IIE Transactions (2010) **42**, 881–896

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/0740817X.2010.491501

Transient analysis of queues for peer-based multimedia content delivery

YOUNG MYOUNG KO and NATARAJAN GAUTAM*

Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843, USA
E-mail: gautam@tamu.edu

Received May 2009 and accepted March 2010

Consider a firm that sells online multimedia content. In order to manage costs and quality of service, this firm maintains a peer network that allows new users to download files from their peers who have previously downloaded the required files. The scenario can be modeled as a queueing system where the number of servers varies over time. Analytical models are developed that are based on fluid and diffusion approximations and allow analysis of transient system performance. The same approximations are used to analyze the steady-state behavior of this network. It is shown that the existing fluid and diffusion approximations are inaccurate for transient analysis. To address this shortcoming, a novel Gaussian-based adjustment is proposed and it significantly improves the accuracy of the approximations. Furthermore, the models used in this research can be extended seamlessly to the case of time-varying system parameters (e.g., arrival rates and service rates). Several numerical examples are provided that show how the proposed adjusted models work for the analysis of transient phenomena.

Keywords: Transient analysis, peer-to-peer networks, fluid and diffusion approximations, asymptotic analysis, online multimedia business

1. Introduction

The online multimedia market is growing at an unprecedented rate. This growth creates increasing demand for network resources (e.g., bandwidth, servers, storage, etc.) and forces a service provider, which we will call a *company* for the remainder of this article, to provide enough resources to produce an adequate Quality of Service (QoS) for its customers. Currently, the market is limited to music files which do not impose a 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 significantly 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 created 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, a peer-to-peer (P2P) architecture can be a viable alternative for a company in that it allows the company to "outsource" resources to peers instead of purchasing all the required resources itself. In other words, the company could redirect customer requests to peers who have downloaded those files in the past. The P2P architecture has already been shown to be stable and scalable in many previous research studies, such as Ge *et al.* (2003), Qiu and Srikant (2004), 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 is continuously increasing (Fraleigh *et al.*, 2003; Gummadi *et al.*, 2003). Despite these benefits (stability and scalability) and the popularity of the P2P architecture, it has not yet been broadly adapted to commercial companies, since it is regarded as a source of illegal content distribution, a perception driven by current free P2P software (e.g., eDonkey, Bittorent, etc.), and thus the company cannot control the distribution of its product. If the content distribution could be brought under the control of companies, then 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 (http://www.pando.com) are operating P2P networks for content distribution. Furthermore, even companies such as Akamai that provide more established content distribution

*Corresponding author

networks by locating caching servers also seem to be interested in P2P architectures (for example, Akamai purchased a P2P content distribution system called "Redswoosh" in 2007).

Having described the merits of peer-based networks we now describe a major drawback that can arise before the peer network is mature. When new content (e.g., a 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 a P2P architecture. That is the objective of this article.

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 are focusing on the performance analysis of peer network transient behavior, rather than steady-state behavior.

For the transient analysis, we adopt methods derived from fluid and diffusion approximations. Fluid and diffusion approximations of Markov processes based on so-called strong approximations have been established by Kurtz (1978) and are summarized in Ethier and Kurtz (1986). Mandelbaum *et al.* (1998) applied strong approximations to time-varying-rate cases and establish the framework to analyze Markovian queueing networks (also see Mandelbaum *et al.* (2002) and Massey (2002)). In addition, they extend the results in Kurtz (1978) to apply the strong approximations to non-differentiable rate functions of the system state by defining a new derivative called a *scalable Lipschitz derivative*. The theorems used in the strong approximation are functional extensions of the well known "Strong Law of Large Numbers" and "Central Limit Theorem." In fact, there are several ways other than strong approximations to obtain limit processes in different limiting schemes. Methodologies to obtain limit processes are well summarized in Billingsley (1998) and Whitt (2002). Recently, these methods have been used for transient analysis and control of online rental systems such as Netflix (Bassamboo *et al.*, 2009; Bassamboo and Randhawa, 2009). By their nature, methodologies utilizing limit processes are appropriate for modeling large-scale systems. As a result, they have gained popularity for the analysis and design of call-center-like systems (Whitt, 2004, 2006). Specifically, Hampshire *et al.* (2009) utilized strong approximation to solve a call center design problem under a time-varying environment.

Strong approximations have been used in the context of peer networks (Qiu and Srikant, 2004) to analyze steady-state behavior. They show weak convergence to the Ornstein–Uhlenbeck (OU) process in steady-state. As mentioned before, however, their research does not focus on the transient analysis of a peer network in which the network is evolving and shows dynamic behaviors. As described in the previous paragraph, our approach in this research is based on the results in Kurtz (1978). Our model, however, has non-differentiable rate functions of the system state. Although the results of Mandelbaum *et al.* (1998) provide rigorous mathematical models to deal with these non-differentiable rate functions, their model cannot be applied to our scenario because it involves difficulties in computing the covariance matrix entries. For example, one of the differential equations to obtain the variance of the number of customers in a multi-server queueing system with abandonments and retrials in Mandelbaum *et al.* (1998) is of the form:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Var}\big[Q_1^{(1)}(t)\big] \\
&\quad = 2\Big(\beta_t \mathbf{1}_{Q_1^{(0)}(t)>n_t} + \mu_t^1 \mathbf{1}_{Q_1^{(0)}(t)\le n_t}\Big)\mathrm{Cov}\big[Q_1^{(1)}(t),\, Q_1^{(1)}(t)^-\big] \\
&\qquad - 2\Big(\beta_t \mathbf{1}_{Q_1^{(0)}(t)\ge n_t} + \mu_t^1 \mathbf{1}_{Q_1^{(0)}(t)<n_t}\Big)\mathrm{Cov}\big[Q_1^{(1)}(t),\, Q_1^{(1)}(t)^+\big] \\
&\qquad + \lambda_t + \beta\big(Q_1^{(0)}(t) - n_t\big)^+ + \mu_t^1\big(Q_1^{(0)}(t) \wedge n_t\big) + \mu_t^2 Q_2^{(0)}(t).
\end{aligned} \tag{1}$$

On the right-hand side of Equation (1), the term $\mathrm{Cov}[Q_1^{(1)}(t),\, Q_1^{(1)}(t)^-]$ (or $\mathrm{Cov}[Q_1^{(1)}(t),\, Q_1^{(1)}(t)^+]$) makes it impossible to solve the differential equation unless we know the functional relationship between $\mathrm{Cov}[Q_1^{(1)}(t),\, Q_1^{(1)}(t)]$ and $\mathrm{Cov}[Q_1^{(1)}(t),\, Q_1^{(1)}(t)^-]$ (or $\mathrm{Cov}[Q_1^{(1)}(t),\, Q_1^{(1)}(t)^+]$). Thus, in this article, we provide a new way to: (i) cope with the inaccuracy of existing approximations and (ii) achieve computational feasibility.

The rest of the article is organized as follows. In Section 2, we explain the system we are considering and establish mathematical models in detail. In Section 3, we analyze our system with fluid and diffusion approximations based on the results of Kurtz (1978) and Mandelbaum *et al.* (2002). We call these *standard* fluid and diffusion models in the rest of the article to distinguish them from our adjusted models. It turns out that both fluid and diffusion approximations work well in steady-state conditions. We, however, show significant inaccuracy in both fluid and diffusion approximations during a transient period. In Section 4, we explain our new adjustment approach and show the improved approximations. In order to validate our adjusted model and to see the effects of parameters, several numerical examples are provided in Section 5. We also show, through numerical

experiments, that our adjusted model gives precise results under time-varying rate functions. In Section 6, we provide concluding remarks and suggest extensions for future research.

2. Problem description

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

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 the media content and customers access and purchase this content via the company's Web site. The company operates a peer network consisting of peers who have purchased from the content company and are given authorization to pass this content on to new customers. The company manages a single queue for waiting customers and allocates a customer at the head of the queue to a peer when the peer becomes available. Figure 1 is a simplified illustration of our target system. When new content is created, the company prepares an 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 these customers become new peers as soon as they complete the download of the content and can then share the content with customers arriving in the future. Peers can move between an active peer pool and an inactive peer pool as they 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 a peer leaves or moves to the inactive pool while serving a customer, that customer is either allocated to an available peer in the active peer pool or is sent back to the queue. The peer network grows when a new peer joins and shrinks when a peer leaves. Throughout this article, 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 is a straightforward extension that we show later in this article. 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 article, 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.

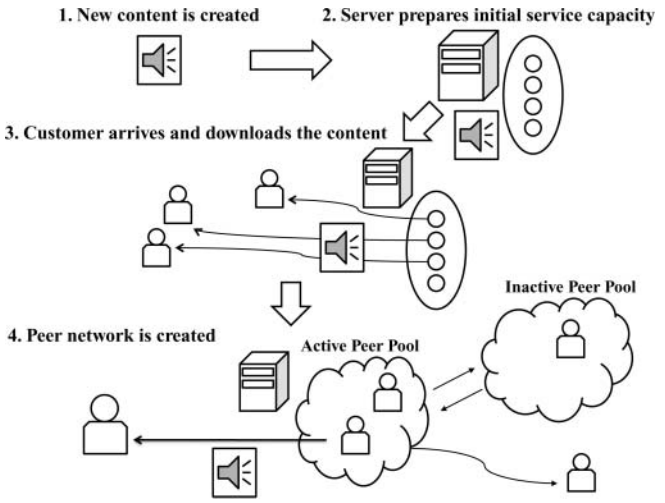

Fig. 1. System illustration.

2.2. *Mathematical model*

Let $\mathbf{X}(t) = (x(t), y(t), z(t))^{\mathrm{T}}$ denote the state of the system at time t where $x(t)$ is the number of customers in the system, i.e., those who are waiting in the queue or are downloading the content, $y(t)$ is the number of peers in the active peer pool and $z(t)$ 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 2 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 a single queue, which is managed by the company. Therefore, this process can be characterized as an $M/M/y(t)$-type queue with server vacations in which the number of servers changes over time. Here, we use a 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.

2.3. *Objective*

Figure 3 illustrates a typical evolution of peer networks. From Fig. 3, we can define three stages based on the

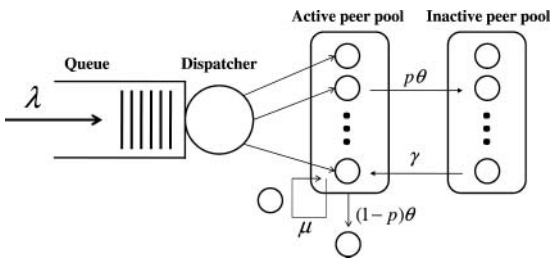

Fig. 2. Simplified system model.

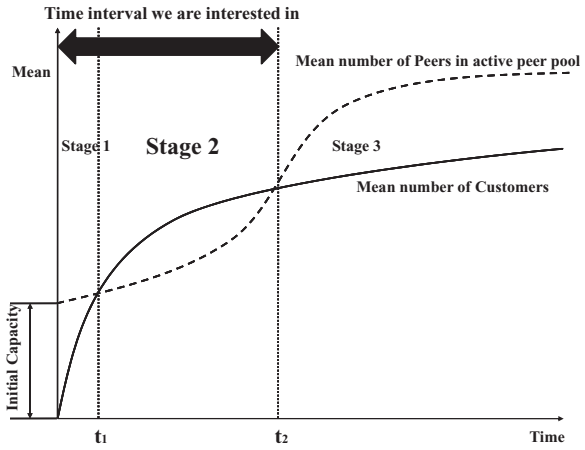

Fig. 3. Typical evolution of peer networks on average.

number of customers and peers. At the beginning of stage 1 (i.e., $t = 0$), the company prepares its 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 formed. 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 (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 accurately 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 3.

3. Fluid and diffusion approximations

In this section, we extend fluid and diffusion approximations using the method provided by Kurtz (1978) and Mandelbaum *et al.* (2002) for our problem. After developing the results, we will show the inadequacy of these approximations. Fluid and diffusion approximations are used in several previous studies (Mandelbaum *et al.*, 1998; Mandelbaum *et al.*, 2002; Qiu and Srikant, 2004; Whitt, 2004, 2006). The first step of this approach is to define a sequence of stochastic processes and to obtain the fluid model by taking the limit of the sequence. A fluid model takes the role of the expected value for each time point. The second step is to obtain a diffusion model by taking the limit to the centered process multiplied by some adequate scaling factor. In Markovian networks, this centered process converges to a Gaussian process under certain conditions that are described later.

Consider $\mathbf{X}(t) = (x(t), y(t), z(t))^{\mathrm{T}}$ as defined in Section 2.2. Assume that there is no customer and the company prepares C service units at time $t = 0$; i.e., $\mathbf{X}(0) = (0, C, 0)^{\mathrm{T}}$. Then, for our model, the sample path can be constructed using the following equation:

$$\begin{aligned}\mathbf{X}(t) &= \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ C \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} A_1\left(\int_0^t \lambda \mathrm{d}s \right) \\ &\quad + \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} A_2\left(\int_0^t \mu \min(x(s), y(s)) \mathrm{d}s \right) \\ &\quad + \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} A_3\left(\int_0^t p\theta y(s) \mathrm{d}s \right) + \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} A_4 \\ &\quad \left(\int_0^t (1-p)\theta y(s) \mathrm{d}s \right) \\ &\quad + \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} A_5\left(\int_0^t \gamma z(s) \mathrm{d}s \right), \end{aligned} \tag{2}$$

where $A_1(\cdot)$, $A_2(\cdot)$, $A_3(\cdot)$, $A_4(\cdot)$, and $A_5(\cdot)$ are independent Poisson processes corresponding to customer arrival, service, peer up, peer leaving, and peer down, respectively. To apply fluid and diffusion approximations to Equation (2), consider a sequence of stochastic processes $\{\mathbf{X}_n(t)\}_{n \geq 1}$ so that $\mathbf{X}_n(t)$ is the solution to the following equation:

$$\begin{aligned}\mathbf{X}_n(t) = \begin{pmatrix} x_n(t) \\ y_n(t) \\ z_n(t) \end{pmatrix} &= \begin{pmatrix} 0 \\ C \\ 0 \end{pmatrix} + \frac{1}{n} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} A_1\left(n \int_0^t \lambda \mathrm{d}s \right) \right. \\ &\quad + \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} A_2\left(n \int_0^t \mu \min(x_n(s), y_n(s)) \mathrm{d}s \right) \\ &\quad + \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} A_3\left(n \int_0^t p\theta y_n(s) \mathrm{d}s \right) \\ &\quad + \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} A_4\left(n \int_0^t (1-p)\theta y_n(s) \mathrm{d}s \right) \\ &\quad \left. + \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} A_5\left(n \int_0^t \gamma z_n(s) \mathrm{d}s \right) \right\}. \end{aligned} \tag{3}$$

Note that n is a scaling factor so that we obtain the fluid approximation model by letting $n \to \infty$ for $\{\mathbf{X}_n(t)\}$. That is described in the following theorem.

Theorem 1. *(Deterministic fluid model.) Let* $\bar{\mathbf{X}}(t)$ *denote the deterministic fluid model corresponding to* $\mathbf{X}_n(t)$ *that satisfies:*

$$\begin{aligned}\bar{\mathbf{X}}(t) &= \begin{pmatrix}\bar{x}(t)\\ \bar{y}(t)\\ \bar{z}(t)\end{pmatrix}\\ &= \begin{pmatrix}0\\ C\\ 0\end{pmatrix} + \int_0^t \left[\begin{pmatrix}1\\0\\0\end{pmatrix}\lambda + \begin{pmatrix}-1\\1\\0\end{pmatrix}\mu\min(\bar{x}(s),\bar{y}(s))\right.\\ &\quad + \begin{pmatrix}0\\-1\\1\end{pmatrix}p\theta\bar{y}(s) + \begin{pmatrix}0\\-1\\0\end{pmatrix}(1-p)\theta\bar{y}(s)\\ &\quad \left. + \begin{pmatrix}0\\1\\-1\end{pmatrix}\gamma\bar{z}(s)\right]\mathrm{d}s. \end{aligned} \tag{4}$$

Then, $\lim_{n\to\infty}\mathbf{X}_n(t) = \bar{\mathbf{X}}(t)$ *a.s.*

Proof. Let $\mathbf{X} = (x, y, z)^{\mathrm{T}}$ and define $f_1(\mathbf{X}) = \lambda$, $f_2(\mathbf{X}) = \mu\min(x, y)$, $f_3(\mathbf{X}) = \theta p y$, $f_4(\mathbf{X}) = \theta(1-p)y$, and $f_5(\mathbf{X}) = \gamma z$. Then, Equation (3) can be written as

$$\mathbf{X}_n(t) = \begin{pmatrix}0\\C\\0\end{pmatrix} + \sum_{i=1}^{5}\frac{1}{n}\mathbf{l}_i A_i\left(n\int_0^t f_i(\mathbf{X}_n(s))\mathrm{d}s\right),$$

where

$$\mathbf{l}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix},\quad \mathbf{l}_2 = \begin{pmatrix}-1\\1\\0\end{pmatrix},\quad \mathbf{l}_3 = \begin{pmatrix}0\\-1\\1\end{pmatrix},$$

$$\mathbf{l}_4 = \begin{pmatrix}0\\-1\\0\end{pmatrix} \text{ and } \quad \mathbf{l}_5 = \begin{pmatrix}0\\1\\-1\end{pmatrix}.$$

Then, it is easy to verify that the $f_i(\cdot)$s are Lipschitz and there exist ϵ_is such that $|f_i(\mathbf{X})| \le \epsilon_i(1 + |\mathbf{X}|)$. Since $\sum|\mathbf{l}_i|^2\epsilon_i < \infty$, by Theorem 2.1 and 2.2 in Kurtz (1978), $\lim_{n\to\infty}\mathbf{X}_n(t) = \bar{\mathbf{X}}(t)$ a.s. ■

Before moving to the diffusion approximation model, we investigate the graph of the fluid model over time since the fluid model is closely related to the diffusion model, which will be explained in Theorem 2. The fluid model is deterministic and typically its graph is similar to Fig. 3. In the original process (i.e., $\mathbf{X}(t)$), 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 fluid approximation results:

$$\begin{aligned}t_1 &= \inf\{t : \bar{x}(t) = \bar{y}(t), t \ge 0\},\\ t_2 &= \inf\{t : \bar{x}(t) = \bar{y}(t), t > t_1\}.\end{aligned}$$

Notice t_1 and t_2, depicted in Fig. 3, for further clarification. The switching times t_1 and t_2 can be obtained directly by solving Equation (4). Defining t_1 and t_2 using the fluid model is reasonable since the queue is empty at t_2 on average.

Now we move our attention to the diffusion model. For the diffusion model, we apply central limit theorem by defining the scaled centered process.

Theorem 2. *(Diffusion approximation.) Let* $\mathbf{D}_n(t)$ *be the scaled centered process; i.e.,* $\mathbf{D}_n(t) = \sqrt{n}\big(\mathbf{X}_n(t) - \bar{\mathbf{X}}(t)\big)$ *and assume measure zero at* t_1 *and* t_2*. Then, we can define the diffusion approximation model as*

$$\mathbf{D}(t) = (d_1(t), d_2(t), d_3(t))^{\mathrm{T}} = \lim_{n\to\infty}\sqrt{n}(\mathbf{X}_n(t) - \bar{\mathbf{X}}(t)).$$

Define the matrices $\mathbf{K}_1$, $\mathbf{K}_2$, *and* $\mathbf{L}(t)$ *as follows:*

$$\mathbf{K}_1 = \begin{pmatrix}-\mu & 0 & 0\\ \mu & -\theta & \gamma\\ 0 & p\theta & -\gamma\end{pmatrix},$$

$$\mathbf{K}_2 = \begin{pmatrix}0 & -\mu & 0\\ 0 & \mu-\theta & \gamma\\ 0 & p\theta & -\gamma\end{pmatrix},$$

and

$$\mathbf{L}(t) = \begin{pmatrix}\sqrt{\lambda} & -\sqrt{\mu\min(\bar{x}(t),\bar{y}(t))} & 0 & 0 & 0\\ 0 & \sqrt{\mu\min(\bar{x}(t),\bar{y}(t))} & -\sqrt{p\theta\bar{y}(t)} & -\sqrt{(1-p)\theta\bar{y}(t)} & \sqrt{\gamma\bar{z}(t)}\\ 0 & 0 & \sqrt{p\theta\bar{y}(t)} & 0 & -\sqrt{\gamma\bar{z}(t)}\end{pmatrix}.$$

Then, $\mathbf{D}(t)$ *is the solution of the following integral equation: for* $0 \le t < t_1$*:*

$$\mathbf{D}(t) = \int_0^t \mathbf{L}(s)\begin{pmatrix}\mathrm{d}B_1(s)\\ \mathrm{d}B_2(s)\\ \mathrm{d}B_3(s)\\ \mathrm{d}B_4(s)\\ \mathrm{d}B_5(s)\end{pmatrix} + \int_0^t \mathbf{K}_1\cdot\mathbf{D}(s)\mathrm{d}s, \tag{5}$$

for $t_1 \le t < t_2$*:*

$$\begin{aligned}\mathbf{D}(t) &= \begin{pmatrix}d_1(t_1)\\ d_2(t_1)\\ d_3(t_1)\end{pmatrix} + \int_{t_1}^t \mathbf{L}(s)\begin{pmatrix}\mathrm{d}B_1(s)\\ \mathrm{d}B_2(s)\\ \mathrm{d}B_3(s)\\ \mathrm{d}B_4(s)\\ \mathrm{d}B_5(s)\end{pmatrix}\\ &\quad + \int_{t_1}^t \mathbf{K}_2\cdot\mathbf{D}(s)\mathrm{d}s,\end{aligned} \tag{6}$$

and for $t \geq t_2$*:*

$$\mathbf{D}(t) = \begin{pmatrix} d_1(t_2) \\ d_2(t_2) \\ d_3(t_2) \end{pmatrix} + \int_{t_2}^{t} \mathbf{L}(s) \begin{pmatrix} \mathrm{d}B_1(s) \\ \mathrm{d}B_2(s) \\ \mathrm{d}B_3(s) \\ \mathrm{d}B_4(s) \\ \mathrm{d}B_5(s) \end{pmatrix} + \int_0^t \mathbf{K}_1 \cdot \mathbf{D}(s)\mathrm{d}s, \quad (7)$$

where $B_1(t)$*,* $B_2(t)$*,* $B_3(t)$*,* $B_4(t)$*, and* $B_5(t)$ *are independent standard Brownian motions.*

Proof. With the same definition of $\mathbf{X}$, $\mathbf{l_i}$ and $f_i(\cdot)$ as in the proof of Theorem 1, define $F(\mathbf{X})$ as follows:

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

Then, by Kurtz (1978), the centered process $\mathbf{D}(t)$ satisfies the following integral equation:

$$\mathbf{D}(t) = \sum_{i=1}^{5} \mathbf{l}_i \int_0^t \sqrt{f_i(\bar{\mathbf{X}}(s))}\, dB(s) + \int_0^t \partial F(\bar{\mathbf{X}}(s)) \cdot \mathbf{D}(s)\mathrm{d}s,$$

where $\partial F(\bar{\mathbf{X}}(t))$ is the gradient of $F(\bar{\mathbf{X}}(t))$. For $0 \leq t < t_1$, Equation (5) is straightforward. However, according to Kurtz (1978), the drift matrix of Equation (5) and Equation (6) requires differentiability at any time point. In our model, we fail to satisfy differentiability at times t_1 and t_2. We can resolve this problem by assuming measure zero at t_1 and t_2 similar to what Mandelbaum *et al.* (2002) considers. Then, we can obtain Equation (6) for $t_1 \leq t < t_2$ and Equation (7) for $t \geq t_2$. ■

Note that the diffusion model in Equations (5), (6), and (7) turns out to be a Gaussian process and is closely related to the fluid model ($\bar{\mathbf{X}}(t)$). Depending on the fluid model, the diffusion model changes its behavior at time points t_1 and t_2.

Theorem 2 indicates that the diffusion model is a linear model. Therefore, we could obtain the expectation and covariance matrix of $\mathbf{D}(t)$ in the following way.

Theorem 3. *(Expectation and covariance matrix.) Let* $\mathbf{m}(t)$ *denote* $E[\mathbf{D}(t)]$ *and* $\mathbf{\Sigma}(t)$ *denote* $\mathrm{Cov}[\mathbf{D}(t), \mathbf{D}(t)]$*. Then, with the same definition of* $\mathbf{K}_1$*,* $\mathbf{K}_2$*, and* $\mathbf{L}(t)$ *as in Theorem 2,* $\mathbf{m}(t)$ *is the solution to the following differential equation: for* $0 \leq t < t_1$ *or* $t \geq t_2$*:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{m}(t) = \mathbf{K}_1 \cdot \mathbf{m}(t), \quad (8)$$

and for $t_1 \leq t < t_2$*:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{m}(t) = \mathbf{K}_2 \cdot \mathbf{m}(t). \quad (9)$$

Moreover, $\mathbf{\Sigma}(t)$ *is the unique symmetric semi-positive definite solution to the following differential equation: for* $0 \leq t < t_1$ *or* $t \geq t_2$*:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{\Sigma}(t) = \mathbf{K}_1 \cdot \mathbf{\Sigma}(t) + \mathbf{\Sigma}(t) \cdot \mathbf{K}_1^{\mathrm{T}} + \mathbf{L}(t) \cdot \mathbf{L}(t)^{\mathrm{T}}, \quad (10)$$

and for $t_1 \leq t < t_2$*.*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{\Sigma}(t) = \mathbf{K}_2 \cdot \mathbf{\Sigma}(t) + \mathbf{\Sigma}(t) \cdot \mathbf{K}_2^{\mathrm{T}} + \mathbf{L}(t) \cdot \mathbf{L}(t)^{\mathrm{T}}. \quad (11)$$

Proof. For $0 \leq t < t_1$, we know that $E[\mathbf{D}(0)] = 0 < \infty$ since $\mathbf{D}(0) = 0$. Then, by Theorem 8.2.6 in Arnold (1992), $\mathbf{m}(t)$ and $\mathbf{\Sigma}(t)$ satisfy Equation (8) and Equation (10). From Equation (8), we also have $E[\mathbf{D}(t_1)] < \infty$. Therefore, we can also apply Theorem 8.2.6 in Arnold (1992) and obtain Equations (9) and (11). Since $E[\mathbf{D}(t_2)] < \infty$, we obtain Equations (8) and (10) for $t \geq t_2$. ■

Summarizing, we established the fluid and diffusion models. We found that the diffusion model is a Gaussian process and that the mean vector and covariance matrix can be obtained by solving the ordinary differential equations from (8) to (11). Once we build the fluid and diffusion models, we need to define the approximation for our original process. Based on the definition of $\mathbf{D}(t)$, we use $\bar{\mathbf{X}}(t) + \mathbf{D}(t)$ as an approximation of $\mathbf{X}(t)$ (i.e., $\mathbf{X}(t) \approx \bar{\mathbf{X}}(t) + \mathbf{D}(t)$). By Theorem 3, we obtain $E[\mathbf{D}(t)] = \mathbf{m}(t) = \mathbf{0}$ for all $t \geq 0$ since $\mathbf{m}(0) = E[\mathbf{D}(0)] = E[\lim_{n\to\infty} \sqrt{n}(\mathbf{X}_n(0) - \bar{\mathbf{X}}(0))] = E[\lim_{n\to\infty} \sqrt{n}(\mathbf{x}_0 - \mathbf{x}_0)] = \mathbf{0}$. Therefore,

$$E[\mathbf{X}(t)] \approx E[\bar{\mathbf{X}}(t)] + E[\mathbf{D}(t)] = \bar{\mathbf{X}}(t) + \mathbf{0},$$

and

$$\mathrm{Cov}[\mathbf{X}(t), \mathbf{X}(t)] \approx \mathrm{Cov}[\mathbf{D}(t), \mathbf{D}(t)].$$

Figure 4 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 Fig. 4(a) is for $\bar{\mathbf{X}}(t)$ and Fig. 4(b) for $\mathbf{\Sigma}(t)$. The simulation result is obtained by averaging 5000 simulation runs. We see that the fluid and diffusion models are close to the simulation results when t is small. We, however, notice that the fluid and diffusion models show big differences, especially in covariance matrix entries around t_2. We find two significant problems in the fluid and diffusion models from Fig. 4. Let t_2' denote the switching time between stages 2 and 3 in the simulation result. Then, the following points can be made.

1. The fluid model shows some error near the time t_2'. From the experiments with different parameters, we see that the fluid model 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 model at time t_2. Moreover, our diffusion model shows a significant difference from the simulation result around t_2. These spikes come from the sudden change of the drift matrix from $\mathbf{K}_2$ and $\mathbf{K}_1$ at time t_2 in Theorem 2 and this

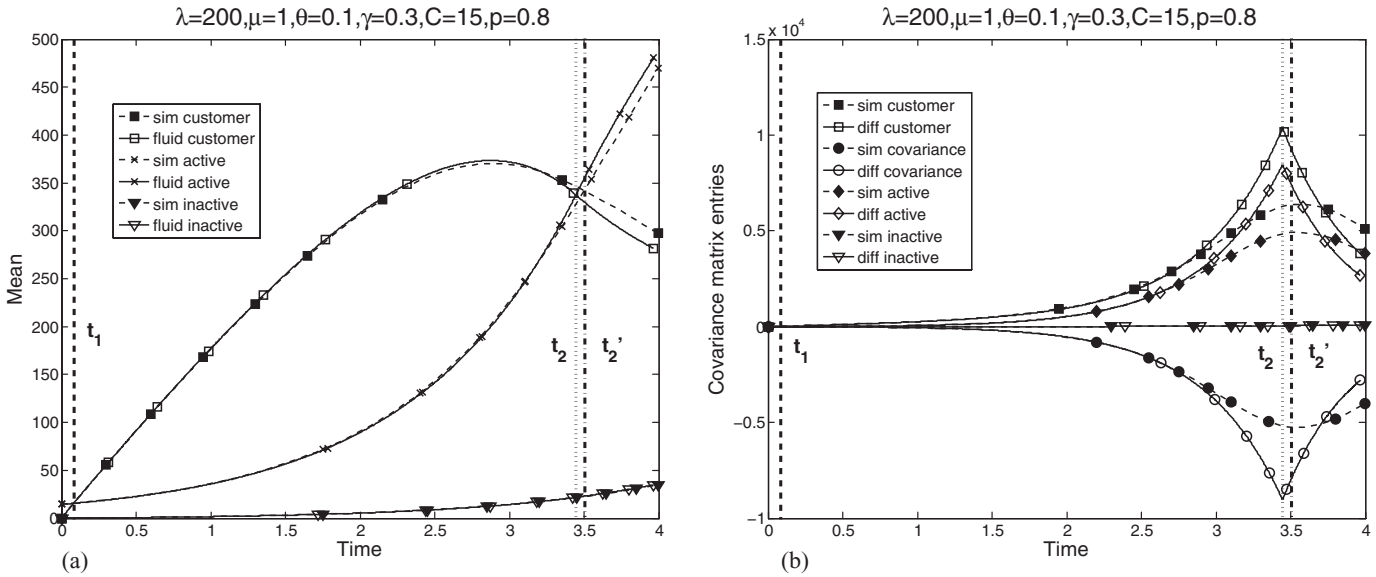

Fig. 4. Standard fluid and diffusion approximations: (a) mean number of customers and peers and (b) covariance matrix entries.

switching is caused by the non-differentiability of the $\min(\bar{x}(t), \bar{y}(t))$ in the fluid model.

Remark 1. 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 propose a Gaussian-based adjustment for the fluid and diffusion models and will explain it in the next section.

However, before moving to the next section, we provide the steady-state behavior of the diffusion model since fluid and diffusion approximations work well in steady-state. From Theorems 1 and 2, we notice that $\min(\bar{x}(t), \bar{y}(t)) = \bar{x}(t)$ for $t > t_2$ and this implies that the non-differentiability of $\min(\bar{x}(t), \bar{y}(t))$ 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 model for our original process is also an OU process in steady-state.

Theorem 4. *(Steady-state behavior.) Let* $\mathbf{D}(\infty)$ *be the scaled centered process* $\mathbf{D}(t)$ *defined in Theorem 2 when* $t \to \infty$. *Then, for* $0 \le p < 1$, $\mathbf{D}(\infty)$ *is a three-dimensional OU process with the drift matrix given by*

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

and the diffusion coefficient matrix given by

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

Proof. When $t > t_2$, the drift matrix is given by $\mathbf{K}$. By solving differential equations in Equation (4) for $t > t_2$ and taking $t \to \infty$, we obtain:

$$\lim_{t\to\infty} \bar{x}(t) = \frac{\lambda}{\mu}, \tag{12}$$

$$\lim_{t\to\infty} \bar{y}(t) = \frac{\lambda}{(1-p)\theta}, \tag{13}$$

$$\lim_{t\to\infty} \bar{z}(t) = \frac{\lambda p}{(1-p)\gamma}. \tag{14}$$

Then, by Theorem 2 and Equations (12) to (14), we have:

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

■

Remark 2. Notice that the steady-state number of customers, active peers and inactive peers via Equations (12) to (14) are respectively λ/μ, $\lambda/((1-p)\theta)$ and $\lambda p/((1-p)\gamma)$. The simulations also converge to the same values.

4. Adjusting the fluid and diffusion models

In the previous section, we saw that spikes in the diffusion model are caused by the non-differentiability of the

"min" function in the fluid model. In addition to non-differentiability, notice that the "min" function causes error in the fluid model itself. From the following simple lemma, we can explain the error in the fluid model.

Lemma 1. *Let X and Y be random variables such that $E[X] < \infty$ and $E[Y] < \infty$. Then,*

$$E[\min(X, Y)] \leq \min(E(X), E(Y)).$$

Recall that when solving Equation (4) in Theorem 1, we actually solve the following differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{x}(t) = \lambda - \mu \min(\bar{x}(t), \bar{y}(t)), \tag{15}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{y}(t) = \mu \min(\bar{x}(t), \bar{y}(t)) - \theta\bar{y}(t) + \gamma\bar{z}(t), \tag{16}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{z}(t) = p\theta\bar{y}(t) - \gamma\bar{z}(t).$$

In Section 3, for any time point t, we regard $E[\mathbf{X}(t)]$ as $\bar{\mathbf{X}}(t)$ (i.e., $\min(\bar{x}(t), \bar{y}(t)) = \min(E[x(t)], E[y(t)])$). We, however, observe $E[\min(x(t), y(t))]$ rather than $\min(E[x(t)], E[y(t)])$ in simulations and from Lemma 1, we have $E[\min(x(t), y(t))] \leq \min(E[x(t)], E[y(t)])\ \forall t \in [0, \infty)$. Therefore, we can verify that the increasing rate of $\bar{x}(t)$ is less than the increasing rate of $E[x(t)]$ in simulations, and the increasing rate of $\bar{y}(t)$ is greater than the increasing rate of $E[y(t)]$ in simulations from Equations (15) and (16). This implies that the fluid model should underestimate the switching time between stages 2 and 3 and shows the error compared with the simulation results. To fix this problem, we use the following theorem.

Theorem 5. *Let $\mathbf{X}(t)$ be the stochastic process satisfying the following equation:*

$$\mathbf{X}(t) = \mathbf{x}_0 + \sum_{\mathbf{l}} \mathbf{l} A_{\mathbf{l}}\left(\int_0^t f_{\mathbf{l}}(\mathbf{X}(s))\mathrm{d}s\right), \tag{17}$$

where $\mathbf{l} \in Z^d$, $\mathbf{x}_0 = \mathbf{X}(0)$ which is constant, as described in Section 3, the $A_{\mathbf{l}}$ are independent Poisson processes, and $f_{\mathbf{l}}$ are non-negative and satisfy the conditions defined in Kurtz (1978). Then, $E[\mathbf{X}(t)]$ is the solution to the following equation:

$$E[\mathbf{X}(t)] = \mathbf{x}_0 + \sum_{\mathbf{l}} \mathbf{l}\int_0^t E[f_{\mathbf{l}}(\mathbf{X}(s))]\mathrm{d}s. \tag{18}$$

Proof. Take expectation on both sides of Equation (17). Then,

$$\begin{aligned} E\big[\mathbf{X}(t)\big] &= E\left[\mathbf{x}_0 + \sum_{\mathbf{l}} \mathbf{l} A_{\mathbf{l}}\left(\int_0^t f_{\mathbf{l}}(\mathbf{X}(s))\mathrm{d}s\right)\right] \\ &= \mathbf{x}_0 + \sum_{\mathbf{l}} \mathbf{l} E\left[A_{\mathbf{l}}\left(\int_0^t f_{\mathbf{l}}\big(\mathbf{X}(s)\big)\mathrm{d}s\right)\right] \\ &= \mathbf{x}_0 + \sum_{\mathbf{l}} \mathbf{l} E\left[\int_0^t f_{\mathbf{l}}(\mathbf{X}(s))\mathrm{d}s\right] \end{aligned}$$

due to Poisson process's expected value

$$= \mathbf{x}_0 + \sum_{\mathbf{l}} \mathbf{l}\int_0^t E\Big[f_{\mathbf{l}}\big(\mathbf{X}(s)\big)\Big]\mathrm{d}s$$

by the conditions in Kurtz (1978) and Fubini theorem. ■

Corollary 1. *If the $f_{\mathbf{l}}(\mathbf{X})$ are constant or a linear combination of the components of $\mathbf{X}$, then,*

$$E[\mathbf{X}(t)] = \bar{\mathbf{X}}(t),$$

where $\mathbf{X}(t)$ is the solution to Equation (17) and $\bar{\mathbf{X}}(t)$ is the deterministic fluid model from Theorem 1.

Proof. Using the linearity of expectation, the Equation (18) is the same as that for the fluid model. ■

Remark 3. In many situations, the $f_{\mathbf{l}}$ are constant or linear combinations of components of $\mathbf{X}$. In these cases, Theorem 5 and Corollary 1 imply that standard fluid model would be a good approximation for the expected value of the system state.

If we use the solution of Equation (18) as a fluid approximation model instead of the solution of Equation (4) in Theorem 1, we expect to obtain more accurate results. However, to solve Equation (18), we encounter a fundamental problem. To obtain $E[\min(x(t), y(t))]$, we need to know the joint distribution of $x(t)$ and $y(t)$ for any time point t. Unfortunately, there is no explicit way to obtain the joint distribution of them and hence we need to assume it in a reasonable way. Recall that in Section 3, we saw that $\mathbf{X}(t)$ is approximated by the Gaussian process, and mean and variance were obtained from $\bar{\mathbf{X}}(t)$ and $\mathbf{D}(t)$, respectively. In addition, from previous research studies such as Mandelbaum and Pats (1998) and Mandelbaum *et al.* (2002), we notice that empirical densities of original processes are well matched with Gaussian density in several applications, even if the rate functions are non-differentiable. Therefore, it could be reasonable to use a Gaussian density function to calculate $E[\min(x(t), y(t))]$. Then, we can rewrite Equation (18) as a differential equation form to fit our model as follows:

$$\begin{aligned} \frac{\mathrm{d}\bar{x}(t)}{\mathrm{d}t} &= \lambda - \mu\{q(t)\bar{x}(t) + (1 - q(t))\bar{y}(t) - \sigma^2(t)\phi(0, \bar{x}(t) \\ &\quad - \bar{y}(t), \sigma(t))\}, \end{aligned} \tag{19}$$

$$\begin{aligned} \frac{\mathrm{d}\bar{y}(t)}{\mathrm{d}t} &= \mu\big\{q(t)\bar{x}(t) + (1 - q(t))\bar{y}(t) - \sigma^2(t)\phi\big(0, \bar{x}(t) \\ &\quad - \bar{y}(t), \sigma(t)\big)\big\} - \theta\bar{y}(t) + \gamma\bar{z}(t), \end{aligned} \tag{20}$$

$$\frac{\mathrm{d}\bar{z}(t)}{\mathrm{d}t} = p\theta\bar{y}(t) - \gamma\bar{z}(t), \tag{21}$$

where $q(t) = P(x(t) - y(t) \leq 0)$, $\sigma^2(t)$ is the variance of $x(t) - y(t)$ and $\phi(a, b, c)$ is the value at a of the probability density function of the Gaussian distribution with mean b and standard deviation c. Note that, since for any t, $(x(t), y(t))$ follows bivariate normal distribution,

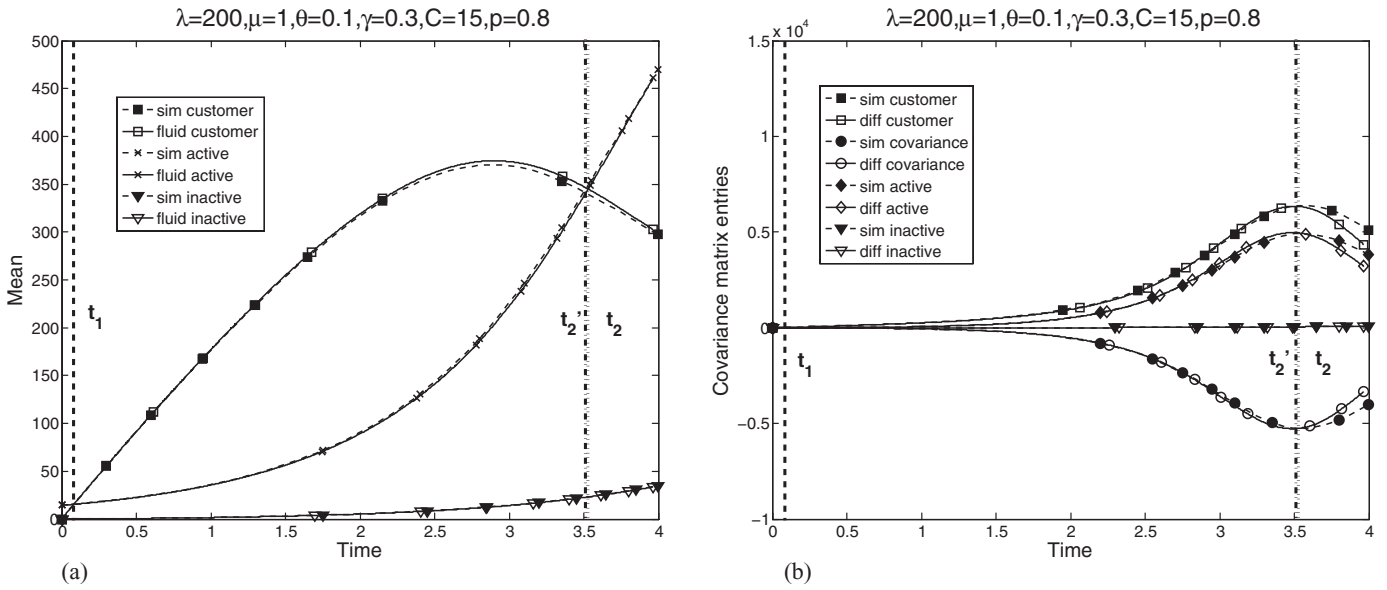

Fig. 5. Adjusted fluid and diffusion approximations with adjustment: (a) mean number of customers and peers and (b) covariance matrix entries.

$x(t) - y(t)$ is also a normal random variable, and mean and variance can be obtained from the mean and covariance matrix of $x(t)$ and $y(t)$ obtained from the diffusion model.

Remark 4. For distinguishing purposes, we call the fluid and diffusion models in Section 3 the standard fluid and diffusion models, and the fluid and diffusion models in this section the adjusted fluid and diffusion models.

As mentioned in Section 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 model obtained from Equations (19) to (21), we can eliminate the non-differentiability of the rate functions and obtain a new drift matrix $\mathbf{K}(t)$ and

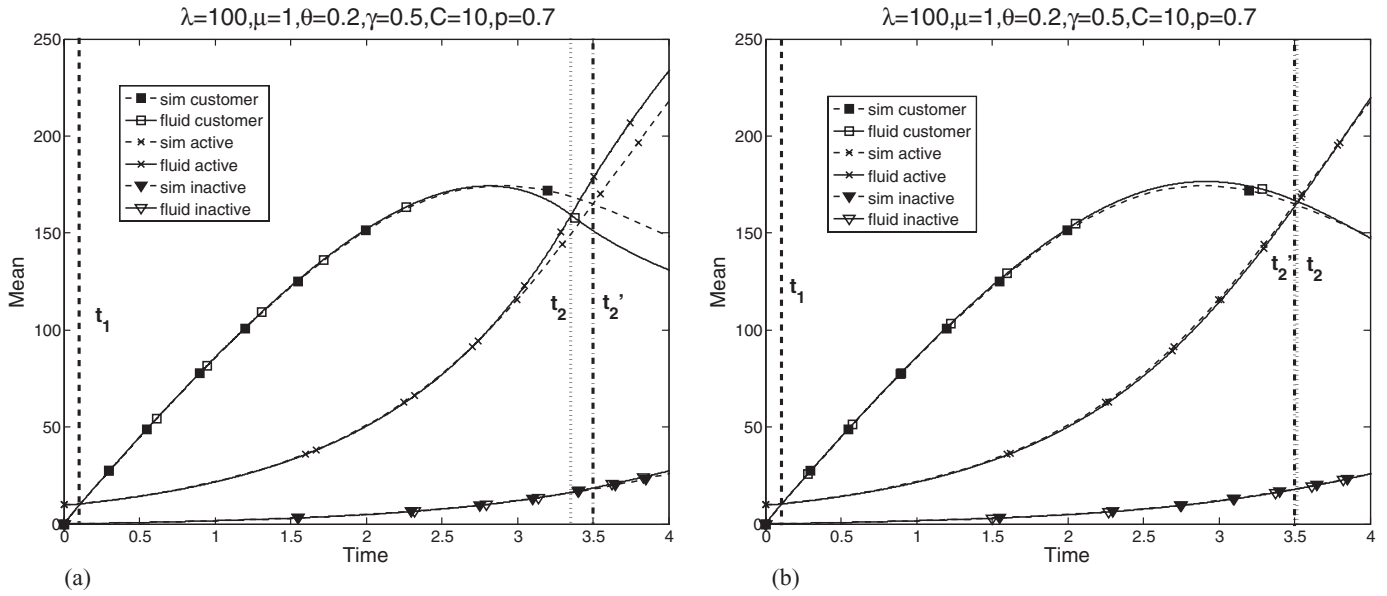

Fig. 6. Comparison of mean numbers of customers and peers for (a) standard and (b) adjusted models in Example 1.

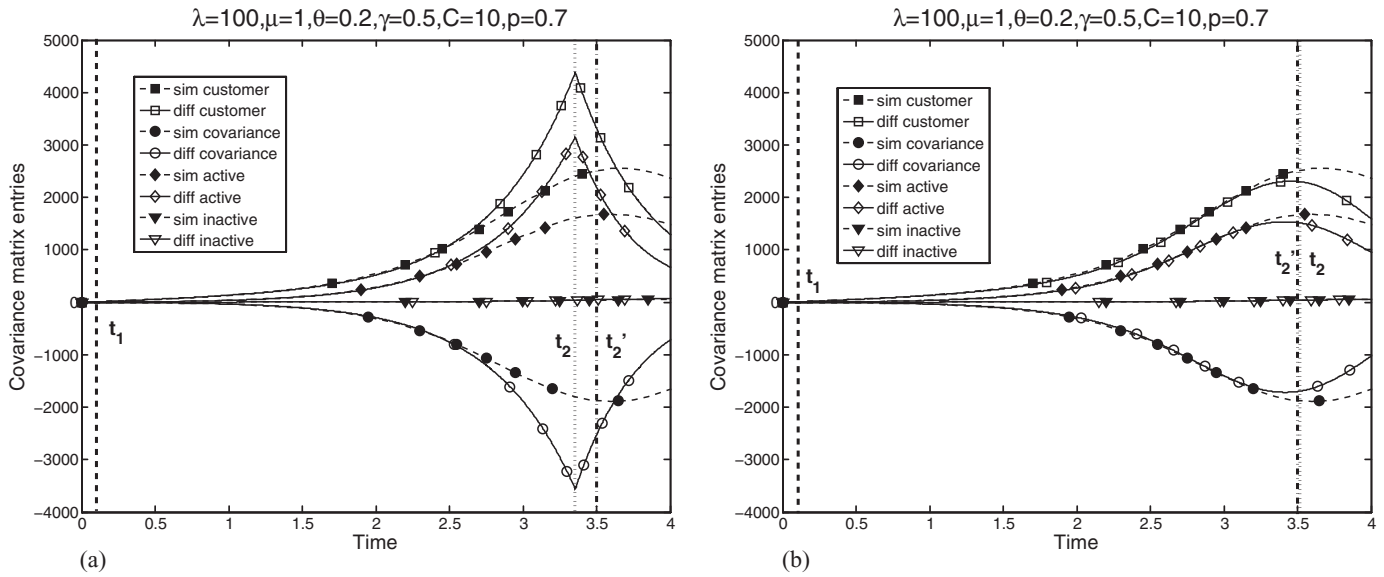

Fig. 7. Comparison of covariance matrices for (a) standard and (b) adjusted models in Example 1.

a diffusion coefficient matrix $\mathbf{L}(t)$ as follows:

$$\mathbf{K}(t)=\begin{pmatrix} -\mu\times q(t) & -\mu\times(1-q(t)) & 0 \\ \mu\times q(t) & \mu\times(1-q(t))-\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix},$$

$$\mathbf{L}(t)=\begin{pmatrix} \sqrt{\lambda} & -\sqrt{\mu\alpha(t)} & 0 & 0 & 0 \\ 0 & \sqrt{\mu\alpha(t)} & -\sqrt{p\theta\bar{y}(t)} & -\sqrt{(1-p)\theta\bar{y}(t)} & \sqrt{\gamma\bar{z}(t)} \\ 0 & 0 & \sqrt{p\theta\bar{y}(t)} & 0 & -\sqrt{\gamma\bar{z}(t)} \end{pmatrix},$$

where $\alpha(t)=q(t)\bar{x}(t)+(1-q(t))\bar{y}(t)-\sigma^2(t)\phi(0,\bar{x}(t)-\bar{y}(t),\sigma(t))$.

From the definition of $q(t)$, it is a Gaussian distribution function and is differentiable with respect to $\bar{x}(t)$ and $\bar{y}(t)$. Hence both $q(t)$ and $\alpha(t)$ are differentiable with respect to $\bar{x}(t)$ and $\bar{y}(t)$, and we get rid of the differentiability issue in $\mathbf{K}(t)$ and $\mathbf{L}(t)$. With the newly obtained $\mathbf{K}(t)$ and $\mathbf{L}(t)$, we have an additional differential equation from Theorem 3.

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{\Sigma}(t)=\mathbf{K}(t)\cdot\mathbf{\Sigma}(t)+\mathbf{\Sigma}(t)\cdot\mathbf{K}^{\mathrm{T}}(t)+\mathbf{L}(t)\cdot\mathbf{L}(t)^{\mathrm{T}}, \tag{22}$$

where $\mathbf{\Sigma}(t)$ is the covariance matrix defined in Theorem 3.

By solving the system of ordinary differential Equations (19) to (22), we can obtain the adjusted fluid and diffusion models.

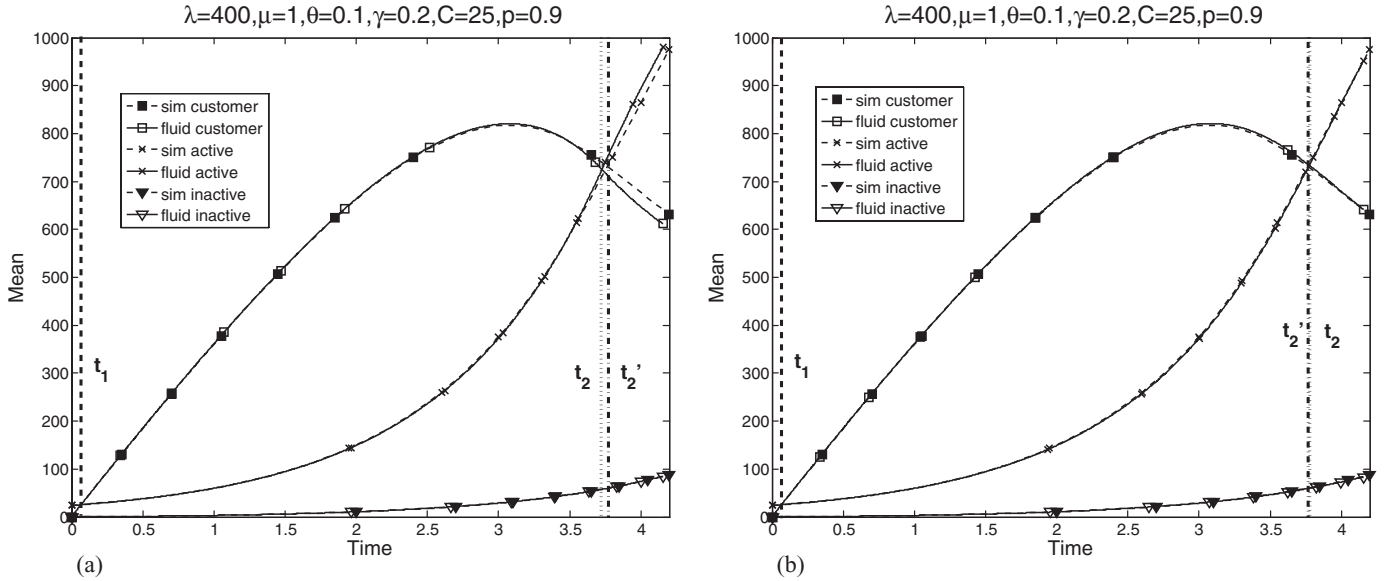

Fig. 8. Comparison of mean numbers of customers and peers for (a) standard and (b) adjusted models in Example 2.

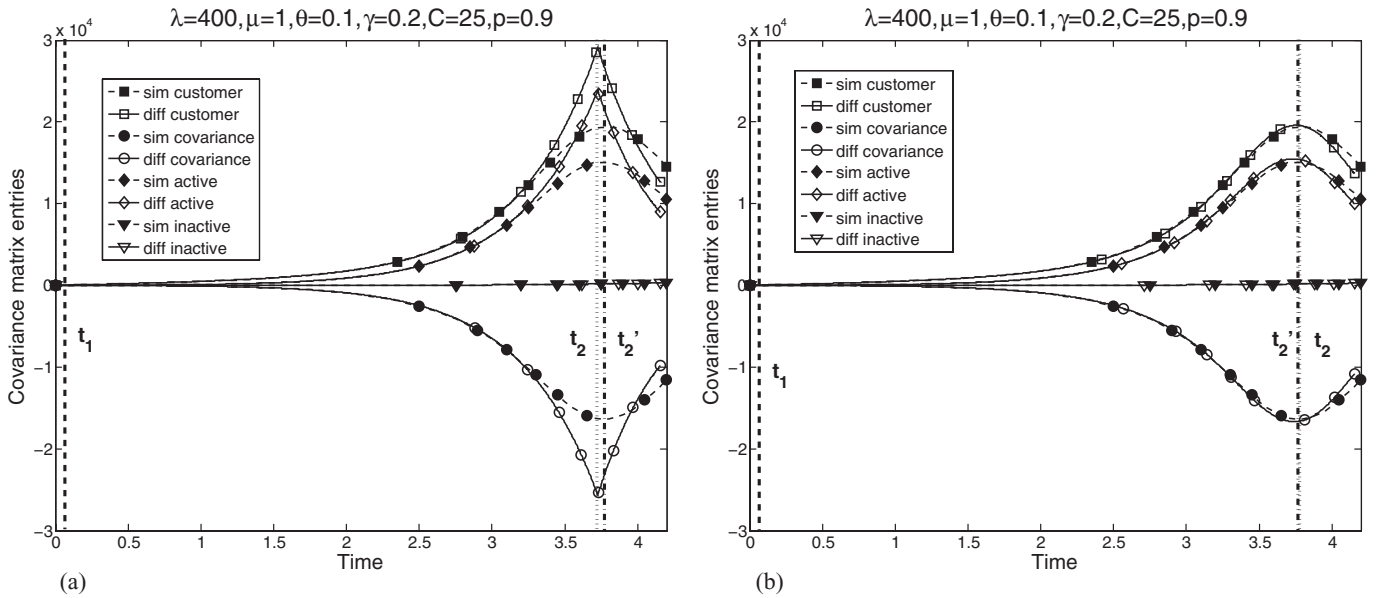

Fig. 9. Comparison of covariance matrices for (a) standard and (b) adjusted models in Example 2.

Figure 5 shows the results from the adjusted fluid and diffusion models with same parameters as in Fig. 4. From Fig. 5, we see that the fluid model is almost the same as the simulation results. For the covariance matrix entries, the sharp spikes disappear and the accuracy is also improved. In fact, the accuracy of the covariance matrix entries is not always significantly improved for all $t > 0$, but they are quite accurate before t_2. The fluid model, however, shows great accuracy regardless of the values of parameters.

Remark 5. We consider the constant rates for arrival, service, peer up and peer down times. However, the fluid and diffusion models can be extended to time-varying rates by substituting λ, μ, θ, and γ with $\lambda(t)$, $\mu(t)$, $\theta(t)$, and $\gamma(t)$ since Theorems 1 to 3 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.

5. Numerical results

In this section, we provide numerical examples to verify our results obtained in Sections 3 and 4. We report

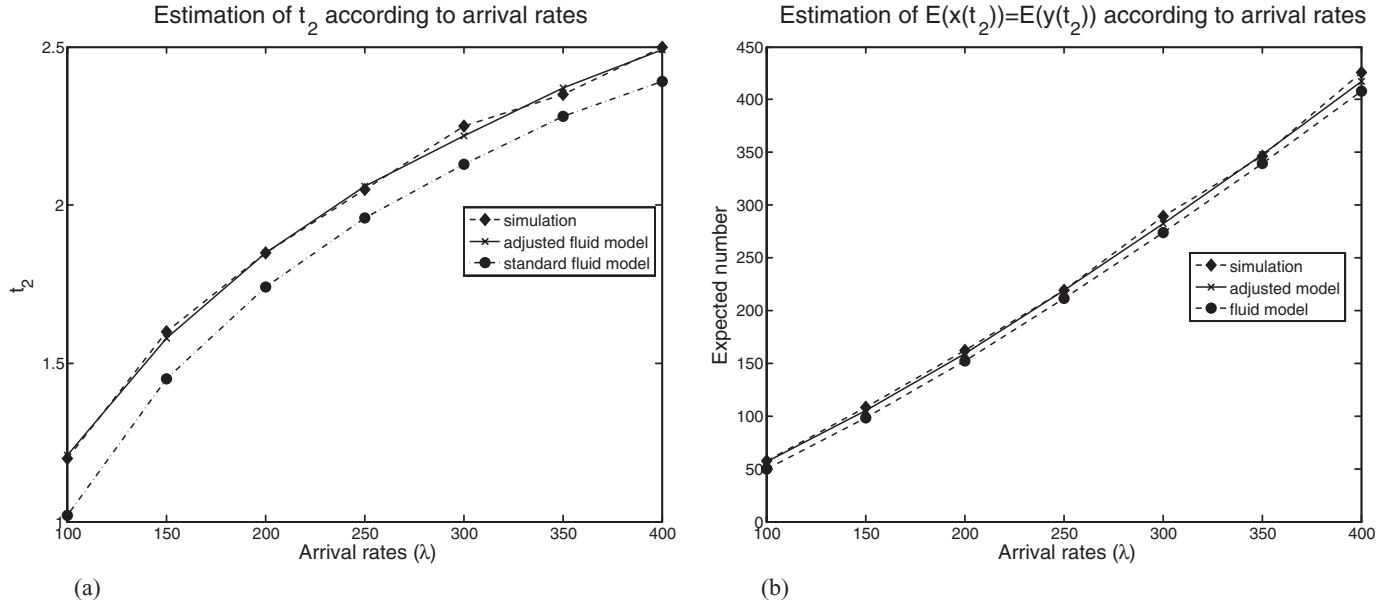

Fig. 10. Estimation of (a) t_2 and (b) $E(x(t_2))$ as a function of λ.

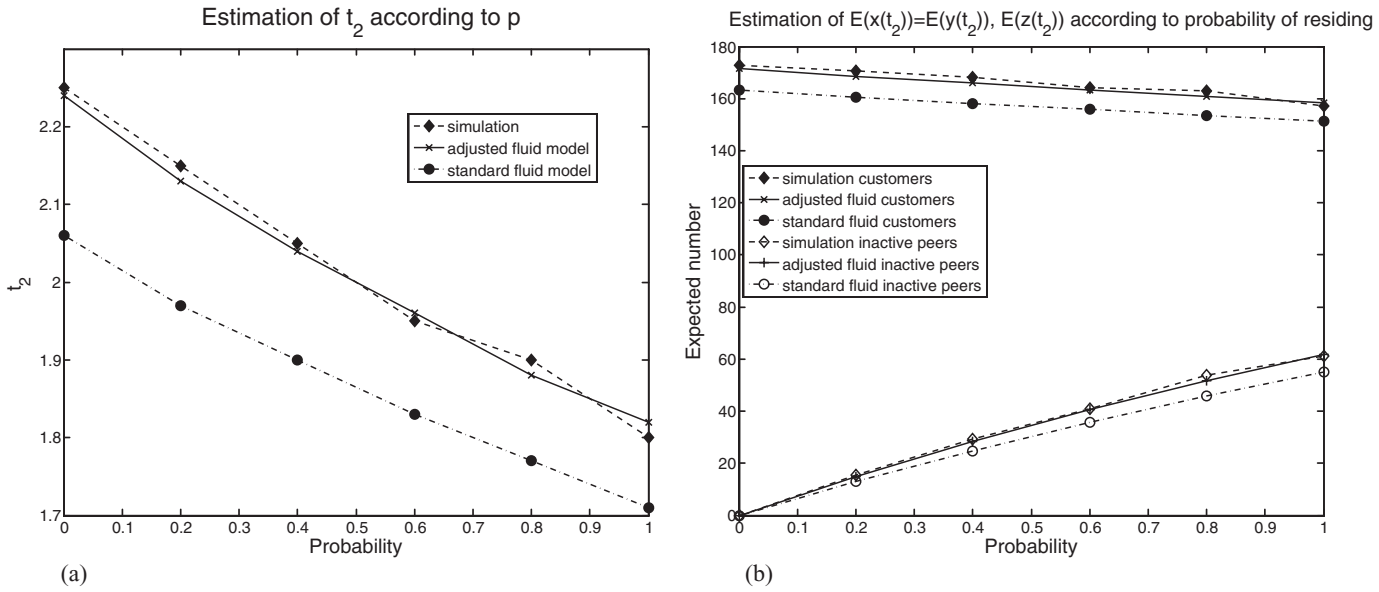

Fig. 11. Estimation of (a) t_2 and (b) $E[\mathbf{X}(t_2)]$ and $E[z(t_2)]$ as a function of p.

numerical experiments to compare the adjusted fluid and diffusion models (described in Section 4) with the standard fluid and diffusion models (described in Section 3) in Section 5.1. In addition to this, we provide some numerical experiments when the rate functions vary over time in Section 5.2.

5.1. *Comparison between the standard and adjusted models*

Table 1 summarizes the key characteristics of the standard and adjusted models so as to highlight the differences between the two models. We provide two examples to demonstrate that the adjusted model outperforms the standard model on $[0, t_2]$. The parameters we use in the examples are summarized in Table 2. 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 needed to cover the traffic, and thus outsourcing traffic does not make sense. We assume a large peer network (more than 100 peers).

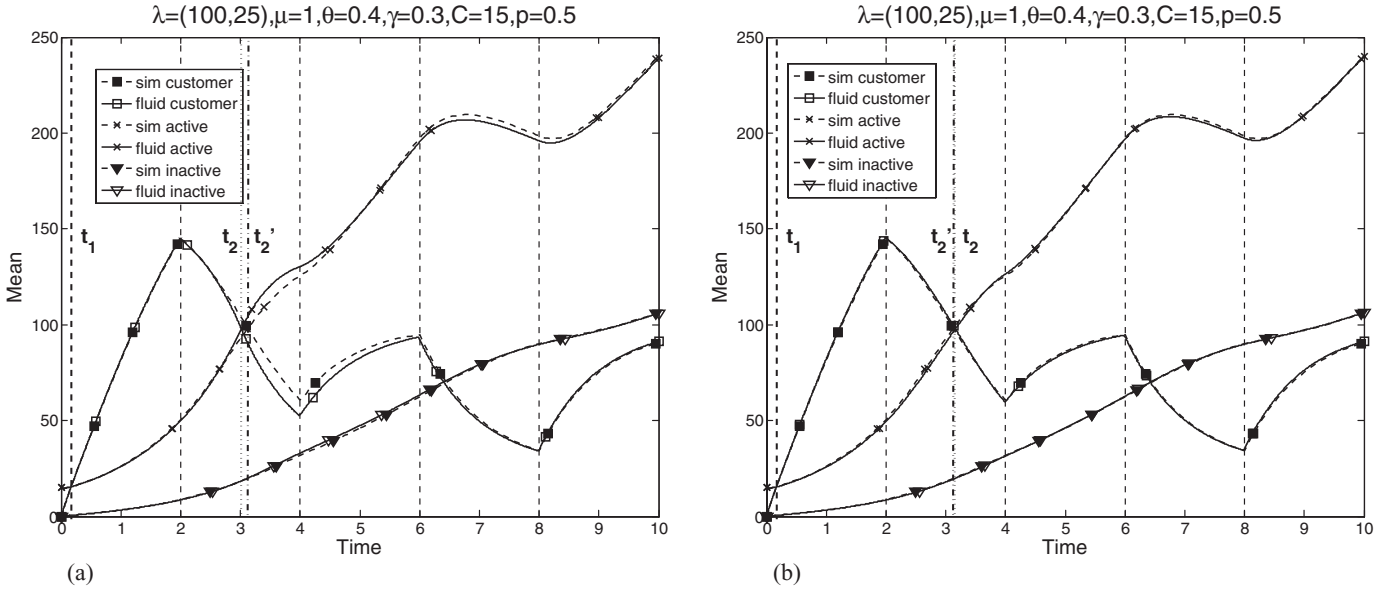

Fig. 12. Mean number of customers and peers with time-varying arrival rate: (a) standard model and (b) adjusted model.

Table 1. Comparison between standard and adjusted models

	Standard model	*Adjusted model*
Rate functions	$f_i(\cdot,\cdot)$	$E[f_i(\cdot,\cdot)]$
Fluid model	Obtained independently	Obtained simultaneously
Diffusion model	Obtained using fluid model	Obtained simultaneously
Assumption	Measure zero at non-differentiable points	Gaussian density
Limitation	Inaccuracy in both fluid and diffusion models around t_2	Inaccuracy in diffusion model after t_2

Table 2. Parameters used in the two examples

Examples	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

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

Parameter values were selected arbitrarily based on the above conditions. We conducted 5000 simulation runs for each example and compared the simulation results with the results of the standard and adjusted models to see the accuracy of each model. Figures 6 and 7 illustrate the comparison of mean numbers and covariance matrix entries with the setting of Example 1. Figures 8 and 9 show the results for Example 2. In both examples, the standard models show inaccuracy in estimating both expected values and covariance matrix entries. As mentioned in Section 3, we see that the standard models always underestimate t_2. For covariance matrix entries, the standard models show more than 100% errors at $t = t_2$ in both examples. In contrast, the adjusted models estimate t_2 reasonably well, especially as the arrival rate becomes higher, which is desirable for real applications. Although the adjusted models 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 models are more suitable for a transient analysis than the standard models. 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 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.

1. The arrival rate implies the demand for the content. When operating a peer network, preparing for a burst in 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.
2. The probability of residing in the system determines the current and potential service capacity. If $p = 0$, there are no peers in the inactive peer pool. In this case, service capacity depends solely 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.

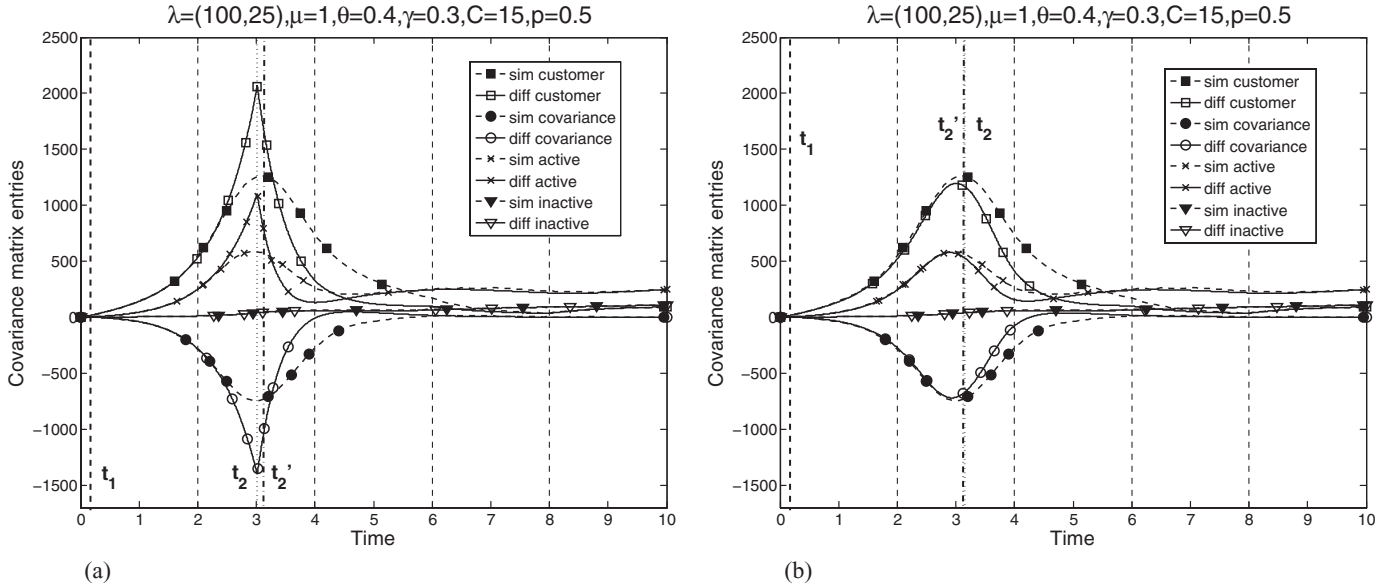

Fig. 13. Covariance matrix entries with time-varying arrival rate: (a) standard model and (b) adjusted model.

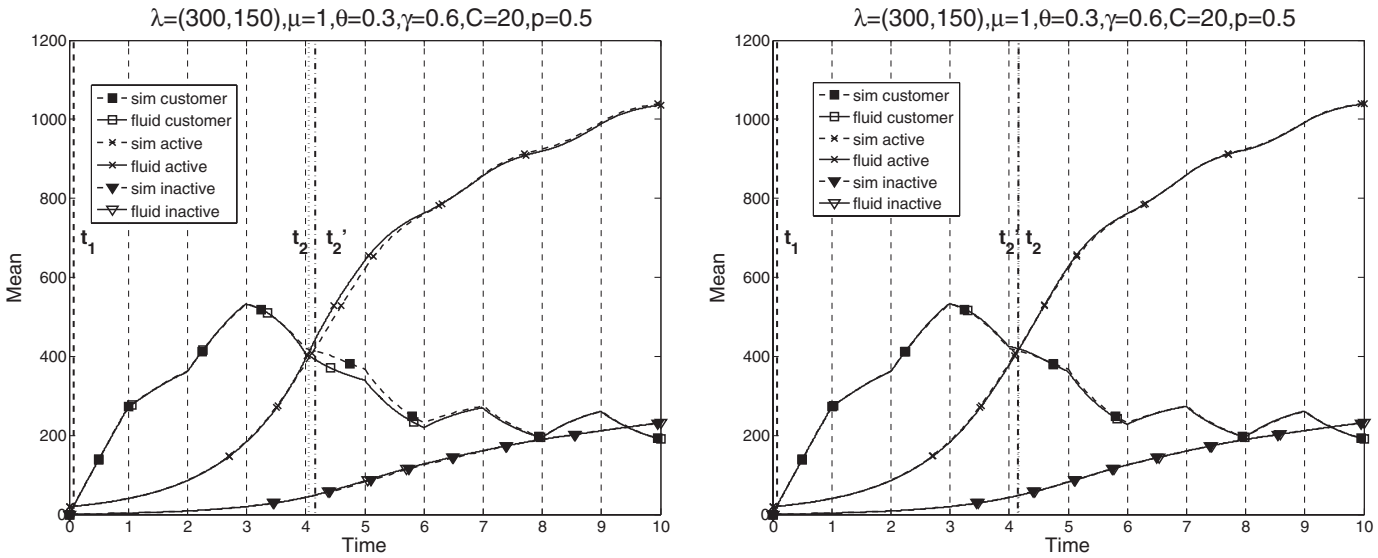

Fig. 14. Mean number of customers and peers with time-varying arrival rate: (a) standard model and (b) adjusted model.

Figures 10 and 11 show the changes of t_2 and $E[\mathbf{X}(t_2)]$ over λ and p, respectively. As seen in Fig. 10, t_2 and $E[\mathbf{X}(t_2)]$ 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 $E[x(t_2)](= E[y(t_2)])$ decrease according to p, whereas $E[z(t_2)]$ increases. This implies that increasing the potential service capacity (i.e., number of inactive peers) accelerates the rate of increase in the number of peers which enables our system to reach stage 3 earlier. In addition to these observations, we see that the adjusted fluid model provides more accurate t_2 and $E[\mathbf{X}(t_2)]$ than the standard fluid model.

5.2. *Time-varying rate functions*

In Remark 5, we mentioned that fluid and diffusion approximations can be extended to time-varying rate functions; i.e., the arrival rate is $\lambda(t)$, the service rate is $\mu(t)$, and the peer up and peer down times are $1/\theta(t)$ and $1/\gamma(t)$ on average, respectively. In this section, we show two numerical examples in which the arrival rate changes over time (μ, θ, and γ are held constant over time only for illustration purposes).

Figures 12 and 13 show the mean and covariance matrix entries of the number of customers and peers with the arrival rate alternating between 100 and 25 every

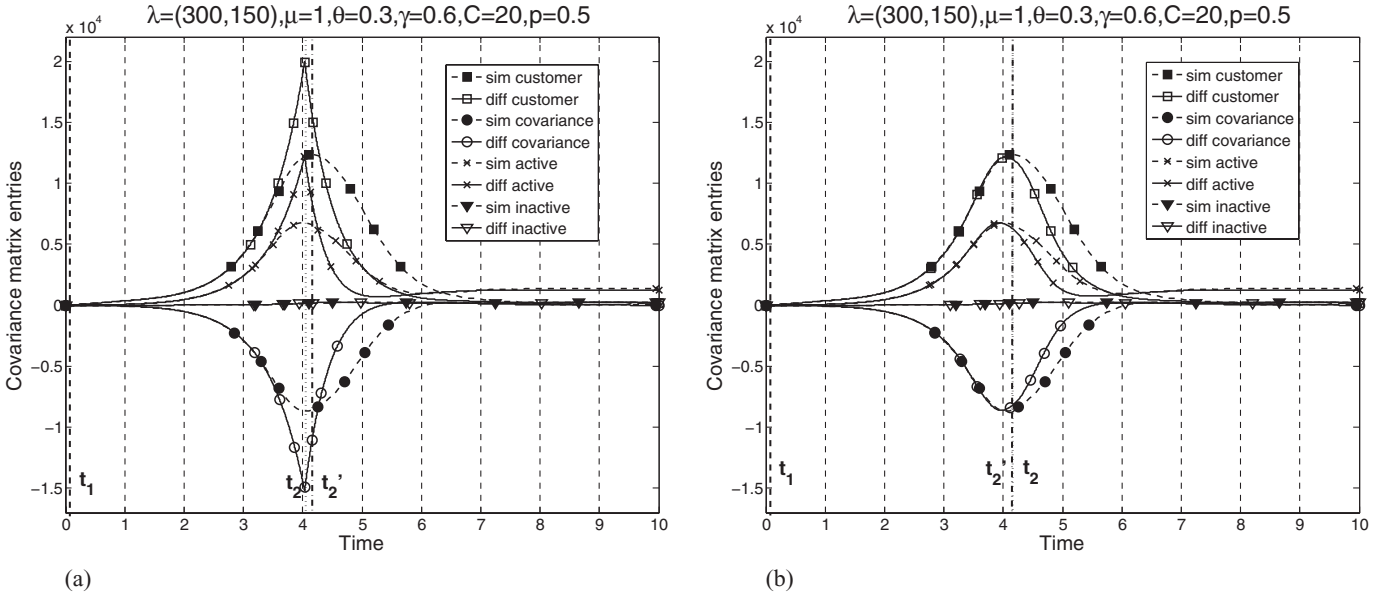

Fig. 15. Covariance matrix entries with time-varying arrival rate: (a) standard model and (b) adjusted model.

two time units. We apply both the standard and adjusted models and compare them with simulation results. As seen in Fig. 12, the adjusted model gives quite accurate results in all ranges of time intervals, whereas the standard model shows some error around $t \in [2.5, 5]$ and gives accurate results after $t > 6$. For the standard fluid model, note that $\min(\bar{x}(t), \bar{y}(t))$ changes the value from $\bar{y}(t)$ to $\bar{x}(t)$ near $t = 3$ and after that it remains in $\bar{x}(t)$. Therefore, we can explain this phenomenon using Theorem 4 and Lemma 1, similar to the case of constant rate functions. For the covariance matrix entries, both the standard and adjusted models show shapes similar to the case of constant rates functions. Although the adjusted diffusion model also shows errors, we can see that the accuracy is significantly improved compared with the standard model, especially before t_2 (recall the definition of t_2 in Section 3). In this example, we use a 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 in each time unit. As shown in Figs 14 and 15, we observe results similar to the first example. The standard fluid model shows inaccuracy around $t \in [3.7, 6]$ whereas the adjusted fluid model provides an excellent estimation. The adjusted diffusion model is almost exact for $t < t_2$ but shows inaccuracy after t_2 just like the first example. From the examples, we can state that our adjusted fluid and diffusion models can be used successfully in the time interval we are interested in, i.e., $0 \leq t \leq t_2$.

6. Conclusions

In this paper, 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 Fig. 3) 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. We, however, observe that standard fluid and diffusion approximations are highly inaccurate around t_2 which is the result of 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 that show that the adjusted models outperform the standard models in terms of accuracy, especially before t_2 as desired. Moreover, we provide several numerical examples that show 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 a 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 analyses. For example, we see that a higher sojourn probability decreases t_2 values, whereas the expected number of customers does not decrease much. For time-varying rate functions, we consider discrete arrival rate functions. From the examples provided, the increase (or decrease) in the rate of the number of customers is immediately affected by changes in the 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.

There are several extensions to this paper that can be considered in the future. First, we assume that the $\mathbf{X}(t)$ process is Gaussian. This assumption, however, is broken around the switching time between stages 2 and 3 (i.e., around time t_2) in simulation and it might cause inaccuracy of covariance matrix entries during the early part of stage 3. To overcome this, studies on how to obtain an asymptotic distribution of $\mathbf{X}(t)$ are required. Empirically, we observe that the distribution of $\mathbf{X}(t)$ shows extreme value type distribution near the switching time. Second, we assume that all the times follow exponential distributions. In some situations, this assumption is not realistic. Therefore, the relaxation of this assumption could be considered in future model formulations.

Acknowledgements

The authors thank the reviewers, associate editor and department editor for their comments and suggestions that led to considerable improvements in the content and presentation of this paper. This research was partially supported by NSF grant CMMI-0946935.

References

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.

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

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.

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.

Ethier, S.N. and Kurtz, T.G. (1986) *Markov Processes: Characterization and Convergence*, 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.

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.

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.

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**(2), 231–259.

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

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

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

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

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

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

Whitt, W. (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, **50**(10), 1449–1461.

Whitt, W. (2006) 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*, IEEE, Piscataway, NJ, **4**, pp. 2242–2252.

Biographies

Young Myoung Ko is a Ph.D. candidate in Industrial Engineering at Texas A & M University. He received B.S. and M.S. degrees in Industrial Engineering from Seoul National University, Seoul, Korea. His research focuses on the analysis of complex stochastic systems, and covers the domains of online service operations, communication networks and energy-aware system design.

Natarajan Gautam is an Associate Professor in the Department of Industrial Systems Engineering at Texas A&M University with a courtesy appointment in the Department of Electrical and Computer Engineering. He received his M.S. and Ph.D. degrees in Operations Research from the University of North Carolina at Chapel Hill. His research interests are in the areas of modeling, analysis and performance evaluation of stochastic systems with special emphasis on computer, telecommunication, and information systems. He is an Associate Editor for the *INFORMS Journal on Computing* and *Omega*.