TRANSIENT ANALYSIS OF LARGE-SCALE STOCHASTIC SERVICE SYSTEMS

A Dissertation

by

YOUNG MYOUNG KO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Major Subject: Industrial Engineering

Transient Analysis of Large-scale Stochastic Service Systems

TRANSIENT ANALYSIS OF LARGE-SCALE STOCHASTIC SERVICE SYSTEMS

A Dissertation

by

YOUNG MYOUNG KO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Natarajan Gautam
Committee Members,	Yu Ding
	Lewis Ntaimo
	Jean-Francois Chamberland-Tremblay
Head of Department,	Brett A. Peters

May 2011

Major Subject: Industrial Engineering

ABSTRACT

Transient Analysis of Large-scale Stochastic Service Systems. (May 2011)

Young Myoung Ko, B. S., Seoul National University;

M.S., Seoul National University

Chair of Advisory Committee: Dr. Natarajan Gautam

The transient analysis of large-scale systems is often difficult even when the systems belong to the simplest $M/M/n$ type of queues. To address analytical difficulties, previous studies have been conducted under various *asymptotic* regimes by suitably accelerating parameters, thereby establishing some useful mathematical frameworks and giving insights into important characteristics and intuitions. However, some studies show significant limitations when used to approximate real service systems: (i) they are more relevant to steady-state analysis; (ii) they emphasize proofs of convergence results rather than numerical methods to obtain system performance; and (iii) they provide only one set of limit processes regardless of actual system size.

Attempting to overcome the drawbacks of previous studies, this dissertation studies the transient analysis of large-scale service systems with time-dependent parameters. The research goal is to develop a methodology that provides accurate approximations based on a technique called *uniform acceleration*, utilizing the theory of *strong approximations*. We first investigate and discuss the possible inaccuracy of limit processes obtained from employing the technique. As a solution, we propose adjusted fluid and diffusion limits that are specifically designed to approximate large, finite-sized systems. We find that the adjusted limits significantly improve the quality of approximations and hold asymptotic exactness as well. Several numerical results provide evidence of the effectiveness of the adjusted limits. We study both a call

center which is a canonical example of large-scale service systems and an emerging peer-based Internet multimedia service network known as P2P.

Based on our findings, we introduce a possible extension to systems which show non-Markovian behavior that is unaddressed by the uniform acceleration technique. We incorporate the denseness of phase-type distributions into the derivation of limit processes. The proposed method offers great potential to accurately approximate performance measures of non-Markovian systems with less computational burden.

To my family

ACKNOWLEDGMENTS

First and foremost, I express my heartfelt appreciation to my Ph.D. advisor, Dr. Natarajan Gautam. His academic advice and support encouraged me to find and solve the interesting and challenging problems which have culminated in this dissertation. My committee members, Dr. Yu Ding, Dr. Lewis Ntaimo, and Dr. Jean-Franccois Chamberland-Tremblay, inspired me by exhibiting deep interest in my work. Thank you for your valuable comments especially at critical moments during the course my study.

It was a pleasure to study with so many distinguished professors at Texas A&M University. I especially thank my colleagues, Chiwoo Park, Eunshin Byon, Chaehwa Lee, Soondo Hong, Sangdo Choi, and Jung Jin Cho, for sharing the ups and downs of the dissertation process. Texas A&M University is fortunate to have Judy Meeks, whose kindness and expert advice helped me complete several important administrative procedures.

Finally, I could not have come so far without the support of my family. My wife, Hae Jong Yu, deserves my deepest gratitude. Her wholehearted love enabled me to achieve my academic goal. Our beloved twins, Jessica and Rebecca, always made me happy and motivated me to work hard. I knew I could count on my sisters, Yun Kang and Ye Kang, and my parents and my parents-in-law for unstinting support and encouragement. Thank you for your unshakable belief in my ability and for the love you have for our family.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

I.1. Motivation

Popular applications of large-scale service systems include call centers, Internet-based services and mobile networks. These and similar service systems exhibit time-varying characteristics that are sometimes meaningful only in a finite time interval. For example, call centers which operate 24/7 in today's global economy as well as those which "open" and "close" in a shorter timeframe have a daily transient period, in which the arrival rate of customers is highly time-varying (Zeltyn and Mandelbaum (2005)), often changing day by day, week by week, or month by month. Analyzing the operations of these service systems requires examining their transient behaviors, e.g., expected queue length at 3 pm, the probability that customer's virtual waiting time is greater than 2 minutes during peak periods, etc. Another type of service system, Internet-based multimedia services, i.e. Apple iTunes Services (2011), also has a transient period, e.g., each time a user creates a new music or video file.

Transient analysis is critically important for optimal service operations. However, even assuming stationarity and Markovian properties, obtaining accurate performance measures is not trivial. For this reason, asymptotic analysis with fluid and diffusion limits is gaining in popularity (Iglehart, 1965, Mandelbaum *et al.*, 2002, Mandelbaum and Pats, 1998, Whitt, 2006a,b). Typically, an analyst first obtains fluid and diffusion limits by utilizing Functional Law of Large Numbers (FLLN) and Functional Central Limit Theorem (FCLT). Knowing the limits allows the analyst to investigate asymptotic characteristics of the systems under the conditions specified, such as heavy

---

The journal model is *IIE Transactions*.

traffic, large number of servers, etc. Although the use of FLLN and FCLT is a common feature of the literature mentioned above, these studies have different regimes, aims, scenarios, and assumptions. Asymptotic approaches are particularly useful when analyzing large-scale stochastic systems, because many complicated aspects disappear in the asymptotic realm. Performance measures like average queue lengths and virtual waiting times can only require the first two moments of the arrival and service distributions. However, three major limitations occur when applying asymptotic approaches to approximate the real service systems.

First, most real systems are not in the asymptotic realm even if their parameters are fairly large; in fact, asymptotic analysis assumes that some parameters increase to infinity. In other words whatever the parameters of the real systems, the analysis can only use the same limit processes. Suppose there are two different multi-server queues. In the first queue, the customer arrival rate is 10 and the number of servers is 15. In the second queue, the customer arrival rate is 100 and the number of servers is 150. Applying existing asymptotic methods will provide only one set of limit processes to approximate both queues. Note, however, that the performance of the two queues is not only a matter of scaling of the same limit processes. Second, transient analysis may be inappropriate for the two queues. In a steady state, the limit processes usually require only the first two moments of arrival and service distributions. In a transient state, the shape of distributions heavily affects the performance measures, i.e. the first two moments may not be enough. As a result, asymptotic methods taking advantage of the first two moments may fail to approximate the transient dynamics of the systems well enough. Third, most previous studies focus on establishing a convergent sequence and show the existence and uniqueness of limit processes under certain regimes. Only a few studies consider computational methods to obtain the numerical values of performance measures. Consequently, even if one can find a

satisfactory asymptotic model, it may not be possible to obtain explicit numerical values of performance measures.

Furthermore, numerous new types of service systems, for instance, peer-to-peer (P2P) network services, cloud computing services, and social networks, are rapidly emerging. Their asymptotic analysis may not be helpful, unless the drawbacks described above can be overcome. Hence, the objectives of this research are to:

- Accurately approximate time-varying stochastic service systems;
- Derive new limit processes specifically designed to approximate large, finite-size systems;
- Provide efficient computational methods to obtain performance measures;
- Conduct analyses of several large-scale service systems.

The following section states the mathematical expressions of the problem.

I.2. Problem description

We consider a probability space $(\Omega, \mathscr{F}, P)$. Let $\{X(t), t \geq 0\}$ denote a d-dimensional stochastic process which is the state of a stochastic service system and the solution to the following equation:

$$X(t) = x_0 + \sum_{i=1}^{k} l_i R_i(t), \tag{1}$$

where x_0 is the initial value of the process, $l_i \in \mathbf{R}^d$, and $R_i(t)$'s are $\mathscr{F}_t$-adapted counting processes.

It would help understand equation (1) to consider each $R_i(t)$ as cumulated number of arrivals or departures of customers. For example, in $G/G/s$ type service systems, $R_i(\cdot)$'s represent arrival and departure processes. However, they are not

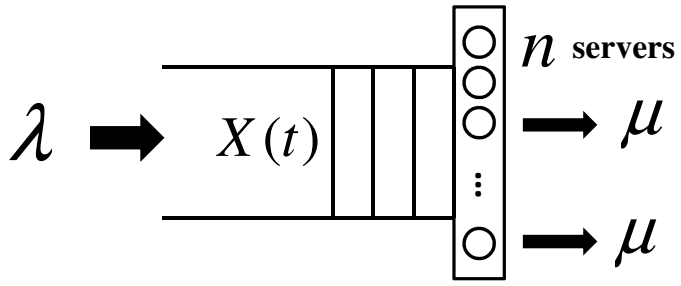


Fig. 1. $M/M/n$ queueing system

restricted to the arrivals or departure processes: e.g., in the epidemic model, they denote the number of people infected or cured by time t.

Specifically, this dissertation focuses on the case where each $R_i(t)$ corresponds to a non-homogeneous Poisson process which is denoted by $Y_i(f_i(t,\cdot))$ as described in Kurtz (1978) and Mandelbaum *et al.* (1998). Then, equation (1) can be rewritten as follows:

$$X(t) = x_0 + \sum_{i=1}^{k} l_i Y_i\Big(\int_0^t f_i\big(s, X(s)\big) ds \Big), \tag{2}$$

where Y_i's are independent rate-1 Poisson processes, and $f_i(t,\cdot)$'s are continuous Lipschitz rate functions.

To help understand equation (2), consider a $M/M/n$ queue. There are n number of servers. Customers arrive to the system with rate λ and the service rate of each server is μ_t (see Figure 1). Then, the system can be expressed as follows:

$$X(t) = Y_1\bigg(\int_0^t \lambda ds \bigg) - Y_2\bigg(\int_0^t \mu\Big(X(s) \wedge n \Big) ds \bigg)$$

where $Y_1(\cdot)$ and $Y_2(\cdot)$ represent the cumulative number of arrivals and departures respectively, $f_1(t,x) = \lambda$, $f_2(t,x) = \mu(x \wedge n)$, $l_1 = 1$, and $l_2 = -1$.

With the representation of the process (equation (2)), performance measures of interest are $E[X(t)]$ and $Cov[X(t), X(t)]$ on any compact time intervals.

I.3. Organization of the dissertation

The organization of the dissertation is as follows. Chapter II summarizes previous research studies on transient analysis and asymptotic methodologies. We explain one of the asymptotic analysis techniques, called uniform acceleration in detail since it provides the basis of this study. Chapter III starts with describing the limited applicability of the uniform acceleration technique when used to approximate real service systems having finite values of parameters. To overcome this limitation, we derive new fluid and diffusion limits, the main results of this dissertation. We discuss how newly derived limits contribute to estimation accuracy and investigate the connection to the existing (or standard) limits. Chapter IV shows an application of adjusted limits to the call-center model, a canonical example of multi-server service systems. Chapter V introduces an emerging peer-based multimedia service systems to deliver contents via Internet and show how adjusted limits successfully approximate this non-traditional service systems. Chapter VI suggests a possible extension to non-Markovian systems by utilizing phase-type approximations and provides an example in wireless mobile sensor networks. Chapter VII recapitulates the key results of this dissertation and lists some future research topics.

CHAPTER II

LITERATURE REVIEW AND BACKGROUND RESEARCH

In this chapter, we explain previous research studies on transient analysis and asymptotic approaches. This chapter focuses on the literature from a methodological perspective. Many other research studies for application-specific topics will be referred to in Chapters IV-VI. Section II.1 summarizes previous research studies regarding transient analysis in queueing systems. Section II.2 explains asymptotic analysis techniques and introduces two popular applications. There are several ways to construct a sequence of processes depending on how to accelerate parameters. This research pays more attention to the technique called uniform acceleration utilizing the theory of strong approximations since it is adequate for the transient analysis of large-scale systems. We visit the details of this technique in Section II.3.

II.1. Transient analysis

Most of the studies on transient analysis have been conducted for queueing systems. For single server queues, Grassmann (1977) and Abate and Whitt (1989) develop algorithms to obtain transient queue length and waiting time distributions under a Markovian setting. In van de Coevering (1995), he reviews several studies on the transient analysis of $M/M/1$ queues especially focusing on numerical computations for practitioners. There are also studies on single server queues under a processor sharing scheme. Chen *et al.* (1997) and Jean-Marie and Robert (1994) provide fluid approximation results by utilizing Functional Law of Large Numbers. Hampshire *et al.* (2006) derives fluid and diffusion limits for $M_t/M_t/1/PS$ utilizing uniform acceleration techniques. For infinite server queues, Collings and Stoneman (1976), Foley (1982), and Eick *et al.* (1993) provide mathematical results for time-dependent

Markovian queues. Nelson and Taaffe (2004a) and Nelson and Taaffe (2004b) develop numerically exact solution procedures to obtain moments of the $Ph_t/Ph_t/\infty$ queues using partial-moment differential equations, which can be used to approximate $G/G/\infty$. Since the transient analysis itself is challenging, many studies concerning it rely on asymptotic approaches by increasing some parameters to infinity. Therefore, in the following sections, we will more closely look at the literature on asymptotic analysis and its applications.

II.2. Asymptotic analysis

Obtaining limit processes is a procedure to establish convergence in a certain functional space (e.g. space C and D). Mathematical procedures to do it are well summarized in Billingsley (1999) and Whitt (2002). As the terms "limit" and "asymptotic" imply, stochastic process limits provide good approximation results under certain conditions such as a large number of servers, large population, heavy traffic, etc.

One of the popular applications of asymptotic analysis is an epidemic model to examine spreading mechanism of a disease (Ball and Neal (2004), Sellke (1983), Ball and Barbour (1990), Ball *et al.* (1997), Andersson and Djehiche (1994), Andersson (1999), Reinert (1995)). Specifically, the literature listed above provides asymptotic analytical (mathematical) results when the number of nodes (or population) increases to infinity along with other parameters suitably. Since the spread of a disease is considered on a city or a nationwide scale, i.e. large population size, the asymptotic approach is adequate for the analysis of those systems and actually it performs well in some numerical studies. However, most of the studies do not provide computational methods or use Markovian assumptions to obtain numerical values of performance measures.

Another application that is more popular and directly related to service systems is a call center model. The call center model is a canonical example of service systems. For that model, methodologies to obtain fluid and diffusion limits, as described in Halfin and Whitt (1981), have been developed in the literature using two different ways in terms of the traffic intensity.

The first approach is to consider the convergence of a sequence of traffic intensities to a certain value. Depending on the value to which the sequence converges, there are three different operational regimes: efficiency driven (ED), quality and efficiency driven (QED), and quality driven (QD). Roughly speaking, if the traffic intensity (ρ) of the limit process is strictly greater than 1, it is called ED regime. If $\rho = 1$, then that is QED, otherwise QD. Many research studies have been done under the ED and QED regimes for multi-server queues like call centers (Halfin and Whitt (1981), Puhalskii and Reiman (2000), Garnet *et al.* (2002), Whitt (2006a), Whitt (2006b), Pang and Whitt (2009)). Recently, the QED regime, also known as Halfin-Whitt regime, has received a lot of attention; this is because it actually achieves both high utilization of servers and quality of service (Zeltyn and Mandelbaum (2005)), and is a favorable operational regime for call centers with strict performance constraints (Mandelbaum and Zeltyn (2009)).

The second way to obtain limit processes is to accelerate parameters keeping the traffic intensity fixed. An effective methodology called "uniform acceleration" which is based on the theory of strong approximations enables the analysis of time-dependent queues (Kurtz (1978), Mandelbaum and Pats (1995), Mandelbaum and Massey (1995), Mandelbaum *et al.* (1998), Mandelbaum and Pats (1998), Massey (1985), Massey and Whitt (1993), Massey and Whitt (1998), Whitt (1990), Hampshire *et al.* (2006), Hampshire *et al.* (2009)) and in fact is the basis of this dissertation. The advantage of accelerating parameters based on strong approximations as de-

scribed in Kurtz (1978) is that it can be applied to a wide class of stochastic processes including various telecommunication systems (Massey (2002)) and can be nicely extended to time-dependent systems by combining with the results in Mandelbaum *et al.* (1998). However, it might not be applied to multi-server queues directly since the rate functions (e.g. net arrival rates and service rates) considered in Kurtz (1978) are assumed to be differentiable everywhere. But some rate functions in multi-server queues are not differentiable everywhere since they are of the forms, $\min(\cdot, \cdot)$ or $\max(\cdot, \cdot)$.

To extend the theory to non-smooth rate functions, Mandelbaum *et al.* (1998) proves weak convergence by introducing a new derivative called "scalable Lipschitz derivative" and provides models for several queueing systems such as Jackson networks, multi-server queues with abandonment and retrials, multi-class preemptive priority queues, etc. In addition, several sets of ordinary differential equations are also provided to obtain the mean value and covariance matrix of the limit processes. It, however, turns out that the resulting sets of differential equations are numerically unsolvable in general. In a follow-on paper, Mandelbaum *et al.* (2002) provides numerical results for queue lengths and waiting times in multi-server queues with abandonment and retrials by adding a constraint to deal with computational intractability. Specifically, the authors restrict their attention to the cases where the time periods when the fluid limit is the same as the number of servers have measure zero. Note that when the fluid limit has the same value as the number of servers, we say the queue is in the critically loaded phase. By doing so, they were able to apply the diffusion limit in Kurtz (1978) to multi-server queues since in this case *scalable Lipschitz derivatives are essentially the same as ordinary derivatives ignoring a set of non-differentiable points.* Adding this constraint seems restrictive in theory. However, in practice, it is quite reasonable. For example, the number

of servers is usually piecewise constant, and the fluid limit is a continuous function of time including non-linear terms. Therefore, the fluid limit possibly stays close to the number of servers but is not likely to stay there on any compact time interval. Nevertheless, as pointed out in Mandelbaum *et al.* (2002), if the queues stay close to the critically loaded phase (called *lingering* in Mandelbaum *et al.* (2002)) for a long time, their approach actually causes significantly inaccurate results despite the fact that it is asymptotically true.

In addition to the area described above, asymptotic analysis has been used for the analysis and control of emerging online video rental systems such as Netflix (Bassamboo *et al.* (2009), Bassamboo and Randhawa (2009)).

II.3. Background research

In this section, we recapitulate the fluid and diffusion limits in Kurtz (1978) and Mandelbaum *et al.* (1998). We will call them *standard* fluid and diffusion limits in the rest of the dissertation. The fluid and diffusion limits are obtained by increasing parameters according to the uniform acceleration technique. For the better fit to service systems, we follow the notation in Mandelbaum *et al.* (1998) instead of that in Kurtz (1978). However, basically those two notations are the same if we adjust the definition of the system state suitably. Moreover, it is worthwhile to note that for $\eta \in \mathbf{N}$, the state of the system $X^{\eta}(t)$ includes jumps but the limit process is continuous. Therefore, the weak convergence result that is presented is with respect to uniform topology in Space D (Billingsley (1999) and Whitt (2002)).

Let $\{X^{\eta}(t), t \geq 0\}$ be an arbitrary d-dimensional stochastic process which is the

solution to the following integral equation:

$$X^{\eta}(t) = x_0^{\eta} + \sum_{i=1}^{k} l_i Y_i\Big(\int_0^t \eta f_i\Big(s, \frac{X^{\eta}(s)}{\eta}\Big) ds\Big), \tag{3}$$

where $x_0^{\eta} = X^{\eta}(0)$ is a constant d-dimensional vector, Y_i's are independent rate-1 Poisson processes, $l_i \in \mathbf{Z}^d$ for $i \in \{1, 2, \ldots, k\}$ are constant, and $f_i(t, \cdot)$'s are Lipschitz continuous functions on compact time intervals.

Typically the process $X^{\eta}(t)$ (usually called a scaled process) is obtained by considering η times faster arrival rate and larger number of servers. This type of setting is used in the literature and is denoted as "uniform acceleration" in Massey and Whitt (1998), Mandelbaum *et al.* (1998), and Mandelbaum *et al.* (2002). Then, the following theorem provides the fluid limit to which $\{X^{\eta}(t)\}_{\eta \geq 1}$ converges almost surely as $\eta \to \infty$. For that, we first define

$$F(t, x) = \sum_{i=1}^{k} l_i f_i(t, x). \tag{4}$$

Theorem 1 (Fluid limit, Kurtz (1978), Mandelbaum *et al.* (1998))**.** *If there is a constant $M < \infty$ such that $|F(t, x) - F(t, y)| \leq M|x - y|$ for all $t \in [0, T]$ and $T < \infty$. Then, $\lim_{\eta \to \infty} \frac{X^{\eta}(t)}{\eta} = \bar{X}(t)$ a.s. where $\bar{X}(t)$ is the solution to the following integral equation:*

$$\bar{X}(t) = x_0 + \sum_{i=1}^{k} l_i \int_0^t f_i\big(s, \bar{X}(s)\big) ds.$$

Note that $\bar{X}(t)$ is a deterministic time-varying quantity. We will connect $\bar{X}(t)$ and $X^{\eta}(t)$ defined in equation (3) via equation (5), but before that we provide the following result. Once we have the fluid limit, we can obtain the diffusion limit from the scaled centered process ($D^{\eta}(t)$). Define $D^{\eta}(t)$ to be $\sqrt{\eta}\Big(\frac{X^{\eta}(t)}{\eta} - \bar{X}(t)\Big)$. Then, the limit process of $D^{\eta}(t)$ is provided by the following theorem.

Theorem 2 (Diffusion limit, Kurtz (1978), Mandelbaum *et al.* (1998))**.** *Suppose F is differentiable almost everywhere with respect to x and the Lebesgue measure of the set $\{t : \partial F(t, \bar{X}(t))$ does not exist.$\}$ is zero. For some $M < \infty$ and $i \in \{1, \ldots, k\}$, if f_i's and F satisfy.*

$$|f_i(t,x) - f_i(t,y)| \leq M|x-y| \quad \text{and} \quad \left|\frac{\partial}{\partial x_i}F(t,x)\right| \leq M, \qquad \text{for almost all } t \in [0,T],$$

then $\lim_{\eta\to\infty} D^\eta(t) = D(t)$ *where* $D(t)$ *is the solution to*

$$D(t) = \sum_{i=1}^{k} l_i \int_0^t \sqrt{f_i\big(s, \bar{X}(s)\big)}dW_i(s) + \int_0^t \partial F\big(s, \bar{X}(s)\big)D(s)ds,$$

$W_i(\cdot)$'s are independent standard Brownian motions, and $\partial F(t,x)$ is the gradient matrix of $F(t,x)$ with respect to x.

Remark 1. *According to Ethier and Kurtz (1986), if $D(0)$ is a constant or a Gaussian random vector, then $D(t)$ is a Gaussian process.*

Now, we have the fluid and diffusion limits for $X^\eta(t)$. Therefore, for a large η, $X^\eta(t)$ is approximated by

$$X^\eta(t) \approx \eta\bar{X}(t) + \sqrt{\eta}D(t). \tag{5}$$

If we follow this approximation, we can also approximate the mean and covariance matrix of $X^\eta(t)$ denoted by $E\big[X^\eta(t)\big]$ and $Cov\big[X^\eta(t), X^\eta(t)\big]$ respectively as

$$E\big[X^\eta(t)\big] \approx \eta\bar{X}(t) + \sqrt{n}E\big[D(t)\big], \text{ and} \tag{6}$$

$$Cov\big[X^\eta(t), X^\eta(t)\big] \approx \eta Cov\big[D(t), D(t)\big]. \tag{7}$$

In equations (6) and (7), only $\bar{X}(t)$ is known. Therefore, in order to get approximated values of $E\big[X^\eta(t)\big]$ and $Cov\big[X^\eta(t), X^\eta(t)\big]$, we need to obtain $E\big[D(t)\big]$ and $Cov\big[D(t), D(t)\big]$. The following theorem provides a methodology to obtain them.

Theorem 3 (Mean and covariance matrix of linear stochastic systems, Arnold (1992))**.** *Let $Y(t)$ be the solution to the following linear stochastic differential equation.*

$$dY(t) = A(t)Y(t)dt + B(t)dW(t), \quad Y(0) = 0,$$

where $A(t)$ is a $d \times d$ matrix, $B(t)$ is a $d \times k$ matrix, and $W(t)$ is a k-dimensional standard Brownian motion. Suppose $A(t)$ and $B(t)$ are measurable and bounded on a compact time interval $[0, T]$. Let $M(t) = E\big[Y(t)\big]$ and $\Sigma(t) = Cov\big[Y(t), Y(t)\big]$. Then, $M(t)$ and $\Sigma(t)$ can be obtained as the unique solution to the following ordinary differential equations:

$$\begin{aligned} \frac{d}{dt}M(t) &= A(t)M(t), \text{ and} \\ \frac{d}{dt}\Sigma(t) &= A(t)\Sigma(t) + \Sigma(t)A(t)' + B(t)B(t)'. \end{aligned} \tag{8}$$

Corollary 1. *If $M(0) = 0$, then $M(t) = 0$ for $t \geq 0$.*

By Corollary 1, if $D(0) = 0$, then $E\big[D(t)\big] = 0$ for $t \geq 0$. Therefore, if $\bar{X}(0) = X(0) = x_0$, then we can rewrite approximate equation (6) to be

$$E\big[X^\eta(t)\big] \approx \eta\bar{X}(t). \tag{9}$$

In many cases, $D(0)$ is set to be zero for the sake of analytical convenience. Then, for a large η, the fluid limit is regarded as an approximation of $E[X^\eta(t)]$ (equation (9)) and the diffusion limit is used to approximate $Cov[X^\eta(t), X^\eta(t)]$ (equation (7)). However, when we closely look at equations (9) and (7), notice that only one set of fluid and diffusion limits is available for all η. Scaling is the only way to adjust approximations among different η values. Will the scaling of fluid and diffusion limits be enough to describe the difference in the dynamics of systems with different η's? In the following chapters, we will answer the question both theoretically and numerically.

CHAPTER III

ADJUSTED FLUID AND DIFFUSION LIMITS

III.1. Inaccuracy of the fluid and diffusion limits as approximations

In this section, we explain the possible inaccuracy of both fluid and diffusion limits when approximating target systems. Consider the following equation to get the exact value of $E\big[X^\eta(t)\big]$ by the following theorem.

Theorem 4 (Expected value of $X^\eta(t)$)**.** *Consider $X^\eta(t)$ defined in equation (1). Then, for $t \in [0,T]$, $E\big[X^\eta(t)\big]$ is the solution to the following equation.*

$$E\big[X^\eta(t)\big] = x_0^\eta + \sum_{i=1}^{k} l_i \int_0^t \eta E\bigg[f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big)\bigg] ds. \tag{10}$$

Proof. Take expectation on both sides of equation (1). Then,

$$\begin{aligned} E\big[X^\eta(t)\big] &= x_0^\eta + \sum_{i=1}^{k} l_i E\bigg[Y_i\Big(\int_0^t \eta f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big) ds\Big)\bigg] \\ &= x_0^\eta + \sum_{i=1}^{k} l_i E\bigg[\int_0^t \eta f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big) ds\bigg] \\ &= x_0^\eta + \sum_{i=1}^{k} l_i \int_0^t \eta E\bigg[f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big)\bigg] ds \\ &\qquad \text{by Fubini theorem in Folland (1999).} \end{aligned}$$

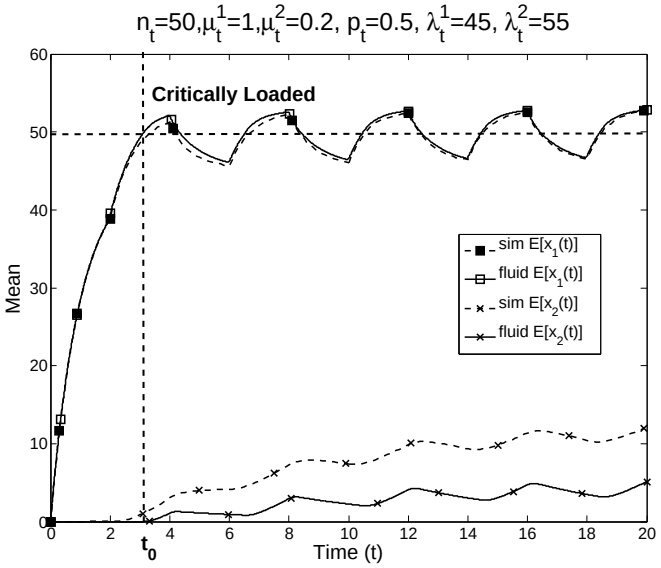
□

Comparing Theorems 1 and 4, notice that we cannot conclude that $\eta\bar{X}(t)$ in Theorem 1 is a good approximation of $E\big[X^\eta(t)\big]$ since $E\Big[f_i\big(t, \frac{X^\eta(t)}{\eta}\big)\Big]$ is not the same as $f_i\Big(t, E\big[\frac{X^\eta(t)}{\eta}\big]\Big)$. They will eventually become identical when η goes to infinity. However, the problem lies in the fact that we have no choice but to use the same $\bar{X}(t)$ (with scaling) as approximations no matter which η values (number of servers)

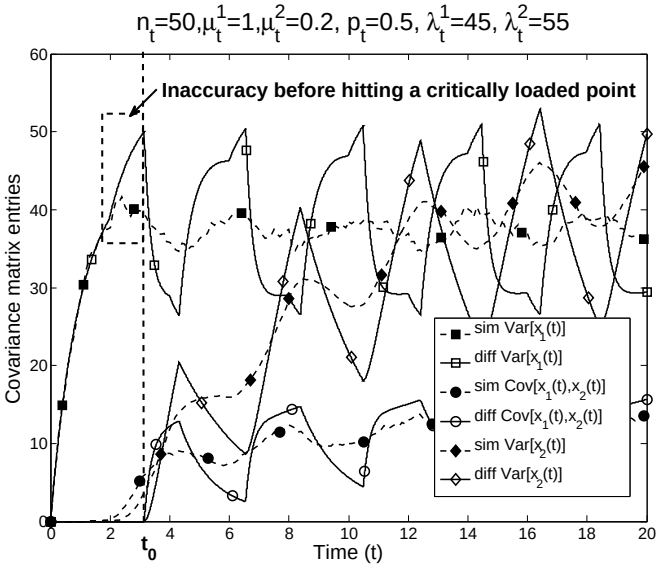
the real systems (e.g. call centers) have, i.e. the same $\bar{X}(t)$ is always used for $\eta = 50$, $\eta = 500$, and $\eta = 5,000$. Therefore, if one derives a new fluid limit specifically designed for each η value, it would provide more improved results, and that is the adjusted fluid limit that this dissertation proposes in Section III.2. Now, recall the diffusion limit in Theorem 2 as follows:

$$D(t) = \sum_{i=1}^{k} l_i \int_0^t \sqrt{f_i\big(s, \bar{X}(s)\big)} dW_i(s) + \int_0^t \partial F\big(s, \bar{X}(s)\big) D(s) ds. \tag{11}$$

Theorem 2 still holds if the set $\{t : \partial F(t, \bar{X}(t))$ does not exist.$\}$ has measure zero. However, non-differentiability of function $F(t, \cdot)$ implies that the drift matrix $(\partial F(t, \cdot))$ can be discontinuous whenever $\bar{X}(t)$ hits the non-differentiable points. This discontinuity may result in sharp spikes in the estimation of the covariance matrix. We will show how significantly the quality of approximation is affected from a numerical example of a call center. Figures 2 (a) and (b) show the estimation of the mean value and covariance matrix of the multi-server queues respectively; we will visit this application in detail in Chapter IV. Since the number of servers is 50, as shown in Figure 2 (a), the mean value of $x_1(t)$, the number of customer in the system, is fluctuating close to the number of servers. From the figure, we also confirm that the fluid limit is quite inaccurate for the estimation of the mean value of $x_2(t)$, the number of customer in the retrial queue. For the covariance matrix, as shown in Figure 2 (b), the diffusion limit brings about immense estimation inaccuracy (sharp spikes) in the nearly critically loaded phase. Recall that under the parameters in Figure 2 *the fluid and diffusion limits using the scalable Lipschitz derivatives in Mandelbaum* et al. *(1998) are virtually the same as those in Kurtz (1978) and Mandelbaum* et al. *(2002).* Therefore, the methodology in Mandelbaum *et al.* (1998) also results in the sharp spikes. As mentioned before, it turns out that the sharp spikes arise from the discontinuity of the drift matrix in the diffusion limits at the non-differentiable points



(a) Simulation vs fluid limit



(b) Simulation vs diffusion limit

Fig. 2. Simulation vs fluid and diffusion limits

of rate functions. We will revisit and explain it in Section III.4.

In the next section, we describe our approach to the above issues in both fluid and diffusion limits. Instead of accelerating parameters, we keep η fixed and construct a new sequence $\{Z^{\eta,\nu}(t)\}_{\nu\geq 1}$ which converges to $E[X^{\eta}(t)]$ almost surely. We derive fluid and diffusion limits for the new sequence and show that they are asymptotically identical to the standard fluid and diffusion limits in Kurtz (1978) and Mandelbaum *et al.* (1998).

III.2. Adjusted fluid and diffusion limits

The basic idea of our approach is to derive new fluid and diffusion limits *for a fixed η (fixed number of servers)*. In this approach, we want to approximate multi-server queues having *a finite number of servers*. To do so, for a fixed η, we first define new rate functions $g_i^{\eta}(t,x)$'s from the existing $f_i(t,x)$'s as follows:

$$g_i^{\eta}(t,x) \;=\; \eta E\Big[f_i\Big(t, \frac{X^{\eta}(t)}{\eta} - \frac{E\big[X^{\eta}(t)\big]}{\eta} + \frac{x}{\eta}\Big)\Big].$$

With above new rate functions, we construct a new sequence of stochastic processes, $\{Z^{\eta,\nu}(t)\}_{\nu\geq 1}$ such that $Z^{\eta,\nu}(t)$ is the solution to the following integral equations:

$$Z^{\eta,\nu}(t) = \nu x_0^{\eta} + \sum_{i=1}^{k} l_i Y_i\Big(\int_0^t \nu g_i^{\eta}\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big). \tag{12}$$

Notice that once we show that $g_i^{\eta}(t,x)$'s are Lipschitz functions on $[0,T]$, we can apply Theorems 1 and 2, and are able to obtain new fluid and diffusion limits for $Z^{\eta,\nu}(t)$. From the following lemmas, we prove that the functions $g_i^{\eta,\nu}(t,\cdot)$'s are actually Lipschitz functions.

Lemma 1. *For a fixed η and $i \in \{1,2,\ldots,k\}$, if $|f_i(t,x)| \leq C_i(1+|x|)$ on $[0,T]$,*

then $g_i^\eta(t,x)$*'s satisfy*

$$|g_i^\eta(t,x)| \;\leq\; D_i(1+|x|) \quad \textit{for some } D_i < \infty.$$

Proof.

$$\begin{aligned}
|g_i^\eta(t,x)| &= \left|\eta E\left[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big)\right]\right| \\
&\leq \left|\eta E\left[C_i\Big(1 + \Big|\frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big|\Big)\right]\right| \\
&\leq \left|\eta C_i\Big(1 + E\Big[\Big|\frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta}\Big|\Big] + \Big|\frac{x}{\eta}\Big|\Big)\right| \\
&\leq D_i(1+|x|),
\end{aligned}$$

where

$$D_i = \eta C_i \sup_{t\leq T}\left(1 + E\Big[\Big|\frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta}\Big|\Big]\right).$$

□

For the next lemma, we would like to define

$$G^\eta(t,x) = \sum_{i=1}^{k} l_i g_i^\eta(t,x). \tag{13}$$

Lemma 2. *For a fixed* η *and* $i \in \{1,2,\ldots,k\}$*, if* $|f_i(t,x) - f_i(t,y)| \leq M|x-y|$ *on* $[0,T]$*, then* $g_i^\eta(t,x)$*'s satisfy*

$$|g_i^\eta(t,x) - g_i^\eta(t,y)| \leq M|x-y|,$$

and if $|F(t,x) - F(t,y)| \leq M|x-y|$*, then* $G^\eta(t,x)$ *satisfies*

$$|G^\eta(t,x) - G^\eta(t,y)| \leq M|x-y|.$$

Proof. For any $t \in [0,T]$,

$$\begin{aligned}
|g_i^\eta(t,x) - g_i^\eta(t,y)| &= \eta\Big|E\Big[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big)\Big] \\
&\quad -E\Big[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{y}{\eta}\Big)\Big]\Big| \\
&= \eta\Big|E\Big[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{x}{\eta}\Big) \\
&\quad -f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + \frac{y}{\eta}\Big)\Big]\Big| \\
&\leq M|x-y|.
\end{aligned}$$

Since M does not depend on t, $|f_i(t,x) - f_i(t,y)| \leq M|x-y|$ on $[0,T]$. □

Hence, with the results in Lemmas 1, and 2, we now derive the adjusted fluid limit.

Theorem 5 (Adjusted fluid limit)**.** *Under the same assumptions in Theorem 1, i.e., for all* $t \in [0,T]$

$$\big|f_i(t,x)\big| \leq C_i\big(1+|x|\big) \quad \textit{for } i \in \{1,\ldots,k\}, \tag{14}$$

$$\big|F(t,x) - F(t,y)\big| \leq M|x-y|, \tag{15}$$

for a fixed η*,*

$$\lim_{\nu\to\infty} \frac{Z^{\eta,\nu}(t)}{\nu} = \bar{Z}^\eta(t) \quad a.s., \tag{16}$$

where $\bar{Z}^\eta(t)$ *is the solution to the following integral equation:*

$$\bar{Z}^\eta(t) = x_0^\eta + \sum_{i=1}^k l_i \int_0^t g_i^\eta\big(s, \bar{Z}^\eta(s)\big)ds, \tag{17}$$

and furthermore

$$\bar{Z}^\eta(t) = E\big[X^\eta(t)\big] = x_0^\eta + \sum_{i=1}^k l_i \int_0^t \eta E\Big[f_i\Big(s, \frac{X^\eta(s)}{\eta}\Big)\Big]ds. \tag{18}$$

Proof. From Lemmas 1 and 2, (14) and (15) imply

$$|g_i^\eta(t,x)| \le D_i(1+|x|) \quad \text{and} \quad |G^\eta(t,x) - G^\eta(t,y)| \le M|x-y|.$$

Therefore, by Theorem 1, we have equation (17), and by the definition of $g_i^\eta(t,x)$'s, we have equation (18). □

Comparing equation (18) with equation (10) in Theorem 4, we notice that Theorem 5 via equation (18) could provide the exact estimation of $E\big[X^\eta(t)\big]$. Once we have the adjusted fluid limit, we can derive the adjusted diffusion limit from it. The following theorem explains the adjusted diffusion limit.

Theorem 6 (Adjusted diffusion limit)**.** *Under the same settings in Theorem 2, for a fixed η, suppose the Lebesgue measure of the set $\{t : \partial G^\eta(t, \bar{Z}^\eta(t))$ does not exist.$\}$ is zero. Define a sequence of scaled centered processes $\{V^{\eta,\nu}(t)\}$ on a time interval $[0,T]$ to be*

$$V^{\eta,\nu}(t) = \sqrt{\nu}\Big(\frac{Z^{\eta,\nu}(t)}{\nu} - \bar{Z}^\eta(t)\Big),$$

where $Z^{\eta,\nu}(t)$ and $\bar{Z}^\eta(t)$ are solutions to equations (12) and (17) respectively. If $f_i(t,x)$'s and $F(t,x)$ satisfy equations (14) and (15) respectively, then $\lim_{\nu\to\infty} V^{\eta,\nu}(t) = V^\eta(t)$, where

$$V^\eta(t) = \sum_{i=1}^k l_i \int_0^t \sqrt{g_i^\eta\big(s, \bar{Z}^\eta(s)\big)}dW_i(s) + \int_0^t \partial G^\eta\big(s, \bar{Z}^\eta(s)\big)V^\eta(s)ds,$$

$W_i(\cdot)$'s are independent standard Brownian motions, and $\partial G^\eta\big(t, \bar{Z}^\eta(t)\big)$ is the gradient matrix of $G^\eta\big(t, \bar{Z}^\eta(t)\big)$ with respect to $\bar{Z}^\eta(t)$. Furthermore, $V^\eta(t)$ is a Gaussian process.

Proof. From definition of $G^\eta(t,x)$ in (13), we can easily verify that $G^\eta(t,x)$ is differ-

entiable almost everywhere, and hence $|G^\eta(t,x) - G^\eta(t,y)| \le M|x-y|$ implies

$$\left|\frac{\partial}{\partial x_i} G^\eta(t,x)\right| \le M_i \quad \text{for some } M_i < \infty, \text{ almost all } t \le T, \text{ and } i \in \{1,\ldots,d\}.$$

Therefore, by Theorem 2, we prove this theorem. □

From Theorems 5 and 6, we obtain the adjusted fluid and diffusion limits from $Z^{\eta,\nu}(t)$. Recall that we call the fluid and diffusion limits in Section II.3 as *standard* and the ones derived in this section as *adjusted* limits. Now one may ask a natural question. What is the relationship between the standard and adjusted limits? The following theorem suggests that the adjusted limits are asymptotically identical to the standard fluid and diffusion limits.

Theorem 7 (Relationship between standard and adjusted limits)**.** *For $t \in [0,T]$, if $\eta f_i(t, x/\eta) = f_i(t,x)$ for $i \in \{1,2,\ldots,k\}$, then,*

$$\lim_{\eta\to\infty} \frac{\bar{Z}^\eta(t)}{\eta} = \bar{X}(t), \quad \textit{and} \tag{19}$$

$$\lim_{\eta\to\infty} \frac{V^\eta(t)}{\sqrt{\eta}} = D(t). \tag{20}$$

Proof. It is enough to show that $\lim_{\eta\to\infty} g^\eta(t,\eta x)/\eta = f_i(t,x)$ for $t \in [0,T]$, which results in the integral equations for LHS of equations (19) and (20) identical to those for RHS. If $\eta f_i(t,x/\eta) = f_i(t,x)$ for $i \in \{1,2,\ldots,k\}$,

$$\begin{aligned}
\lim_{\eta\to\infty} \frac{g^\eta(t,\eta x)}{\eta} &= \lim_{\eta\to\infty} E\left[f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + x\Big)\right] \\
&= E\left[\lim_{\eta\to\infty} f_i\Big(t, \frac{X^\eta(t)}{\eta} - \frac{E\big[X^\eta(t)\big]}{\eta} + x\Big)\right] \\
&= f_i(t,x).
\end{aligned}$$

□

Now, we turn our attention to the solution of the adjusted fluid and diffusion

limits for a fixed η. *Theoretically, Theorem 5 guarantee the exact estimation to* $E\big[X^\eta(t)\big]$. However, the functions $g_i^\eta(\cdot,\cdot)$'s, in fact, cannot be identified unless we know the distribution of $X^\eta(t)$, which forces us to develop a methodology to approximate $g_i^\eta(\cdot,\cdot)$'s for the sake of computational feasibility. Nonetheless, when applying the adjusted fluid limit to the multi-server queues with abandonment and retrials, we have a good candidate distribution to obtain $g_i^\eta(\cdot,\cdot)$'s. So, the following section will describe a computational methodology to get approximated adjusted limits.

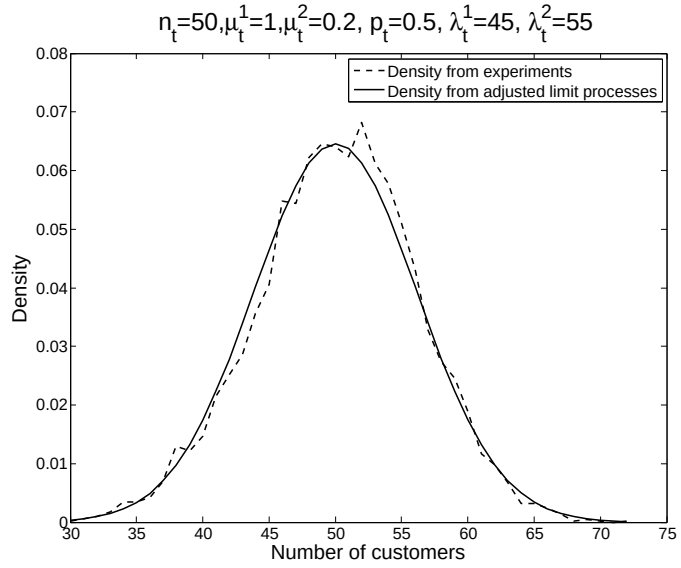
III.3. Approximation of adjusted limits with Gaussian density

In general, there is no clear way to find the distribution of $X^\eta(t)$. Without knowledge of it, it is not possible to obtain $g_i^\eta(\cdot,\cdot)$. However, we could approximate its distribution based on the asymptotic distribution. As mentioned in Section II.3, equation (5) implies that the distribution of $X^\eta(t)$ becomes closer to Gaussian distribution as η increases. Also, for the multi-server queueing systems like call centers, it was experimentally shown that empirical density is actually quite close to the Gaussian density even if the number of servers are not very large: see left graph in Figure 3. Similar empirical results are found in Mandelbaum *et al.* (1998) and Mandelbaum *et al.* (2002). Therefore, using Gaussian distribution to approximate the distribution of $X^\eta(t)$ is reasonable especially when the values of accelerated parameters are large.

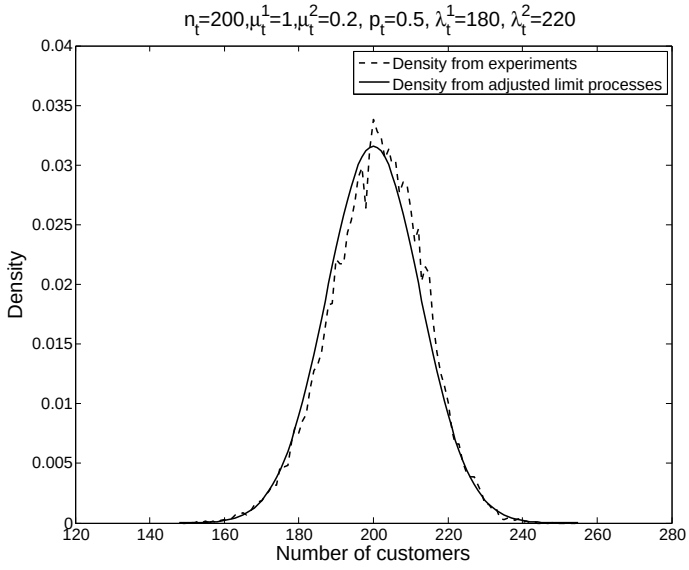
Once we decide to use the Gaussian density, it provides the following two additional benefits:

1. Gaussian distribution can be completely characterized by the mean and covariance matrix which can be obtained from the fluid and diffusion limits.

2. By using Gaussian density, $g_i^\eta(t,\cdot)$'s will be smooth even if $f_i(t,\cdot)$'s are not,



(a)



(b)

Fig. 3. Empirical density vs Gaussian density

which enables us to apply Theorem 6 without measure-zero assumption.

The second benefit is not obvious and hence we provide a proof of that.

Lemma 3. *For any fixed $t > 0$, assume that $X^\eta(t)$ follows a multivariate normal distribution with the mean $\mu = (\mu_1, \ldots, \mu_d)'$ and covariance matrix Σ. Then, $g_i^\eta(t,x)$'s are differentiable everywhere with respect to x.*

Proof. WLOG, we prove this lemma for $\eta = 1$. Define

$$\phi(x,y) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(y-\mu+x)'\Sigma^{-1}(y-\mu+x)}{2} \right).$$

Using Gaussian density,

$$g_i^\eta(t,x) = \int_{\mathbf{R}^d} f_i(t,y)\phi(x,y)dy.$$

For $j \in \{1,\ldots,d\}$, since $\phi(x,y)$ is differentiable with respect to x_j and $|f_i(t,y)\frac{d}{dx_j}\phi(x,y)|$ is integrable,

$$\begin{aligned} \frac{d}{dx_j} g_i^\eta(t,x) &= \frac{d}{dx_j}\int_{\mathbf{R}^d} f_i(t,y)\phi(x,y)dy \\ &= \int_{\mathbf{R}^d} f_i(t,y)\frac{d}{dx_j}\phi(x,y)dy \text{ by Theorem 2.27 in (Folland, 1999),} \end{aligned}$$

where x_j is j^{th} component of x. Therefore, $g_i^\eta(t,x)$ is differentiable with respect to x_j. □

Now, we have $g_i^\eta(t,\cdot)$'s which are differentiable. Then, we can apply Theorem 6 to obtain the diffusion limit for $Z^{\eta,\nu}(t)$. Finally, we approximate the adjusted fluid and diffusion limits by utilizing Gaussian density. Therefore, we compare the adjusted limits with the empirical mean and covariance matrix. Note when we explain Theorem 5, we do not consider $\Sigma^\eta(t)$, the covariance matrix of $X^\eta(t)$. However, in order to obtain $g_i^\eta(t,\cdot)$'s from Gaussian density, we should consider $\Sigma^\eta(t)$. In order

to reflect that, we rewrite $g_i^\eta(t,x)$'s as follows:

$$g_i^\eta\big(t,x\big) \;\rightarrow\; g_i^\eta\big(t,x,u\big) \quad \text{for } i \in \{1,\ldots,k\} \text{ and} \tag{21}$$

$$G^\eta\big(t,x\big) \;\rightarrow\; G^\eta\big(t,x,u\big). \tag{22}$$

Note that the u term in equations (21) and (22) represent the covariance matrix of $X^\eta(t)$.

Proposition 1 (Estimation of mean and covariance matrix using adjusted limits)**.** *The quantities $\bar{Z}^\eta(t)$ and $\Sigma^\eta(t)$ are obtained by solving the following simultaneous ordinary differential equations with initial values given by $\bar{Z}^\eta(0) = x_0^\eta$ and $\Sigma^\eta(0) = 0$:*

$$\frac{d}{dt}\bar{Z}^\eta(t) \;=\; \sum_{i=1}^{k} l_i g_i^\eta\big(t,\bar{Z}^\eta(t),\Sigma^\eta(t)\big), \text{ and} \tag{23}$$

$$\frac{d}{dt}\Sigma^\eta(t) \;=\; A(t)\Sigma^\eta(t) + \Sigma^\eta(t)A(t)' + B(t)B(t)', \tag{24}$$

where $A(t)$ is the gradient matrix of $G^\eta\big(t,\bar{Z}^\eta(t),\Sigma^\eta(t)\big)$ with respect to $\bar{Z}^\eta(t)$, and $B(t)$ is the $d \times k$ matrix such that its i^{th} column is $l_i\sqrt{g_i^\eta\big(t,\bar{Z}^\eta(t),\Sigma^\eta(t)\big)}$.

Proof. By rewriting (17) in Theorem 5 as a differential equation form, we have (23), and by Theorem 3, we have (24). Note that since both $\bar{Z}^\eta(t)$ and $\Sigma^\eta(t)$ are variables, we should solve (23) and (24) simultaneously. □

Although we obtain new rate functions for our adjusted limits, we need some intuition regarding how they contribute to increasing accuracy especially in the critically loaded phases. Thus, in the next section, we revisit the inaccuracy in the previous approaches and explain how our adjusted limits treat this from simple $M_t/M_t/n_t$ queues. The notation, $M_t/M_t/n_t$, is a generalization of Kendall's notation to add time-varying features to $M/M/n$ queues. The arrival and service processes are non-homogeneous Poisson processes, and the number of servers changes over time.

III.4. Intuition behind the functions $g_i^\eta(\cdot,\cdot)$'s

In this section, we explain some intuition regarding the functions $g_i^\eta(t,\cdot)$'s. For the sake of clarity, we consider a simple $M_t/M_t/n_t$ queue. We use $\eta = 1$ and remove the superscript η, i.e., we use $g_i(t,\cdot)$ instead of $g_i^\eta(t,\cdot)$. Let $X(t)$ denote the number of customers in the system at time t. Then, $X(t)$ is the solution to the following integral equation:

$$X(t) = X(0) + Y_1\Big(\int_0^t \lambda_s ds\Big) - Y_2\Big(\int_0^t \big(X(s)\wedge n_s\big)\mu_s ds\Big).$$

Here, for convenience, define $f_1(t,x) = \lambda_t$, $f_2(t,x) = \big(x\wedge n_t\big)\mu_t$, and $F(t,x) = \lambda_t - \big(x\wedge n_t\big)\mu_t$. Applying theorems in Section II.3, we have the fluid and diffusion limits, $\bar{X}(t)$ and $U(t)$ respectively, from the following integral equations:

$$\begin{aligned}
\bar{X}(t) &= X(0) + \int_0^t \lambda_s - \big(\bar{X}(s)\wedge n_s\big)\mu_s ds, \text{ and} \\
U(t) &= U(0) + \int_0^t \Big(\sqrt{\lambda_s}, \sqrt{\big(\bar{X}(s)\wedge n_s\big)\mu_s}\Big)\begin{pmatrix} dW_1(s) \\ dW_2(s)\end{pmatrix} + \int_0^t \partial F(s,\bar{X}(s))U(s)ds,
\end{aligned}$$

where

$$\partial F(t,\bar{X}(t)) = \begin{cases} -\mu_t & \text{if } \bar{X}(t)\le n_t, \\ 0 & \text{otherwise.}\end{cases}$$

Notice that the drift part $\partial F(t,\bar{X}(t))$ of the diffusion limit is completely determined by the fluid limit, and here we notice a possibility that the diffusion limit could produce the sharp spikes described in Section III.1. When the $\bar{X}(t)$ is much smaller than the number of server n_t (underloaded phase), then $Pr[X(t)\ge n_t]$ is likely to be very small, i.e. if we run several independent realization of the process, only small fraction of them are in overloaded or critically loaded phases. In this case, the drift part of the diffusion limit is $-\mu_t$. Now, suppose that $\bar{X}(t)$ is smaller than but fairly

close to n_t (still underloaded phase). Then, $Pr[X(t) \geq n_t]$ would be relatively large. However, the drift part is still the same, $-\mu_t$. The drift part which significantly affects the covariance matrix structure does not reflect $Pr[X(t) \geq n_t]$ at all, while $Pr[X(t) \geq n_t]$ is closely related to the covariance matrix. Furthermore, consider the case where $\bar{X}(t)$ becomes slightly larger than n_t. Then, the drift part suddenly changes to zero. As a result, if $\bar{X}(t)$ is fluctuating close to n_t, then the drift part of the diffusion limit show repeated jumps between the values $-\mu_t$ and 0 although the state of the queue itself does not changes much. Undoubtedly, it produces sharp spikes in the covariance matrix as shown in Figure 2 and make the quality of the approximation worse. Now, we turn our attention to the functions $g_i(t, \cdot)$'s. Let us follow the procedure to obtain $g_2(t, x)$. Note that $g_1(t, x) = f_1(t, x)$.

Define $G(t, x) = g_1(t, x) - g_2(t, x) = \lambda_t - g_2(t, x)$. For a fixed t_0, let $X = X(t_0)$, $\mu = \mu_{t_0}$ and $n = n_{t_0}$ and $x = E[X]$. Then,

$$g_2(t_0, x) = E\big[\mu(X \wedge n)\big] = \mu\Big\{E[X\mathbb{I}_{X \leq n}] + nPr[X > n]\Big\}. \tag{25}$$

From equation (25), notice the following characteristics of the function $g_2(t, x)$:

1. If $Pr[X > n]$ becomes closer to 1, $\partial g_2(t_0, x)/\partial x$ approaches 0;

2. If $Pr[X > n]$ gets closer to 0, $\partial g_2(t_0, x)/\partial x$ gets to be closer to μ.

Notice that the drift part $\partial G(t, \cdot)$ determines its value between $-\mu_t$ and 0 according to $Pr[X(t) > n_t]$. Therefore, even if the queue makes phase transitions frequently, the drift part of the adjusted diffusion limit does not make sudden changes.

Summarizing this chapter, Table 1 compares key characteristics of the standard and adjusted limits. One can see the difference between two limits at a glance. In the following chapters, we actually analyze several service systems and show how the adjusted limits outperform the standard limits in the analyses.

Table 1. Comparison between standard and adjusted limits

	Standard limit	Adjusted limit
Rate functions	$f_i(\cdot,\cdot)$'s	$g_i^\eta(\cdot,\cdot)$'s
Fluid limit	obtained independently	obtained simultaneously
Diffusion limit	obtained using fluid limit	
Assumption	measure zero at non-differentiable points	Gaussian density
Advantage	easy to compute	accurate approximation for a fixed η
Limitation	inaccuracy in both fluid and diffusion limits	inaccuracy when Gaussian assumption fails

CHAPTER IV

MULTI-SERVER QUEUES WITH ABANDONMENT, RETRIAL, AND TIME-VARYING PARAMETERS

In this chapter we apply the adjusted fluid and diffusion limits to service systems having many servers represented by call centers. We provide the result of applying standard fluid and diffusion limits as well and show how dramatically estimation accuracy is improved.

IV.1. Critically loaded multi-server queues

According to Mandelbaum *et al.* (1998) and Mandelbaum *et al.* (2002), time-dependent queues make transitions among three phases: underloaded, critically loaded, and overloaded. The phase of the system is determined by its fluid limit. The limit process in strong approximations does not require any regimes such as QD, QED, or ED. However, from Section 1.4 in Zeltyn and Mandelbaum (2005), we could find a rough correspondence between the operational regimes (QD, QED, and ED) and the phases in time-varying queues (underloaded, critically loaded, and overloaded).

Explaining it briefly, Zeltyn and Mandelbaum (2005) models operational regimes from tracing data of real call centers. They say the call center is working in the ED regime when the occupancy is 100% with higher abandonment rates. Similarly, they associate the time slots with the other operational regimes such as QED and QD. The tracing table used for the explanation of the operational regimes, indeed, represents the dynamics of the time-varying multi-server queues. Therefore, the ED, QED and QD regimes could correspond to the overloaded, critically loaded and underloaded phases respectively. From the tracing data in Zeltyn and Mandelbaum (2005), we recognize the importance of the critically loaded phase as nearly 100% utilization and

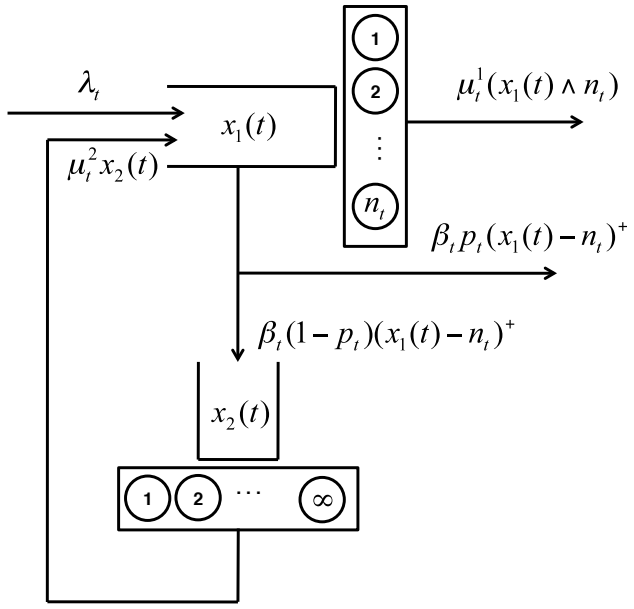


Fig. 4. Multi-server queue with abandonment and retrials, Mandelbaum *et al.* (2002)

low abandonment rates which most of companies want are achieved in that phase.

Therefore, capturing the dynamics of multi-server queues near the critically loaded phase is also of significant importance. In this chapter, we reveal how standard fluid and diffusion limits are inaccurate especially in the critically loaded phase. Therefore, we will focus more on the critically loaded phase and show how this inaccuracy is improved by the adjusted limits.

IV.2. Problem description

Consider Figure 4 that illustrates a multi-server queue with abandonment and retrials as described in Mandelbaum *et al.* (1998) and Mandelbaum *et al.* (2002). There are n_t number of servers in the service node at time t. Customers arrive to the service node according to a non-homogeneous Poisson process at rate λ_t. The service time of each customer follows a distribution having a memoryless property at rate μ_t^1. Customers

in the queue are served under the FCFS policy and the abandonment rate of customers is β_t with exponentially distributed time to abandon. Abandoning customers leave the system with probability p_t or go to a retrial queue with probability $1-p_t$. The retrial queue is equivalent to an infinite-server-queue and hence each customer in the retrial queue waits there for a random amount of time with mean $1/\mu_t^2$ and returns to the service node.

Let $X(t) = \big(x_1(t), x_2(t)\big)$ be the system state where $x_1(t)$ is the number of customers in the service node and $x_2(t)$ is the number of customers in the retrial queue at time t. Then, $X(t)$ is the unique solution to the following integral equations:

$$\begin{aligned} x_1(t) &= x_1(0) + Y_1\Big(\int_0^t \lambda_s ds\Big) + Y_2\Big(\int_0^t x_2(s)\mu_s^2 ds\Big) \\ &\quad -Y_3\Big(\int_0^t \big(x_1(s)\wedge n_s\big)\mu_s^1 ds\Big) - Y_4\Big(\int_0^t \big(x_1(s)-n_s\big)^+\beta_s(1-p_s)ds\Big) \\ &\quad -Y_5\Big(\int_0^t \big(x_1(s)-n_s\big)^+\beta_s p_s ds\Big), \text{ and} \qquad (26) \\ x_2(t) &= x_2(0) + Y_4\Big(\int_0^t \big(x_1(s)-n_s\big)^+\beta_s(1-p_s)ds\Big) - Y_2\Big(\int_0^t x_2(s)\mu_s^2 ds\Big), \qquad (27) \end{aligned}$$

where Y_i's are independent rate-1 Poisson processes.

In the following sections, we provide analytical models to obtain standard and adjusted fluid and diffusion limits.

IV.3. Standard fluid and diffusion limits, Mandelbaum *et al.* (1998)

Deriving standard fluid and diffusion limits is done by Mandelbaum *et al.* (1998). Therefore, in this section, we summarize their results to compare them with our results. In order to obtain the limit processes, define $X^\eta(t) = (x_1^\eta(t), x_2^\eta(t))'$ by

accelerating the arrival rate and the number of servers by the factor of η as follows:

$$\begin{aligned}
x_1^\eta(t) &= x_1^\eta(0) + Y_1\Big(\int_0^t \eta\lambda_s ds\Big) + Y_2\Big(\int_0^t x_2^\eta(s)\mu_s^2 ds\Big) - Y_3\Big(\int_0^t \big(x_1^\eta(s)\wedge \eta n_s\big)\mu_s^1 ds\Big) \\
&\quad -Y_4\Big(\int_0^t \big(x_1^\eta(s)-\eta n_s\big)^+\beta_s(1-p_s)ds\Big) - Y_5\Big(\int_0^t \big(x_1^\eta(s)-\eta n_s\big)^+\beta_s p_s ds\Big), \\
&= x_1^\eta(0) + Y_1\Big(\int_0^t \eta\lambda_s ds\Big) + Y_2\Big(\int_0^t \eta\Big(\frac{x_2^\eta(s)}{\eta}\mu_s^2\Big)ds\Big) \\
&\quad -Y_3\Big(\int_0^t \eta\Big(\frac{x_1^\eta(s)}{\eta}\wedge n_s\Big)\mu_s^1 ds\Big) - Y_4\Big(\int_0^t \eta\Big(\frac{x_1^\eta(s)}{\eta}-n_s\Big)^+\beta_s(1-p_s)ds\Big) \\
&\quad -Y_5\Big(\int_0^t \eta\Big(\frac{x_1^\eta(s)}{\eta}-n_s\Big)^+\beta_s p_s ds\Big), \text{ and} \qquad (28) \\
x_2^\eta(t) &= x_2^\eta(0) + Y_4\Big(\int_0^t \big(x_1^\eta(s)-\eta n_s\big)^+\beta_s(1-p_s)ds\Big) - Y_2\Big(\int_0^t x_2^\eta(s)\mu_s^2 ds\Big), \\
&= x_2^\eta(0) + Y_4\Big(\int_0^t \eta\Big(\frac{x_1^\eta(s)}{\eta}-n_s\Big)^+\beta_s(1-p_s)ds\Big) \\
&\quad -Y_2\Big(\int_0^t \eta\Big(\frac{x_2^\eta(s)}{\eta}\mu_s^2\Big)ds\Big). \qquad (29)
\end{aligned}$$

We now obtain the standard fluid limit by taking $\eta\to\infty$.

Theorem 8 (Fluid limit, Mandelbaum *et al.* (1998))**.** *Let $\bar{X}(t) = (\bar{x}_1(t), \bar{x}_2(t))'$ be the solution to the following integral equation:*

$$\begin{aligned}
\bar{x}_1(t) &= \bar{x}_1(0) + \int_0^t \lambda_s + \bar{x}_2(s)\mu_s^2 - \big(\bar{x}_1(s)\wedge n_s\big)\mu_s^1 \\
&\quad -\big(\bar{x}_1(s)-n_s\big)^+\beta_s(1-p_s) - \big(\bar{x}_1(s)-n_s\big)^+\beta_s p_s ds, \text{ and} \\
\bar{x}_2(t) &= \bar{x}_2(0) + \int_0^t \big(\bar{x}_1(s)-n_s\big)^+\beta_s(1-p_s) - \big(\bar{x}_2(s)\mu_s^2\big)ds.
\end{aligned}$$

If $\lim_{\eta\to\infty} X^\eta(0)/\eta = \bar{X}(0)$, then, on any compact time intervals,

$$\lim_{\eta\to\infty}\frac{X^\eta(t)}{\eta} = \bar{X}(t) \quad \textit{almost surely.}$$

Define matrices, $K_1(t)$, $K_2(t)$, and $L(t)$ as follows:

$$
\begin{aligned}
K_1(t) &= \begin{pmatrix} -\mu_t^1 - \beta_t & \mu_t^2 \\ \beta_t(1-p_t) & -\mu_t^2 \end{pmatrix}, \\
K_2(t) &= \begin{pmatrix} 0 & \mu_t^2 \\ 0 & -\mu_t^2 \end{pmatrix}, \text{ and} \\
L(t) &= \begin{pmatrix} \sqrt{\lambda_t} & 0 \\ \sqrt{\mu_t^2 \bar{x}_2(t)} & -\sqrt{\mu_t^2 \bar{x}_2(t)} \\ -\sqrt{\mu_t^1(\bar{x}_1(t) \wedge n_t)} & 0 \\ -\sqrt{(\bar{x}_1(t) - n_t)^+ \beta_t(1-p_t)} & \sqrt{(\bar{x}_1(t) - n_t)^+ \beta_t(1-p_t)} \\ -\sqrt{(\bar{x}_1(t) - n_t)^+ \beta_t p_t} & 0 \end{pmatrix}'.
\end{aligned}
$$

With the matrices above, we derive the diffusion limit.

Theorem 9 (Diffusion limit, Mandelbaum *et al.* (1998))**.** *Let* $D^\eta(t) = \sqrt{\eta}(X^\eta/\eta - \bar{X}(t))$*. Then,*

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

and $D(t)$ *is the solution to the following integral equation:*

$$D(t) = D(0) + \int_0^t L(s) dB(s) + \int_0^t K(s) D(s) ds,$$

where, $B(t)$ *is a 5-dimensional standard Brownian motion and*

$$K(t) = \begin{cases} K_1(t) & \textit{if } \bar{x_1}(t) \leq n_t, \\ K_2(t) & \textit{otherwise.} \end{cases}.$$

In Theorem 9, we notice that the drift matrix of the diffusion limit makes sudden changes between $K_1(t)$ and $K_2(t)$ depending on the fluid limit. This means that if $\bar{x_1}(t)$ stays less (or greater) than n_t, we say the diffusion limit has a somewhat *stable*

drift matrix. On the other hand, if $\bar{x_1}(t)$ stays close to n_t and frequently crosses it (i.e., nearly critically loaded), the diffusion limit experiences frequent changes in the drift matrix. The drift matrix of the diffusion limit significantly affects the value of the covariance matrix. Sudden changes in the drift matrix causes sharp spikes in the covariance matrix, which makes the estimation accuracy worse. We observe this phenomenon by numerical examples in Section IV.5.

IV.4. Adjusted fluid and diffusion limits

In this section, we derive the adjusted fluid and diffusion limits for the call center models. In order to maintain numerical tractability, we use the Gaussian-based approximation for $g_i^\eta(\cdot,\cdot)$'s. For a fixed η (or a fixed number of servers), we obtain approximated $g_i^\eta(\cdot,\cdot)$'s as follows:
For $x = (x_1, x_2)'$ and $\sigma_1^\eta(t)^2 = Var[x_1^\eta(t)]$,

$$
\begin{aligned}
g_1^\eta(t,x) &= \eta\lambda_t, \\
g_2^\eta(t,x) &= \mu_t^2 x_2, \\
g_3^\eta(t,x) &= \mu_t^1\big(\eta n_t + (x_1 - \eta n_t)\Phi(\eta n_t, x_1, \sigma_1^\eta(t)) - \sigma_1^\eta(t)^2(t)\phi(\eta n_t, x_1, \sigma_1^\eta(t))\big), \\
g_4^\eta(t,x) &= \beta_t(1-p_t)\Big((x_1 - \eta n_t)\big(1 - \Phi(\eta n_t, x_1, \sigma_1^\eta(t))\big) + \sigma_1^\eta(t)^2(t)\phi(\eta n_t, x_1, \sigma_1^\eta(t))\Big),
\end{aligned}
$$

and

$$
g_5^\eta(t,x) = \beta_t p_t\Big((x_1 - \eta n_t)\big(1 - \Phi(\eta n_t, x_1, \sigma_1^\eta(t))\big) + \sigma_1^\eta(t)^2(t)\phi(\eta n_t, x_1, \sigma_1^\eta(t))\Big),
$$

where $\Phi(a,b,c)$ and $\phi(a,b,c)$ are function values at point a of the Gaussian CDF and PDF respectively with mean b and standard deviation c.

Since $f_1(t,x)$ and $f_2(t,x)$ are constant and linear with respect to x respectively, $g_1(t,x) = \eta f_1(t, x/\eta)$ and $g_2(t,x) = \eta f_2(t, x/\eta)$. The derivation of other $g_i(\cdot,\cdot)$'s is straightforward but requires some computational efforts and hence we provide the

details in Appendix A.1.

Then, we can define a new sequence of stochastic processes, $\{Z^{\eta,\nu}(t)\}_{\nu\geq 1}$ as described in equation 12 and derive the adjusted limits as follows:

Let $\sigma^\eta(t)$ be the standard deviation of $z_1^\eta(t)$. Define

$$\begin{aligned}
\Phi^\eta(t) &= \Phi(\eta n_t, \bar{z}_1^\eta(t), \sigma^\eta(t)), \\
\phi^\eta(t) &= \phi(\eta n_t, \bar{z}_1^\eta(t), \sigma^\eta(t)), \\
\alpha_1^\eta(t) &= \eta n_t + (\bar{z}_1^\eta(t) - \eta n_t)\Phi^\eta(t) - \sigma^\eta(t)^2\phi^\eta(t), \\
\alpha_2^\eta(t) &= (\bar{z}_1^\eta(t) - \eta n_t)(1 - \Phi^\eta(t)) + \sigma^\eta(t)^2\phi^\eta(t), \\
K^\eta(t) &= \begin{pmatrix} -\mu_t^1\Phi^\eta(t) - \beta_t(1-\Phi^\eta(t)) & \mu_t^2 \\ \beta_t(1-p_t)(1-\Phi^\eta(t)) & -\mu_t^2 \end{pmatrix}, \text{ and} \\
L^\eta(t) &= \begin{pmatrix} \sqrt{\eta\lambda_t} & 0 \\ \sqrt{\mu_t^2\bar{z}_2^\eta(t)} & -\sqrt{\mu_t^2\bar{z}_2^\eta(t)} \\ -\sqrt{\mu_t^1\alpha_1^\eta(t)} & 0 \\ -\sqrt{\beta_t(1-p_t)\alpha_2^\eta(t)} & \sqrt{\beta_t(1-p_t)\alpha_2^\eta(t)} \\ -\sqrt{\beta_t p_t\alpha_2^\eta(t)} & 0 \end{pmatrix}'.
\end{aligned}$$

Theorem 10 (Adjusted fluid limit)**.** *Suppose* $\lim_{\nu\to\infty} Z^{\eta,\nu}(0)/\nu = \bar{Z}^\eta(0)$*. Then, on any compact time interval*

$$\lim_{\nu\to\infty} \frac{Z^{\eta,\nu}(t)}{\nu} = \bar{Z}^\eta(t) \quad a.s. \tag{30}$$

where $\bar{Z}^\eta(t)$ *is the solution to the following integral equation:*

$$\begin{aligned}
\bar{z}_1^\eta(t) &= \bar{z}_1^\eta(0) + \int_0^t \lambda_s + \mu_s^2\bar{z}_2^\eta(s) - \mu_s^1\alpha_1^\eta(s) - \beta_s\alpha_2^\eta(s)ds, \text{ and} \\
\bar{z}_2^\eta(t) &= \bar{z}_2^\eta(0) + \int_0^t -\mu_s^2\bar{z}_2^\eta(s) + \beta_s(1-p_s)\alpha_2^\eta(s)ds.
\end{aligned}$$

Proof. Since $f_i(t,\cdot)$'s are Lipschitz functions, so are $g_i^\eta(t,\cdot)$'s. Therefore, by Theo-

rem 5, we prove this theorem. □

Once we derive the adjusted fluid limit, we now obtain the adjusted diffusion limit from the scaled centered processes.

Theorem 11 (Adjusted diffusion limit)**.** *Let* $V^{\eta,\nu}(t) = \sqrt{\nu}(Z^{\eta,\nu}(t) - \bar{Z}^{\eta}(t))$*. Then*

$$\lim_{\nu\to\infty} V^{\eta,\nu}(t) = V^{\eta}(t) \quad \textit{in distribution,} \tag{31}$$

and $V^{\eta}(t)$ *is the solution to the following integral equation:*

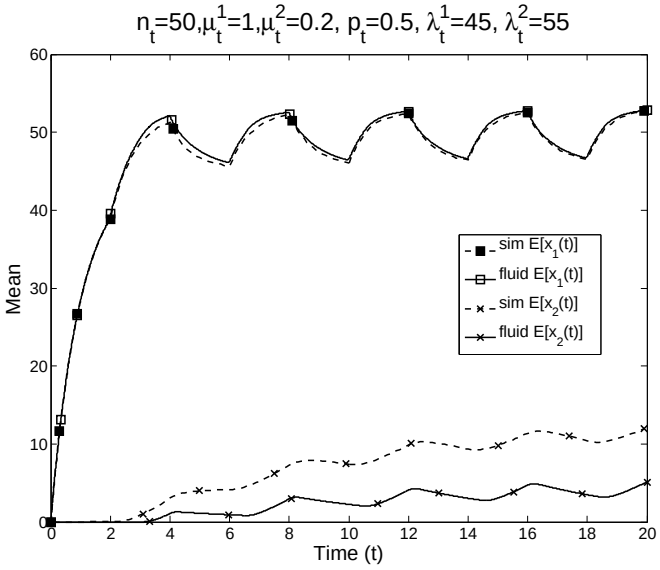
$$V^{\eta}(t) = V^{\eta}(0) + \int_0^t L^{\eta}(s)dB(s) + \int_0^t K^{\eta}(s)V^{\eta}(s)ds,$$

where $B(t)$ *is a 5-dimensional standard Brownian motion.*

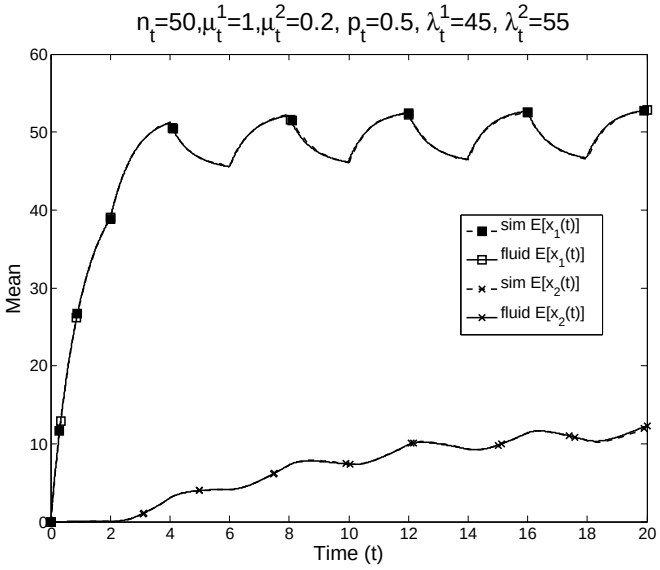
Proof. Since $g_i^{\eta}(t,\cdot)$'s are obtained from Gaussian density, by Lemma 3, $g_i^{\eta}(t,\cdot)$'s are differentiable everywhere. Therefore, by Theorem 6, we have this theorem. □

IV.5. Numerical results

In this section, we show several numerical results to justify how effectively the adjusted limits approximate the multi-server queues with abandonment and retrials. We compare our adjusted limits against the standard ones. Under the similar settings in Mandelbaum *et al.* (2002), we use 5,000 independent simulation runs and use them as a reference model. We use constant rates for all parameters except the arrival rate. The arrival rate alternates between 45 and 55 every two time units. Figures 5 and 6 show the estimation of mean values from one experiment. The number of servers (n_t) is 50 and the service rate of each server is 1 for all $t \geq 0$. As seen in Figure 5, the multi-server queue is nearly critically loaded, i.e. $\bar{x}_1(t) \approx n_t$. As Mandelbaum *et al.* (2002) points out, the standard fluid limit shows significant inaccuracy for $E\big[x_2(t)\big]$. On the other hand, the adjusted fluid limit provides an excellent approximation



(a) Mean numbers by the fluid limit



(b) Mean numbers by the adjusted fluid limit

Fig. 5. Comparison of mean values, $E\big[X(t)\big]$

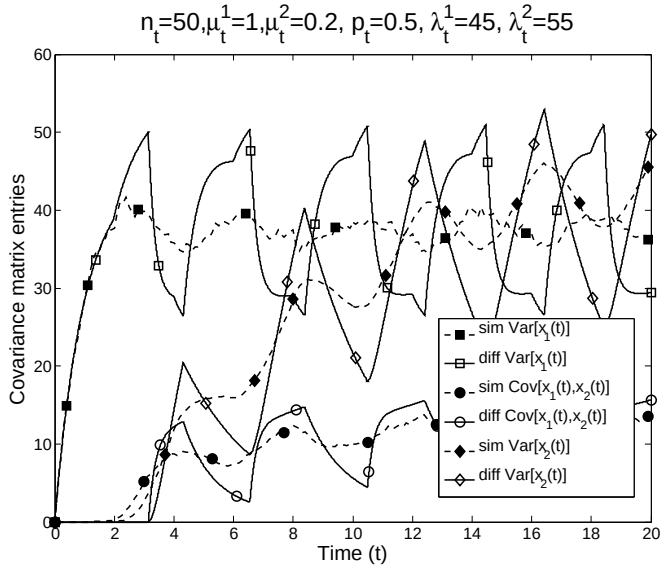
result. Especially, one can recognize remarkable improvement in the estimation of $E\big[x_2(t)\big]$. For the mean value of $x_1(t)$, the adjusted fluid limit provides a lot better approximation result.

When we see the estimation of the covariance matrix, we also notice the adjusted diffusion limit shows dramatic improvement. As seen in Figure 6, the standard diffusion limit causes "spikes" as also pointed out in Section III.1. The adjusted diffusion limit, however, provides excellent accuracy without spikes.

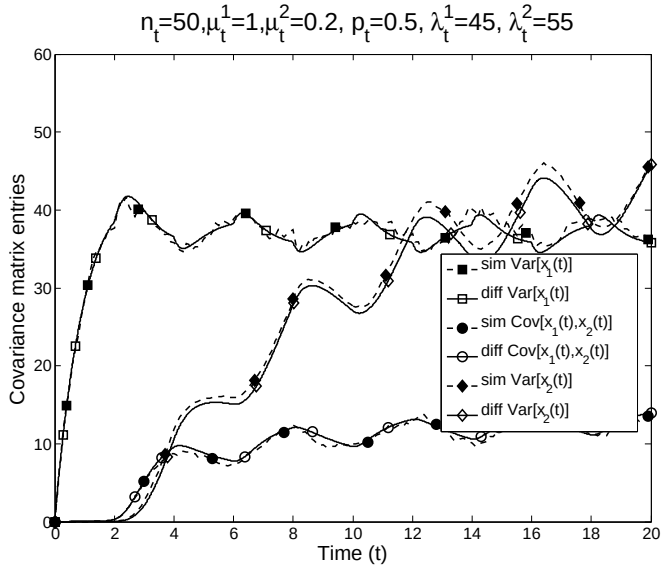
Besides this specific example, in order to verify the effectiveness of our methodology, we conduct several experiments with different parameter combinations. Table 2 describes the setting of each experiment. In Table 2, "svrs" is the number of servers (n_t), "alter" is the time length for which each arrival rate lasts, and "time" is the end time of our analysis. We already recognize that the standard fluid and diffusion limits work well when it *does not linger* too long close to the critically loaded phase. For comparison, therefore, our experiments contain several cases where the system

Table 2. Experiment setting

exp	svrs	λ_1	λ_2	μ_1	μ_2	β	prob	alter	time
1	50	45	55	1	0.2	2.0	0.5	2	20
2	200	180	220	1	0.2	2.0	0.5	2	20
3	300	270	330	1	0.2	2.0	0.5	2	20
4	400	360	440	1	0.2	2.0	0.5	2	20
5	400	390	410	1	0.2	2.0	0.5	2	20
6	50	45	55	1	2.0	2.0	0.5	2	20
7	50	45	55	1	0.2	5.0	0.5	2	20
8	50	45	55	1	0.2	2.0	0.9	2	20



(a) Covariance matrix by the diffusion limit

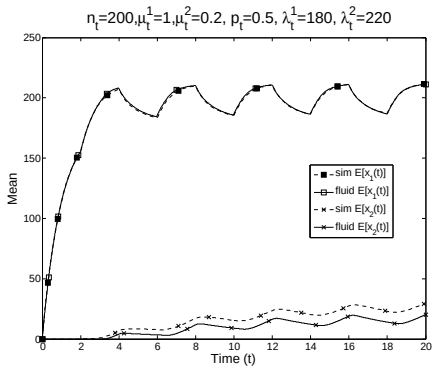


(b) Covariance matrix by the adjusted diffusion limit

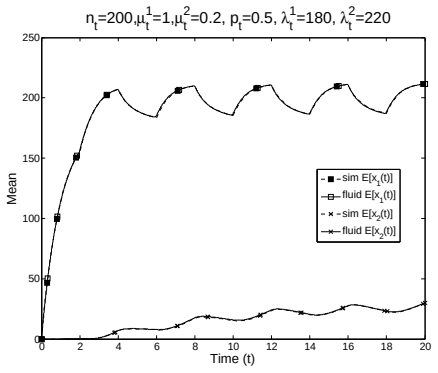
Fig. 6. Comparison of covariance matrix entries, $Cov\big[X(t), X(t)\big]$

does linger relatively longer. Experiments 1-4 are intended to see how the estimation accuracy would be improved as we increase the number of servers along with the arrival rates. Experiment 5 is set in order to observe the effect of *lingering* in the nearly critically loaded phase even if we the number of servers is fairly large. We change parameters other than number of servers and arrival rates in experiments 6-8 to see the effects of them. In fact, from a large number of experiments not listed in Table 2, we observe that they do not affect estimation accuracy significantly.

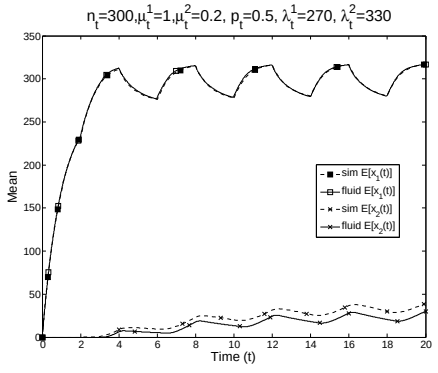
Here we explain the overall results: for the details of numerical results, see Tables 4-8 in Appendix A.2. As seen in Figures 7 and 8, the standard fluid and diffusion limits show improvement in estimation accuracy when we scale the number of servers along with the arrival rates since the standard limits are asymptotically true. In fact, the improvement when using the adjusted limits does not seem obvious. However, it is because they already provide excellent estimation results even when the number of servers is relatively small, and the adjusted limits always outperform the standard fluid and diffusion limits. We also see the effect of lingering near the critically loaded phase in Figures 9 and 10. Although the number of servers is fairly large, it does debase the quality of approximations significantly when we use the standard fluid and diffusion limits. On the other hand, the adjusted ones provide excellent accuracy for both mean and covariance matrix in all cases. Figure 11 illustrates the average percentile difference of both approaches against simulation. Figure 11 (a) is obtained by averaging all difference across time. From Figure 11 (a), we notice that the adjusted limits show promise relative to the standard ones. In Figure 11 (b), we graph the differences especially at the time points when the queues are critically loaded. It turns out that the adjusted limits are still accurate, while the standard ones show even worse accuracy as expected. Note that, in Figure 11 (b), huge estimation difference, more than 300%, is observed when estimating



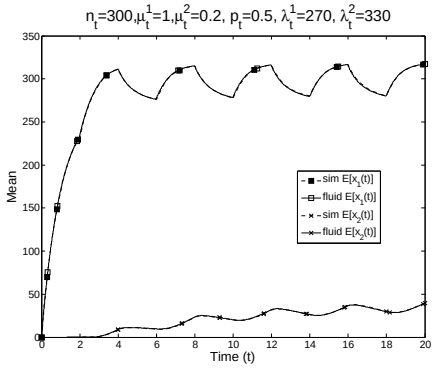
(a) Standard fluid limit of exp. 2



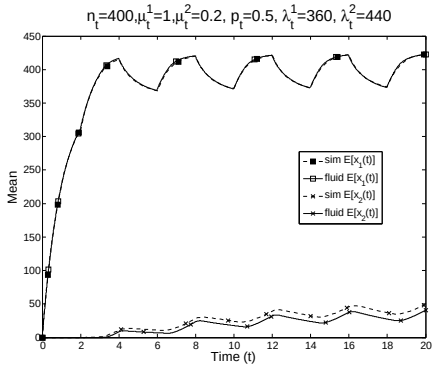
(b) Adjusted fluid limit of exp. 2



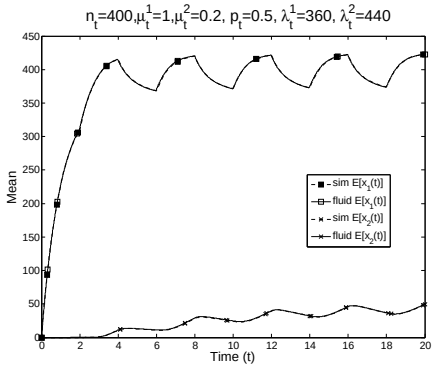
(c) Standard fluid limit of exp. 3



(d) Adjusted fluid limit of exp. 3

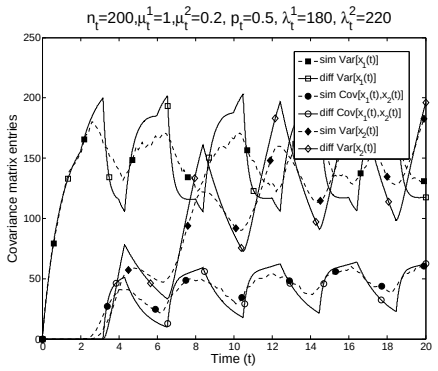


(e) Standard fluid limit of exp. 4

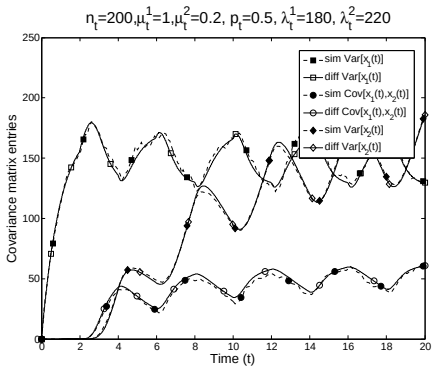


(f) Adjusted fluid limit of exp. 4

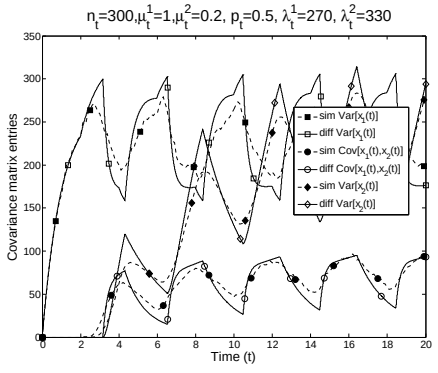
Fig. 7. Comparison of mean values, $E\big[X(t)\big]$



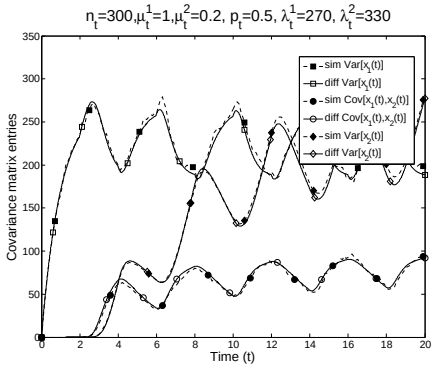
(a) Standard diffusion limit of exp. 2



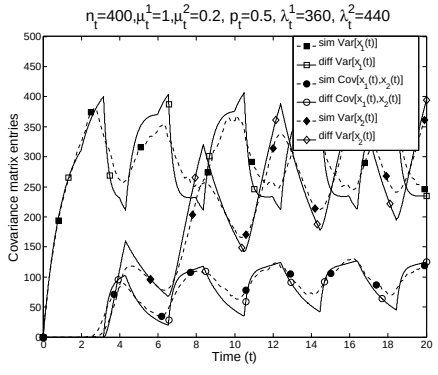
(b) Adjusted diffusion limit of exp. 2



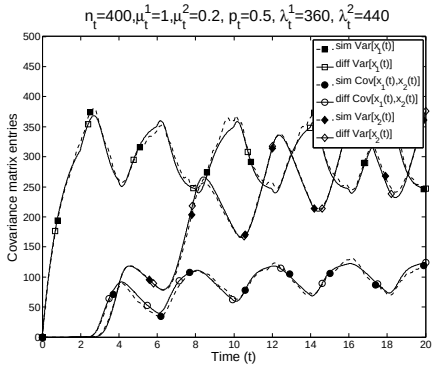
(c) Standard diffusion limit of exp. 3



(d) Adjusted diffusion limit of exp. 3

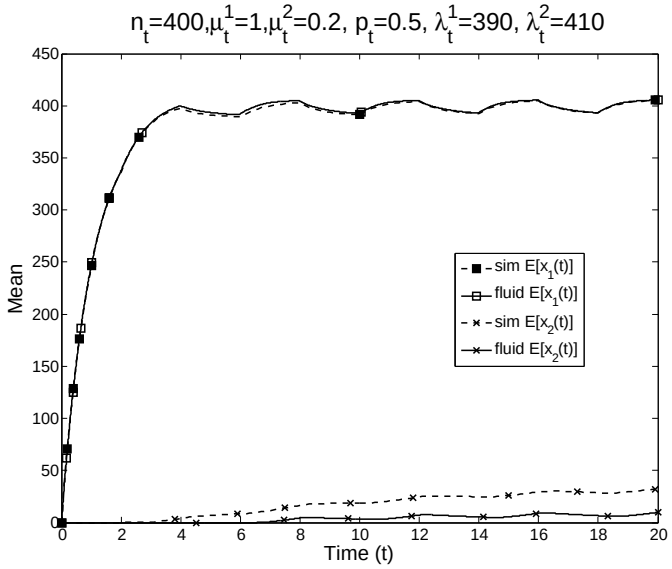


(e) Standard diffusion limit of exp. 4

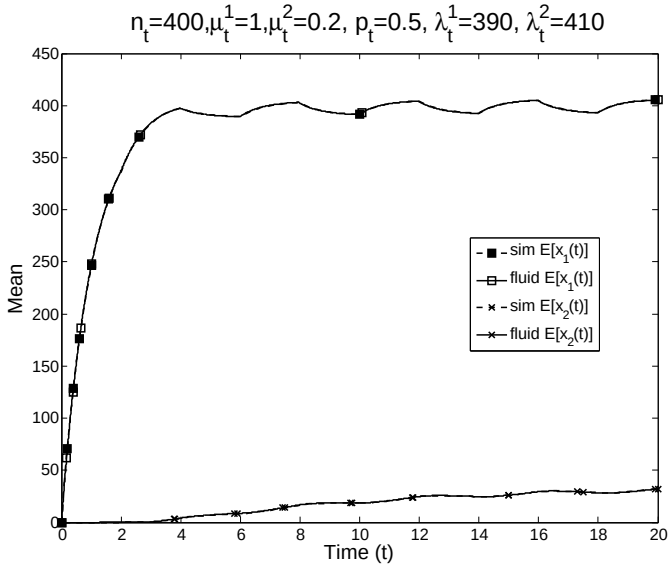


(f) Adjusted diffusion limit of exp. 4

Fig. 8. Comparison of covariance matrices, $Cov\big[X(t), X(t)\big]$

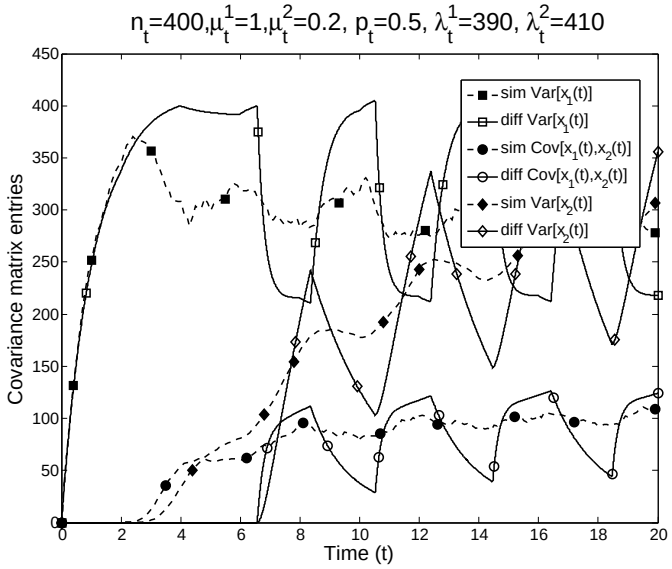


(a) Fluid limit of exp. 5

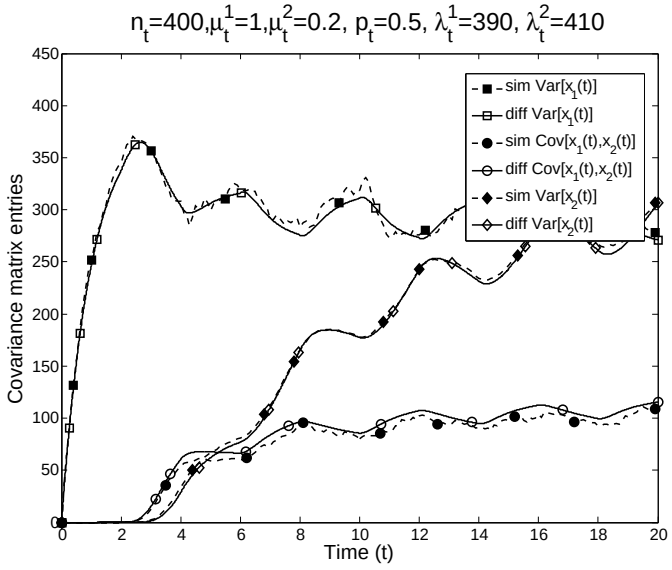


(b) Adjusted fluid limit of exp. 5

Fig. 9. Fluid limits in the nearly critically loaded phase

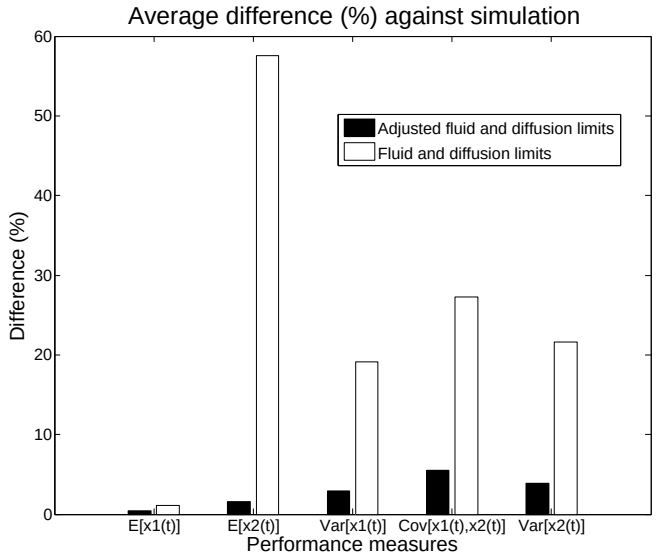


(a) Diffusion limit of exp. 5

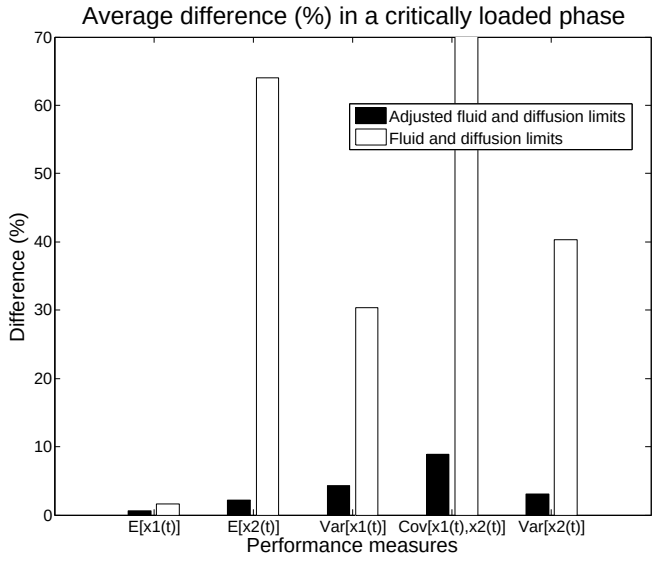


(b) Adjusted diffusion limit of exp. 5

Fig. 10. Diffusion limits in the nearly critically loaded phase



(a) Average difference for all experiments



(b) Average difference at a critically loaded point

Fig. 11. Average difference against simulation

$Cov[x_1(t), x_2(t)]$ using the standard fluid and diffusion limits. However, the graph is cropped at the 70% level for the illustration purpose. We know that those results are from our limited experiments and hence do not make an absolute conclusion about two methodologies. Nonetheless, we could not deny the fact that the adjusted limits provide accurate estimation results consistently, while the standard fluid and diffusion limits result in inconsistent accuracy.

IV.6. Chapter summary

In this chapter, we explain the fluid and diffusion limits used in the analysis of time-varying multi-server queues with abandonment and retrials and show potential problems that one faces in obtaining accuracy in the nearly critically loaded phase. To address those problems, we applied adjusted fluid and diffusion limits *specifically designed for the approximation of the multi-server queues with finite number of servers*. It turns out that the adjusted limits achieve great improvement in approximation accuracy of performance measures, which was verified by a number of numerical experiments.

CHAPTER V

PEER-BASED MULTIMEDIA SERVICE NETWORKS

In this chapter, we conduct the transient analysis of Internet-based multimedia services utilizing peer-to-peer networks. To do so, we apply both standard and adjusted limits, and compare them to emphasize the effectiveness of adjusted limits. We also provide additional analyses to obtain important intuitions on a peer network itself.

V.1. Transient analysis of peer-based service networks

The online multimedia market is growing at an unprecedented rate. This growth accompanies increasing demand for network resources (e.g. bandwidth, servers, storages, etc.) and forces a service provider, which we call a "company" for the remainder of this chapter, to equip enough resources to satisfy adequate quality of service (QoS). Currently, the market is limited to music files which do not impose significant overhead for the companies even though they require many more resources than simple web pages. The market, however, is now moving to video content (e.g. movies, dramas, online lectures, user created content, etc.) that is 10 to 100 times larger than music files. This implies that the volume of multimedia content is increasing tremendously as the market grows. In addition to the increase in volume, the demand for multimedia content tends to fluctuate according to their popularity; when popular content is created, a burst of traffic may be brought on by the demand. Therefore, under these circumstances, maintaining enough resources to serve multimedia content with a satisfactory QoS level becomes a major problem that companies must solve.

To address this problem, peer-to-peer (P2P) architecture can be a viable alternative for a company to "outsource" resources to peers instead of purchasing all the resources by itself. In other words, the company could redirect requests to peers

who have downloaded those files in the past. P2P architecture has already been proven to be stable and scalable in many previous research studies, such as Qiu and Srikant (2004), Ge *et al.* (2003), and Yang and de Veciana (2004). Furthermore, P2P applications have become some of the most dominant applications in terms of network traffic, and P2P traffic volume keeps increasing (Fraleigh *et al.* (2003), Gummadi *et al.* (2003)). Despite these benefits (stability and scalability) and the popularity of P2P architecture, it has not yet been broadly adapted to commercial companies, since it is regarded as a source of illegal content distribution attributed to current free P2P software, e.g., BitTorrent (2011), in that the company cannot control the distribution of the content. If the content's distribution could be under the control of companies, they could not only distribute network bandwidth, but also reduce the number of servers with a satisfactory service level, by adopting P2P architecture. In fact, a few companies such as Pando (2011) are operating P2P networks for content-distribution. Furthermore, even companies such as Akamai (2011) that provide more established content distribution networks by locating caching servers, also seem to be interested in P2P architecture (for example, Akamai purchased a P2P content distribution system called "Redswoosh" in 2007).

Having described the merits of peer-based networks, when operating a peer network to utilize the benefits of P2P architecture, a significant problem, however, can arise before the peer network is mature. When new content (e.g. movie) comes out, only the company's servers have the content. If not enough service capacity is prepared and the demand is large, then the company could suffer from a large queue of customers. Since new content continues to be created, the company would encounter this problem whenever new content is provided. Therefore, it is important to study the behavior of a peer network during a transient period, especially for companies that utilize P2P architecture. That is the objective of this chapter.

For peer networks, most research focuses on modeling and performance analysis of steady state behavior (Ge *et al.* (2003), Clévenot and Nain (2004), Qiu and Srikant (2004)), or optimal peer search and selection (Adler *et al.* (2005)) of a peer network itself. The literature typically deals with peer networks in a completely decentralized fashion, such as in Bittorent; they do not consider peer networks operated by commercial companies. However, our system is a hybrid scheme with a centralized dispatcher much like Napster. In addition, other research studies have not focused on transient behavior of peer networks, which is crucial for commercial companies as mentioned before. Therefore, this research is different from that in the literature, in that we focus more on the precise performance analysis of peer network transient behavior, rather than steady state behavior. In fact, Shakkottai and Johari (2010) deals with the initial burst of traffic of P2P networks. However they emphasize the insight regarding delay experienced in P2P networks using a simple ordinary differential equation which is somewhat away from the complex stochastic dynamics of the real P2P networks.

V.2. Problem description

In this section, we explain the system we consider and the mathematical model. Based on the mathematical model, we subsequently define the problem and objective.

V.2.1. System description

We consider an online entertainment company that sells digital media content via the World Wide Web. The company's servers store media content through which customers access and purchase content via the company's web site. The company operates a peer network consisting of peers who purchased these content before and are given the authorization for delivering content to new customers. The company

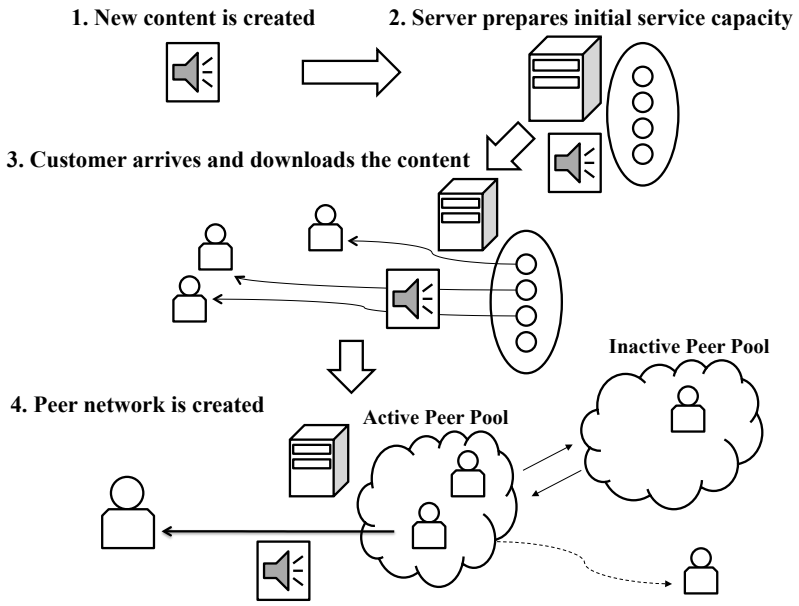


Fig. 12. System illustration

manages one queue for waiting customers and allocates a new customer in the queue to a peer when the peer becomes available. Figure 12 is a simplified illustration of our target system. When new content is created, the company prepares the initial service capacity (in terms of number of servers) to serve that content. Initially, arriving customers download the content from the company's servers. All the customers become new peers as soon as they complete download of the content so that they can share the content with customers arriving in the future. Since peers consist of users' computers, peers can move between an active peer pool and an inactive peer pool as users turn their computers on and off. Only peers in the active peer pool can serve new customers. Peers can also leave the peer network after serving a random amount of time. If the leaving peer or the peer just moving to the inactive pool is serving a customer, the customer is allocated to an available peer in the active peer

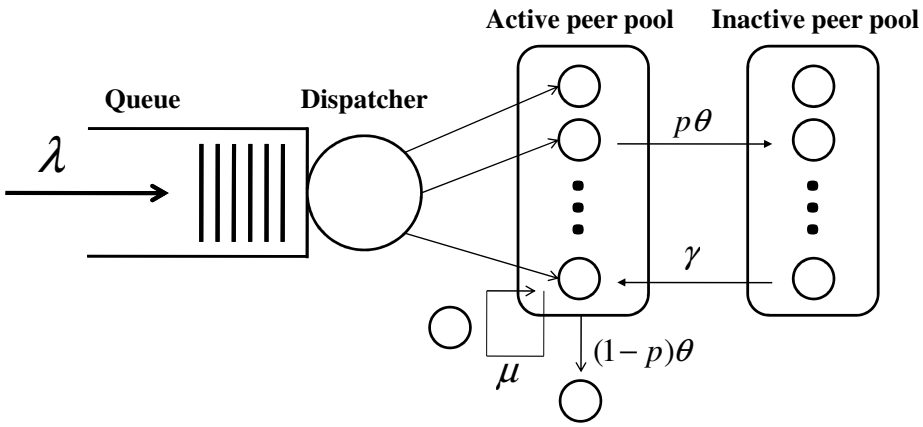


Fig. 13. Simplified system model

pool or pushed back to the queue. Notice that the peer network grows when a new peer joins and shrinks when a peer leaves. Throughout this chapter, we assume that customers arrive to the system with average rate λ per unit time, the mean service rate for each customer is μ per unit time, the on and off times of each peer are $1/\theta$ and $1/\gamma$ time units on average, respectively. When a peer leaves the active peer pool, he/she leaves the system with probability p and moves to the inactive peer pool with probability $1-p$. Note that time-varying rates would be a straightforward extension that we will show later in the chapter. We assume for mathematical tractability that the service units initially prepared by the company act like peers.

Note that we use the term "content" instead of "file" or "chunk" to indicate multimedia data. In fact, many P2P software programs divide a file into several chunks for the sake of transmission efficiency. The objective of this chapter, however, is not to analyze a specific P2P software, but to provide a methodology to model a class of queues having P2P architecture. Therefore, the content can be a file in one application and can be a chunk in another application.

V.2.2. Mathematical model

Let $X(t) = (x_1(t), x_2(t), x_3(t))'$ denote the state of the system at time t where $x_1(t)$ is the number of customers in the system, i.e. those who are waiting in the queue or are downloading the content, $x_2(t)$ is the number of peers in the active peer pool, and $x_3(t)$ is the number of peers in the inactive peer pool. We assume that all times (i.e. inter-arrival time, service time, on time and off time) follow exponential distributions with parameters λ, μ, θ, and γ, respectively. Figure 13 shows an abstract system model. We can think of peers in the active peer pool as working servers and peers in the inactive peer pool as servers on vacation. Note that waiting customers are located in one queue, which is managed by the company. Therefore, this process can be characterized as a $M/M/x_2(t)$ type queue with server vacations in which the number of servers changes over time. Here, we use Markovian assumption, i.e. Poisson arrival and exponential service time. This assumption has been used and verified in Qiu and Srikant (2004) and Yang and de Veciana (2004) by comparing real trace data from a BitTorrent network.

V.2.3. Objective

Figure 14 illustrates a typical evolution of peer networks. From Figure 14, we can define three stages based on the number of customers and peers. At the beginning of stage 1 (i.e. $t = 0$), the company prepares initial service capacity, and customers begin to arrive. All service capacity becomes full in a short time if the arrival rate is high. In this stage, the queue remains empty (as all customers are at servers). Stage 2 begins when the queue is about to be filled. Due to high arrival rates, the number of customers in the queue increases for some time. However, since the number of peers also increases rapidly, the number of peers catches up with the number of customers

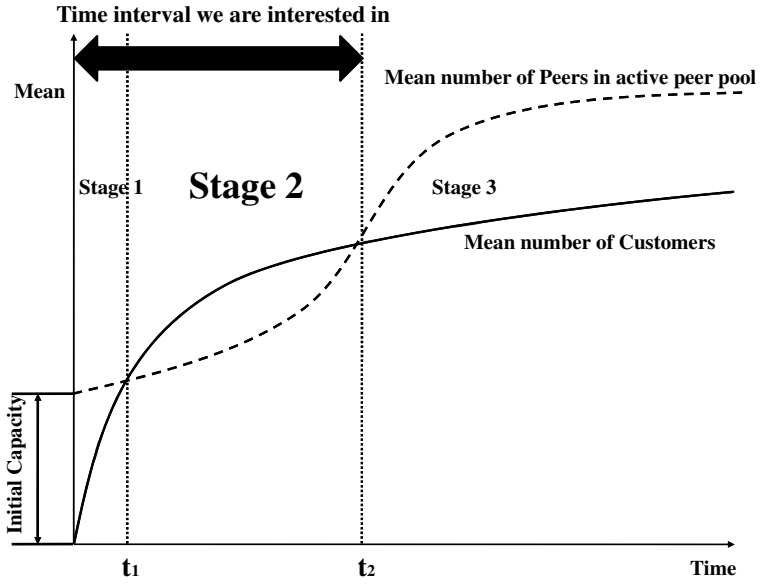


Fig. 14. Typical evolution of peer networks on average

(i.e. the queue becomes empty again) and stage 2 ends. In stage 3, the number of peers is greater than the number of customers and some peers remain idle. Once the peer network is in stage 3, we can say that the peer network is mature or stable. From the company's perspective, stage 2 is the most important stage, since queue length could grow extensively during stage 2, potentially causing significant delay to the customers and breaking the QoS conditions. In that light, the objective of this research is to characterize the dynamics of the system (the number of customers and peers) by establishing an analytical model for the transient period especially focusing on stage 1 and stage 2 rather than stage 3. Therefore, we are interested in the time interval $[0, t_2]$ provided that t_1 and t_2 are the end time points of stages 1 and 2 respectively. Understandably, because of the stochastic aspect of the system, there is some ambiguity in the definition of t_1 and t_2 which we will clarify in Section V.3.

To analyze our model, there are several approaches to consider. We choose fluid and diffusion approximations because of the following shortcomings of other methods such as Continuous Time Markov Chains (CTMC) and simulation:

- CTMC is not appropriate when the state space is more than two-dimensional, especially when the CTMC is not reversible, and the state space is unbounded. As the dimension becomes higher, it becomes extremely difficult to analyze the CTMC and hence is not scalable. In addition, even if we come up with the infinitesimal generator matrix by utilizing techniques such as Matrix Geometric Methods, it is usually good for steady state analysis but not for transient analysis. Further, if transition rates are time-varying, it makes the analysis intractable. In this research, we have three dimensions of state space and unbounded state space. If we add a reneging feature to our model, the dimension of state space would increase. Furthermore, we can consider time-varying transition rates for arrival, service, and peer's up and down. Considering these, we do not think CTMC is appropriate for the transient analysis of our model.

- Although simulation has been extensively used for performance analysis, the most common shortcoming of simulation is computational time. Usually, simulation requires a long time, even to run a single parameter combination. Therefore, to see the relationship among parameters, a significant amount of time is required. This makes it difficult to perform "what-if" analysis. Like the CTMC approach, simulation also has a scalability issue when adding more features to the existing model since whenever we add a feature, the modification of simulation code is not a trivial task. Therefore, considering time and scalability, in our opinion, simulation is also inadequate for analyzing our model.

Fluid and diffusion approximations have great advantages compared to CTMC

and simulation approaches. They are not restricted by dimensional problems and easily extend to time varying rate functions. In addition, they can be used for transient analysis. Since these approximations have analytical results, computation is faster than simulation runs and it is possible to intuitively see the effects of parameters. Thus we utilize fluid and diffusion approximations and will explain them in Sections V.3 and V.4.

V.3. Fluid and diffusion approximations

In this section, we derive fluid and diffusion limits in by Kurtz (1978) and Mandelbaum *et al.* (2002) for our problem. After developing the results, we will show the inadequacy of these approximations. The first step of this approach is to define a sequence of stochastic processes and to obtain the fluid limit by taking limit of the sequence. Fluid limit takes the role of the expected value for each time point. The second step is to obtain a diffusion limit by taking limit to the centered process multiplied by some adequate scaling factor. In Markovian networks, this centered process converges to Gaussian process under certain conditions that are described later.

Consider $X(t) = (x_1(t), x_2(t), x_3(t))'$ as defined in Section V.2.2. Assume that there is no customer and the company prepares C service units at time $t = 0$; i.e. $X(0) = (0, C, 0)'$. Then, for our model, the sample path can be constructed using the

following integral equation:

$$\begin{aligned} x_1(t) &= Y_1\Big(\int_0^t \lambda ds\Big) - Y_2\Big(\int_0^t \mu(x_1(s)\wedge x_2(s))ds\Big), \\ x_2(t) &= C + Y_2\Big(\int_0^t \mu(x_1(s)\wedge x_2(s))ds\Big) - Y_3\Big(\int_0^t p\theta x_2(s)ds\Big) \\ &\quad -Y_4\Big(\int_0^t (1-p)\theta x_2(s)ds\Big) + Y_5\Big(\int_0^t \gamma x_3(s)ds\Big), \text{ and} \\ x_3(t) &= Y_3\Big(\int_0^t p\theta x_2(s) - Y_5\Big(\int_0^t \gamma x_3(s)ds\Big), \end{aligned} \tag{32}$$

where $Y_1(\cdot)$, $Y_2(\cdot)$, $Y_3(\cdot)$, $Y_4(\cdot)$, and $Y_5(\cdot)$ are independent Poisson processes corresponding to customer arrival, service, peer's up, peer's leaving, and peer's down respectively. To apply fluid and diffusion approximations to equation (32), we accelerate the arrival rate λ and the initial service capacity C by multiplying a scaling factor η. Consider a sequence of stochastic processes $\{X^\eta(t)\}_{t\geq 0}$ so that $X^\eta(t) = (x_1^\eta(t), x_2^\eta(t), x_3^\eta(t))'$ is the solution to the following integral equation:

$$\begin{aligned} x_1^\eta(t) &= Y_1\Big(\int_0^t \eta\lambda ds\Big) - Y_2\Big(\int_0^t \eta\mu\Big(\frac{x_1^\eta(s)}{\eta}\wedge\frac{x_2^\eta(s)}{\eta}\Big)ds\Big), \\ x_2^\eta(t) &= \eta C + Y_2\Big(\int_0^t \eta\mu\Big(\frac{x_1^\eta(s)}{\eta}\wedge\frac{x_2^\eta(s)}{\eta}\Big)ds\Big) - Y_3\Big(\int_0^t \eta p\theta\frac{x_2^\eta(s)}{\eta}ds\Big) \\ &\quad -Y_4\Big(\int_0^t \eta(1-p)\theta\frac{x_2^\eta(s)}{\eta}ds\Big) + Y_5\Big(\int_0^t \eta\gamma\frac{x_3^\eta(s)}{\eta}ds\Big), \text{ and} \\ x_3^\eta(t) &= Y_3\Big(\int_0^t \eta p\theta\frac{x_2^\eta(s)}{\eta}\Big) - Y_5\Big(\int_0^t \eta\gamma\frac{x_3^\eta(s)}{\eta}ds\Big). \end{aligned} \tag{33}$$

Note that η is a scaling factor so that we obtain the fluid approximation limit by letting $\eta\to\infty$ for $\{X^\eta(t)\}$. That is described in the next theorem.

Theorem 12 (Fluid limit)**.** *Let $\bar{X}(t) = (\bar{x}_1(t), \bar{x}_2(t), \bar{x}_3(t))'$ denote the deterministic*

fluid limit corresponding to $X^\eta(t)$ *that satisfies*

$$\begin{aligned}
\bar{x}_1(t) &= \int_0^t \lambda - \mu(\bar{x}_1(s) \wedge \bar{x}_2(s))ds, \\
\bar{x}_2(t) &= C + \int_0^t \mu(\bar{x}_1(s) \wedge \bar{x}_2(s)) - \theta \bar{x}_2(s) + \gamma \bar{x}_3(s)ds, \ \textit{and} \\
\bar{x}_3(t) &= \int_0^t p\theta \bar{x}_2(s) - \gamma \bar{x}_3(s)ds
\end{aligned} \tag{34}$$

Then,

$$\lim_{\eta \to \infty} \frac{X^\eta(t)}{\eta} = \bar{X}(t) \quad a.s..$$

Proof. Let $X = (x, y, z)'$ and define $f_1(X) = \lambda$, $f_2(X) = \mu(x \wedge y)$, $f_3(X) = \theta p y$, $f_4(X) = \theta(1-p)y$, and $f_5(X) = \gamma z$. Then, equation (33) can be written as

$$X^\eta(t) = (0, C, 0)' + \sum_{i=1}^{5} l_i Y_i\Big(\int_0^t \eta f_i\Big(\frac{X^\eta(s)}{\eta}\Big)ds\Big),$$

where $l_1 = (1,0,0)'$, $l_2 = (-1,1,0)'$, $l_3 = (0,-1,1)'$, $l_4 = (0,-1,0)'$, and $l_5 = (0,1,-1)'$. Then, it is easy to verify that the $f_i(\cdot)$'s are Lipschitz and there exist ϵ_i's such that $|f_i(X)| \leq \epsilon_i(1 + |X|)$. Since $\sum |l_i|^2 \epsilon_i < \infty$, by Theorem 1, $\lim_{\eta\to\infty} X^\eta/\eta = \bar{X}(t)$ a.s.. □

Before moving to the diffusion limit, we investigate the graph of the fluid limit over time since it is closely related to the diffusion limit, which will be explained in Theorem 13. The fluid limit is deterministic and its graph is identical to Figure 14. In the original process (i.e. $X(t)$), the end time of stage 2 (denoted by t_2) is random and hard to obtain from any stopping time of stochastic process since defining the stopping time itself is ambiguous. For example, it is not possible to define the first or second time when the number of peers exceeds the number of customers as a stopping time since the number of peers and customers can meet several times around the end time of stage 1 (denoted by t_1). Therefore, without hurting our objective significantly,

we define t_1 and t_2 via the fluid limit;

$$\begin{aligned} t_1 &= \inf\{t : \bar{x}_1(t) = \bar{x}_2(t), t \geq 0\} \\ t_2 &= \inf\{t : \bar{x}_1(t) = \bar{x}_2(t), t > t_1\} \end{aligned}$$

Notice t_1 and t_2, depicted in Figure 14, for further clarification. The switching times t_1 and t_2 can be obtained directly by solving the integral equation (34). Defining t_1 and t_2 using the fluid limit is reasonable since the queue is empty at t_2 on average. Now we move our attention to the diffusion limit. For the diffusion limit, we apply Central Limit Theorem by defining the scaled centered process.

Theorem 13 (Diffusion limit)**.** *Let $D^\eta(t)$ be the scaled centered process; i.e. $D^\eta(t) = \sqrt{\eta}\big(X^\eta(t)/\eta - \bar{X}(t)\big)$. Then, we derive the diffusion limit as*

$$D(t) = \big(d_1(t), d_2(t), d_3(t)\big)' = \lim_{\eta\to\infty} D^\eta(t).$$

Define the matrices K_1, K_2, and $L(t)$ as follows:

$$\begin{aligned} K_1 &= \begin{pmatrix} -\mu & 0 & 0 \\ \mu & -\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix}, \\ K_2 &= \begin{pmatrix} 0 & -\mu & 0 \\ 0 & \mu-\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix}, \text{ and} \\ L(t) &= \begin{pmatrix} \sqrt{\lambda} & 0 & 0 \\ -\sqrt{\mu(\bar{x}_1(t)\wedge\bar{x}_2(t))} & \sqrt{\mu(\bar{x}_1(t)\wedge\bar{x}_2(t))} & 0 \\ 0 & -\sqrt{p\theta\bar{x}_2(t)} & \sqrt{p\theta\bar{x}_2(t)} \\ 0 & -\sqrt{(1-p)\theta\bar{x}_2(t)} & 0 \\ 0 & \sqrt{\gamma\bar{x}_3(t)} & -\sqrt{\gamma\bar{x}_3(t)} \end{pmatrix}'. \end{aligned}$$

Then, $D(t)$ is the solution of the following integral equation:

for $0 \le t < t_1$,

$$D(t) = \int_0^t L(s)dB(s) + \int_0^t K_1 \cdot D(s)ds, \tag{35}$$

for $t_1 \le t < t_2$,

$$D(t) = (d_1(t_1), d_2(t_1), d_3(t_1))' + \int_{t_1}^t L(s)dB(s) + \int_{t_1}^t K_2 \cdot D(s)ds, \tag{36}$$

and for $t \ge t_2$,

$$D(t) = (d_1(t_2), d_2(t_2), d_3(t_2))' + \int_{t_2}^t L(s)dB(s) + \int_0^t K_1 \cdot D(s)ds, \tag{37}$$

where $B(t)$ is a 5-dimensional standard Brownian motion.

Proof. With the same definition of X, l_i's and $f_i(\cdot)$'s as in the proof of Theorem 12, define $F(X)$ as follows:

$$F(X) = \sum_{i=1}^{5} l_i f_i(X).$$

Then, by Theorem 2, the centered process $D(t)$ satisfies the following integral equation:

$$D(t) = \sum_{i=1}^{5} l_i \int_0^t \sqrt{f_i\big(\bar{X}(s)\big)}dB(s) + \int_0^t \partial F\big(\bar{X}(s)\big) \cdot D(s)ds,$$

where $\partial F\big(\bar{X}(t)\big)$ is the gradient of $F\big(\bar{X}(t)\big)$. For $0 \le t < t_1$, (35) is straightforward. Since t_1 and t_2 have measure zero similar to what Mandelbaum *et al.* (2002) considers, we can obtain (36) for $t_1 \le t < t_2$ and (37) for $t \ge t_2$. □

Note that the diffusion limit in (35), (36), and (37) turns out to be a Gaussian process and is closely related to the fluid limit ($\bar{X}(t)$). Depending on the fluid limit, the diffusion limit changes its behavior at time points t_1 and t_2.

Theorem 13 indicates that the diffusion limit is a linear model. Therefore, we could obtain the expectation and covariance matrix of $D(t)$ in the following way.

Theorem 14 (Expectation and Covariance matrix)**.** *Let $m(t)$ denote $E\big(D(t)\big)$ and $\Sigma(t)$ denote $Cov\big(D(t), D(t)\big)$. Then, with the same definition of K_1, K_2, and $L(t)$ as in Theorem 13, $m(t)$ is the solution to the following differential equation: for $0 \le t < t_1$ or $t \ge t_2$,*

$$\frac{d}{dt}m(t) = K_1 \cdot m(t), \tag{38}$$

and for $t_1 \le t < t_2$,

$$\frac{d}{dt}m(t) = K_2 \cdot m(t). \tag{39}$$

Moreover, $\Sigma(t)$ is the unique symmetric semi-positive definite solution to the following differential equation:
for $0 \le t < t_1$ or $t \ge t_2$,

$$\frac{d}{dt}\Sigma(t) \;=\; K_1 \cdot \Sigma(t) + \Sigma(t) \cdot K_1' + L(t) \cdot L(t)', \tag{40}$$

and for $t_1 \le t < t_2$,

$$\frac{d}{dt}\Sigma(t) \;=\; K_2 \cdot \Sigma(t) + \Sigma(t) \cdot K_2' + L(t) \cdot L(t)'. \tag{41}$$

Proof. For $0 \le t < t_1$, we know that $E\big(D(0)\big) = 0 < \infty$ since $D(0) = 0$. Then, by Theorem 3, $m(t)$ and $D(t)$ satisfy (38) and (40). From (38), we also have $E\big(D(t_1)\big) < \infty$. Therefore, we can also apply Theorem 8.2.6 in Arnold (1992) and obtain (39) and (41). Since $E\big(D(t_2)\big) < \infty$, we obtain (38) and (40) for $t \ge t_2$. □

Summarizing, we established the fluid and diffusion limits. We found that the diffusion limit is a Gaussian process and that the mean vector and covariance matrix

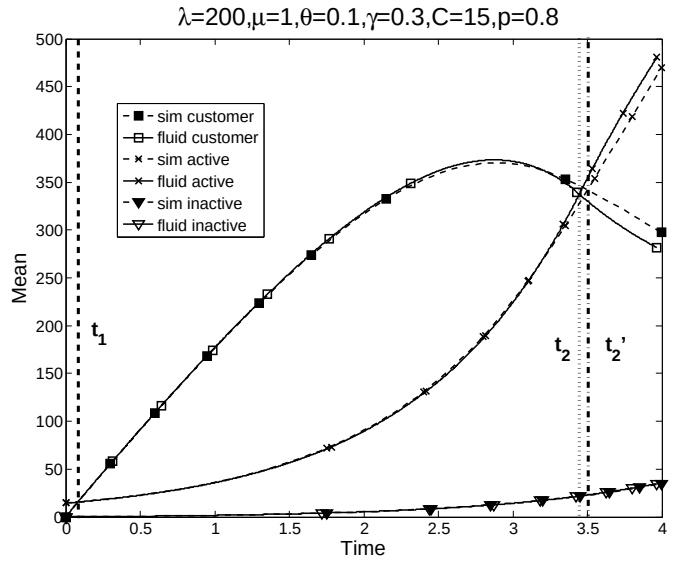
can be obtained by solving the ordinary differential equations from (38) to (41). Once we build the fluid and diffusion limits, we need to define the approximation for our original process for a fixed η. Based on the definition of $D(t)$, we use $\eta\bar{X}(t) + \sqrt{\eta}D(t)$ as an approximation of $X^\eta(t)$; see equations (5)-(7) in Chapter II. By Theorem 14, we obtain $E[D(t)] = m(t) = 0$ for all $t \geq 0$ since $m(0) = E[D(0)] = E[\lim_{n\to\infty} \sqrt{\eta}(X^\eta(0)/\eta - \bar{X}(0))] = E[\lim_{\eta\to\infty} \sqrt{\eta}(x_0 - x_0)] = 0$. Therefore,

$$\begin{aligned} E[X^\eta(t)] &\approx \eta E[\bar{X}(t)] + \eta E[D(t)] = \eta\bar{X}(t) + 0 \quad \text{and} \\ Cov[X^\eta(t), X^\eta(t)] &\approx \eta Cov[D(t), D(t)]. \end{aligned}$$

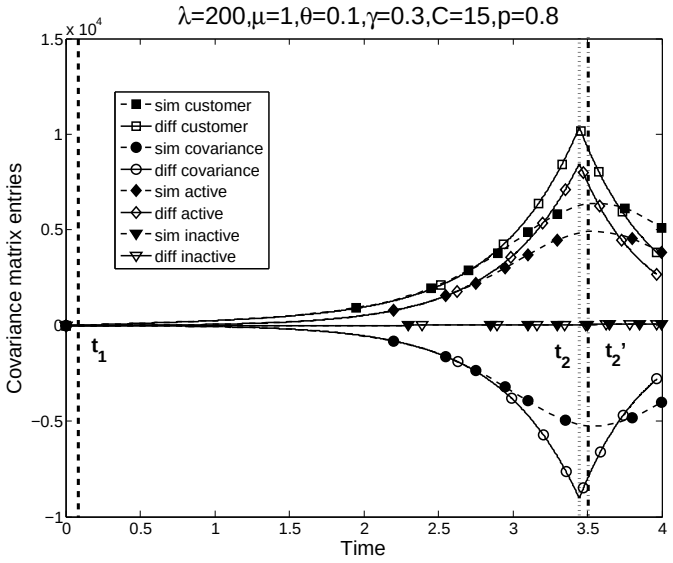
Figure 15 shows the fluid and diffusion approximation results compared with the simulation results when $\lambda = 200$, $\mu = 1$, $\theta = 0.1$, $\gamma = 0.3$, $p = 0.8$ and the initial service units is 15 ($C = 15$). Note that Figure 15 (a) is for $\bar{X}(t)$ and Figure 15 (b) for $\Sigma(t)$. The simulation result is obtained by averaging 5,000 simulation runs. We see that the fluid and diffusion limits are close to the simulation results when t is small. We, however, notice that the fluid and diffusion limits show a big difference, especially in covariance matrix entries around t_2. We find two significant problems in the fluid and diffusion limits from Figure 15. Let t_2' denote the switching time between stages 2 and 3 in the simulation result. Then,

1. The fluid limit shows some estimation error near the time t_2'. From the experiments with different parameters, we see that the fluid limit always underestimates the switching time between stages 2 and 3, i.e. $t_2 < t_2'$. This implies that at time t_2, the average number of customers is greater than the average number of active peers in the simulation results.

2. Sharp spikes are always observed in the diffusion limit at time t_2. Moreover, our diffusion limit shows significant difference from the simulation result around t_2.



(a) Mean number of customers and peers



(b) Covariance matrix entries

Fig. 15. Standard fluid and diffusion approximations

These spikes come from the sudden change of the drift matrix from K_2 and K_1 at time t_2 in Theorem 13 and this switching is caused by the non-differentiability of the $\big(\bar{x}_1(t) \wedge \bar{x}_2(t)\big)$ in the fluid limit.

Remark 2. *These problems also occur at time t_1. The process, however, starts with deterministic initial values and the time t_1 is close to the time zero. Thus, the effect of these problems is insignificant.*

To resolve these two problems, we apply the adjusted fluid and diffusion limits and will explain it in the next section.

However, before moving to the next section, we provide the steady state behavior of the P2P networks using fluid and diffusion limits since they provide accurate estimation results in steady state. From Theorem 12 and 13, we notice that $(\bar{x}_1(t) \wedge \bar{x}_2(t)) = \bar{x}_1(t)$ for $t > t_2$ and this implies that the non-differentiability of $(\bar{x}_1(t) \wedge \bar{x}_2(t))$ disappears as $t \to \infty$. Qiu and Srikant (2004) use fluid and diffusion approximations for a similar scenario and mention that their process converges to the OU process in steady state. Since they do not provide the proof for this convergence, we provide the proof (for our scenario) to show that the diffusion limit for our original process is also an OU process in steady state.

Theorem 15 (Steady State Behavior)**.** *Let $D(\infty)$ be the scaled centered process $D(t)$ defined in Theorem 13 when $t \to \infty$. Then, for $0 \leq p < 1$, $D(\infty)$ is a three-dimensional OU process with the drift matrix given by*

$$K = \begin{pmatrix} -\mu & 0 & 0 \\ \mu & -\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix}$$

and the diffusion coefficient matrix given by

$$L = \begin{pmatrix} \sqrt{\lambda} & -\sqrt{\lambda} & 0 & 0 & 0 \\ 0 & \sqrt{\lambda} & -\sqrt{\lambda p/(1-p)} & -\sqrt{\lambda} & \sqrt{\lambda p/(1-p)} \\ 0 & 0 & \sqrt{\lambda p/(1-p)} & 0 & -\sqrt{\lambda p/(1-p)} \end{pmatrix}.$$

Proof. When $t > t_2$, the drift matrix is given by K. By solving differential equations in (34) for $t > t_2$ and taking $t \to \infty$, we obtain

$$\lim_{t\to\infty} \bar{x}_1(t) = \frac{\lambda}{\mu}, \tag{42}$$

$$\lim_{t\to\infty} \bar{x}_2(t) = \frac{\lambda}{(1-p)\theta}, \text{ and} \tag{43}$$

$$\lim_{t\to\infty} \bar{x}_3(t) = \frac{\lambda p}{(1-p)\gamma}. \tag{44}$$

Then, by Theorem 13 and equations (42)-(44), we have

$$L = \begin{pmatrix} \sqrt{\lambda} & -\sqrt{\lambda} & 0 & 0 & 0 \\ 0 & \sqrt{\lambda} & -\sqrt{\lambda p/(1-p)} & -\sqrt{\lambda} & \sqrt{\lambda p/(1-p)} \\ 0 & 0 & \sqrt{\lambda p/(1-p)} & 0 & -\sqrt{\lambda p/(1-p)} \end{pmatrix}.$$

□

Remark 3. *Notice that the steady state number of customers, active peers, and inactive peers via equations (42)-(44) are respectively* λ/μ, $\lambda/((1-p)\theta)$, *and* $\lambda p/((1-p)\gamma)$. *The simulations also converge to the same values.*

V.4. Adjusted fluid and diffusion limits

In the previous section, we saw that spikes in the diffusion limit are caused by the non-differentiability of the "min" function ($\wedge$) in the fluid limit. In addition to non-differentiability, notice that the "min" function causes underestimation of t_2 in the fluid limit. From the following simple lemma, we explain why it occurs.

Lemma 4. *Let X and Y be random variables such that $E(X) < \infty$ and $E(Y) < \infty$. Then,*

$$E[(X \wedge Y)] \leq (E(X) \wedge E(Y)).$$

Recall that when solving equation (34) in Theorem 12, we actually solve the following differential equations:

$$\begin{aligned} \frac{d}{dt}\bar{x}_1(t) &= \lambda - \mu(\bar{x}_1(t) \wedge \bar{x}_2(t)), & (45) \\ \frac{d}{dt}\bar{x}_2(t) &= \mu(\bar{x}_1(t) \wedge \bar{x}_2(t)) - \theta\bar{x}_2(t) + \gamma\bar{x}_3(t), \text{ and} & (46) \\ \frac{d}{dt}\bar{x}_3(t) &= p\theta\bar{x}_2(t) - \gamma\bar{x}_3(t). \end{aligned}$$

In Section V.3, for any time point t, we regard $E\big[X^\eta(t)\big]$ as $\eta\bar{X}(t)$ (i.e. $\eta(\bar{x}_1(t) \wedge \bar{x}_2(t))$ as $\big(E[x_1^\eta(t)] \wedge E[x_2^\eta(t)]\big)$). We, however, observe $E\big[(x_1^\eta(t) \wedge x_2^\eta(t))\big]$ rather than $\big(E[x_1^\eta(t)] \wedge E[x_2^\eta(t)]\big)$ in simulations and from Lemma 4, $E\big[(x_1^\eta(t) \wedge x_2^\eta(t))\big]$ is less than or equal to $\big(E[x_1^\eta(t)] \wedge E[x_2^\eta(t)]\big)$ $\forall t \in [0, \infty)$. Therefore, we can verify that the increasing rate of $\eta\bar{x}_1(t)$ is less than the increasing rate of $E[x_1^\eta(t)]$ in simulations, and the increasing rate of $\eta\bar{x}_2(t)$ is greater than the increasing rate of $E[x_2^\eta(t)]$ in simulations from (45) and (46). In order to improve estimation accuracy, we apply the adjusted fluid and diffusion limits to this problem. For a fixed η (or a fixed number of servers), we obtain $g_i^\eta(\cdot,\cdot)$'s using Gaussian adjustment described in Chapter III as follows:

Define

$$\begin{aligned} \Phi^\eta(t) &= \Phi(0, E[x_1^\eta(t) - x_2^\eta(t)], \sigma^\eta(t)) \text{ and} \\ \phi^\eta(t) &= \phi(0, E[x_1^\eta(t) - x_2^\eta(t)], \sigma^\eta(t)), \end{aligned}$$

where $\sigma^\eta(t)^2$ is the variance of $x_1^\eta(t) - x_2^\eta(t)$, and $\Phi(a,b,c)$ and $\phi(a,b,c)$ are function

values at point a of the Gaussian CDF and PDF respectively with mean b and standard deviation c. For $x = (x_1, x_2, x_3)'$,

$$
\begin{aligned}
g_1^\eta(t,x) &= \eta\lambda, \\
g_2^\eta(t,x) &= \mu\big(\Phi^\eta(t)x_1 + (1-\Phi^\eta(t))x_2 - \sigma^\eta(t)^2\phi^\eta(t)\big), \\
g_3^\eta(t,x) &= p\theta y, \\
g_4^\eta(t,x) &= (1-p)\theta x_2, \text{ and} \\
g_5^\eta(t,x) &= \gamma x_3.
\end{aligned}
$$

The derivation of $g_2^\eta(t,x)$ is provided in Appendix B.1.

Let $Z^{\eta,\nu}(t) = (z_1^{\eta,\nu}(t), z_2^{\eta,\nu}(t), z_3^{\eta,\nu}(t))'$. With $g_i^\eta)(\cdot,\cdot)$'s above, define a new sequence, $\{Z^{\eta,\nu}(t)\}_{\nu\geq 1}$, as follows:

$$
\begin{aligned}
z_1^{\eta,\nu}(t) &= Y_1\Big(\int_0^t \nu g_1^\eta\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big) - Y_2\Big(\int_0^t \nu g_2^\eta\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big), \\
z_2^{\eta,\nu}(t) &= \nu\eta C + Y_2\Big(\int_0^t \nu g_2^\eta\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big) - Y_3\Big(\int_0^t \nu g_3^\eta\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big) \\
&\quad -Y_4\Big(\int_0^t \nu g_4^\eta\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big), \text{ and} \qquad (47) \\
z_3^{\eta,\nu}(t) &= Y_3\Big(\int_0^t \nu g_3^\eta\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big) - Y_5\Big(\int_0^t \nu g_5^\eta\Big(s, \frac{Z^{\eta,\nu}(s)}{\nu}\Big)ds\Big).
\end{aligned}
$$

Then, by taking $\nu \to \infty$, we derive the adjusted fluid limit.

Theorem 16 (Adjusted fluid limit)**.** *If* $\lim_{\nu\to\infty} Z^{\eta,\nu}(0)/\nu = \bar{Z}^\eta(0)$*, then*

$$
\lim_{\nu\to\infty} \frac{Z^{\eta,\nu}(t)}{\nu} = \bar{Z}^\eta(t) \quad a.s. \qquad (48)
$$

where $\bar{Z}^\eta(t)$ is the solution to the following integral equation:

$$\begin{aligned}
\bar{z}_1^\eta(t) &= \int_0^t \lambda - \mu\big\{\Phi^\eta(s)\bar{z}_1^\eta(s) + (1-\Phi^\eta(s))\bar{z}_2^\eta(s) - \sigma^\eta(s)^2\phi^\eta(s)\big\}ds, \qquad (49)\\
\bar{z}_2^\eta(t) &= \eta C + \int_0^t \mu\big\{\Phi^\eta(s)\bar{z}_1^\eta(s) + (1-\Phi^\eta(s))\bar{z}_2^\eta(s) - \sigma^\eta(s)^2\phi^\eta(s)\big\}\\
&\quad -\theta\bar{z}_2^\eta(s)ds, \ \textit{and} \qquad (50)\\
\bar{z}_3^\eta(t) &= \int_0^t p\theta\bar{z}_2^\eta(s) - \gamma\bar{z}_3^\eta(s)ds. \qquad (51)
\end{aligned}$$

This convergence holds uniformly on any compact time intervals.

As mentioned in Section V.3, sharp spikes in covariance matrix entries are caused by the sudden change of the drift matrix such as the change

$$\begin{pmatrix} 0 & -\mu & 0 \\ 0 & \mu-\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix} \to \begin{pmatrix} -\mu & 0 & 0 \\ \mu & -\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix}.$$

If we use the adjusted fluid limit obtained from equations (49)-(51), we can eliminate the non-differentiability of rate functions and obtain a new drift matrix $K^\eta(t)$ and a diffusion coefficient matrix $L^\eta(t)$ as follows:

$$\begin{aligned}
K^\eta(t) &= \begin{pmatrix} -\mu\cdot\Phi^\eta(t) & -\mu\cdot\big(1-\Phi^\eta(t)\big) & 0 \\ \mu\cdot\Phi^\eta(t) & \mu\cdot\big(1-\Phi^\eta(t)\big)-\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix},\\
L^\eta(t) &= \begin{pmatrix} \sqrt{\lambda} & -\sqrt{\mu\alpha^\eta(t)} & 0 & 0 & 0 \\ 0 & \sqrt{\mu\alpha^\eta(t)} & -\sqrt{p\theta\bar{z}_2^\eta(t)} & -\sqrt{(1-p)\theta\bar{z}_2^\eta(t)} & \sqrt{\gamma\bar{z}_3^\eta(t)} \\ 0 & 0 & \sqrt{p\theta\bar{z}_2^\eta(t)} & 0 & -\sqrt{\gamma\bar{z}_3^\eta(t)} \end{pmatrix},
\end{aligned}$$

where $\alpha^\eta(t) = \Phi^\eta(t)\bar{z}_1^\eta(t) + (1-\Phi^\eta(t))\bar{z}_2^\eta(t) - \sigma^\eta(t)^2\phi^\eta(t)$.

From the definition of $\Phi^\eta(t)$, it is a Gaussian distribution function and is differentiable

with respect to $\bar{z}_1^\eta(t)$ and $\bar{z}_2^\eta(t)$. Hence both $\Phi^\eta(t)$ and $\alpha^\eta(t)$ are differentiable with respect to $\bar{z}_1^\eta(t)$ and $\bar{z}_2^\eta(t)$, and we get rid of the differentiability issue in $K(t)$ and $L(t)$. Now, we are ready to derive the adjusted diffusion limit for the P2P networks by considering a sequence of scaled centered processes, $\{V^{\eta,\nu}(t)\}_{\nu\geq 1}$.

Theorem 17 (Adjusted diffusion limit)**.** *Define $V^{\eta,\nu}(t)$ as $\sqrt{\nu}(Z^{\eta,\nu}(t)-\bar{Z}^\eta(t))$. Then*

$$\lim_{\nu\to\infty} V^{\eta,\nu}(t) = V^\eta(t) \quad \textit{in distribution,} \tag{52}$$

and $V^\eta(t)$ satisfies the following integral equation:

$$V^\eta(t) = V^\eta(0) + \int_0^t L^\eta(s)dB(s) + \int_0^t K^\eta(s)V^\eta(s)ds,$$

where $B(t)$ is a 5-dimensional standard Brownian motion.

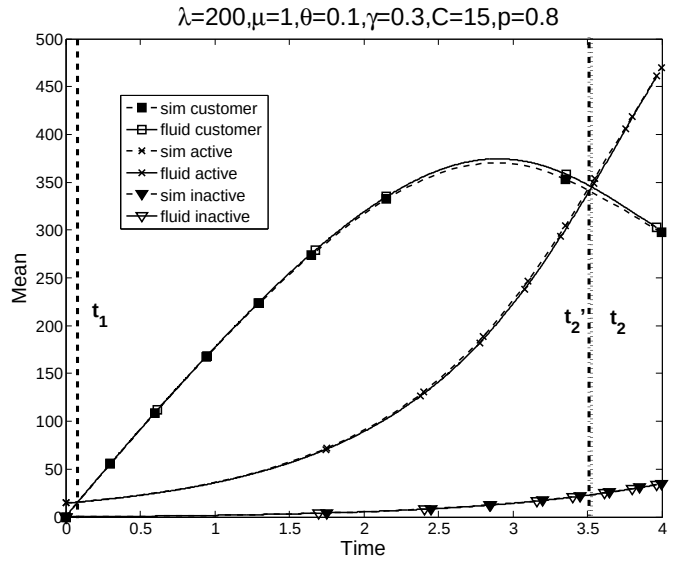
We obtain the covariance matrix of the diffusion limit by solving equation (53) from Theorem 14.

$$\frac{d}{dt}\Sigma(t) \;=\; K(t)\cdot\Sigma(t) + \Sigma(t)\cdot K'(t) + L(t)\cdot L(t)', \tag{53}$$

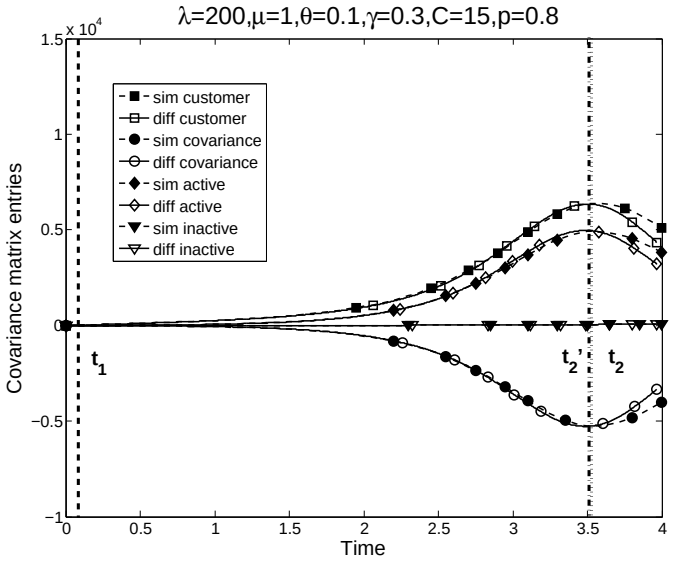
where $\Sigma(t)$ is the covariance matrix defined in Theorem 14.

Figure 16 shows the results from the adjusted fluid and diffusion limits with same parameters in Figure 15. From Figure 16, we see that the fluid limit is almost the same as the simulation results. For the covariance matrix entries, sharp spikes disappear and the accuracy is also improved. In fact, the accuracy of covariance matrix entries is not always improved much for all $t > 0$, but they are quite accurate before t_2. The fluid limit, however, shows great accuracy regardless of the values of parameters.

Remark 4. *We consider the constant rates for arrival, service, peer's up and down times. However, the fluid and diffusion limits can extend to time-varying rates by*



(a) Mean number of customers and peers



(b) Covariance matrix entries

Fig. 16. Adjusted fluid and diffusion approximations with adjustment

Table 3. Parameters used in two examples

No.	λ	μ	θ	γ	p	C
Example 1	100	1	0.2	0.5	0.7	10
Example 2	400	1	0.1	0.2	0.9	25

substituting λ, μ, θ, and γ with $\lambda(t)$, $\mu(t)$, $\theta(t)$, and $\gamma(t)$ since Theorem 12-14 do not require λ, μ, θ, and γ to be constant functions of t. Furthermore, in Markovian queueing systems, most of the non-differentiabilities of the rate functions are from the use of "min" function. Therefore, we can apply this Gaussian-based adjustment to more general Markovian applications.

V.5. Numerical results

In this section, we provide numerical examples to verify our results obtained through Sections V.3 and V.4. We show more numerical experiments to compare the adjusted fluid and diffusion limits (described in Section V.4) with the standard fluid and diffusion limits (described in Section V.3) in Section V.5.1. In addition to this, we provide some numerical experiments when the rate functions vary over time in Section V.5.2.

V.5.1. Comparison between the standard and adjusted limits

We provide two examples to demonstrate that the adjusted limits outperform the standard limits on $[0, t_2]$. The parameters we use in the examples are summarized in Table 3. We have a criterion to determine parameter values for our problem. In order for a company to take advantage of peer-based networks, the following conditions should be met.

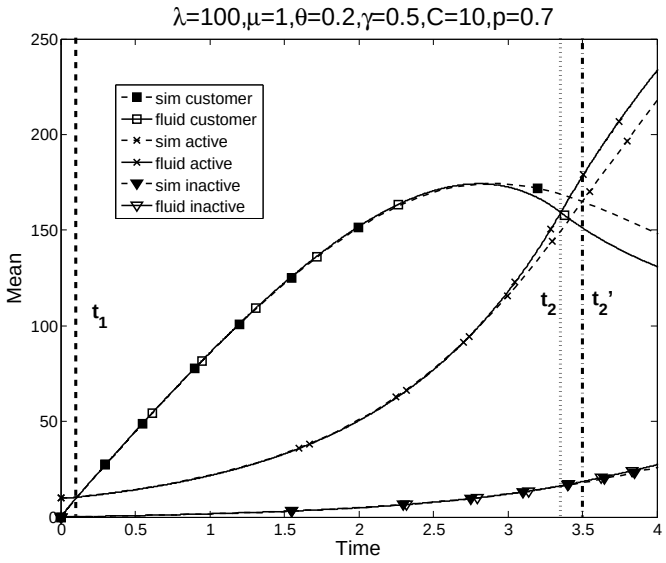
1. Customer arrival rates should be fairly large. If not, there is no need to outsource network traffic.

2. The service rate of each peer is much smaller than the customer arrival rate. If not, only a few peers are enough to cover the traffic, and then outsourcing traffic does not make sense. We assume a large peer-network (more than 100 peers).

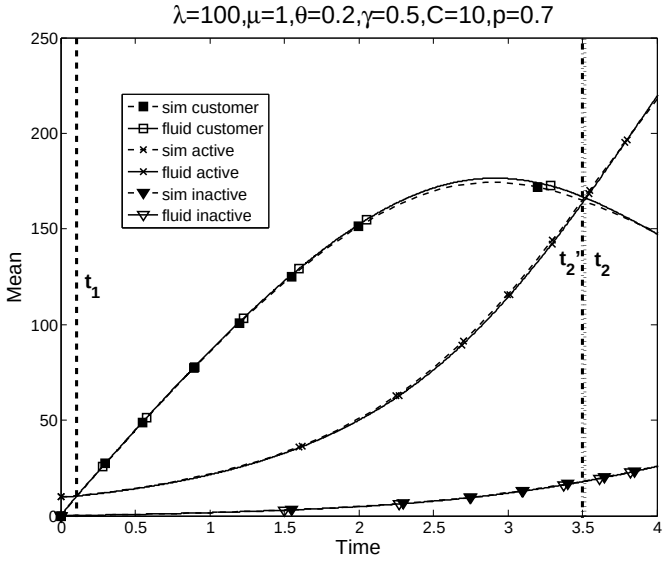
3. Each peer stays relatively long time to serve other customers, i.e. each peer serves more than 3-5 customers. If not, managing contents delivery becomes hard, and it reduces the benefit of outsourcing.

However, parameter values are selected arbitrarily based on the above conditions. We conducted 5,000 simulation runs for each example and compared the simulation results with the results of the standard and adjusted limits to see how accurate each limit is. Figures 17 and 18 illustrate the comparison of mean numbers and covariance matrix entries with the setting of example 1. Figures 19 and 20 show the results for example 2. In both examples, the standard limits show inaccuracy in estimating both expected values and covariance matrix entries. As mentioned in Section V.3, we see that the standard limits always underestimate t_2. For covariance matrix entries, the standard limits show more than 100% errors at $t = t_2$ in both examples. In contrast, the adjusted limits reasonably well estimate t_2, especially as the arrival rate becomes higher, which is desirable for the real applications. Although the adjusted limits show some errors in covariance matrix entries, the errors are less than 25% in example 1 and less than 5% in example 2. Therefore, from these two examples, we can verify that the adjusted limits are more suitable for the transient analysis than the standard limits. We obtained similar results for all the numerical experiments we performed.

Now, we move to the effects of parameters λ and p. Although the other parameters

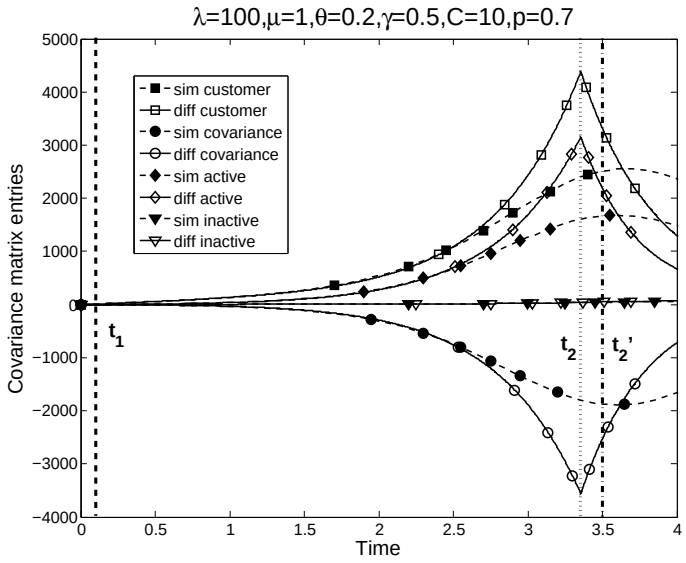


(a) Mean number of customers and peers of standard limit

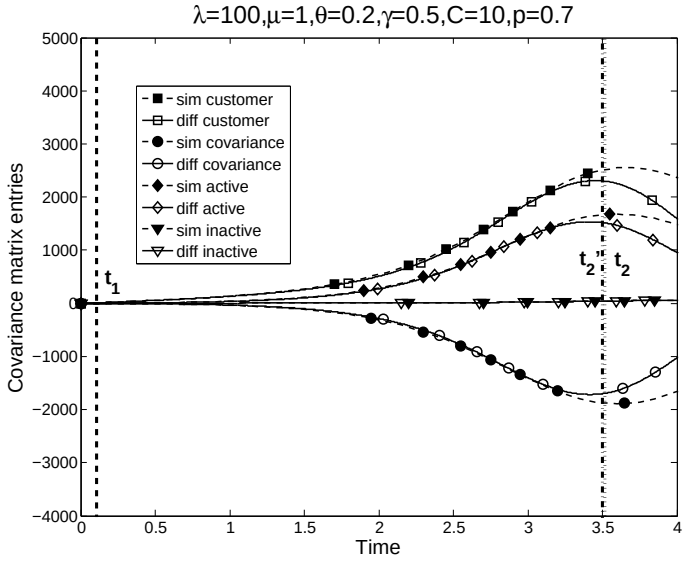


(b) Mean number of customers and peers of adjusted limit

Fig. 17. Comparison of mean numbers between standard and adjusted limits in Example 1

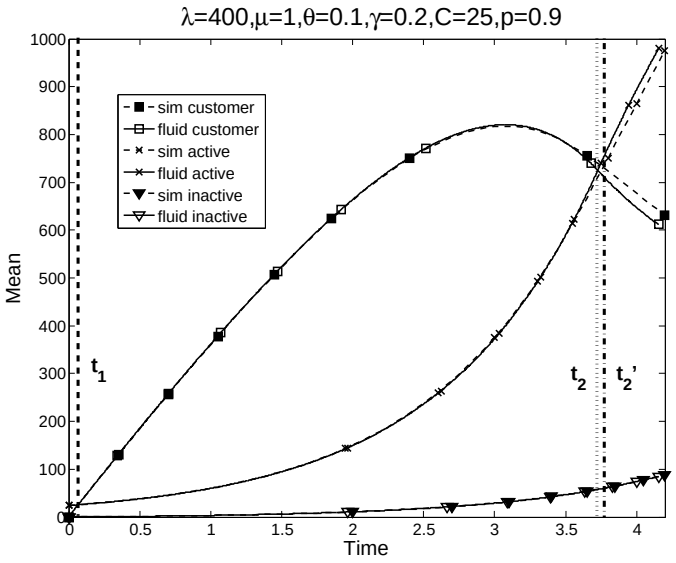


(a) Covariance matrix entries of standard limit

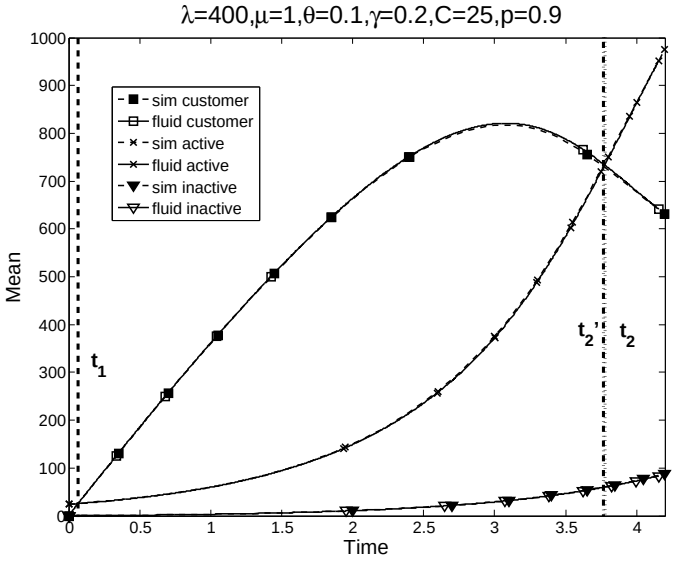


(b) Covariance matrix entries of adjusted limit

Fig. 18. Comparison of covariance matrices between standard and adjusted limits in Example 1

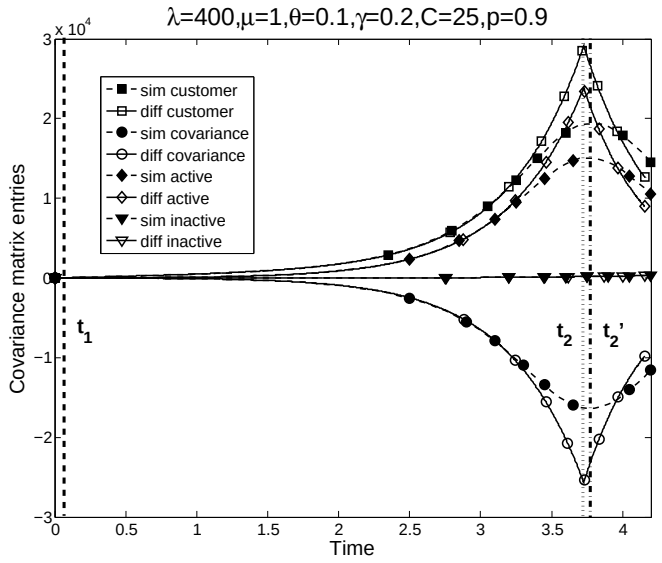


(a) Mean number of customers and peers of standard limit

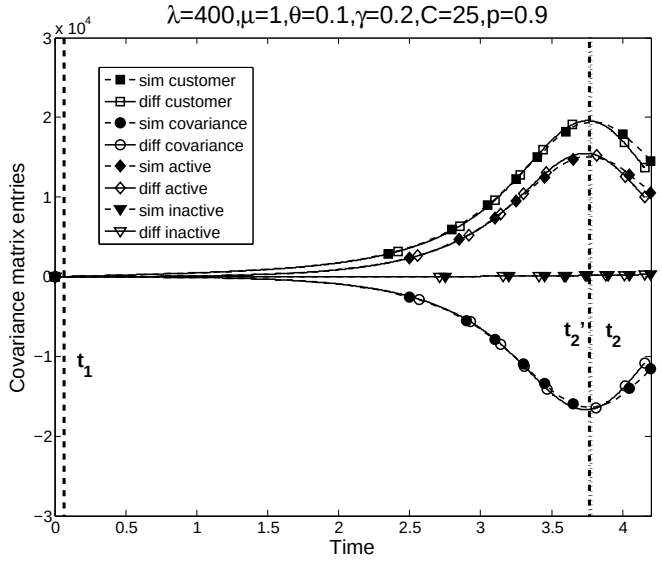


(b) Mean number of customers and peers of adjusted limit

Fig. 19. Comparison of mean numbers between standard and adjusted limits in Example 2



(a) Covariance matrix entries of standard limit



(b) Covariance matrix entries of adjusted limit

Fig. 20. Comparison of covariance matrices between standard and adjusted limits in Example 2

are also important, the arrival rate (λ) and the probability of residing in the system (p), i.e. going to inactive queue, are more interesting due to the following reasons:

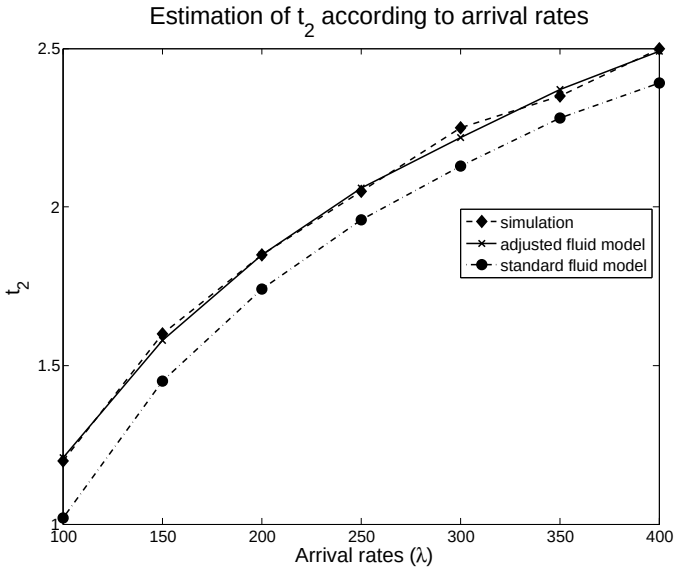
- The arrival rate implies the demand for the content. When operating a peer network, preparing a burst of the demand is crucial. Therefore, it is important to see when to reach stage 3 and how many peers (customers also) reside in the system at the end of stage 2, according to the arrival rates.

- The probability of residing in the system determines the current and potential service capacity. If $p = 0$, there is no peer in the inactive peer pool. In this case, service capacity thoroughly depends on the number of peers in the active peer pool. If $p = 1$, no peer leaves the system and the current and potential service capacity continues to increase.

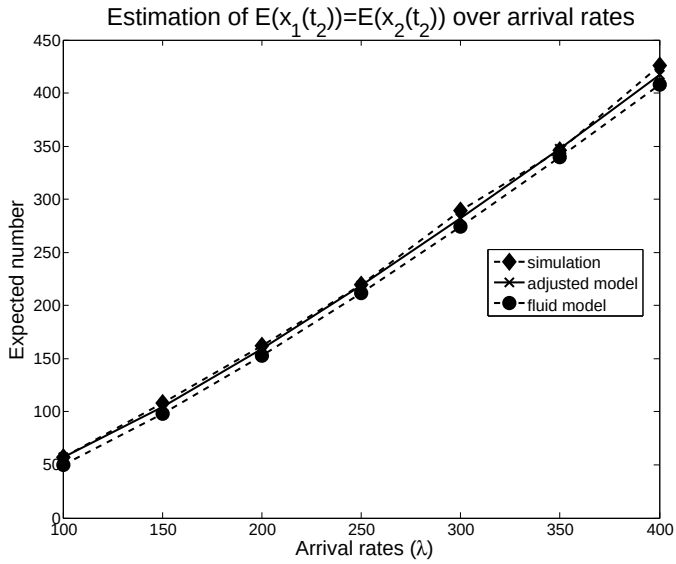
Figures 21 and 22 show the changes of t_2 and $E[X(t_2)]$ over λ and p respectively. As seen in Figure 21, t_2 and $E[X(t_2)]$ increase according to λ. This implies that if a content is popular, more time and peers are required to enter stage 3. For the effect of residing probability p, we can see that t_2 and $E[x_1(t_2)](= E[x_2(t_2)])$ decrease according to p whereas $E[x_3(t_2)]$ increases. This implies that increasing potential service capacity (i.e. number of inactive peers) accelerates the increasing rate of the number of peers so that it enables our system to reach stage 3 earlier. In addition to these observations, we see that the adjusted fluid limit provides more accurate t_2 and $E[X(t_2)]$ than the standard fluid limit.

V.5.2. Time-varying rate functions

In Remark 4, we mentioned that fluid and diffusion approximations can be extended to time varying rate functions; i.e. arrival rate is $\lambda(t)$, service rate is $\mu(t)$, and peer's up and down times are $1/\theta(t)$ and $1/\gamma(t)$ on average, respectively. In this section,

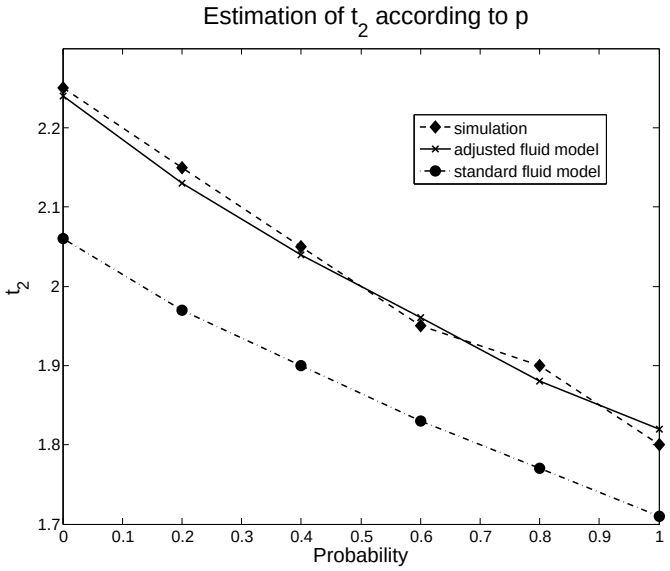


(a) Estimation of t_2 according to λ

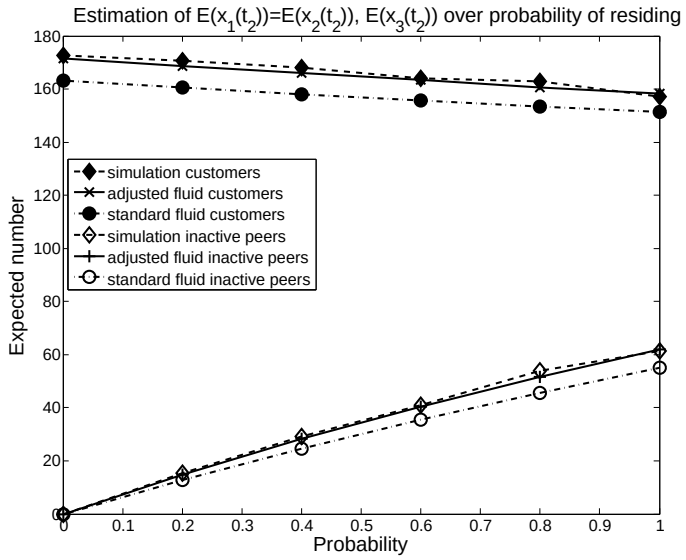


(b) Estimation of $E(x_1(t_2))$ according to λ

Fig. 21. Estimation of t_2 and $E(x_1(t_2))$ according to λ

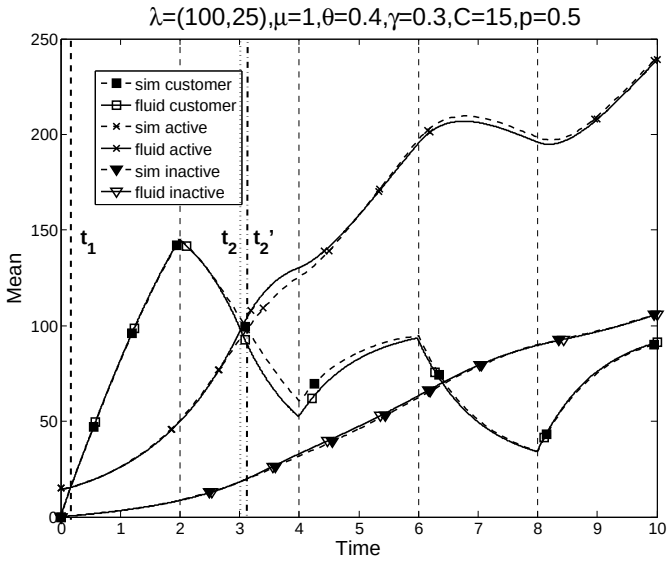


(a) Estimation of t_2 according to p

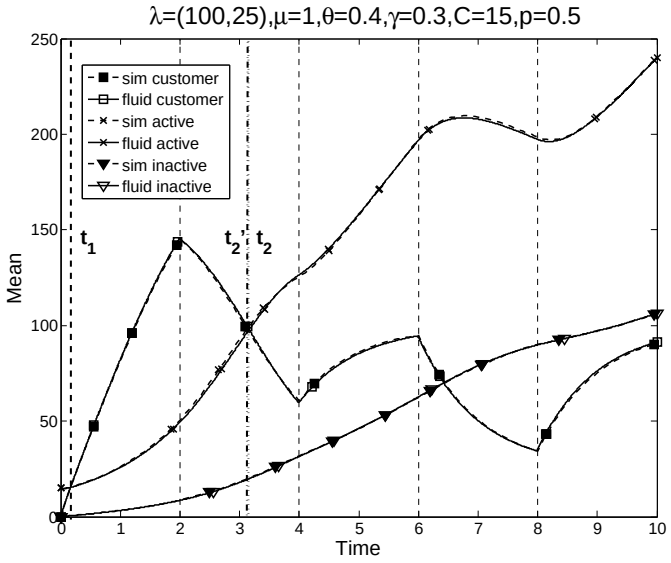


(b) Estimation of $E(x_1(t_2))$ and $E(x_3(t_2))$ according to p

Fig. 22. Estimation of $E(X(t_2))$ according to p

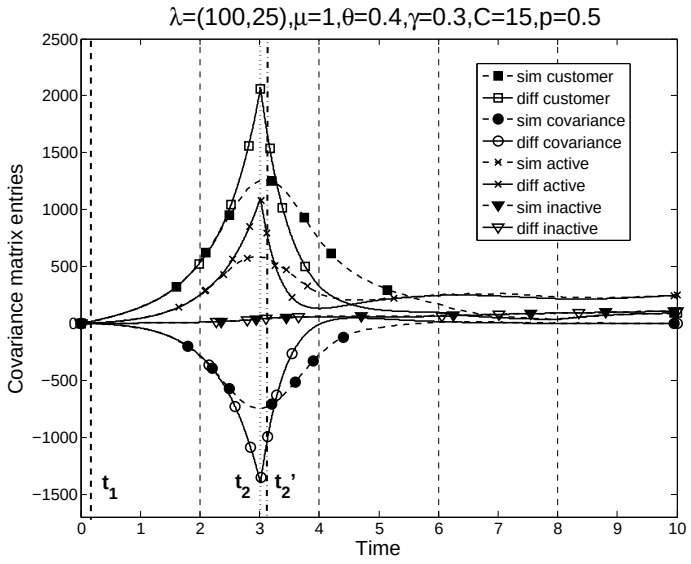


(a) Standard limit with time-varying arrival rate

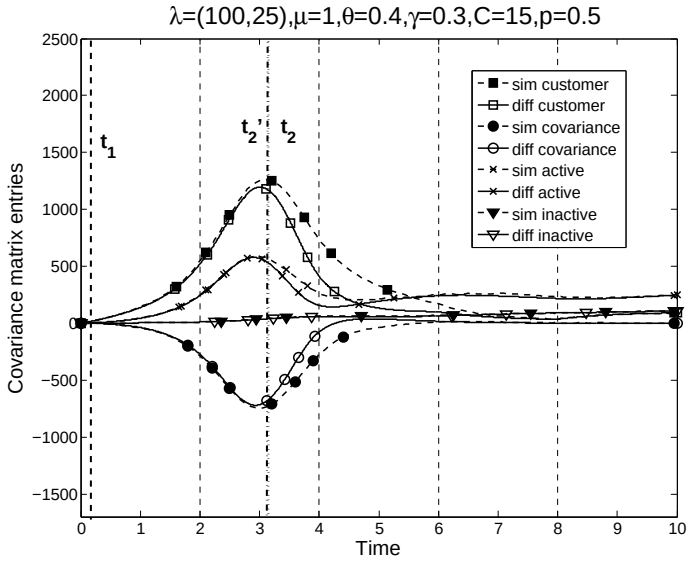


(b) Adjusted limit with time-varying arrival rate

Fig. 23. Mean number with alternating arrival rates between 100 and 25



(a) Standard limit with time-varying arrival rate



(b) Adjusted limit with time-varying arrival rate

Fig. 24. Covariance matrix with alternating arrival rates between 100 and 25

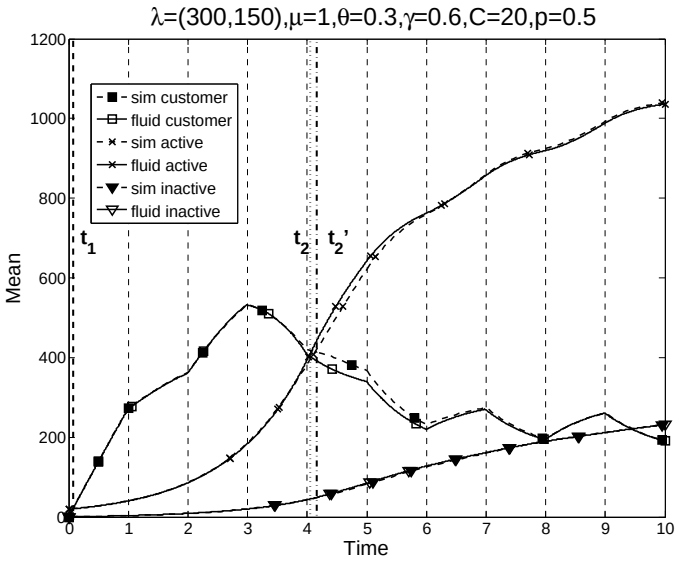
we show two numerical examples in that the arrival rate changes over time (μ, θ, and γ are held constant over time only for illustration purposes). Figures 23 and 24 show the mean and covariance matrix entries of the number of customers and peers with the arrival rate alternating between 100 and 25 every two time units. We apply both the standard and adjusted limits and compare them with simulation results. As seen in Figure 23, the adjusted limit gives quite accurate results in all ranges of time intervals, whereas the standard limit shows some error around $t \in [1.2, 2.5]$ and gives accurate results after $t > 3$. For the standard fluid limit, note that $(\bar{x}_1(t) \wedge \bar{x}_2(t))$ changes the value from $\bar{x}_2(t)$ to $\bar{x}_1(t)$ near $t = 1.5$ and after that it remains in $\bar{x}_1(t)$. Therefore, we can explain the reason for this phenomenon by Theorem 15 and Lemma 4, similar to the case of the constant rate functions. For the covariance matrix entries, both the standard and adjusted limits show shapes similar to the case of constant rates functions. Although the adjusted diffusion limit also shows errors, we can see that the accuracy is significantly improved compared with the standard limit, especially before t_2 (recall the definition of t_2 in Section V.3). In this example, we use the piecewise constant arrival rate function. Vertical dotted lines indicate the times when the arrival rate changes. Note that the change in arrival rate immediately forms the peak point of the mean number of customers, whereas it imposes some delay for the mean number of active peers to reach its peak point. In the second example, we consider heavier traffic and more frequent changes in arrival rates; the arrival rate is alternating between 300 and 150 every one time unit. As shown in Figures 25 and 26, we observe results similar to the first example. The standard fluid limit shows inaccuracy around $t \in [3.7, 6]$ while the adjusted fluid limit provides an excellent estimation. The adjusted diffusion limit is almost exact for $t < t_2$ but shows inaccuracy after t_2 just like the first example. From the examples, we can think that our adjusted fluid and diffusion limits work great during the time

interval we are interested in, i.e. $0 \leq t \leq t_2$.

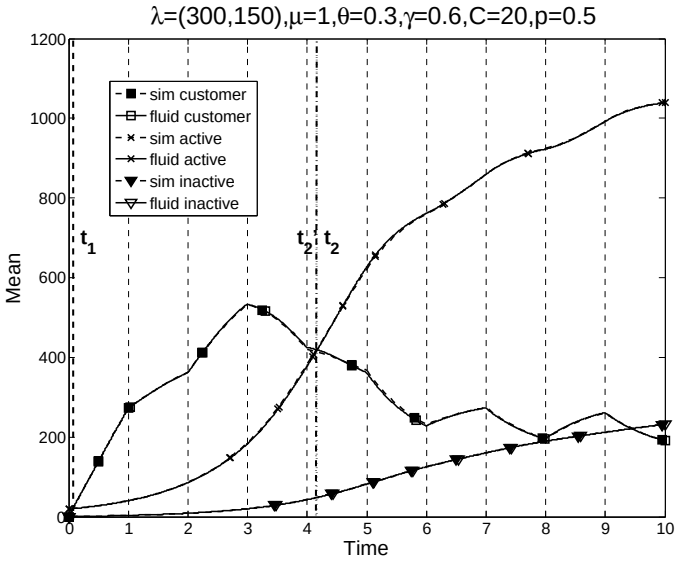
V.6. Chapter summary

In this chapter, we analyze the transient behavior of a peer network that could possibly be operated by a commercial company. We initially utilize standard fluid and diffusion approximations to build a model for peer networks. Using them, we show that the diffusion model turns out to be a three dimensional OU process in steady state. For the transient analysis, we focus on stages 1 and 2 (refer to Figure 14) when the peer network is not mature and the number of customers exceeds the number of peers such that the company is able to satisfy the QoS level; after t_2, when stage 3 begins, the number of customers becomes less than the number of active peers on average, that is, the queue is empty. Yet, we observe that standard fluid and diffusion approximations show great inaccuracy around t_2 which is caused by the non-differentiability of "min" function. To resolve this problem, we apply adjusted fluid and diffusion approximations. We replace the standard fluid model with the adjusted model and it turns out that the non-differentiability of the drift matrix in the diffusion model disappears.

To validate the adjusted models, we provide a number of examples and see the adjusted models outperform the standard models in terms of accuracy, especially before t_2 as desired. Moreover, we provide several numerical examples to see the effects of parameters and also show that the extension to time-varying rate functions is quite straightforward. From the numerical experiments, we see that higher arrival rate causes larger t_2 values and the expected number of customers (peers) at t_2. In addition, we provide other insightful numerical analysis. For example, we see that higher sojourn probability decreases t_2 values, whereas the expected number of cus-

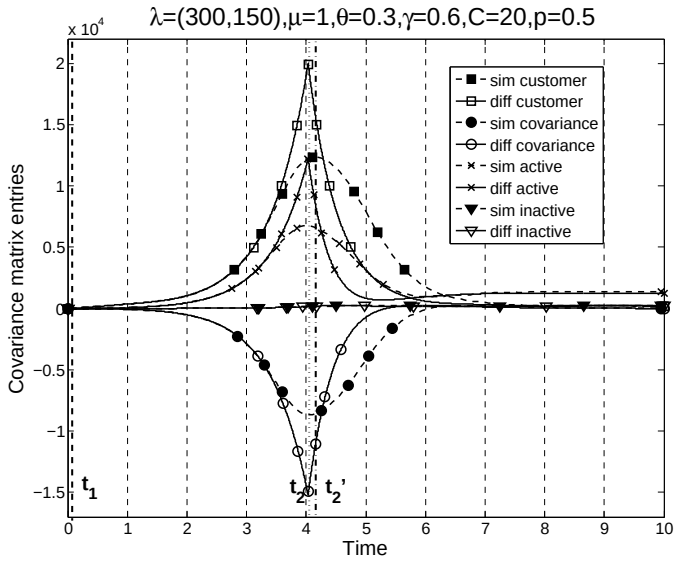


(a) Standard limit with time-varying arrival rate

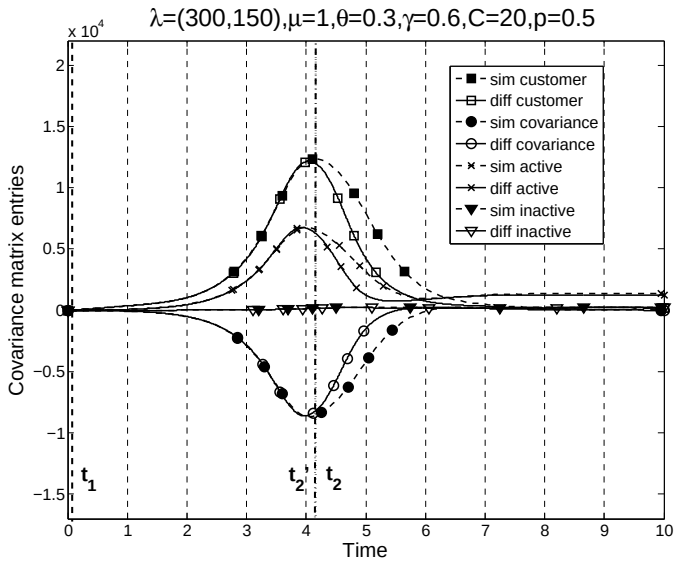


(b) Adjusted limit with time-varying arrival rate

Fig. 25. Mean number with alternating arrival rates between 300 and 150



(a) Standard limit with time-varying arrival rate



(b) Adjusted limit with time-varying arrival rate

Fig. 26. Covariance matrix with alternating arrival rates between 300 and 150

tomers does not decrease much. For time-varying rate functions, we consider discrete arrival rate functions. From the examples provided, the increasing (or decreasing) rate of the number of customers is immediately affected by the changes in arrival rates. We see that the extreme points of the number of active peers appear with some delay, compared to the number of customers, which is due to the service time.

CHAPTER VI

EXTENSION TO NON-MARKOVIAN SYSTEMS

Until now, we explain the adjusted fluid and diffusion limits for the transient analysis of large-scale service systems. In this chapter, we introduce a possible extension of this research to non-Markovian systems. In Section VI.1, we briefly sketch the direction of the research to achieve this work from a simple example of $M_t/G/s$ queues. After that, in Section VI.2, we apply this extension to epidemic-based information dissemination in wireless mobile networks which is one of the popular application of asymptotic analysis.

VI.1. Phase-type approximations

The basic idea is to combine the limit processes with phase-type approximations. For that, we define a fluid limit for a non-Markovian system as the limit of a sequence of fluid limits for Markovian systems.

Suppose $\mathbf{G} = (G_1, G_2, ..., G_K)$ is a distribution set associated with a stochastic process X, e.g. $\mathbf{G} = (G_1, G_2)$ for $G/G/s$ queue. Let $\{\mathbf{PH}_m\}_{m\geq 1}$ be a sequence of phase-type distribution sets jointly converging to $\mathbf{G}$, where for each m $\mathbf{PH}_m = (PH_{1,m}, PH_{2,m}, ..., PH_{K,m})$. For $m \geq 1$, define $\{X_{n,m}\}$ to be a sequence of stochastic processes associated with $\mathbf{PH}_m$ converging to its fluid limits $\bar{X}_m$. Then, we use $\bar{X}_m(t)$ as an approximation of $E[X(t)]$ for a sufficiently large m, i.e. $E[X(t)] \approx \bar{X}_m(t)$. There are two simple numerical examples for $M_t/G/s$ queue: Poisson arrival with alternating rates between 40 and 80 every two time units, 50 servers, and Pareto(2.2,0.83) (and Pareto(1.2,5)) service time distribution (see Figure 27). Note that Pareto(a,b) denote the Pareto distribution of which CCDF is $F^c(t) = (1 + bt)^{-a}$. We use hyper-exponential distributions provided in Feldmann and Whitt (1998) to approximate

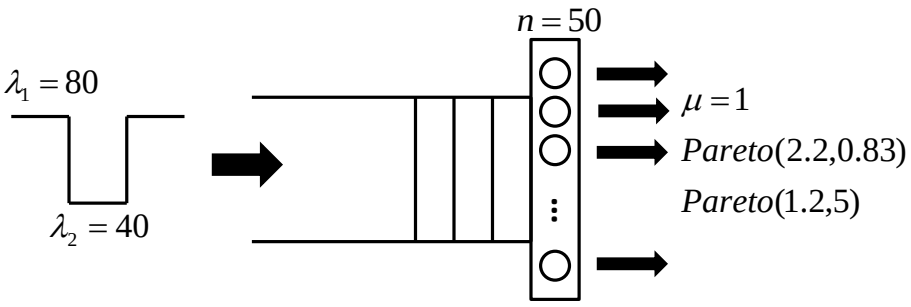
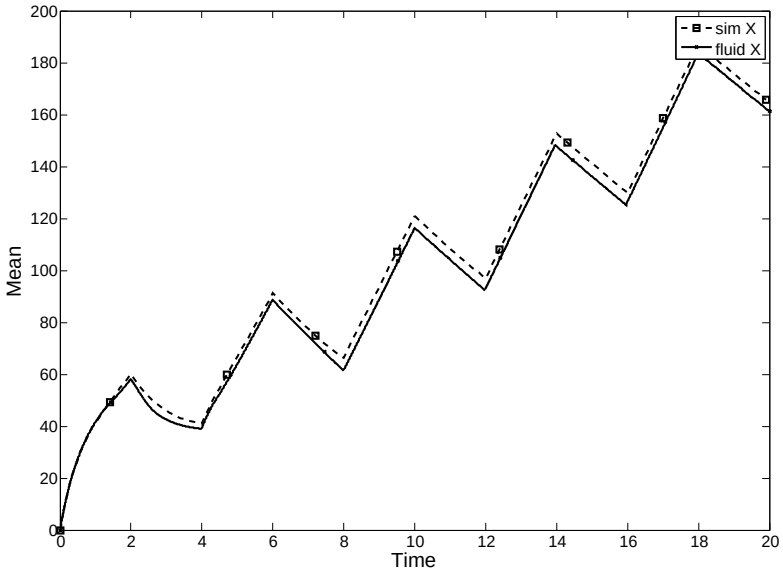


Fig. 27. $M_t/G/s$ queue

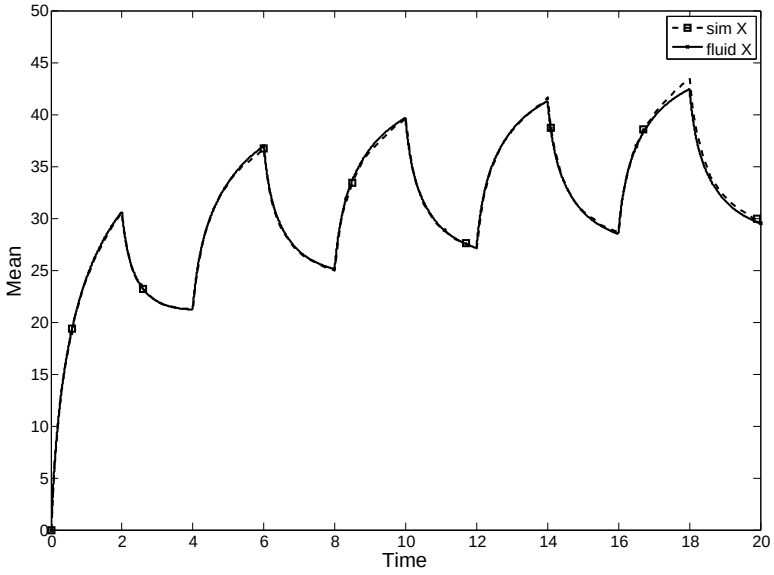
the Pareto distribution. As seen in Figure 28, we notice that proposed approximation is accurate, and it requires only 14-15 ODEs to solve. The number of ODEs in this example increases in $O(K)$ where K is the number of phases used, which implies the proposed method is scalable. To see its potential for broader applications, in the next section, we apply this method to analyze an information transfer model in wireless mobile networks operating under harsh environments.

VI.2. Epidemic-based information dissemination in wireless mobile sensor networks

In the near future, intelligent wireless mobile sensors will be extensively deployed in harsh environments such as military operations, under-sea explorations, hazardous environments, etc. The objectives for the nodes in such wireless mobile sensor networks are to move rapidly, probe, process and transmit information to other nodes. At the end of the "operation" only a subset of the sensor network nodes are recovered (the rest are either lost or severely damaged). Information is retrieved from what is stored in the subset of recovered nodes. Since the sensors have limited computational power, they have to balance between energy conservation and fault-tolerance while transmitting information.



(a) Pareto(2.2,0.83) service time



(b) Pareto(1.2,5) service time

Fig. 28. Mean numbers of $M/G/s$ queues

Epidemic information spreading models lend themselves well to such applications where each node selects at random a neighbor and transmits a quantum of information (referred to as *gossip*). There have been several research studies on epidemic information spreading models. Originally, studies have been conducted to analyze spreading mechanism of a disease. However, due to the scalability and stability of epidemic models, they have been adopted as an effective method to disseminate information over large-scale communication networks. Zhang *et al.* (2007) provides deterministic ODE models that are in fact the asymptotic results of Markovian models for epidemic routing problems. However, instead of analyzing the (asymptotic) phenomena, studies in communication networks are mostly concerned with developing fast and reliable algorithms to transfer information to as many nodes as possible under fixed network topology. Eugster *et al.* (2004) summarizes epidemic models in communication networks and emphasizes the easiness of deployment, robustness and stability. Boyd *et al.* (2006) and Mosk-Aoyama and Shah (2008) apply gossip algorithms to solve distributed computing problems (separable functions) considering both synchronous and asynchronous (with exponential inter-transmission time distributions) models. Weber *et al.* (2006) analyzes the performance of a particular gossip algorithm suitable for fixed networks using copulas. However, besides using fixed networks, the selection of right copulas is still an open question. Haas and Small (2006) and Haas *et al.* (2006) use epidemic models for a routing problem in ad-hoc networks. Some rules of thumb to reduce the number of transmissions are suggested from a number of experiments. Sistla *et al.* (2001) considers layered gossip networks using simulation. Deb *et al.* (2006) looks at the gossip algorithm from an information theoretic standpoint. When multiple messages are transmitted in the network, the number of transmission to deliver messages is reduced by proposed network coding method. In designing and constructing an epidemic information spreading model for

wireless mobile sensor networks, we found that most of previous studies ignore the fact that transmission behavior can be affected by external environment. Although exponential inter-transmission time is justified by Groenevelt *et al.* (2005), it may not be appropriate for harsh environment, where it could be time-varying or have a large variation. Therefore, the objective of this chapter is to conduct the performance analysis, specifically when the number of nodes in the network increases and transmission of information shows non-Markovian behavior.

VI.2.1. Problem description

Given a wireless mobile sensor network with N nodes, consider an information dissemination model for nodes to dissipate sensed information to as many nodes as possible. We specifically concentrate on a scenario where the sensor nodes move around, they sense and process information that they periodically transmit to other nodes. The inter-transmission times are far apart that the nodes would have significantly moved between two transmissions. Between successive transmissions, the nodes sense, process and store information. Unlike the previous studies assuming exponential inter-transmission time distribution (Boyd *et al.* (2005), Karp *et al.* (2000), and Kempe *et al.* (2003)), we consider a general asynchronous time model and investigate whether and how the performance is affected by a heavy-tailed CDF.

Note that the nodes (i) have local knowledge, (ii) have limited computational power, (iii) make distributed decisions, and (iv) move rapidly over time. With those in mind, we arrive at the following information dissemination model. Consider a gossip that was originated at a certain node. During the next transmission time for the node, the node picks its closest neighbor to transmit the gossip. Since the nodes are moving rapidly, we assume that the probability that a certain node is selected (i.e. closest neighbor) is $\frac{1}{N-1}$, even though all nodes may not be candidate neighbors.

At the next transmission time for each of these two nodes that know the gossip, the transmitting node selects a neighboring node at random with probability $\frac{1}{N-1}$. Although not exactly the same but a similar mobility model is provided in Neely and Modiano (2005) known as i.i.d mobility model. They assume a cell-partitioned network and each node chooses a cell randomly to move at next time slot. This implies *infinite mobility* that is reasonable for a network in which nodes are moving quickly relative to inter-transmission times. Moreover, from Theorem III-4 in Grossglauser and Tse (2002), for a large population, our selection mechanism could be justified.

In this manner, during every transmission time of a node, every node that has the gossip (and has not stopped spreading it) selects one of the $N-1$ other nodes to check if it has the gossip. If the selected node already has the gossip, then the transmitting node not only does not transmit the gossip but also stops spreading it; else it continues spreading the gossip. If the size of gossip is small, it would be reasonable to skip the checking procedure and determine whether to continue spreading it based on the acknowledgment. As a result of this explicit stopping criterion, each node stops spreading the gossip and conserves energy. *Note:* stopping the gossip spreading implies stopping the checking procedure as well. Therefore, a node does not spend network resources and energy for the gossip any longer once it decides to stop spreading.

VI.2.2. Fluid approximation

For general inter-transmission time distributions, we adopt techniques that approximate any distribution function having positive support with phase-type distributions since they are proven to be dense in all distributions with positive support (Johnson and Taaffe (1988)) and have *nice* properties. One common method to find a phase-type distribution is matching moments due to its easiness and convenience (Whitt

(1981), Whitt (1982), Altiok (1985), Johnson and Taafe (1990)). It, however, has limitation since it may lose properties of target distribution. Therefore, in this chapter, we use the methodology in Feldmann and Whitt (1998) that approximates the distribution function itself and is applicable even if no moment exists. The main idea of our method is to find a phase-type distribution function which approximates a given distribution function accurately (depending on the problem). For the class of distributions we consider here, i.e. distributions having positive support with decreasing PDFs), the following theorem guarantees that there is a sequence of hyperexponential distributions that converges to an element of that class.

Theorem 18 (Denseness, Feldmann and Whitt (1998))**.** *If F is a CDF with a completely monotone PDF, then there are hyperexponential CDFs $F^{(n)}, n \geq 1$, i.e., CDFs of the form*

$$F^{(n)} = \sum_{i=1}^{k_n} p_{ni}(1 - e^{-\lambda_{ni}t}), \quad t \geq 0,$$

with $\lambda_{ni} \leq \infty$ and $p_{n1} + \cdots + p_{nk_n} = 1$ such that $F^{(n)} \Rightarrow F$ as $n \to \infty$.

Therefore, from Theorem 18, we can choose a hyperexponential distribution function satisfying a desired error bound. After choosing such a hyperexponential distribution function, we combine it with the asymptotic method explained in Chapter II. In fact, approximating a distribution with phase-type distributions is well studied in the literature. However, one crucial limitation commonly raised in the literature is that a Markov chain obtained from phase-type distributions becomes intractable as the number of phases increases for better approximation. In our method, however, we found that only $K + 1$ ODEs are enough to obtain the fluid model of the system.

Suppose the inter-transmission time distribution function of each node is F and

is approximated by hyper-exponential distribution function G where

$$1 - G(u) = \sum_{i=1}^{K} p_i \exp(-\lambda_i u).$$

Note we are not interested in how to find G in this chapter. We apply an existing approximation method proposed by Feldmann and Whitt (1998). Let $X_{n,t} = (x^1_{n,t}, \ldots, x^K_{n,t}, y_{n,t})$, the state of the system and $H_{n,t}$ is the distribution function of inter-transmission time among all nodes at time t with total $n+1$ nodes, then

$$1 - H_{n,t}(u) = \exp\Big(u \sum_{i=1}^{K} \lambda_i x^i_{n,t}\Big).$$

Now consider a split process of Poisson processes. Probability that next event occurs in $x^i_{n,t}$ is

$$\frac{\lambda_i x^i_{n,t}}{\sum_{j=1}^{K} \lambda_j x^j_{n,t}}.$$

Applying minimum of exponential distributions, rate of event is $\sum_{j=1}^{K} \lambda_j x^j_{n,t}$. Therefore, rate of events of each phase is just $\lambda_i x^i_{n,t}$.

Let e_i be $K+1$ dimensional column vector where its i^{th} element is 1 and other elements are 0. Then, $X_{n,t}$ is the solution to the following integral equation.

$$\begin{aligned} X_{n,t} &= X_{n,0} \\ &+ \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}(e_j + e_k - e_i - e_{K+1})Y^1_{ijk}\Big(\int_0^t \frac{y_{n,s}}{n}\lambda_i x^i_{n,t} p_j p_k\Big) \\ &- \sum_{i=1}^{K} e_i Y^2_i\Big(\int_0^t \frac{n - y_{n,s}}{n}\lambda_i x^i_{n,t}\Big). \end{aligned} \tag{54}$$

Let $\tilde{X}_{n,t} = X_{n,t}/n$. Then, the equation (54) can be written as follows:

$$\begin{aligned}\tilde{X}_{n,t} &= \tilde{X}_{n,0} \\ &+ \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{1}{n}(e_j + e_k - e_i - e_{K+1})Y^1_{ijk}\Big(n\int_0^t \tilde{y}_{n,s}\lambda_i\tilde{x}^i_{n,t}p_jp_k\Big) \\ &- \sum_{i=1}^{K}\frac{1}{n}e_iY^2_i\Big(n\int_0^t(1-\tilde{y}_{n,s})\lambda_i\tilde{x}^i_{n,t}\Big). \qquad (55)\end{aligned}$$

Theorem 19 (Fluid limit)**.** *Suppose* $\{\tilde{X}_{n,t}\}_{n\geq 1}$ *is the solution to equation (55) and* $\tilde{X}_{n,0}$ *converges to* $\bar{X}_0$ *almost surely. Then,* $\tilde{X}_{n,t}$ *converges to* $\bar{X}_t$ *almost surely where* $\bar{x}^a_t$ *and* $\bar{y}_t$*, the* a^{th} *and* $(K+1)^{th}$ *components of* $\bar{X}_t$ *respectively, are the solutions to the following differential equations:*

$$\begin{aligned}\frac{d}{dt}\bar{x}^a_t &= \bar{y}_t\Big[2p_a\sum_{i=1,i\neq a}^{K}\lambda_i\bar{x}^i_t + (p_a^2 - (1-p_a)^2)\lambda_a\bar{x}^a_t\Big] \\ &\quad -(1-\bar{y}_t)\lambda_a\bar{x}^a_t \quad \textit{for } a\in\{1,2,\ldots,K\}, \textit{ and} \\ \frac{d}{dt}\bar{y}_t &= -\bar{y}_t\sum_{i=1}^{K}\lambda_i\bar{x}^i_t\end{aligned}$$

Proof. We will derive the a^{th} component of $\bar{X}_t$. We first consider

$$\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{1}{n}(e_j + e_k - e_i - e_{K+1})Y^1_{ijk}\Big(n\int_0^t \tilde{y}_{n,s}\lambda_i\tilde{x}^i_{n,t}p_jp_k\Big)$$

of equation (55). The a^{th} component, $\bar{x}^a_t$ is

1. increasing by 2 when $j = k = a$ and $i \neq a$ in equation (55) with rate

$$\bar{y}_t p_a^2 \sum_{i=1,i\neq a}^{K}\lambda_i, \bar{x}^i_t.$$

2. increasing by 1

   - when $j = k = i = a$ with rate $\bar{y}_t\lambda_a\bar{x}^a_t p_a^2$.

- when $j = a, k \neq a$, and $i \neq a$ with rate

$$\bar{y}_t \sum_{i=1, i\neq a}^{K} \sum_{k=1, k\neq a}^{K} \lambda_i \bar{x}_t^i p_a p_k = \bar{y}_t \sum_{i=1, i\neq a}^{K} \lambda_i \bar{x}_t^i p_a (1 - p_a).$$

- when $k = a, j \neq a$, and $i \neq a$ with rate

$$\bar{y}_t \sum_{i=1, i\neq a}^{K} \sum_{j=1, j\neq a}^{K} \lambda_i \bar{x}_t^i p_a p_j = \bar{y}_t \sum_{i=1, i\neq a}^{K} \lambda_i \bar{x}_t^i p_a (1 - p_a).$$

3. decreasing by 1 when $j \neq a, k \neq a$, and $i = a$ with rate

$$\bar{y}_t \sum_{j=1, j\neq a}^{K} \sum_{k=1, k\neq a}^{K} \lambda_a \bar{x}_t^a p_j p_k = \bar{y}_t \lambda_a \bar{x}^a (1 - p_a)^2.$$

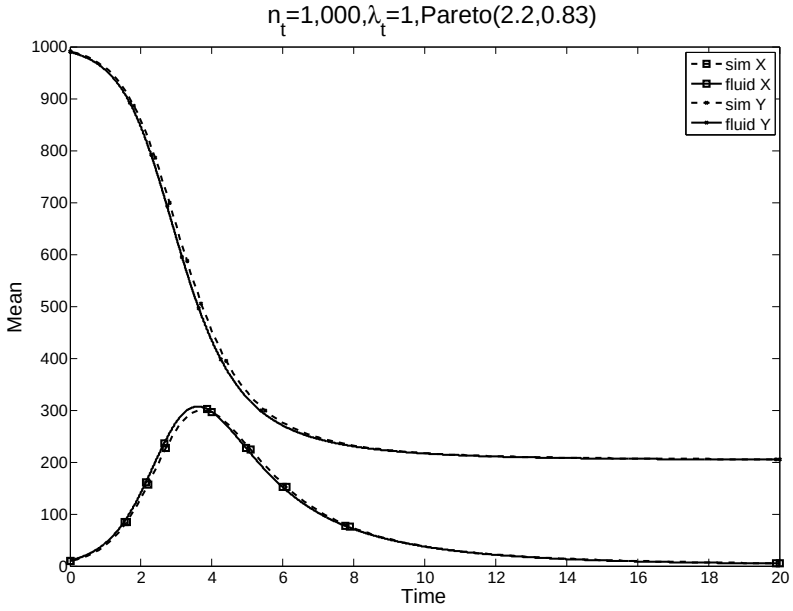
Note that $\bar{y}_t$ is decreasing by 1 in any case. Second, we consider

$$-\sum_{i=1}^{K} \frac{1}{n} e_i Y_i^2 \Big(n \int_0^t (1 - \tilde{y}_{n,s}) \lambda_i \tilde{x}_{n,t}^i \Big)$$

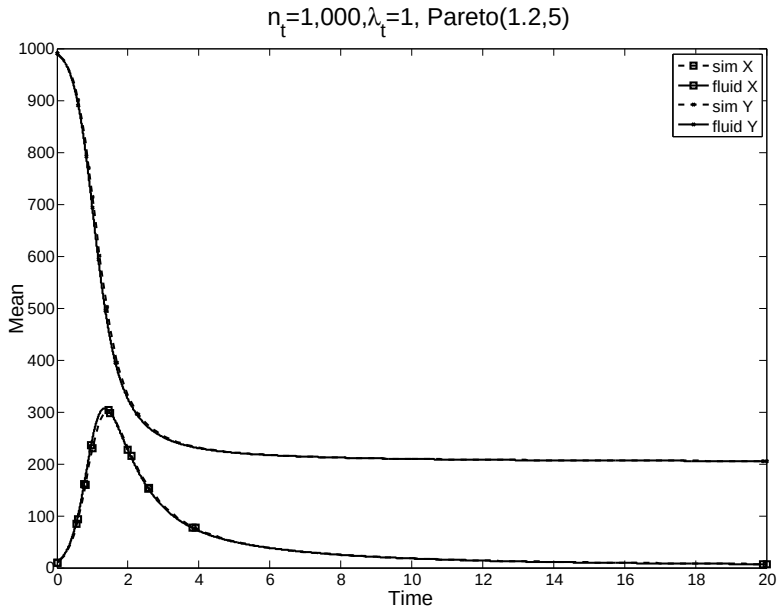
of equation (55). The a^{th} component, $\bar{x}_t^a$ is decreasing only when $i = a$ with rate $(1 - \bar{y}_t)\lambda_a \bar{x}_t^a$ and $\bar{y}_t$ does not change in all cases. □

The number of ODEs in Theorem 19 is just $K + 1$. The resulting phase-type distributions in Feldmann and Whitt (1998) have about 15 phases and accurately fit the target distribution function. Therefore, we could think that less than 30 phases are enough for approximation. Even if we have more than 30 phases, solving that number of ODEs is not computationally expensive.

Figure 29 shows the estimation of mean values using a fluid limit with phase-type approximation. In these examples, we have 1,000 nodes and the inter-transmission time distributions are Pareto(2.2,0.83) and Pareto(1.2,5) respectively. As seen in the figure, the fluid limit with phase-type approximations provides excellent estimation results for the transient dynamics of wireless mobile sensor networks.



(a) Pareto(2.2,0.83) inter-transmission time



(b) Pareto(1.2,5) inter-transmission time

Fig. 29. Evolution of wireless mobile networks on average

CHAPTER VII

CONCLUSION

Transient analysis of large-scale service systems is not sufficiently addressed in previous literature, nor do the techniques fully approximate real systems. This dissertation discusses a new technique to enable more accurate analysis by suitably adjusting fluid and diffusion limits with less computational time. This chapter summarizes the main results and contributions of this dissertation.

VII.1. Methodology to approximate large-scale systems

We describe the fluid and diffusion limits obtained from the uniform acceleration since this technique is adequate for transient analysis of time-dependent systems. However, there are two significant problems when using this technique for balancing accuracy and computational tractability. First, the expectation of a function of a random vector X is not equal to the value of the function of the expectation of X. Therefore, unless they are equal or close, the fluid limit may not provide an accurate estimation of mean values of the system state. Second, non-differentiability of rate functions causes discontinuity in the drift matrix of diffusion limits. To resolve these critical issues, we develop a methodology to obtain the exact estimation of mean values of system states and an algorithm to achieve computational tractability.

The basic concept is to construct a new sequence of stochastic processes which converges to the fluid limit exactly as the mean value of the system state. We prove that if the rate functions in the original model satisfy the conditions for the fluid limit, rate functions in the new model also satisfy them. Therefore, we can apply the adjusted fluid limit if we apply the standard fluid limit. Generally speaking, since no computational method can exactly obtain the adjusted fluid limit, we utilize Gaussian

density because it allows us to see that rate functions in the constructed process are differentiable everywhere and in turn we are able to apply the diffusion limit, even if the rate functions in the original process are not differentiable. Although the standard limits show inaccuracy to approximate the system, we find that the adjusted limits are asymptotically identical to the standard limits. Therefore, the adjusted limits achieve both accurate estimation of the system being analyzed and asymptotic exactness.

VII.2. Multi-server retrial queues (call center model)

We select a critically-loaded call center to test our adjusted limits. A call center is considered critically loaded when the number of servers equals the number of customers. Ideally, most companies operating call centers desire this condition, i.e. no waiting customer and no idle server. However, the transient analysis of critically loaded queues is difficult, and the standard fluid and diffusion limits do not give accurate approximations. We find that the use of the adjusted limits achieves accurate approximation of call centers when they are almost critically loaded.

VII.3. Peer-based Internet services

The second application involves a relatively new type of service system, peer-based multimedia services. P2P architecture is a viable alternative to outsource resources for the online multimedia service industry. However, to successfully deploy P2P architecture, the transient analysis of initial build-up periods must be conducted in advance, because P2P networks are extremely unstable at first. Utilizing standard limits causes significant inaccuracy especially in the early stage of the network evolution. We use the adjusted limits to estimate the performance measures, and successfully demonstrate its effectiveness with numerical experiments.

VII.4. Extension to non-Markovian systems

The adjusted fluid and diffusion limits are derived under Markovian settings. An extension to non-Markovian systems using phase-type distributions is also possible since the fluid and diffusion limits are relatively insensitive to the dimension of the system state. Phase-type distributions are known to be dense in all of the distribution functions having positive support. We incorporate this phase-type approximation into the fluid limit. Preliminary work applied to $M_t/G/s$ appears promising. Next, we apply the extension to non-service systems, an epidemic model in wireless mobile networks. In harsh environments like military operations and natural disasters, environmental factors can significantly affect the behavior of wireless sensors. Considering that non-Markovian inter-transmission times should be natural, we find that phase-type approximations provide accurate estimation returns for non-service systems.

VII.5. Future research

Although the methodology presented in this dissertation is versatile, in some other types of systems, e.g. multi-class queues, the approximation quality is still not good enough. We observe that for those systems the empirical density is not close to the Gaussian density even if the values of the parameters are fairly large. We conjecture that, depending on the problems, convergence rates of empirical density to Gaussian density might require significantly large values of parameters. Future research should investigate the properties of the specific rate functions affecting the shape of empirical density in order to devise a new methodology to find the functions $g_i^{\eta}(\cdot,\cdot)$'s from other density functions.

REFERENCES

Abate, J. and Whitt, W. (1989) Calculating time-dependent performance measures for the $M/M/1$ queue. *IEEE Transactions on Communications*, **37**(10), 1102–1104.

Adler, M., Kumar, R., Ross, K., Rubenstein, D., Suel, T. and Yao, D.D. (2005) Optimal peer selection for P2P downloading and streaming, in *Proceedings of the IEEE INFOCOM*, IEEE, Piscetaway, NJ, pp. 1538–1549.

Akamai (2011) http://www.akamai.com/.

Altiok, T. (1985) On the phase-type approximations of general distributions. *IIE Transactions*, **17**(2), 110–116.

Andersson, H. and Djehiche, B. (1994) A functional limit theorem for the total cost of a multitype standard epidemic. *Advances in Applied Probability*, **26**(3), 690–697.

Andersson, M. (1999) The asymptotic final size distribution of multitype chain-binomial epidemic processes. *Advances in Applied Probability*, **31**(1), 220–234.

Apple iTunes Services (2011) http://www.apple.com/itunes.

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

Ball, F. and Barbour, A. (1990) Poisson approximation for some epidemic models. *Journal of Applied Probability*, **27**(3), 479–490.

Ball, F., Mollison, D. and Scalia-Tomba, G. (1997) Epidemics with two levels of mixing. *The Annals of Applied Probability*, **7**(1), 46–89.

Ball, F. and Neal, P. (2004) Poisson approximations for epidemics with two levels of mixing. *Annals of Probability*, **32**(1), 1168–1200.

Bassamboo, A., Kumar, S. and Randhawa, R.S. (2009) Dynamics of new product introduction in closed rental systems. *Operations Research*, **57**(6), 1347–1359.

Bassamboo, A. and Randhawa, R.S. (2009) Optimal control in a Netflix-like closed rental system. *Working Paper*.

Billingsley, P. (1999) *Convergence of Probability Measures*, Wiley, New York.

BitTorrent (2011) http://www.bittorrent.com/.

Boyd, S., Ghosh, A., Prabhakar, B. and Shah, D. (2005) Gossip algorithms: design, analysis and applications, in *Proceedings of the IEEE INFOCOM*, IEEE, Piscataway, NJ, pp. 1653–1664.

Boyd, S., Ghosh, A., Prabhakar, B. and Shah, D. (2006) Randomized gossip algorithms. *IEEE Transactions on Information Theory*, **52**(6), 2508–2530.

Chen, H., Kella, O. and Weiss, G. (1997) Fluid approximations for a processor-sharing queue. *Queueing Systems*, **27**(1-2), 99–125.

Clévenot, F. and Nain, P. (2004) A simple fluid model for the analysis of the squirrel peer-to-peer caching system, in *Proceedings of the IEEE INFOCOM*, IEEE, Piscataway, NJ, pp. 86–95.

Collings, T. and Stoneman, C. (1976) The $M/M/\infty$ queue with varying arrival and departure rates. *Operations Research*, **24**(4), 760–773.

Deb, S., Medard, M. and Choute, C. (2006) Algebraic gossip: a network coding approach to optimal multiple rumor mongering. *IEEE Transactions on Information Theory*, **52**(6), 2486 – 2507.

Eick, S.G., Massey, W.A. and Whitt, W. (1993) The physics of the $M_t/G/\infty$ queue. *Operations Research*, **41**(4), 731–742.

Ethier, S.N. and Kurtz, T.G. (1986) *Markov Processes: Characterization and Convergence*, 1st edition, Wiley, New York.

Eugster, P., Guerraoui, R., Kermarrec, A. and Massoulie, L. (2004) Epidemic information dissemination in distributed systems. *Computer*, **37**(5), 60–67.

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

Foley, R. (1982) The nonhomogeneous $M/G/\infty$ queue. *Operations Research*, **19**(1), 40–48.

Folland, G.B. (1999) *Real Analysis : Modern Techniques and Their Applications*, 2nd edition, Wiley, New York.

Fraleigh, C., Moon, S., Lyles, B., Cotton, C., Khan, M., Moll, D., Rockell, R., Seely, T. and Diot, S. (2003) Packet-level traffic measurements from the Sprint IP backbone. *IEEE Network*, **17**(6), 6 – 16.

Garnet, O., Mandelbaum, A. and Reiman, M.I. (2002) Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, **4**(3), 208–227.

Ge, Z., Figueiredo, D.R., Jaiswal, S., Kurose, J. and Towsley, D. (2003) Modeling peer-to-peer file sharing systems, in *Proceedings of the IEEE INFOCOM*, IEEE, Piscataway, NJ, pp. 2188–2198.

Grassmann, W. (1977) Transient solutions in Markovian queues : an algorithm for finding them and determining their waiting-time distributions. *European Journal of Operational Research*, **1**(6), 396–402.

Groenevelt, R., Nain, P. and Koole, G. (2005) The message delay in mobile ad hoc networks. *Perform Evaluation*, **62**(1-4), 210–228.

Grossglauser, M. and Tse, D. (2002) Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Transactions on Networking*, **10**(4), 477 – 486.

Gummadi, K., Dunn, R., Saroiu, S., Gribble, S., Levy, H. and Zahorjan, J. (2003) Measurement, modeling, and analysis of a peer-to-peer file-sharing workload, in *Proceedings of the ACM SOSP*, ACM, New York, NY, pp. 314–329.

Haas, Z., Halpern, J. and Li;, L. (2006) Gossip-based ad hoc routing. *IEEE/ACM Transactions on Networking*, **14**(3), 479–491.

Haas, Z. and Small, T. (2006) A new networking model for biological applications of ad hoc sensor networks. *IEEE/ACM Transactions on Networking*, **14**(1), 27—40.

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

Hampshire, R.C., Harchol-Balter, M. and Massey, W.A. (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, R.C., Jennings, O.B. and Massey, W.A. (2009) A time-varying call center design via Lagrangian mechanics. *Probability in the Engineering and Informational Sciences*, **23**(02), 231–259.

Iglehart, D.L. (1965) Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability*, **2**(2), 429–441.

Jean-Marie, A. and Robert, P. (1994) On the transient behavior of the processor sharing queue. *Queueing Systems*, **17**(1-2), 129–136.

Johnson, M. and Taaffe, M. (1988) The denseness of phase distributions. *School of Industrial Engineering Purdue University Research Memorandum*, no. 88-20.

Johnson, M.A. and Taafe, M.R. (1990) Matching moments to phase distributions: nonlinear programming approaches. *Communications in Statistics. Stochastic models*, **6**(2), 259–281.

Karp, R., Schindelhauer, C., Shenker, S. and Vocking, B. (2000) Randomized rumor spreading, in *Proceedings of the IEEE Symposium on Foundations of Computer Science*, IEEE, Piscataway, NJ, pp. 565–574.

Kempe, D., Dobra, A. and Gehrke, J. (2003) Gossip-based computation of aggregate information, in *Proceedings of the IEEE Symposium on Foundations of Computer Science*, IEEE, Piscataway, NJ, pp. 482–491.

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

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

Mandelbaum, A., Massey, W.A. and Reiman, M.I. (1998) Strong approximations for Markovian service networks. *Queueing Systems*, **30**(1-2), 149–201.

Mandelbaum, A., Massey, W.A., Reiman, M.I., Stolyar, A. and Rider, B. (2002) Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, **21**(2-4), 149–171.

Mandelbaum, A. and Pats, G. (1995) State-dependent queues: approximations and applications. *Institute for Mathematics and Its Applications*, **71**, 239–282.

Mandelbaum, A. and Pats, G. (1998) State-dependent stochastic networks. Part I: approximations and applications with continuous diffusion limits. *The Annals of Applied Probability*, **8**(2), 569–646.

Mandelbaum, A. and Zeltyn, S. (2009) Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, **57**(5), 1189–1205.

Massey, W.A. (1985) Asymptotic analysis of the time dependent $M/M/1$ queue. *Mathematics of Operations Research*, **10**(2), 305–327.

Massey, W.A. (2002) The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, **21**(2-4), 173–204.

Massey, W.A. and Whitt, W. (1993) Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*, **13**(1-3), 183–250.

Massey, W.A. and Whitt, W. (1998) Uniform acceleration expansions for Markov chains with time-varying rates. *The Annals of Applied Probability*, **8**(4), 1130–1155.

Mosk-Aoyama, D. and Shah, D. (2008) Fast distributed algorithms for computing separable functions. *IEEE Transactions on Information Theory*, **54**(7), 2997–3007.

Neely, M. and Modiano, E. (2005) Capacity and delay tradeoffs for ad hoc mobile networks. *IEEE Transactions on Information Theory*, **51**(6), 1917 – 1937.

Nelson, B.L. and Taaffe, M.R. (2004a) The $Ph_t/Ph_t/\infty$ queueing system: part I-the single node. *INFORMS Journal on Computing*, **16**(3), 266–274.

Nelson, B.L. and Taaffe, M.R. (2004b) The $Ph_t/Ph_t/\infty$ queueing system: part II-the multiclass network. *INFORMS Journal on Computing*, **16**(3), 275–283.

Pando (2011) http://www.pando.com/.

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

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

Qiu, D. and Srikant, R. (2004) Modeling and performance analysis of BitTorrent-like peer-to-peer networks, in *Proceedings of the ACM SIGCOMM*, volume 34, ACM, New York, NY, pp. 367–378.

Reinert, G. (1995) The asymptotic evolution of the general stochastic epidemic. *The Annals of Applied Probability*, **5**(4), 1061–1086.

Sellke, T. (1983) On the asymptotic distribution of the size of a stochastic epidemic. *Journal of Applied Probability*, **20**(2), 390–394.

Shakkottai, S. and Johari, R. (2010) Demand-aware content distribution on the Internet. *IEEE/ACM Transactions on Networking*, **18**(2), 476 – 489.

Sistla, K., George, A., Todd, R. and Tilak, R. (2001) Performance analysis of flat and layered gossip services for failure detection and consensus in scalable heterogeneous

clusters, in *Proceedings of the IEEE International Parallel & Distributed Processing Symposium*, IEEE, Piscataway, NJ, pp. 802 – 809.

van de Coevering, M. (1995) Computing transient performance measures for the $M/M/1$ queue. *OR Spectrum*, **17**(1), 19–22.

Weber, S., Veeraraghavan, V., Kini, A. and Singhal, N. (2006) Analysis of gossip performance with copulas, in *Proceedings of the Conference on Information Sciences and Systems*, IEEE, Piscataway, NJ, pp. 1212–1217.

Whitt, W. (1981) Approximating a point process by a renewal process: the view through a queue, an indirect approach. *Management Science*, **27**(6), 619–636.

Whitt, W. (1982) Approximating a point process by a renewal process, I: two basic methods. *Operations Research*, **30**(1), 125–147.

Whitt, W. (1990) Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems*, **6**(4), 335–351.

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

Whitt, W. (2006a) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, **50**(10), 1449–1461.

Whitt, W. (2006b) Fluid models for multiserver queues with abandonments. *Operations Research*, **54**(1), 37–54.

Yang, X. and de Veciana, G. (2004) Service capacity of peer to peer networks, in *Proceedings of the IEEE INFOCOM*, volume 4, IEEE, Piscataway, NJ, pp. 2242–2252.

Zeltyn, S. and Mandelbaum, A. (2005) call centers with impatient customers: many-server asymptotics of the M/M/n+ G queue. *Queueing Systems*, **51**(3-4), 361–402.

Zhang, X., Neglia, G., Kurose, J. and Towsley, D. (2007) Performance modeling of epidemic routing. *Computer Networks*, **51**(10), 2867–2891.

APPENDIX A

MATHEMATICAL DERIVATION AND NUMERICAL RESULTS FOR CHAPTER 4

A.1. Derivation of functions $g_i^\eta(t,x)$

For a fixed η, suppose $x_1^\eta(t) \sim N(E[x_1^\eta(t)], \sigma_1^{\eta 2}(t))$. For $x = (x_1, x_2)'$, we have

$$
\begin{aligned}
g_3^\eta\big(t,x\big) &= \eta E\Big[\mu_t^1\Big(\frac{x_1^\eta(t)}{\eta} - \frac{E[x_1^\eta(t)]}{\eta} + \frac{x_1}{\eta} \wedge n_t\Big)\Big] \\
&= E\big[\mu_t^1(x_1^\eta(t) - E[x_1^\eta(t)] + x_1 \wedge \eta n_t)\big] \\
&= \mu_t^1\Big\{E\big[(x_1^\eta(t) - E[x_1^\eta(t)] + x_1)\mathbb{I}_{x_1^\eta(t)-E[x_1^\eta(t)]+x_1 \le \eta n_t}\big] \\
&\quad +\eta n_t Pr[x_1^\eta(t) - E[x_1^\eta(t)] + x_1 > \eta n_t]\Big\} \\
&\quad \text{Let } y_1(t) = x_1^\eta(t) - E[x_1^\eta(t)] + x_1. \\
&= \mu_t^1\Bigg[\int_{-\infty}^{\eta n_t} \frac{y_1(t)}{\sqrt{2\pi}\sigma_1^\eta(t)} \exp\Big(-\frac{(y_1(t)-x_1)^2}{2\sigma_1^{\eta 2}(t)}\Big) dy_1(t) + \eta n_t Pr[y_1(t) > \eta n_t]\Bigg] \\
&= \mu_t^1\Bigg[\frac{-\sigma_1^\eta(t)}{\sqrt{2\pi}} \int_{-\infty}^{\eta n_t} -\frac{y_1(t)-x_1}{\sigma_1^{\eta 2}} \exp\Big(-\frac{(y_1(t)-x_1)^2}{2\sigma_1^{\eta 2}}\Big) dy_1(t) \\
&\quad +x_1 Pr[y_1(t) \le \eta n_t] + \eta n_t Pr[y_1(t) > \eta n_t]\Bigg] \\
&= \mu_t^1\Bigg[-\sigma_1^{\eta 2}(t)\frac{1}{\sqrt{2\pi}\sigma_1^\eta(t)} \exp\Big(-\frac{(\eta n - x_1)^2}{2\sigma_1^{\eta 2}(t)}\Big) \\
&\quad +(x_1 - \eta n_t) Pr[y_1(t) \le \eta n_t] + \eta n_t\Bigg].
\end{aligned}
$$

Therefore, we have $g_3^\eta(t,x)$.

Note $g_4^\eta(\cdot,\cdot)$ and $g_5^\eta(\cdot,\cdot)$ are the same except a constant part with respect to x.

Therefore, it is enough to derive $g_5^\eta(\cdot,\cdot)$. We can show that

$$
\begin{aligned}
g_5^\eta(t,x) &= \eta E\Big[\beta_t p_t\Big(\frac{x_1^\eta(t)}{\eta} - \frac{E[x_1^\eta(t)]}{\eta} + \frac{x_1}{\eta} - n_t\Big)^+\Big] \\
&= \beta_t p_t\Big\{E\big[(x_1^\eta(t) - E[x_1^\eta(t)] + x_1 \vee \eta n_t)\big] - \eta n_t\Big\} \\
&= \beta_t p_t\Big\{E\big[(x_1^\eta(t) - E[x_1^\eta(t)] + x_1)\mathbb{I}_{x_1^\eta(t)-E[x_1^\eta(t)]+x_1>\eta n_t}\big] \\
&\quad +\eta n_t Pr[x_1^\eta(t) - E[x_1^\eta(t)] + x_1 \le \eta n_t] - \eta n_t\Big\} \\
&\quad \text{Let } y_1(t) = x_1^\eta(t) - E[x_1^\eta(t)] + x_1. \\
&= \beta_t p_t\Bigg[\int_{\eta n_t}^\infty \frac{y_1(t)}{\sqrt{2\pi}\sigma_1^\eta(t)} \exp\Big(-\frac{(y_1(t)-x_1)^2}{2\sigma_1^{\eta 2}(t)}\Big) dy_1(t) \\
&\quad +\eta n_t Pr[y_1(t) \le \eta n_t] - \eta n_t\Bigg] \\
&= \beta_t p_t\Bigg[\frac{-\sigma_1^\eta(t)}{\sqrt{2\pi}} \int_{\eta n_t}^\infty -\frac{y_1(t)-x_1}{\sigma_1^{\eta 2}} \exp\Big(-\frac{(y_1(t)-x_1)^2}{2\sigma_1^{\eta 2}}\Big) dy_1(t) \\
&\quad +x_1 Pr[y_1(t) > \eta n_t] + \eta n_t Pr[y_1(t) \le \eta n_t] - \eta n_t\Bigg] \\
&= \beta_t p_t\Bigg[\sigma_1^{\eta 2}(t)\frac{1}{\sqrt{2\pi}\sigma_1^\eta(t)} \exp\Big(-\frac{(\eta n - x_1)^2}{2\sigma_1^{\eta 2}(t)}\Big) \\
&\quad +(x_1 - \eta n_t) Pr[y_1(t) > \eta n_t]\Bigg].
\end{aligned}
$$

Therefore, we have $g_5^\eta(t,x)$.

A.2. Numerical results

Table 4. Estimation of $E\big[x_1(t)\big]$ over time; difference from simulation

Exp.		Time									
#	type	6	7	8	9	10	11	12	13	14	15
1	adj.	1.83	0.48	-1.62	-0.63	-0.39	-0.29	0.40	-0.15	-0.82	-0.13
	standard	0.40	-1.04	-2.64	-1.90	-0.80	-1.20	-0.18	-1.01	-0.80	-0.69
2	adj.	0.95	0.27	-1.57	-0.50	-0.79	-0.03	0.23	0.05	-0.37	-0.14
	standard	0.78	-0.27	-1.82	-0.99	-0.61	-0.35	-0.03	-0.34	-0.46	-0.40
3	adj.	1.29	0.24	-1.54	-0.45	-0.54	-0.16	0.40	0.24	0.22	-0.09
	standard	1.01	-0.18	-1.53	-0.77	-0.98	-0.50	0.28	-0.07	0.23	-0.31
4	adj.	1.24	0.15	-1.61	-0.49	-0.07	-0.02	0.37	0.11	-0.56	-0.19
	standard	1.15	-0.19	-1.35	-0.67	-0.20	-0.29	0.32	-0.14	-0.46	-0.36
5	adj.	0.37	0.03	-0.28	-0.06	-0.11	0.03	0.09	0.06	-0.00	-0.02
	standard	-0.40	-0.63	-0.82	-0.57	-0.31	-0.34	-0.16	-0.26	-0.25	-0.28
6	adj.	1.30	0.54	-1.21	-0.36	-0.30	0.11	0.40	-0.01	-0.17	-0.20
	standard	1.49	0.29	-1.39	-0.74	0.17	0.02	0.61	-0.22	0.24	-0.31
7	adj.	1.68	-0.07	-1.64	-0.13	-0.02	-0.14	0.31	0.24	-0.60	-0.35
	standard	-1.62	-3.88	-4.20	-3.66	-2.26	-3.39	-2.25	-3.03	-2.76	-3.28
8	adj.	1.64	0.31	-1.46	-0.58	-0.26	0.21	0.15	0.25	-0.19	0.12
	standard	-0.21	-1.96	-2.95	-2.78	-1.94	-1.96	-1.32	-1.93	-2.17	-2.05

Table 5. Estimation of $E\big[x_2(t)\big]$ over time; difference from simulation

Exp.		Time									
#	type	6	7	8	9	10	11	12	13	14	15
1	adj.	-1.37	2.90	0.89	-2.03	-2.12	-2.34	-1.12	-0.57	-0.71	-1.14
	std.	77.71	78.84	59.99	63.78	68.73	70.44	58.52	62.73	67.10	68.65
2	adj.	-2.37	1.61	0.84	-3.49	-2.56	-2.85	-1.76	-1.17	-1.11	-1.56
	std.	55.00	57.92	33.81	36.87	41.48	44.67	30.90	35.92	39.56	42.01
3	adj.	-3.47	3.93	2.40	-2.23	-1.27	-1.72	-1.23	0.14	0.06	0.01
	std.	43.47	49.71	26.57	29.38	33.16	35.73	22.49	28.26	30.92	34.20
4	adj.	-3.17	3.20	0.70	-4.33	-2.75	-1.89	-1.03	-0.19	-0.25	-1.10
	std.	35.37	42.81	20.08	22.79	26.22	30.16	18.29	23.35	25.38	27.70
5	adj.	-1.01	0.97	0.73	-0.90	-1.00	-0.59	-0.79	-0.47	-0.24	-0.50
	std.	100.00	94.18	73.94	77.12	81.28	82.09	70.79	74.22	78.18	78.78
6	adj.	-6.58	6.35	1.32	-5.83	-2.07	1.60	-0.07	1.68	-1.74	-0.52
	std.	96.64	78.24	40.17	75.11	94.63	76.05	39.36	77.09	94.60	75.69
7	adj.	-1.52	2.10	1.99	-0.12	0.00	0.29	0.08	0.09	-0.42	-1.06
	std.	77.20	75.69	59.18	65.40	70.23	70.38	58.65	64.27	68.43	67.79
8	adj.	-0.14	2.04	1.82	-3.27	-1.61	-1.26	-1.65	0.19	0.22	-0.44
	std.	77.59	78.60	59.36	62.65	67.64	70.43	56.90	62.15	66.04	68.30

Table 6. Estimation of $Var\big[x_1(t)\big]$ over time; difference from simulation

Exp.		Time									
#	type	6	7	8	9	10	11	12	13	14	15
1	adj.	2.86	-3.23	-4.43	-2.63	-3.98	1.41	0.83	1.58	-2.75	0.02
	std.	-17.89	13.73	15.53	-20.73	-27.51	18.09	18.99	-15.62	-26.91	17.15
2	adj.	-1.19	0.90	-5.74	2.86	-0.95	3.53	0.11	2.39	-1.66	1.34
	std.	-12.11	14.36	5.51	-11.40	-13.82	16.35	9.61	-12.33	-16.16	14.45
3	adj.	3.59	-5.48	-1.10	5.68	2.82	2.94	2.13	-0.74	0.93	-1.41
	std.	-4.05	7.69	7.09	-7.00	-7.01	15.00	9.00	-14.60	-9.62	10.53
4	adj.	-0.65	0.70	0.75	4.37	2.61	0.94	4.62	1.76	4.02	1.75
	std.	-6.28	12.07	7.18	-7.71	-4.74	12.09	9.62	-10.58	-3.84	12.25
5	adj.	0.75	-0.50	4.16	4.51	2.29	-5.35	1.04	0.21	0.96	-0.42
	std.	-23.30	16.34	25.03	-14.48	-24.12	11.81	21.36	-20.46	-27.82	16.96
6	adj.	2.21	-0.03	-1.20	-2.66	-4.41	-4.85	2.66	-0.67	-4.27	0.39
	std.	-8.10	16.36	9.60	-20.87	-16.75	12.43	13.27	-18.46	-16.80	16.82
7	adj.	7.93	5.69	9.33	12.52	8.80	9.50	9.22	7.10	7.30	13.31
	std.	-45.85	50.54	49.23	-41.37	-49.97	51.17	47.32	-52.19	-56.04	52.61
8	adj.	0.06	-1.79	-0.17	2.04	0.08	-0.61	2.02	-0.24	0.42	-1.47
	std.	-21.05	14.17	20.78	-14.91	-21.25	15.11	22.32	-17.50	-21.40	15.05

Table 7. Estimation of $Cov\big[x_1(t), x_2(t)\big]$ over adj.; difference from simulation

Exp.		Time									
#	type	6	7	8	9	10	11	12	13	14	15
1	adj.	-12.00	-5.08	-9.57	-18.15	-9.48	-10.91	-5.76	-1.24	-3.10	-2.03
	std.	49.47	-20.54	-28.92	-6.61	35.86	-25.69	-21.25	6.58	36.03	-14.04
2	adj.	-15.19	-4.39	-9.89	-7.40	-10.49	-8.76	-6.09	-7.50	-4.81	0.90
	std.	38.21	-13.72	-16.37	6.60	27.72	-16.08	-10.74	4.81	29.09	-4.71
3	adj.	-2.93	-5.13	-3.47	-4.45	-2.58	-5.58	-1.78	-2.63	0.32	-3.80
	std.	37.28	-14.23	-7.63	9.48	28.60	-12.58	-4.81	9.25	28.44	-8.72
4	adj.	-16.78	-5.06	-0.41	-0.63	5.79	-2.72	3.03	-1.50	5.91	-2.11
	std.	21.08	-13.76	-3.38	13.17	30.93	-8.53	1.24	10.40	29.73	-5.99
5	adj.	-11.21	-7.95	-0.98	-5.45	-7.17	-13.04	-5.12	-5.41	-3.60	-7.30
	std.	100.00	-4.36	-11.86	19.11	52.47	-16.14	-15.25	13.33	46.73	-11.80
6	adj.	-3.61	2.87	-0.40	-3.34	-4.23	-2.08	-0.87	-0.09	-6.08	2.52
	std.	95.91	-30.60	-41.94	26.42	93.38	-36.81	-41.84	28.92	93.14	-30.60
7	adj.	-5.94	-5.85	0.97	-2.30	-2.63	2.14	2.75	-15.02	-12.12	-9.90
	std.	57.03	11.72	22.98	24.55	42.82	20.77	24.91	11.42	33.11	14.63
8	adj.	6.42	0.38	-8.32	-5.35	4.03	6.00	-4.38	5.43	5.73	-15.89
	std.	71.65	-22.00	-30.97	23.00	66.81	-14.74	-25.75	30.19	66.70	-42.77

Table 8. Estimation of $Var\big[x_2(t)\big]$ over time; difference from simulation

Exp.		Time									
#	type	6	7	8	9	10	11	12	13	14	15
1	adj.	5.37	10.58	5.05	-0.43	0.17	-1.26	-0.49	0.22	1.06	3.48
	std.	33.13	29.99	-16.71	-6.54	20.04	14.46	-15.56	-2.42	21.71	18.92
2	adj.	2.43	7.55	2.16	-4.56	-0.80	0.66	-1.39	-2.17	-0.58	1.54
	std.	14.93	12.50	-15.93	-11.18	7.57	4.35	-13.86	-5.86	9.61	5.96
3	adj.	-2.82	9.06	2.16	-2.74	0.72	3.39	1.65	3.42	3.82	3.03
	std.	3.14	8.16	-13.75	-7.76	6.63	2.89	-9.48	0.27	10.46	5.28
4	adj.	-3.45	3.60	0.54	-5.28	-1.15	1.10	-0.41	0.44	0.38	0.65
	std.	-2.85	-0.87	-13.49	-8.48	3.68	0.36	-9.05	-1.20	6.52	1.55
5	adj.	6.80	5.83	2.85	0.28	0.16	0.97	0.05	-0.19	1.20	1.24
	std.	100.00	63.29	-14.34	-1.32	28.92	22.89	-20.40	-5.91	24.08	16.17
6	adj.	10.02	13.30	4.08	4.66	11.23	9.03	1.55	12.40	10.59	9.95
	std.	97.96	3.94	-39.20	55.02	96.58	-1.78	-41.49	59.04	96.51	-0.47
7	adj.	-11.51	-4.29	-5.94	-8.15	-8.22	-8.18	-8.07	-6.98	-8.76	-12.30
	std.	30.39	4.97	-24.06	3.38	27.23	5.82	-16.49	8.18	28.48	3.15
8	adj.	-1.57	2.12	2.95	-2.75	0.43	1.81	0.52	1.54	3.77	0.84
	std.	52.96	53.03	20.08	27.87	43.84	46.93	21.02	31.76	45.23	45.18

APPENDIX B

MATHEMATICAL DERIVATION FOR CHAPTER 5

B.1. Derivation of functions $g_i^\eta(t,x)$'s

Since $f_i(\cdot,\cdot)$'s except $f_2(\cdot,\cdot)$ are linear or constant,

$$g_i^\eta(t,x) = \eta f_i(t, x/\eta) \quad \text{for } i \in \{1,3,4,5\}.$$

Now, we derive $g_2^\eta(t,x)$.

$$\begin{aligned}
g_2^\eta(t,x) &= \eta E\Big[\mu\Big\{\Big(\frac{x_1^\eta(t)}{\eta} - E\Big[\frac{x_1^\eta(t)}{\eta}\Big] + \frac{x_1}{\eta}\Big) \wedge \Big(\frac{x_2^\eta(t)}{\eta} - E\Big[\frac{x_2^\eta(t)}{\eta}\Big] + \frac{x_2}{\eta}\Big)\Big\}\Big] \\
&= E\Big[\mu\big\{\big(x_1^\eta(t) - E[x_1^\eta(t)] + x_1\big) \wedge \big(x_2^\eta(t) - E[x_2^\eta(t)] + x_2\big)\big\}\Big] \\
&\quad \text{Let } y(t) = (x_1^\eta(t) - E[x_1^\eta(t)] + x_1) - (x_2^\eta(t) - E[x_2^\eta(t)] + x_2). \\
&= \mu\Big[E[(y(t) \wedge 0)\mathbb{I}_{y(t)\le 0}]] + x_2\Big] \\
&= \mu\Bigg[(x_1 - x_2)Pr[y(t) \le 0] + x_2 \\
&\quad -\sigma^{\eta 2}(t)\frac{1}{\sqrt{2\pi}\sigma^\eta(t)}\exp\Big(-\frac{(x_1-x_2)^2}{2\sigma^{\eta 2}(t)}\Big)\Bigg] \\
&= \mu\Bigg[x_1 Pr[y(t) \le 0] + x_2 Pr[y(t) > 0] \\
&\quad -\sigma^{\eta 2}(t)\frac{1}{\sqrt{2\pi}\sigma^\eta(t)}\exp\Big(-\frac{(x_1-x_2)^2}{2\sigma^{\eta 2}(t)}\Big)\Bigg].
\end{aligned}$$

VITA

Young Myoung Ko received B.S. and M.S. degrees in Industrial Engineering from Seoul National University, Seoul, Korea. His research focuses mainly on the mathematical analysis of transient stochastic systems, and covers the domains of online service operations, communication networks and energy-aware system design.

He may be reached at his department address:

Department of Industrial and Systems Engineering
Texas A&M University, 3131 TAMU
College Station, TX. 77843-3131

This document was typeset in LaTeX by the author.