^D\left(\int_0^t \mu_{10} X_1(s)\, ds\right), \\ X_i(t) = Y_{1i}^S\left(\int_0^t \mu_{1i} X_1(s)\, ds\right) &+ \sum_{l=2, l \neq i}^{m_S} Y_{li}^S\left(\int_0^t \mu_{li} X_l(s)\, ds\right) \\ &- \sum_{l \neq i}^{m_S} Y_{il}^S\left(\int_0^t \mu_{il} X_i(s)\, ds\right) - Y_i^D\left(\int_0^t \mu_{i0} X_i(s)\, ds\right) \\ &\qquad \text{for } 2 \leq i \leq m_S.\end{aligned}$$

The major difference between the finite and infinite server settings is the rate functions for the $X_i$ Poisson processes. In the finite setting, at most $n_t$ customers can be processed at any time $t$, however, in the infinite server setting, this is no longer a limitation. Thus, all of the rate functions with the terms $X_1 \wedge (n - \sum_{r=2}^{m_S} X_r(s))^+$ since the term $(n - \sum_{r=2}^{m_S} X_r(s))^+$ is equal to $\infty$.

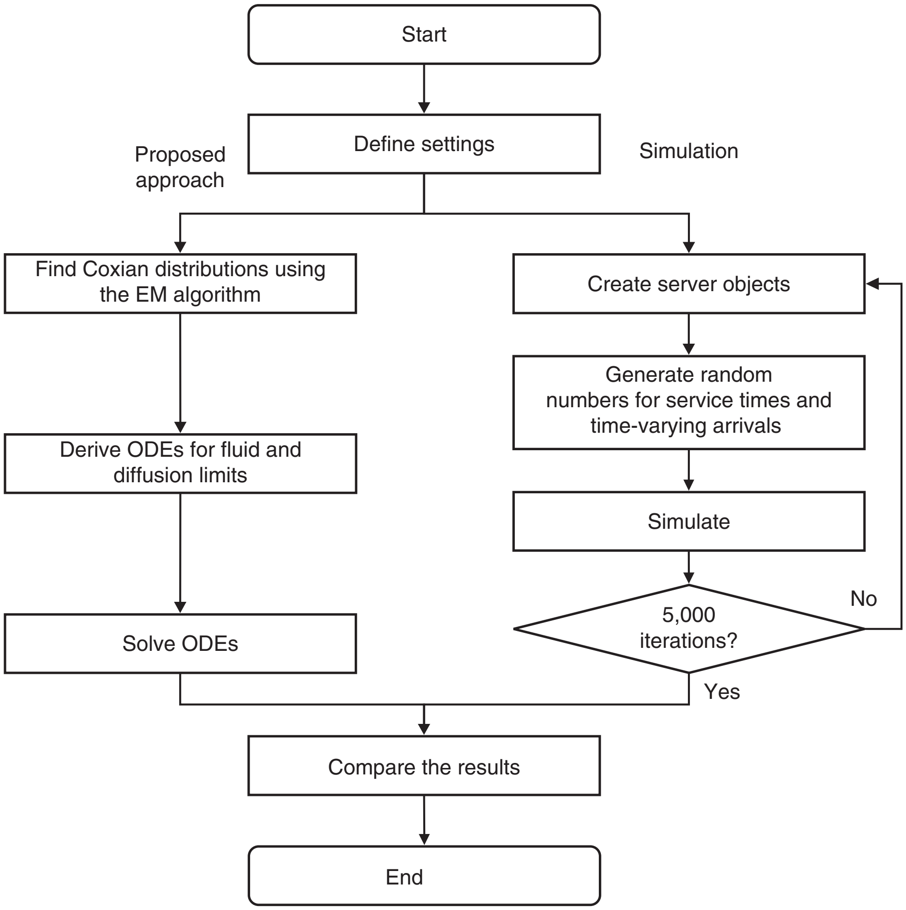
### 5.2. Infinite Server Fluid Limit Theorem

We have the following proposition for the fluid limit for the $Ph_t/Ph_t/\infty$ queue.

**Proposition 2.** *Suppose* $\mathbf{V}^{\eta,\infty}(0)/\eta \to \mathbf{v}^\infty(0)$ *as* $\eta \to \infty$, *then*

$$\lim_{\eta \to \infty} \frac{\mathbf{V}^{\eta,\infty}(t)}{\eta} = \mathbf{v}^\infty(t) \textit{ almost surely,}$$

**Figure 3.** Overall Flow of the Numerical Study


Start
Define settings
Proposed approach
Simulation
Find Coxian distributions using the EM algorithm
Derive ODEs for fluid and diffusion limits
Solve ODEs
Create server objects
Generate random numbers for service times and time-varying arrivals
Simulate
5,000 iterations?
No
Yes
Compare the results
End

*where* $\mathbf{v}^\infty(t)$ *is the solution to the following system of ordinary differential equations*:

$$\frac{d}{dt}\mathbf{v}^\infty(t) = \sum_{j=1}^{m_A}\sum_{k\neq j}^{m_A}\mathbf{d}^A_{jk} f^A_{jk}(t,\mathbf{v}^\infty(t)) + \sum_{j=1}^{m_A}\sum_{k\neq j}^{m_A}\mathbf{d}^I_{jk} f^I_{jk}(t,\mathbf{v}^\infty(t)) + \sum_{i=1}^{m_S}\sum_{l\neq i}^{m_S}\mathbf{d}^S_{il} f^S_{il}(t,\mathbf{v}^\infty(t)) + \sum_{i=1}^{m_S}\mathbf{d}^D_i f^D_i(t,\mathbf{v}^\infty(t)),$$

*where the rate functions correspond to the infinite server representation given in Section* 5.1.

**Proof.** The proof of this result immediately follows from the proof of the finite case and setting $n=\infty$.

### 5.3. Infinite Server Diffusion Limit Theorem

Now that we have the fluid limit, $\mathbf{v}^\infty(t)$, we can derive the diffusion limit as follows:

**Proposition 3.** *Let* $\mathbf{D}^{\eta,\infty}(t) = \sqrt{\eta}(\mathbf{V}^{\eta,\infty}(t)/\eta - \mathbf{v}^\infty(t))$, *then we have that*

$$\lim_{\eta\to\infty}\mathbf{D}^{\eta,\infty}(t) = \mathbf{D}^\infty(t) \textit{ in distribution},$$

*where* $\mathbf{D}^\infty(t)$ *is the solution to the following stochastic differential equation*

$$d\mathbf{D}^\infty(t) = \mathbf{H}(t,\mathbf{v}^\infty(t)) + \partial\mathbf{F}(t,\mathbf{v}^\infty(t))\mathbf{D}^\infty(t)\,dt,$$

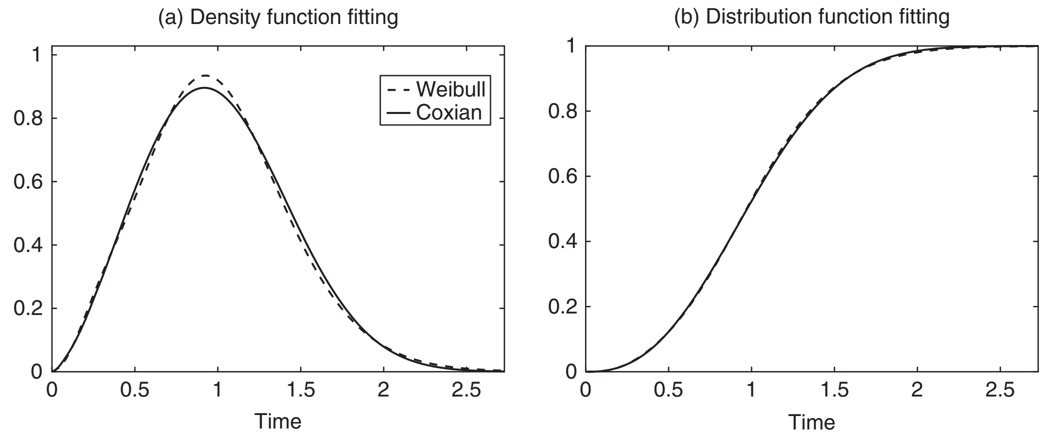
*and* $\partial\mathbf{F}(t,\mathbf{v}^\infty(t))$ *is the gradient matrix of* $\mathbf{F}(t,\mathbf{v}^\infty(t))$ *with respect to* $\mathbf{v}^\infty(t)$, *where the rate functions correspond to the infinite server representation given in Section* 5.1.

**Proof.** The proof of this result immediately follows from the proof of the finite case and setting $n=\infty$.

## 6. Numerical Results

In this section, we provide some numerical results comparing the proposed method with the simulation results. Referring to the flow chart in Figure 3, we choose Coxian distributions to approximate Weibull and lognormal distributions for interarrival and service times. Coxian distributions have a unique initial state that the proposed method requires and the overall fitting quality is known to be good (Asmussen et al. 1996). We use the EM algorithm developed by Asmussen et al. (1996), although other phase-type distributions and fitting algorithms can also be used. Since we want to approximate the distribution itself, we use 8–10 phases to fit the target distributions accurately. Figure 4 illustrates a density and distribution fitting with a Coxian distribution. In this example, we use 10 phases to approximate the Weibull distribution. We

**Figure 4.** Weibull(1.1271, 2.5) and Corresponding Coxian Distributions



(a) Density function fitting

(b) Distribution function fitting

derive the ordinary differential equations (ODEs) from Equations (9) and (11), and solve them using MATLAB. We write the simulation code in C++. To generate a general time-varying arrival process, we implement the algorithm based on the standard equilibrium renewal process (SERP) explained in the longer version of Liu and Whitt (2012). Explaining SERP briefly, we have a stationary renewal process with a general interevent time distribution $G(\cdot)$ which we call a base distribution. Then, we can obtain a time-varying arrival process by applying the change of time technique to the renewal process with a given time-varying rate function. Likewise, we find a phase-type distribution $Ph(\cdot)$ for fitting the base distribution $G(\cdot)$ and obtain a time-varying arrival process by applying the same change of time technique. We, therefore, do not have to run the fitting algorithm multiple times to find one phase-type distribution at time $t$ and another one at time $t'$. We use Weibull distributions with mean 1 as a base distribution to generate time-varying arrival times. We run 5,000 independent instances for each setting and estimate the mean and the variance of the number of customers in the system and the probability of (excessive) delay over time.

We choose two Weibull distributions having the same mean 1 for the arrival processes: the squared coefficient of variation (SCoV) of Weibull(0.79, 0.7) is 2.1387 which is greater than one, and the SCoV of Weibull(1.1271, 2.5) is 0.1831 which is less than one. Time-varying rates are applied to the base distributions for constructing the actual arrival processes. We do not consider the case when the SCoV is 1 since it is an exponential distribution and has been studied extensively in the literature. For the service times, we choose two lognormal distributions with the different SCoV values. Without loss of generality, the means of two service time distributions are 1. Increasing the number of servers makes us expect more accurate estimations since the fluid and diffusion limits are asymptotically exact. Therefore, we compare the cases when the number of servers is 50 and 200. The corresponding time-varying rates to the number of servers are $45 + 30\sin(2\pi t/10)$ and $180 + 120\sin(2\pi t/10)$ respectively. Then, we have eight combinations of experiments: two distributions for arrivals, two distributions for services, two values of the number of servers:

Exp. 1: 50 servers, SCoV of inter-arrival times > 1 and SCoV of service times > 1

—Time-varying rate: $45 + 30\sin(2\pi t/10)$

—Base interarrival time distribution: Weibull(0.79, 0.7), SCoV = 2.1387

—Service time distribution: Lognormal(−0.5, 1), SCoV = 1.7183

Exp. 2: 200 servers, SCoV of interarrival times > 1 and SCoV of service times > 1

—Time-varying rate: $180 + 120\sin(2\pi t/10)$

—Base interarrival time distribution: Weibull(0.79, 0.7), SCoV = 2.1387

—Service time distribution: Lognormal(−0.5, 1), SCoV = 1.7183

Exp. 3: 50 servers, SCoV of interarrival times > 1 and SCoV of service times < 1

—Time-varying rate: $45 + 30\sin(2\pi t/10)$

—Base interarrival time distribution: Weibull(0.79, 0.7), SCoV = 2.1387

—Service time distribution: Lognormal(−0.2027, 0.6368), SCoV = 0.5

Exp.4: 200 servers, SCoV of interarrival times > 1 and SCoV of service times < 1

—Time-varying rate: $180 + 120\sin(2\pi t/10)$

—Base interarrival time distribution: Weibull(0.79, 0.7), SCoV = 2.1387

—Service time distribution: Lognormal(−0.2027, 0.6368), SCoV = 0.5

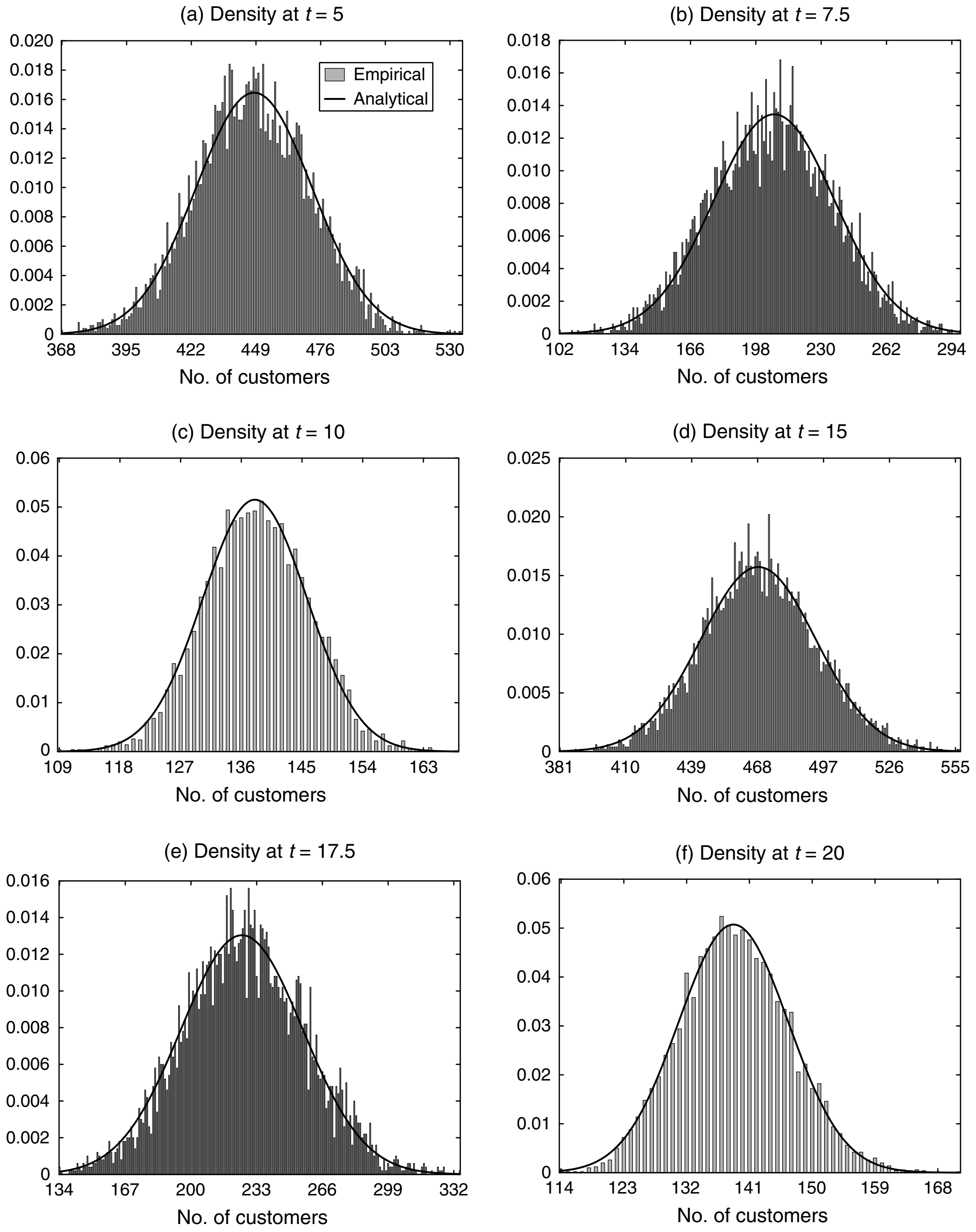
Exp. 5: 50 servers, SCoV of interarrival times < 1 and SCoV of service times > 1

—Time-varying rate: $45 + 30\sin(2\pi t/10)$

—Base interarrival time distribution: Weibull(1.1271, 2.5), SCoV = 0.1831

—Service time distribution: Lognormal(−0.5, 1), SCoV = 1.7183

**Figure 5.** Density of the Number of Customers at Time 5, 7.5, 10, 15, 17.5, and 20



Exp. 6: 200 servers, SCoV of interarrival times > 1 and SCoV of service times > 1

—Time-varying rate: $180 + 120\sin(2\pi t/10)$

—Base interarrival time distribution: $\text{Weibull}(1.1271, 2.5), \text{SCoV} = 0.1831$

—Service time distribution: $\text{Lognormal}(-0.5, 1)$, $\text{SCoV} = 1.7183$

Exp. 7: 50 servers, SCoV of interarrival times < 1 and SCoV of service times < 1

—Time-varying rate: $45 + 30\sin(2\pi t/10)$

—Base interarrival time distribution: $\text{Weibull}(1.1271, 2.5), \text{SCoV} = 0.1831$

—Service time distribution: $\text{Lognormal}(-0.2027, 0.6368), \text{SCoV} = 0.5$

Exp. 8: 200 servers, SCoV of interarrival times > 1 and SCoV of service times > 1

—Time-varying rate: $180 + 120\sin(2\pi t/10)$

—Base interarrival time distribution: $\text{Weibull}(1.1271, 2.5), \text{SCoV} = 0.1831$

—Service time distribution: $\text{Lognormal}(-0.2027, 0.6368), \text{SCoV} = 0.5$.

We mention that the queue length distributions are approximately Gaussian in Section 4. Figure 5 compares the empirical density and the density from the

**Figure 6.** Comparison Between Exp. 1 and Exp. 2

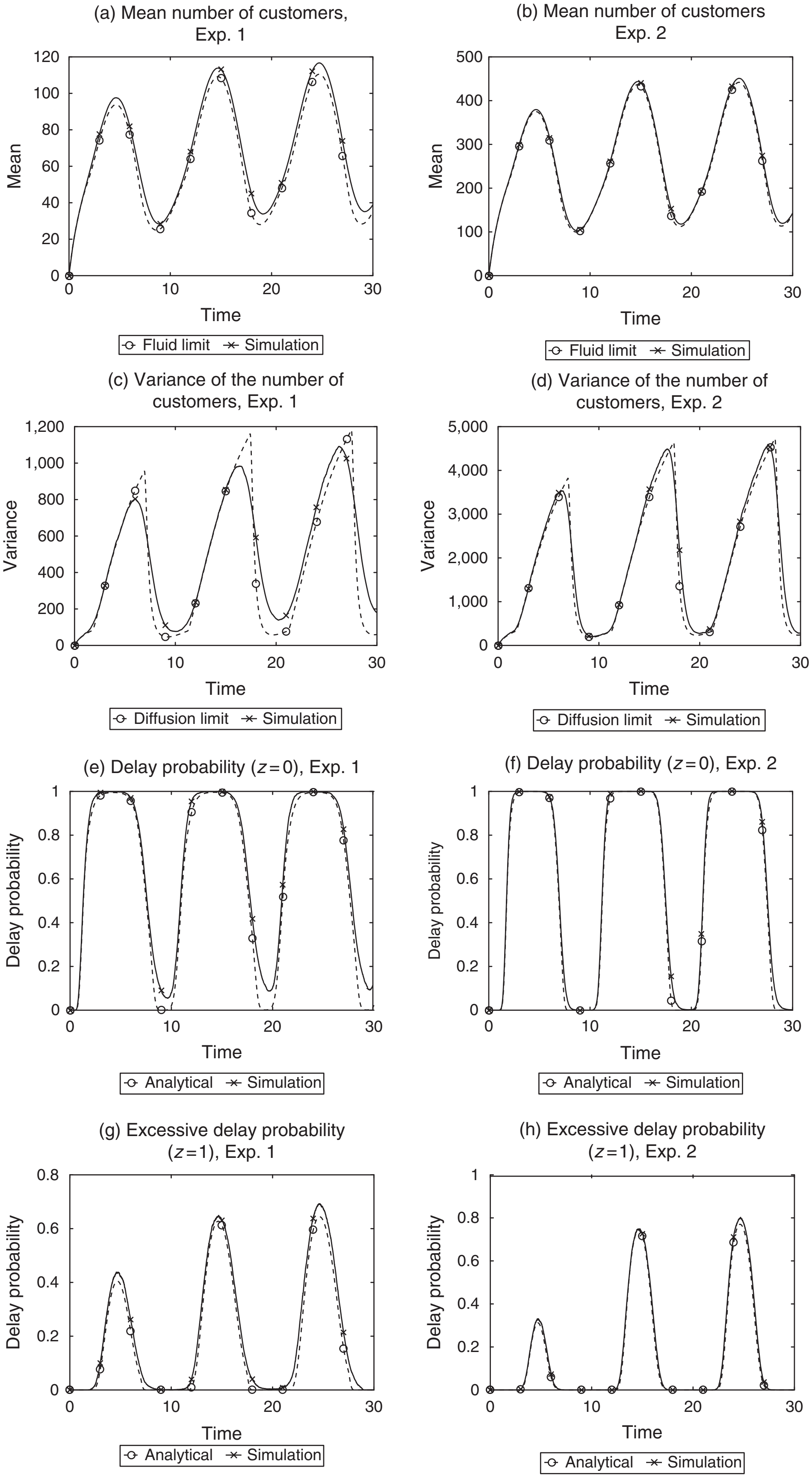
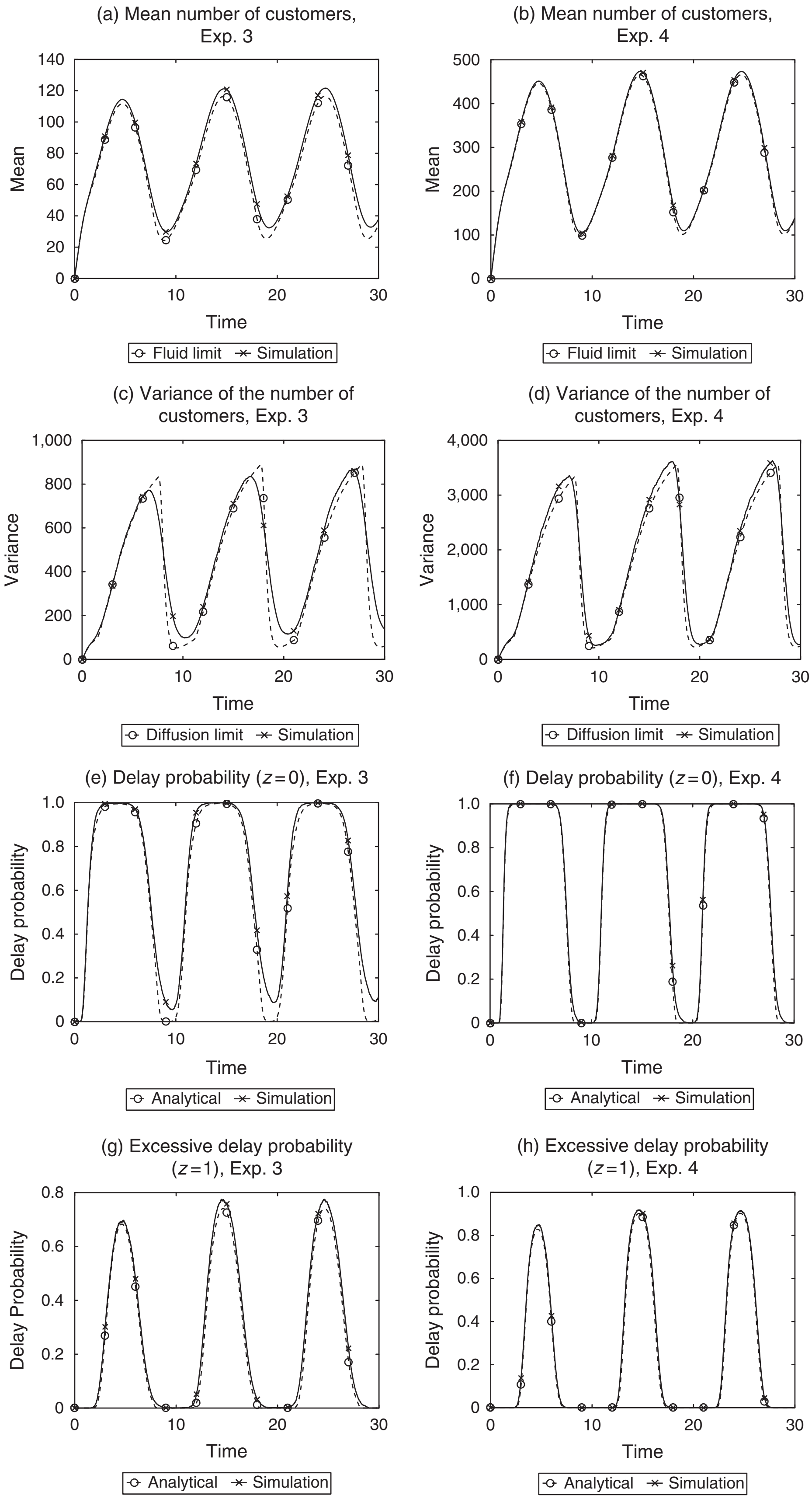


**Figure 7.** Comparison Between Exp. 3 and Exp. 4


(a) Mean number of customers, Exp. 3

(b) Mean number of customers, Exp. 4

(c) Variance of the number of customers, Exp. 3

(d) Variance of the number of customers, Exp. 4

(e) Delay probability ($z = 0$), Exp. 3

(f) Delay probability ($z = 0$), Exp. 4

(g) Excessive delay probability ($z = 1$), Exp. 3

(h) Excessive delay probability ($z = 1$), Exp. 4

**Figure 8.** Comparison Between Exp. 5 and Exp. 6

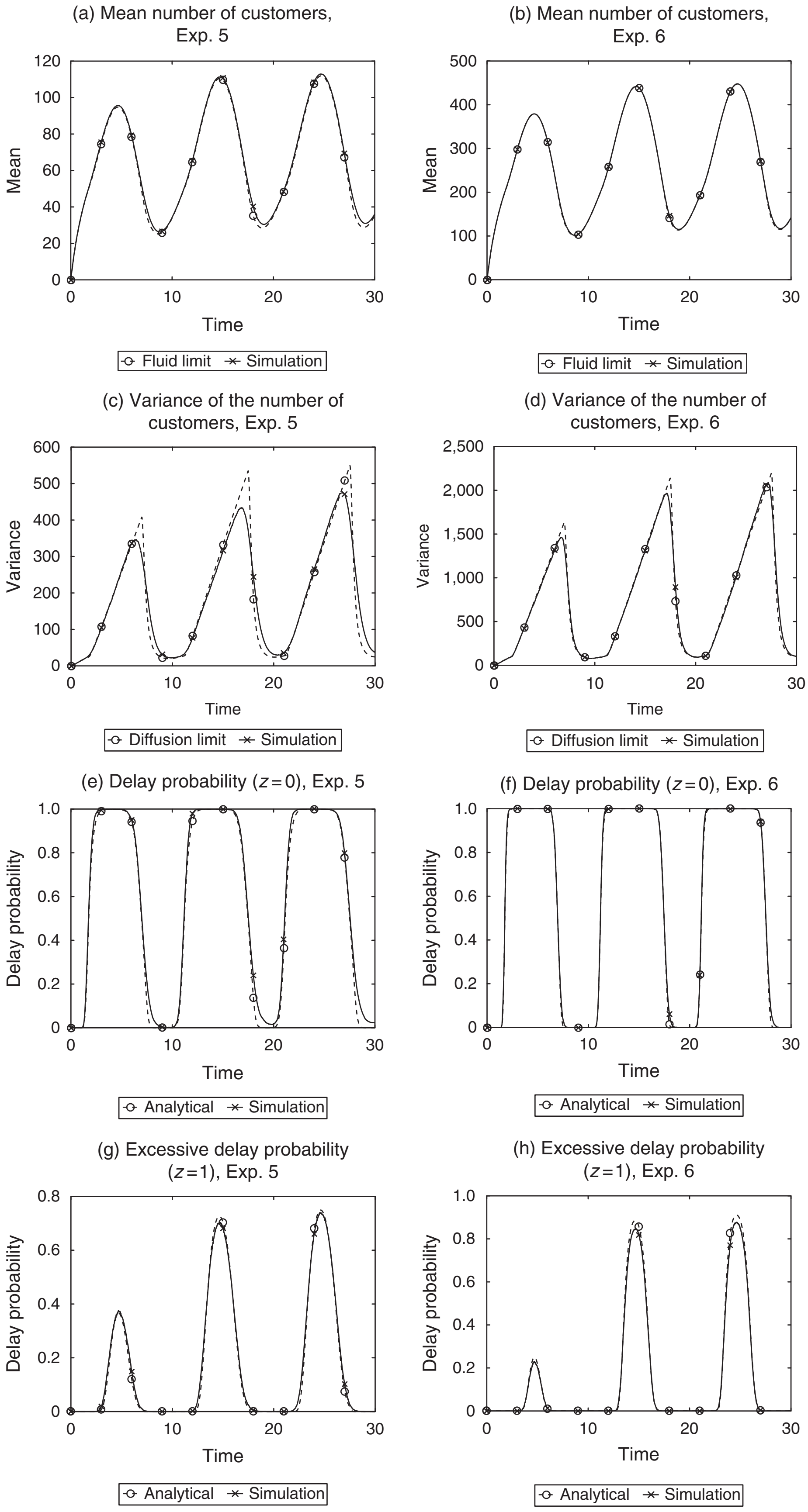
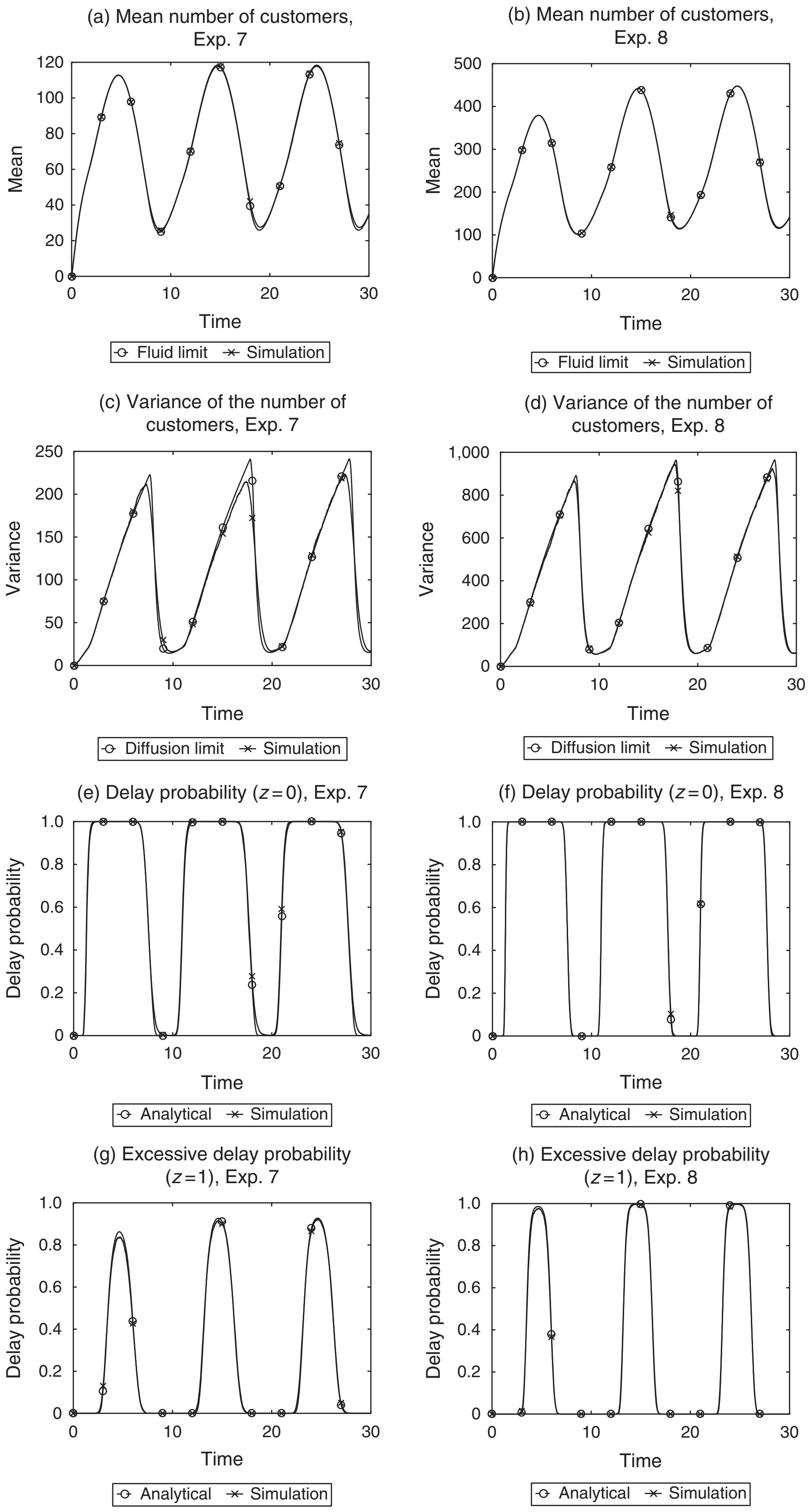


**Figure 9.** Comparison Between Exp. 7 and Exp. 8



diffusion limit at several time points (underloaded times 5 and 10, critically loaded times 7.5 and 17.5, and overloaded times 5 and 15). Although we observe some skewness in the empirical density, the Gaussian approximation seems to work well.

Figures 6–9 plot the mean and the variance of the number of customers and the probability of delay over time comparing the proposed method and the simulation results for the cases of 50 and 200 servers. Each figure represents a different combination of distributions for arrival processes and service times. Overall we observe that the proposed method provides accurate estimations of the mean and the variance of the number of customers and the probability of delay. Comparing Figures 6(a) and 6(b), we observe that increasing the number of servers results in more accurate estimations of the mean as expected. We observe the same result for the variance (Figures 6(c) and 6(d)) and the probability of (excessive) delay (Figures 6(e)–6(h)). The same results hold across different distribution settings (Figures 7–9). The distributions in Figure 6 have the largest SCoV values and those in Figure 9 have the smallest SCoV values. In Figures 6 and 9, we observe that the proposed method works better when the SCoV values are small.

## 7. Conclusion

This paper describes a new methodology to approximate the queue length distributions of large-scale $G_t/G_t/n_t$ queues. Instead of analyzing a $G_t/G_t/n_t$ directly, we study a $Ph_t/Ph_t/n_t$ queue since phase-type distributions can approximate positive-valued distributions in any level of accuracy. Applying the uniform acceleration and strong approximations to $Ph_t/Ph_t/n_t$ queues to obtain fluid and diffusion limits, we encounter the lingering problem in our formulation and cannot obtain the diffusion limit. To resolve the issue, we propose a new formulation with an additional condition that is not quite restrictive. The new formulation works well and we successfully derive the fluid and diffusion limits. We find that the queue length process is approximately a Gaussian process and we derive ordinary differential equations to obtain the mean and variance of the queue length over time.

From the numerical study, we observe that the proposed method works better when the distributions for arrival processes and service times have smaller SCoVs. Because the uniform acceleration method increases the number of servers to infinity, the estimations should become more accurate as the number of servers increases. We exactly observe this phenomenon as expected.

We suggest two directions for future research. For example, to obtain the diffusion limit, we put an additional condition (a unique initial state for phase-type distributions). Although it does not seem to be critical, the method will be improved if the restriction can be removed. Extending the proposed method to multidimensional queueing networks is another possible research direction that we plan to pursue in a follow-up paper.

## References

Arfeen MA, Pawlikowski K, McNickle D, Willig A (2013) The role of the Weibull distribution in Internet traffic modeling. *Proc. 2013 25th Internat. Teletraffic Congress (ITC)* (IEEE, Piscataway, NJ), 1–8.

Arnold L (1992) *Stochastic Differential Equations: Theory and Applications* (Krieger Publishing Company, Malabar, FL).

Asmussen SR, Nerman O, Olsson M (1996) Fitting phase-type distributions via the EM algorithm. *Scandinavian J. Statist.* 23(4): 419–441.

Barbour AD (1976) Networks of queues and the method of stages. *Adv. Appl. Probab.* 8(3):584–591.

Billingsley P (1999) *Convergence of Probability Measures* (John Wiley & Sons, Hoboken, NJ).

Bobbio A, Horváth A, Telek M (2005) Matching three moments with minimal acyclic phase type distributions. *Stochastic Models* 21(2–3):303–326.

Botta RF, Harris CM (1986) Approximation with generalized hyperexponential distributions: Weak convergence results. *Queueing Systems* 1(2):169–190.

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center. *J. Amer. Statist. Assoc.* 100(469):36–50.

Creemers S, Defraeye M, Van Nieuwenhuyse I (2014) G-RAND: A phase-type approximation for the nonstationary $G(t)/G(t)/s(t)+G(t)$ queue. *Performance Evaluation* 80:102–123.

Dai J, He S, Tezcan T, et al. (2010) Many-server diffusion limits for $G/Ph/n+GI$ queues. *Ann. Appl. Probab.* 20(5):1854–1890.

Eick SG, Massey WA, Whitt W (1993) The physics of the $M_t/G/\infty$ queue. *Oper. Res.* 41(4):731–742.

Engblom S, Pender J (2014) Approximations for the moments of nonstationary and state dependent birth-death queues. Working paper, arXiv:1406.6164.

Feldmann A, Whitt W (1998) Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation* 31(3–4):245–279.

Glynn PW (1982) On the Markov property of the $GI/G/\infty$ Gaussian limit. *Adv. Appl. Probab.* 14(1):191–194.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.

Hampshire RC, Massey WA (2005) Variational optimization for call center staffing. *Proc. 2005 Conf. Diversity Comput.* (IEEE, Piscataway, NJ), 4–6.

Hampshire RC, Massey WA (2010) Dynamic optimization with applications to dynamic rate queues. Gray P, ed. *2010 Tutorials on Operations Research: Risk and Optimization in an Uncertain World*, Chap. 10 (INFORMS, Hanover, MD), 208–247.

Hampshire RC, Harchol-Balter M, Massey WA (2006) Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates. *Queueing Systems* 53(1–2):19–30.

Hampshire RC, Jennings OB, Massey WA (2009a) A time-varying call center design via lagrangian mechanics. *Probab. Engrg. Informational Sci.* 23(2):231–259.

Hampshire RC, Massey WA, Wang Q (2009b) Dynamic pricing to control loss systems with quality of service targets. *Probab. Engrg. Informational Sci.* 23(2):357–383.

Johnson MA, Taaffe MR (1991) An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queueing Systems* 8(1):129–147.

Ko YM, Gautam N (2013) Critically loaded time-varying multiserver queues: Computational challenges and approximations. *INFORMS J. Comput.* 25(2):285–301.

Kurtz T (1978) Strong approximation theorems for density dependent Markov chains. *Stochastic Processes Their Appl.* 6(3):223–240.

Liu Y, Whitt W (2012) The $G_t/GI/s_t+GI$ many-server fluid queue. *Queueing Systems* 71(4):405–444.

Liu Y, Whitt W (2014a) Algorithms for time-varying networks of many-server fluid queues. *INFORMS J. Comput.* 26(1):59–73.

Liu Y, Whitt W (2014b) Many-server heavy-traffic limit for queues with time-varying parameters. *Ann. Appl. Probab.* 24(1): 378–421.

Mandelbaum A, Massey W, Reiman M (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1): 149–201.

Mandelbaum A, Massey WA, Reiman MI, Stolyar A (2002) Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecomm. Systems* 21(2–4):149–171.

Massey WA, Pender J (2011) Poster: Skewness variance approximation for dynamic rate multiserver queues with abandonment. *ACM SIGMETRICS Performance Evaluation Rev.* 39(2):74–74.

Massey WA, Pender J (2013) Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* 75(2–4):243–277.

Nelson BL, Taaffe MR (2004a) The $Ph_t/Ph_t/\infty$ queueing system: Part I—The single node. *INFORMS J. Comput.* 16(3):266–274.

Nelson BL, Taaffe MR (2004b) The $[Ph_t/Ph_t/\infty]^K$ queueing system: Part II—The multiclass network. *INFORMS J. Comput.* 16(3): 275–283.

Neuts MF (1981) *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach* (Dover Publication, Inc., Mineola, NY).

Niyirora J, Pender J (2017) Optimal staffing of service systems with constraints. *Naval Res. Logist.* Forthcoming.

Osogami T, Harchol-Balter M (2006) Closed form solutions for mapping general distributions to quasi-minimal PH distributions. *Performance Evaluation* 63(6):524–552.

Ou J, Li J, Özekici S (1997) Approximating a cumulative distribution function by generalized hyperexponential distributions. *Probab. Engrg. Informational Sci.* 11(1):11–18.

Pang G, Whitt W (2009) Heavy-traffic limits for many-server queues with service interruptions. *Queueing Systems* 61(2–3):167–202.

Pender J (2014) Gram charlier expansion for time varying multiserver queues with abandonment. *SIAM J. Appl. Math.* 74(4):1238–1265.

Pender J (2015a) An analysis of nonstationary coupled queues. *Telecomm. Systems* 61(4):1–16.

Pender J (2015b) Nonstationary loss queues via cumulant moment approximations. *Probab. Engrg. Informational Sci.* 29(1):27–49.

Pender J (2015c) The truncated normal distribution: Applications to queues with impatient customers. *Oper. Res. Lett.* 43(1):40–45.

Puhalskii AA (2013) On the $M_t/M_t/K_t+M_t$ queue in heavy traffic. *Math. Methods Oper. Res.* 78(1):119–148.

Puhalskii AA, Reiman MI (2000) The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* 32(2):564–595.

Reed J (2009) The $G/GI/N$ queue in the Halfin–Whitt regime. *Ann. Appl. Probab.* 19(6):2211–2269.

Sasaki Y, Imai H, Tsunoyama M, Ishii I (2004) Approximation of probability distribution functions by coxian distribution to evaluate multimedia systems. *Systems Comput. Japan* 35(2):16–24.

Whitt W (1982) On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Probab.* 14(1):171–190.

Whitt W (2002) *Stochastic Process Limits*, 1st ed. (Springer-Verlag, New York).

Whitt W (2006) Fluid models for multiserver queues with abandonments. *Oper. Res.* 54(1):37–54.

Yu K, Huang M-L, Brill PH (2012) An algorithm for fitting heavy-tailed distributions via generalized hyperexponentials. *INFORMS J. Comput.* 24(1):42–52.