An approximation algorithm for the queue length distributions of time-varying many-server queues

Young Myoung Ko[1]

[1]Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Gyeongbuk, 790-784 South Korea

Abstract

This paper proposes an approximation algorithm for estimating 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), with large n_t values. The algorithm uses phase-type distributions to approximate inter-arrival/service times and apply fluid and diffusion approximations developed for Markovian systems. We develop an alternative model in order to bypass the lingering problem in the diffusion model. Numerical experiments demonstrate the effectiveness of the proposed method.

1 Introduction

Real-world applications of large-scale queueing systems such as data centers and call centers show time-varying behavior, and their arrival/service processes are not Markovian in general (Brown et al. [7], Arfeen et al. [1], Nelson and Taaffe [23]). Many of the recent studies on large-scale non-Markovian queues rely on the asymptotic approach utilizing fluid and diffusion limits as described in Billingsley [4] and Whitt [30]. Research on non-Markovian systems has progressed to the point of analyzing underloaded systems (a.k.a. the offered-load model, infinite-server queues) due to their analytical or numerical tractability (Whitt [29], Glynn [11], Eick et al. [9], Nelson and Taaffe [23, 22]). 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 [12], Puhalskii and Reiman [28], Pang and Whitt [27], Whitt [31], Liu and Whitt [16, 18, 17]).

This paper uses the *uniform acceleration* method which is coupled with the strong approximations theory and accelerates parameters while keeping the traffic intensity constant (Kurtz [15], Mandelbaum et al. [19, 20]). Kurtz [15] establishes strong approximation theorems for state-dependent continuous time Markov chains (CTMCs) having differentiable rate functions. Extending Kurtz [15], Mandelbaum et al. [19] 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. [20] show that the result in Kurtz [15] can be directly applied when the fluid limit stays at the non-differentiable points of rate functions for a measure-zero amount of time. Ko and Gautam [14] propose a Gaussian-based approximation method that achieves better approximation quality when the fluid limit lingers around the non-differentiable points. Massey and Pender [21] improve the result of Ko and Gautam [14] by introducing a new Gaussian skewness approximation. Liu and Whitt [16] propose a fluid limit for $G_t/GI/s_t + GI$ queues and extend the work of Mandelbaum et al. [19] in the sense that they consider non-Markovian inter-arrival, service and abandonment times. In a follow-up paper, Liu and Whitt [17] assume a Gaussian process as a limit process and provide a heavy-traffic diffusion limit for $G_t/M/s_t + GI$ queues. As shown in Mandelbaum et al. [20], Ko and Gautam [14], Liu and Whitt [17], 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.

Using phase-type distributions for approximating general distributions in queueing analysis is not new. The matrix-geometric method (MGM) described in Neuts [24] 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 [23] develop a method based on the partial-moment differential equations (PMDEs) 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 inter-arrival 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. [8] devise an accurate phase-type approximation algorithm for small-tomedium-sized $G_t/G_t/s_t + G_t$ queues using two-moment matching procedures.

The contributions of this study can be summarized with the following points. First, we derive fluid and diffusion limits for a $Ph_t/Ph_t/n_t$ queue (an approximation of a $G_t/G_t/n_t$ queue) using the *uniform acceleration* technique. When we 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 the non-differentiable points during some intervals having positive measure. This prevents us from deriving the diffusion limit. We propose an alternative formulation that enables us to successfully obtain the diffusion limit. Second, we use phase-type distributions to approximate the general distributions themselves, which may require many phases for accurate approximation. The number of 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. The number of phases used for approximating inter-arrival 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) under a regular PC environment. The proposed method is scalable in terms of the number of servers and the number of phases compared with the results in Nelson and Taaffe [23] and Creemers et al. [8].

The remainder of this paper is organized as follows. Section 2 describes the $G_t/G_t/n_t$ queueing system and the problem settings. Section 3 builds a mathematical model for describing the dynamics of the system from the $Ph_t/Ph_t/n_t$ queue. We explain the lingering problem and introduce an alternative model for resolving it. Section 4 explains fluid and diffusion approximations. Section 5 discusses the numerical examples used to validate the effectiveness of our proposed approach. Section 6 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.. $\bar{x}(t) > n_t$) phases and hits the critically loaded phase (i.e. $\bar{x}(t) = n_t$) at only the countably many time points. 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 $0 \leq t \leq T$ and $T < \infty$.

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 (Whitt [29] and Asmussen et al. [3]). 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 $\mathbf{s} = -\mathbf{Se}$ 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 our scope, and we refer to the reader to [5, 13, 32, 6, 10, 26, 3, 25]. We explain the fitting algorithm we use in Section 5.

We assume phase-type distributions with 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:

$$\mathbf{S_A} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1m_A} \\ \vdots & \vdots & \vdots \\ \lambda_{m_A1} & \cdots & \lambda_{m_Am_A} \end{pmatrix}, \quad \mathbf{s_A} = (\lambda_{10}, \ldots, \lambda_{m_A0})' \tag{1}$$

$$\mathbf{S_S} = \begin{pmatrix} \mu_{11} & \cdots & \mu_{1m_S} \\ \vdots & \vdots & \vdots \\ \mu_{m_S1} & \cdots & \mu_{m_Sm_S} \end{pmatrix}, \quad \mathbf{s_S} = (\mu_{10}, \ldots, \mu_{m_S0})', \tag{2}$$

where λ_{jk}'s and μ_{il}'s accord 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 λ_{jk} and μ_{il} with $\lambda_{jk}(t)$ and

$\mu_{il}(t)$.

3 Mathematical model

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

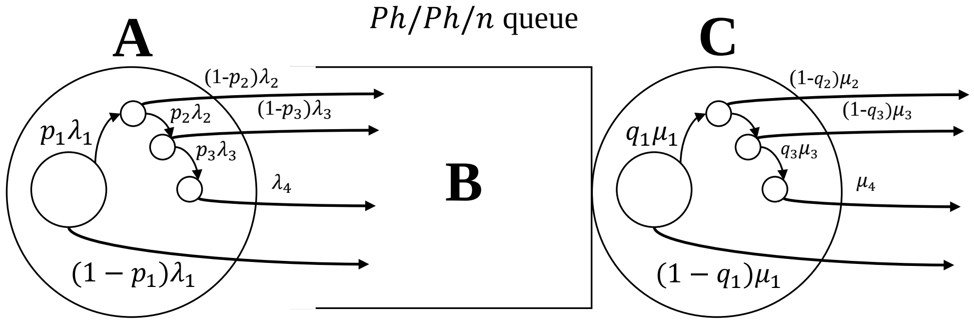


Figure 1: $Ph/Ph/n$ queue with Coxian distributions

Figure 1 illustrates an example of $Ph/Ph/n$ queue with Coxian inter-arrival and service times. In order 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 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, $X_i(t)$ be the number of customers being served in phase j of the service process, and $Z(t)$ be the number of customers in the system. Note that the number of customers in the waiting space is $Z(t) - \sum_{i=1}^{m_S} X_i(t)$. Then, the state of the system

$\mathbf{V}(t) = (U_1(t), \dots, U_{m_A}, X_1(t), \dots, X_{m_S}, Z(t))'$ is the solution to the following integral equations:

$$
\begin{aligned}
U_j(t) =& U_j(0) + \sum_{k=1,k\neq j}^{m_A} Y^A_{kj}\Big(\int_0^t \lambda_{kj} U_k(s)ds\Big) - \sum_{k=1,k\neq j}^{m_A} Y^A_{jk}\Big(\int_0^t \lambda_{jk} U_j(s)ds\Big) \\
&- \sum_{k=1,k\neq j}^{m_A} \sum_{l=1}^{m_S} Y^I_{jkl}\Big(\int_0^t \lambda_{j0}\alpha_k\beta_l U_j(s)\mathbf{1}_{\{Z(s)\le n\}} ds\Big) \\
&- \sum_{k=1,k\neq j}^{m_A} Y^Q_{jk}\Big(\int_0^t \lambda_{j0}\alpha_k U_j(s)\mathbf{1}_{\{Z(s)>n\}} ds\Big) \\
&+ \sum_{k=1,k\neq j}^{m_A} \sum_{l=1}^{m_S} Y^I_{kjl}\Big(\int_0^t \lambda_{k0}\alpha_j\beta_l U_k(s)\mathbf{1}_{\{Z(s)\le n\}} ds\Big) \\
&+ \sum_{k=1,k\neq j}^{m_A} Y^Q_{kj}\Big(\int_0^t \lambda_{k0}\alpha_j U_k(s)\mathbf{1}_{\{Z(s)>n\}} ds\Big) \text{ for } 1 \le j \le m_A,
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
X_i(t) =& \sum_{j=1}^{m_A}\sum_{k=1}^{m_A} Y^I_{jki}\Big(\int_0^t \lambda_{j0}\alpha_k\beta_i U_j(s)\mathbf{1}_{\{Z(s)\le n\}} ds\Big) + \sum_{l=1,l\neq i}^{m_S} Y^S_{li}\Big(\int_0^t \mu_{li} X_l(s)ds\Big) \\
&- \sum_{l=1,l\neq i}^{m_S} Y^S_{il}\Big(\int_0^t \mu_{il} X_i(s)ds\Big) - Y^D_{i0}\Big(\int_0^t \mu_{i0} X_i(s)\mathbf{1}_{\{Z(s)\le n\}} ds\Big) \\
&- \sum_{l=1,l\neq i}^{m_S} Y^D_{il}\Big(\int_0^t \mu_{i0} X_i(s)\mathbf{1}_{\{Z(s)>n\}}\beta_l ds\Big) \\
&+ \sum_{l=1,l\neq i}^{m_S} Y^D_{li}\Big(\int_0^t \mu_{l0} X_l(s)\mathbf{1}_{\{Z(s)>n\}}\beta_i ds\Big) \text{ for } 1 \le i \le m_S,
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
Z(t) =& \sum_{j=1}^{m_A}\sum_{k=1}^{m_A}\sum_{l=1}^{m_S} Y^I_{jkl}\Big(\int_0^t \lambda_{j0}\alpha_k\beta_l U_j(s)\mathbf{1}_{\{Z(s)\le n\}} ds\Big) \\
&+ \sum_{j=1}^{m_A}\sum_{k=1}^{m_A} Y^Q_{jk}\Big(\int_0^t \lambda_{j0}\alpha_k U_j(s)\mathbf{1}_{\{Z(s)>n\}} ds\Big) - \sum_{i=1}^{m_S} Y^D_{i0}\Big(\int_0^t \mu_{i0} X_i(s)\mathbf{1}_{\{Z(s)\le n\}} ds\Big) \\
&- \sum_{i=1}^{m_S}\sum_{l=1}^{m_S} Y^D_{il}\Big(\int_0^t \mu_{i0} X_i(s)\mathbf{1}_{\{Z(s)>n\}}\beta_l ds\Big).
\end{aligned} \tag{5}
$$

For notational convenience, the 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 λ_{jk}, μ_{il} and n with $\lambda_{jk}(t)$, $\mu_{il}(t)$, and $n(t)$ respectively under mild conditions in Mandelbaum et al. [19]. Poisson processes, $Y^A_{kj}(\cdot)$'s count the number of transitions from phase k to phase j of the arrival process. When the waiting space is empty ($Z(t) \le n$), Poisson processes, $Y^I_{jkl}(\cdot)$'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, $Y^Q_{jk}(\cdot)$'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, $Y^S_{li}(\cdot)$'s, count the internal transitions from phase l to phase j of the service process. When the waiting space is empty, Poisson processes, $Y^D_{i0}(\cdot)$'s, count the number of departures from phase i of the service process. When the waiting space is not empty, Poisson processes, $Y^D_{il}(\cdot)$'s, count the number of departures from phase i and a new customer enters phase l from the the waiting space. Note that the Poisson processes explained above 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 whether the time during which the fluid limit stays at the non-differentiable points has measure zero or not.

Let $\bar{\mathbf{v}}(t) = (\bar{u}_1(t), \ldots, \bar{u}_{m_A}(t), \bar{x}_1(t), \ldots, \bar{x}_{m_S}(t), \bar{z}(t))'$ be the fluid limit of $\mathbf{V}(t)$. We check the Poisson process, $Y^D_{il}(\cdot)$ in equation (4). The fluid limit for $Y^D_{il}(\cdot)$ 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 $\{t : \bar{z}(t) > n\}$ 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 $(\bar{x}_1(t), \ldots, \bar{x}_{m_S}(t))'$ stays at the non-differential point during the overloaded period and we cannot obtain the diffusion limit from the result of Kurtz [15] and Mandelbaum et al. [20]. 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. 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. [3], the Coxian distribution provides almost the same quality of fit as the general phase-type distribution with the same number of phases. The additional assumption, 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. We write the formulation as follows:

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

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

$$\begin{aligned} X_i(t) =& Y_{1i}^S\Big(\int_0^t \mu_{1i}\Big[\mathbf{1}_{\{\sum_{r=1}^{m_S} X_r(s)\leq n\}}X_1(s) + \mathbf{1}_{\{\sum_{r=1}^{m_S} X_r(s)>n\}}\Big(n - \sum_{r=2}^{m_S} X_r(s)\Big)^+\Big]ds\Big) \\ & + \sum_{l=2,l\neq i}^{m_S} Y_{li}^S\Big(\int_0^t \mu_{li}X_l(s)ds\Big) - \sum_{l=1,l\neq i}^{m_S} Y_{il}^S\Big(\int_0^t \mu_{il}X_i(s)ds\Big) - Y_i^D\Big(\int_0^t \mu_{i0}X_i(s)ds\Big) \\ & \text{for } 2 \leq i \leq m_S. \end{aligned} \tag{8}$$

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. Poisson processes, $Y_i^D(\cdot)$'s count departures from phase i of the service process. Note that the Poisson processes explained above 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 explain the fluid and diffusion approximations.

4 Fluid and diffusion approximations

First, we provide some definitions for notational convenience.

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

$\mathbf{v} = (u_1, \ldots, u_{m_A}, x_1, \ldots, x_{m_S})'$.

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

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

$\mathbf{d}_i^D$: $(m_A + m_S) \times 1$ vector, 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.

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

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

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

With the definitions above, we rewrite equations (6)-(8) in a vector form as follows:

$$\begin{aligned}\mathbf{V}(t) =&\mathbf{V}(0) + \sum_{j=1}^{m_A}\sum_{k=1,k\neq j}^{m_A} \mathbf{d}^A_{jk} Y^A_{jk}\Big(\int_0^t f^A_{jk}(s,\mathbf{V}(s))ds\Big) + \sum_{j=1}^{m_A}\sum_{k=1}^{m_A} \mathbf{d}^I_{jk} Y^I_{jk}\Big(\int_0^t f^I_{jk}(s,\mathbf{V}(s))ds\Big)\\ &+ \sum_{i=1}^{m_S}\sum_{l=1,l\neq i}^{m_S} \mathbf{d}^S_{il} Y^S_{il}\Big(\int_0^t f^S_{il}(s,\mathbf{V}(s))ds\Big) + \sum_{i=1}^{m_S} \mathbf{d}^D_{i} Y^D_{i}\Big(\int_0^t f^D_{i}(s,\mathbf{V}(s))ds\Big).\end{aligned}$$

Following the procedure of the uniform acceleration in Mandelbaum et al. [19] and Kurtz [15], we define a sequence of processes $\{\mathbf{V}^\eta(t), \eta \geq 1, t \geq 0\}$, where

$$\begin{aligned}\mathbf{V}^\eta(t) =&\mathbf{V}^\eta(0) + \sum_{j=1}^{m_A}\sum_{k=1,k\neq j}^{m_A} \mathbf{d}^A_{jk} Y^A_{jk}\Big(\eta\int_0^t f^A_{jk}(s,\mathbf{V}^\eta(s)/\eta)ds\Big)\\ &+ \sum_{j=1}^{m_A}\sum_{k=1}^{m_A} \mathbf{d}^I_{jk} Y^I_{jk}\Big(\eta\int_0^t f^I_{jk}(s,\mathbf{V}^\eta(s)/\eta)ds\Big) + \sum_{i=1}^{m_S}\sum_{l=1,l\neq i}^{m_S} \mathbf{d}^S_{il} Y^S_{il}\Big(\eta\int_0^t f^S_{il}(s,\mathbf{V}^\eta(s)/\eta)ds\Big)\\ &+ \sum_{i=1}^{m_S} \mathbf{d}^D_{i} Y^D_{i}\Big(\eta\int_0^t f^D_{i}(s,\mathbf{V}^\eta(s)/\eta)ds\Big).\end{aligned}$$

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

Proposition 1 (Fluid limit, Mandelbaum et al. [19], Kurtz [15])**.** *Suppose* $\mathbf{V}^\eta(0) = \mathbf{V}(0)$*. Then,*

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

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

$$\begin{aligned}\frac{d}{dt}\bar{\mathbf{v}}(t) =&\sum_{j=1}^{m_A}\sum_{k=1,k\neq j}^{m_A} \mathbf{d}^A_{jk} f^A_{jk}(t,\bar{\mathbf{v}}(t)) + \sum_{j=1}^{m_A}\sum_{k=1,k\neq j}^{m_A} \mathbf{d}^I_{jk} f^I_{jk}(t,\bar{\mathbf{v}}(t))\\ &+ \sum_{i=1}^{m_S}\sum_{l=1,l\neq i}^{m_S} \mathbf{d}^S_{il} f^S_{il}(t,\bar{\mathbf{v}}(t)) + \sum_{i=1}^{m_S} \mathbf{d}^D_{i} f^D_{i}(t,\bar{\mathbf{v}}(t)).\end{aligned} \tag{9}$$

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

Proposition 2 (Diffusion limit, Mandelbaum et al. [19] and Kurtz [15])**.** *Let* $\mathbf{D}^\eta(t) = \sqrt{\eta}(\mathbf{V}(t)/\eta - \bar{\mathbf{v}}(t))$*. Then,*

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

where $\mathbf{D}(t)$ *is the solution to*

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

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,* $\{\mathbf{D}(t), t \geq 0\}$ *is a Gaussian process (Arnold [2]).*

Therefore, for a large η,

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

Note that increasing η indeed means increasing the number of servers along with other parameters (Mandelbaum et al. [20]). 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:

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

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 3 (Mean and covariance matrix of $\mathbf{D}(t)$, Arnold [2])**.** *Let* $\mathbf{M}(t) = E[\mathbf{D}(t)]$ *and* $\mathbf{\Sigma}(t) = Cov[\mathbf{D}(t), \mathbf{D}(t)]$*. Then,* $\mathbf{M}(t)$ *and* $\mathbf{\Sigma}(t)$ *are the unique solution to the following ordinary equations:*

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

$$\frac{d}{dt}\mathbf{\Sigma}(t) = \partial\mathbf{F}(t, \bar{\mathbf{v}}(t))\mathbf{\Sigma}(t) + \mathbf{\Sigma}(t)\partial\mathbf{F}(t, \bar{\mathbf{v}}(t))' + \mathbf{G}(t, \bar{\mathbf{v}}(t)). \tag{11}$$

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

Recall that we start with the empty queue, which means 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 E$[\mathbf{V}(t)]$ and Cov$[\mathbf{V}(t), \mathbf{V}(t)]$ as follows:

$$\begin{aligned}\mathrm{E}[\mathbf{V}(t)] &\approx \bar{\mathbf{v}}(t),\\ \mathrm{Cov}[\mathbf{V}(t), \mathbf{V}(t)] &\approx \mathbf{\Sigma}(t).\end{aligned}$$

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 a Gaussian process and we can obtain the mean and variance of $X(t)$ as follows:

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

5 Numerical results

In this section, we provide some numerical results comparing the proposed method with the simulation results. Figure 2 shows the overall flow of the numerical study and Table 1 explains the settings of the experiments.

Referring to the flow chart in Figure 2, we choose Coxian distributions to approximate Weibull and lognormal distributions in Table 1. 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. [3]). We use the EM algorithm developed by Asmussen et al. [3], 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 3 illustrates a density and distribution fitting with a Coxian distribution. In this example, we use 10 phases to approximate the Weibull distribution. We derive the ordinary differential equations (ODEs) from equations (9) and (11), and solve them using MATLAB. We write the simulation code in C++. In order to generate a general time-varying arrival process, we implement the algorithm based on the standard equilibrium

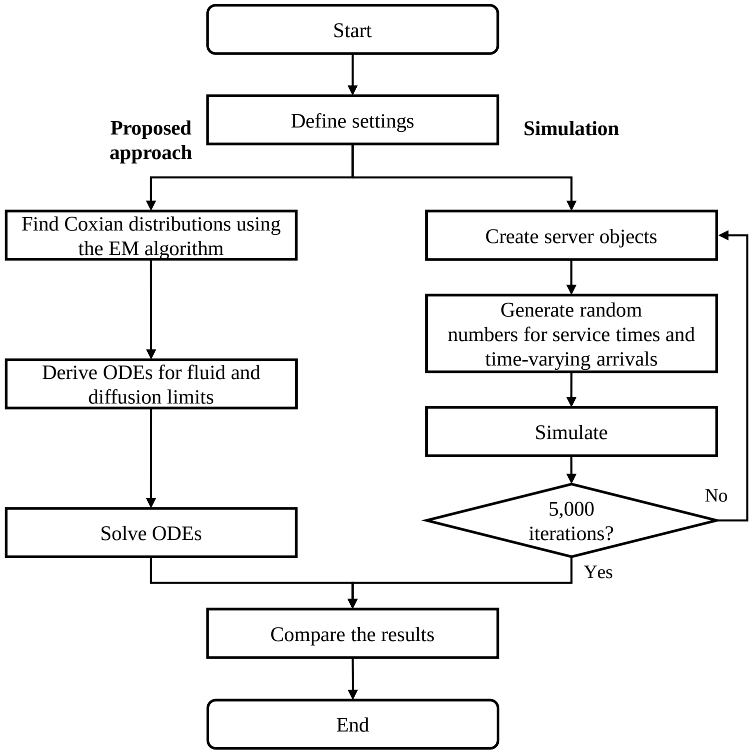

Start
Define settings
Proposed approach
Simulation
Find Coxian distributions using the EM algorithm
Derive ODEs for fluid and diffusion limits
Solve ODEs
Create server objects
Generate random numbers for service times and time-varying arrivals
Simulate
5,000 iterations?
No
Yes
Compare the results
End

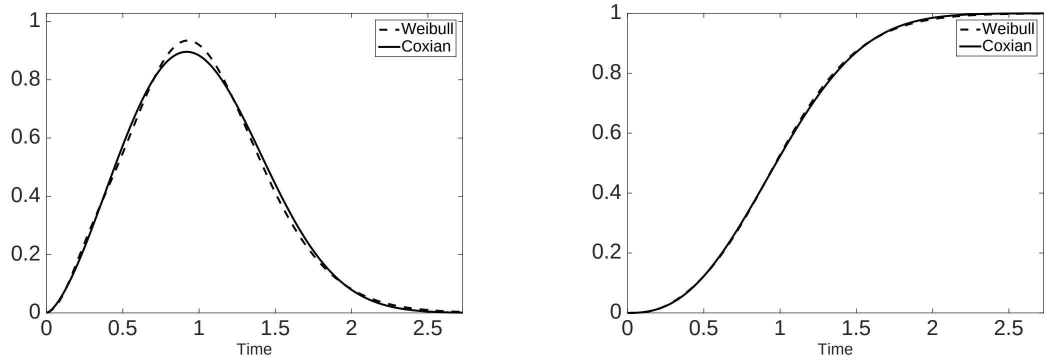
Figure 2: Overall flow of the numerical study

renewal process (SERP) explained in the longer version of Liu and Whitt [16]. We use Weibull distributions with mean 1 (see Table 1) 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 over time.

As in Table 1, we choose two Weibull distributions 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. We do not consider the case when the SCoV is

Table 1: Settings for the numerical study

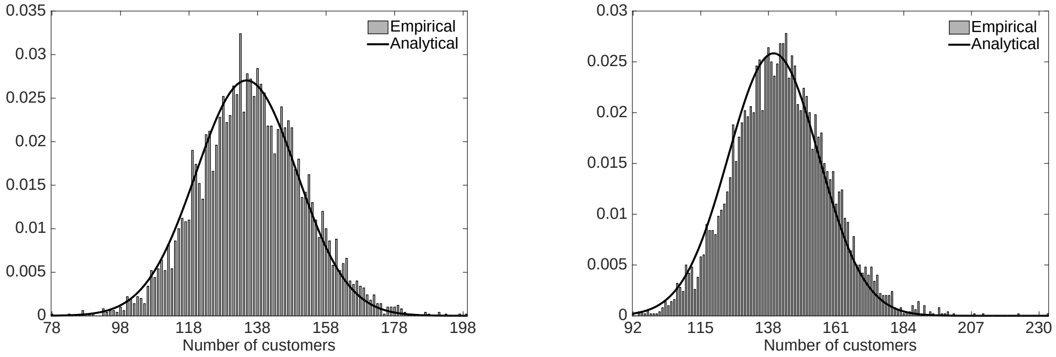
Setting	Description
# of servers	50/200
# of iterations	5,000
Time-varying rate	$45 + 30\sin(2\pi t/10)/180 + 120\sin(2\pi t/10)$
Inter-arrival time	$\text{Weibull}(0.79, 0.7), \text{SCoV} = 2.1387/\text{Weibull}(1.1271, 2.5), \text{SCoV} = 0.1831$
Service time	$\text{Lognormal}(-0.5, 1), \text{SCoV} = 1.7183/\text{Lognormal}(-0.2027, 0.6368), \text{SCoV} = 0.5$



(a) Density function fitting

(b) Distribution function fitting

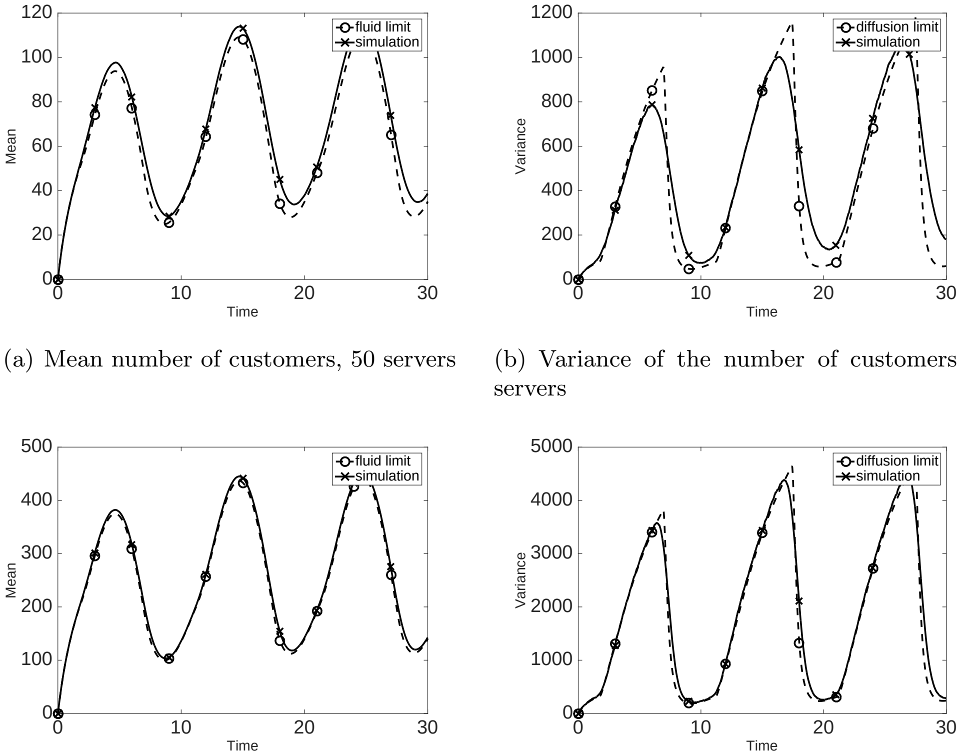
Figure 3: Weibull(1.1271, 2.5) and corresponding Coxian distributions



(a) Density at $t = 10$

(b) Density at $t = 20$

Figure 4: Density of the number of customers at time 10 and 20



(a) Mean number of customers, 50 servers

(b) Variance of the number of customers, 50 servers

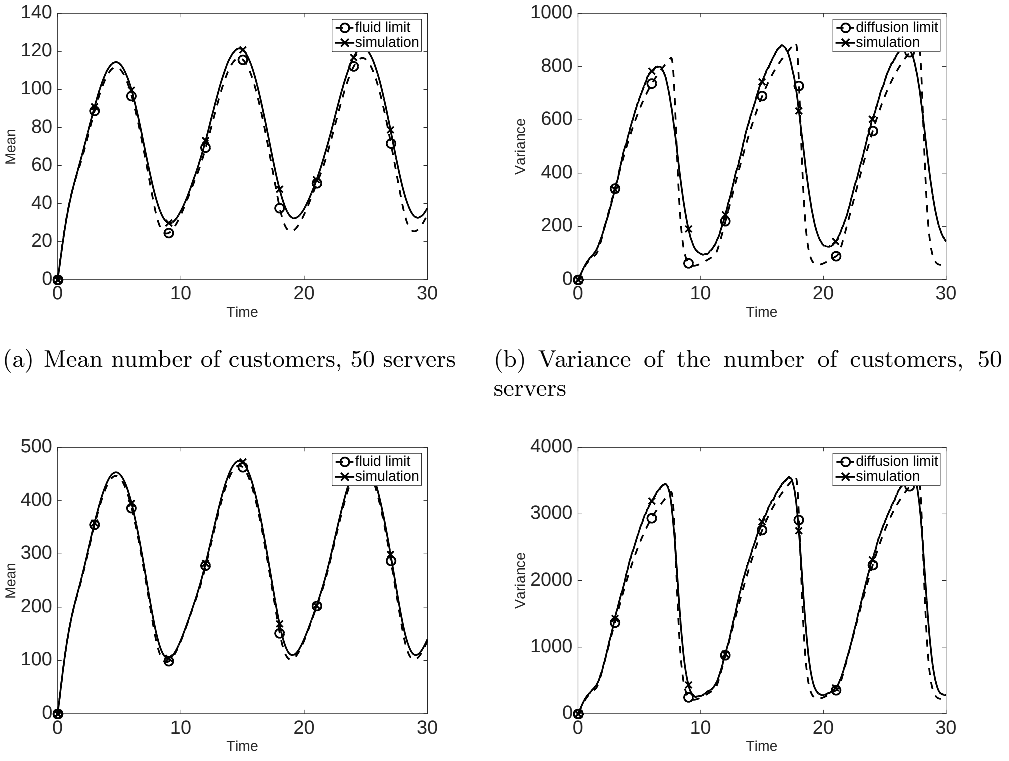
(c) Mean number of customers, 200 servers

(d) Variance of the number of customers, 200 servers

Figure 5: Weibull(0.79,0.7)(SCoV=2.1387) and Lognormal(-0.5,1)(SCoV=1.7183)

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. 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 8 combinations of experimental settings: two distributions for arrivals, two distributions for services, two values of the number of servers.

We mention that the queue length distributions are approximately Gaussian in Section 4. Figure 4 compares the empirical density and the density from the diffusion limit at time 10 and 20. Although we observe some skewness in the empirical density, the Gaussian approximation seems to work well. Figures 5-8 plot the mean and the variance of the number of customers over time comparing the proposed method and the simulation results for the cases of 50 and 200 servers. Each figure



(a) Mean number of customers, 50 servers

(b) Variance of the number of customers, 50 servers

(c) Mean number of customers, 200 servers

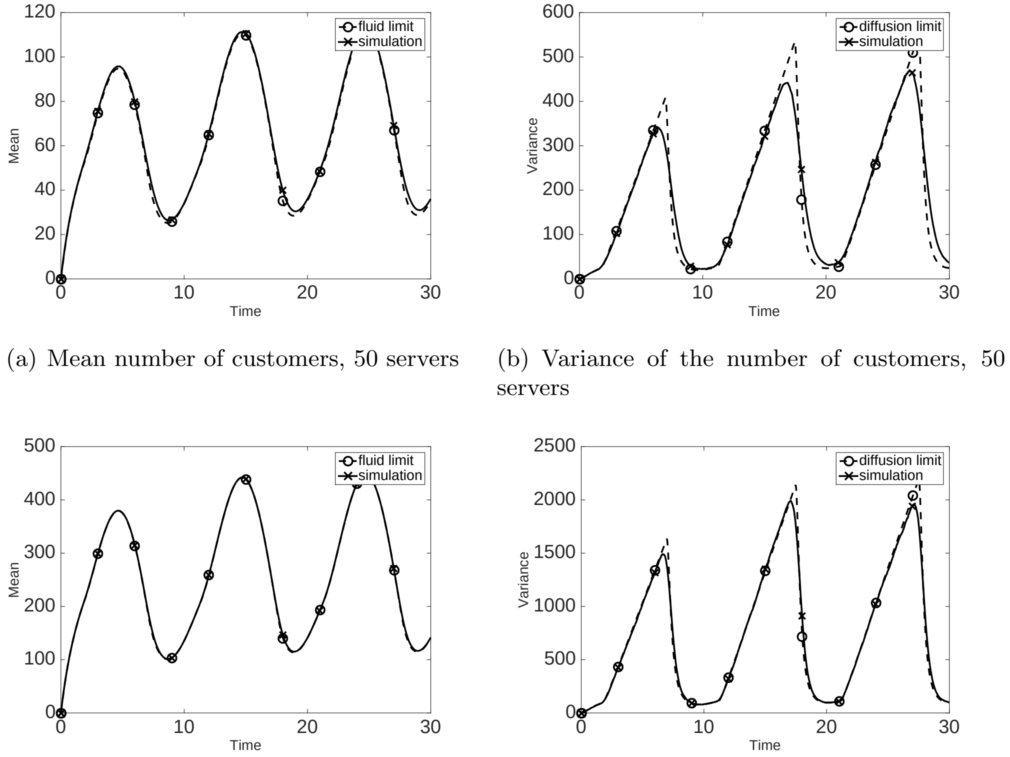
(d) Variance of the number of customers, 200 servers

Figure 6: Weibull(0.79,0.7)(SCoV=2.1387) and Lognormal(-0.2027,0.6368)(SCoV=0.5)

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 in the system. Comparing Figures 5 (a) and (c), 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 (see Figures 5 (b) and (d)). The same results hold across different distribution settings (Figures 6-8). The distributions in Figure 5 have the largest SCoV values and those in Figure 8 have the smallest SCoV values. In Figures 5 and 8, we observe that the proposed method works better when the SCoV values are small.

6 Conclusion

This paper describes a computational method 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 $Ph_t/Ph_t/n_t$



(a) Mean number of customers, 50 servers

(b) Variance of the number of customers, 50 servers

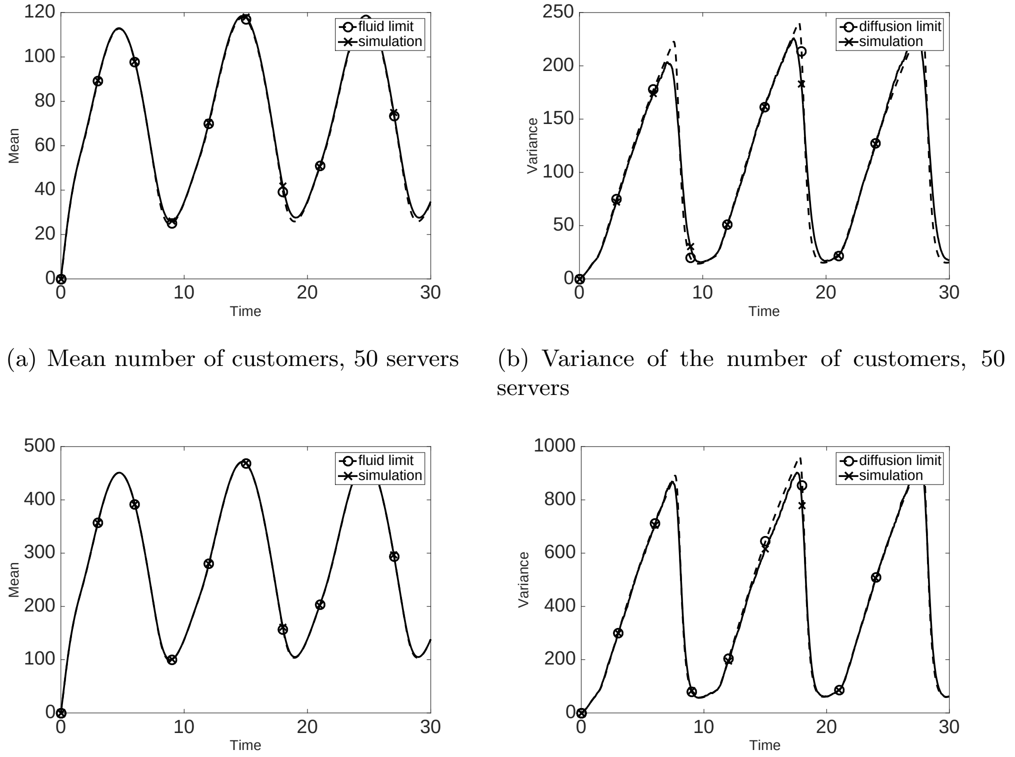
(c) Mean number of customers, 200 servers

(d) Variance of the number of customers, 200 servers

Figure 7: Weibull(1.1271,2.5)(SCoV=0.1831) and Lognormal(-0.5,1)(SCoV=1.7183)

queue since phase-type distributions can approximate positive-valued distributions in any level of accuracy. Applying the uniform acceleration method 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. Since 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.



(a) Mean number of customers, 50 servers

(b) Variance of the number of customers, 50 servers

(c) Mean number of customers, 200 servers

(d) Variance of the number of customers, 200 servers

Figure 8: Weibull(1.1271,2.5)(SCoV=0.1831) and Lognormal(-0.2027,0.6368)(SCoV=0.5)

We suggest two directions for future research. For example, in order 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 queueing networks is another possible research direction.

References

[1] Muhammad Asad Arfeen, K. Pawlikowski, D. McNickle, and A. Willig. The role of the Weibull distribution in Internet traffic modeling. In *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*, pages 1–8. IEEE, September 2013.

[2] Ludwig Arnold. *Stochastic Differential Equations: Theory and Applications.* Krieger Publishing Company, 1992.

[3] Sø ren Asmussen, Olle Nerman, and Marita Olsson. Fitting Phase-Type Distributions via the EM Algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441, 1996.

[4] Patrick Billingsley. *Convergence of Probability Measures.* A John Wiley & Sons, Inc., Publication, 1999.

[5] A. Bobbio, A. Horváth, and M. Telek. Matching Three Moments with Minimal Acyclic Phase Type Distributions. *Stochastic Models*, 21(2-3):303–326, 2005.

[6] R. F. Botta and C. M. Harris. Approximation with generalized hyperexponential distributions: Weak convergence results. *Queueing Systems*, 1(2):169–190, September 1986.

[7] Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. Statistical Analysis of a Telephone Call Center. *Journal of the American Statistical Association*, 100(469):36–50, March 2005.

[8] Stefan Creemers, Mieke Defraeye, and Inneke Van Nieuwenhuyse. G-RAND: A phase-type approximation for the nonstationary $G(t)/G(t)/s(t) + G(t)$ queue. *Performance Evaluation*, 80:102–123, August 2014.

[9] S. G. Eick, W. A. Massey, and W. Whitt. The Physics of the $M_t/G/\infty$ Queue. *Operations Research*, 41(4):731–742, July 1993.

[10] Anja Feldmann and Ward Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3-4):245–279, 1998.

[11] Peter W. Glynn. On the Markov Property of the $GI/G/\infty$ Gaussian Limit. *Advances in Applied Probability*, 14(1):191–194, 1982.

[12] Shlomo Halfin and Ward Whitt. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, 29(3):567–588, May 1981.

[13] Mary A. Johnson and Michael R. Taaffe. An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queueing Systems*, 8(1):129–147, December 1991.

[14] Young Myoung Ko and Natarajan Gautam. Critically Loaded Time-Varying Multiserver Queues: Computational Challenges and Approximations. *INFORMS Journal on Computing*, 25(2):285–301, May 2013.

[15] T Kurtz. Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and their Applications*, 6(3):223–240, February 1978.

[16] Yunan Liu and Ward Whitt. The $G_t/GI/s_t+GI$ many-server fluid queue. *Queueing Systems*, 71(4):405–444, March 2012.

[17] Yunan Liu and Ward Whitt. Many-server heavy-traffic limit for queues with time-varying parameters. *The Annals of Applied Probability*, 24(1):378–421, February 2014.

[18] Yunan Liu and Ward Whitt. Algorithms for Time-Varying Networks of Many-Server Fluid Queues. *INFORMS Journal on Computing*, 26(1):59–73, February 2014.

[19] Avi Mandelbaum, William Massey, and Martin Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30(1):149–201, 1998.

[20] Avi Mandelbaum, William A. Massey, Martin I. Reiman, and Alexander Stolyar. Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials. *Telecommunication Systems*, 21(2-4):149–171, 2002.

[21] William A. Massey and Jamol Pender. Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75(2-4):243–277, February 2013.

[22] Barry L. Nelson and Michael R. Taaffe. The $[Ph_t/Ph_t/\infty]^K$ Queueing System: Part II-The Multiclass Network. *INFORMS Journal on Computing*, 16(3):275–283, August 2004.

[23] Barry L. Nelson and Michael R. Taaffe. The $Ph_t/Ph_t/\infty$ Queueing System: Part I-The Single Node. *INFORMS Journal on Computing*, 16(3):266–274, August 2004.

[24] Marcel F. Neuts. *Matrix-Geometric Solutions In Stochastic Models: An Algorithmic Approach.* Dover Publication, Inc., 1981.

[25] Takayuki Osogami and Mor Harchol-Balter. Closed form solutions for mapping general distributions to quasi-minimal PH distributions. *Performance Evaluation*, 63(6):524–552, June 2006.

[26] Jihong Ou, Jingwen Li, and Süleyman Özekici. Approximating a Cumulative Distribution Function by Generalized Hyperexponential Distributions. *Probability in the Engineering and Informational Sciences*, 11(1):11–18, 1997.

[27] Guodong Pang and Ward Whitt. Heavy-traffic limits for many-server queues with service interruptions. *Queueing Systems*, 61(2-3):167–202, Mar 2009.

[28] Anatolii A. Puhalskii and Martin I. Reiman. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32(2):564–595, Jun 2000.

[29] Ward Whitt. On the Heavy-Traffic Limit Theorem for $GI/G/\infty$ Queues. *Advances in Applied Probability*, 14(1):171–190, 1982.

[30] Ward Whitt. *Stochastic Process Limits.* Springer, 1 edition, 2002.

[31] Ward Whitt. Fluid Models for Multiserver Queues with Abandonments. *Operations Research*, 54(1):37–54, January 2006.

[32] K. Yu, M.-L. Huang, and P. H. Brill. An Algorithm for Fitting Heavy-Tailed Distributions via Generalized Hyperexponentials. *INFORMS Journal on Computing*, 24(1):42–52, March 2011.