Transient Analysis of Queues for Peer-based Multimedia Content Delivery

Young Myoung Ko and Natarajan Gautam

March 24, 2010

Abstract

Consider a firm that sells on-line multimedia content. In order to manage costs and quality of service, this firm maintains a peer-network whereby users download files from other peers that have downloaded them in the past. The scenario can be modeled as a queueing system where the number of servers varies over time. We develop analytical models based on fluid and diffusion approximations to analyze transient system performance. Using the same approximations, we also analyze the steady-state behavior of such a network. We, however, find that existing fluid and diffusion approximations are inaccurate for transient analysis. To address this shortcoming, we provide a novel Gaussian-based adjustment and significantly improve accuracy in approximations. Further, models used in this research can be extended seamlessly to the case where system parameters (e.g. arrival rates and service rates) are time-varying. We provide several numerical examples to show how our adjusted models work for transient analysis.

1 Introduction

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 paper, 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 [17], Ge et al. [9], and Yang and de Veciana [21]. 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. [8], Gummadi et al. [10]). 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. eDonkey, Bittorent, etc), 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 (http://www.pando.com) are operating P2P networks for content-distribution. Furthermore, even companies such as Akamai 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 paper.

For peer networks, most research focuses on modeling and performance analysis of steady state behavior (Ge et al. [9], Clévenot and Nain [6], Qiu and Srikant [17]), or optimal peer search and selection (Adler et al. [1]) 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 [12] and well summarized in Ethier and Kurtz [7]. Mandelbaum et al. [13] applies strong approximations to time-varying-rate cases and establish the framework to analyze Markovian queueing networks (also see Mandelbaum et al. [14], Massey [16]). In addition, they extend the results in Kurtz [12] to apply the strong approximations to non-differentiable rate functions of the system state by defining a new derivative called "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 Whitt [18] and Billingsley [5]. Recently, these methods have been used for transient analysis and control of online rental systems such as Netflix (Bassamboo et al. [3], Bassamboo and Randhawa [4]). 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 [19], Whitt [20]). Specifically, Hampshire et al. [11] utilizes 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 [17]) to analyze steady state behavior. They show weak convergence to the Ornstein-Uhlenbeck (OU) process in steady state. As mentioned before, their research, however, 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 [12]. Our model, however, has non-differentiable rate functions of the system state. Although the results of Mandelbaum et al. [13] 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. [13] is of the form

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

On the right hand side of equation (1), the term $Cov[Q_1^{(1)}(t), Q_1^{(1)}(t)^-]$ (or $Cov[Q_1^{(1)}(t), Q_1^{(1)}(t)^+]$) makes it impossible to solve the differential equation unless we know the functional relationship between $Cov[Q_1^{(1)}(t), Q_1^{(1)}(t)]$ and $Cov[Q_1^{(1)}(t), Q_1^{(1)}(t)^-]$ (or $Cov[Q_1^{(1)}(t), Q_1^{(1)}(t)^+]$). So, in this paper, we provide a new way to 1) cope with the inaccuracy of existing approximations and 2) achieve computational feasibility.

The rest of the paper 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 [12] and Mandelbaum et al. [14]. We call these "standard" fluid and diffusion models in the rest of the paper to distinguish them from our adjusted models. It turns out that both fluid and diffusion approximations work well in steady state. We, however, show significant inaccuracy in both fluid and diffusion approximations during a transient period. In Section 4, we explain our new adjustment and show the improved approximations. In order to validate our adjusted model and to see the effects of parameters, several numerical examples

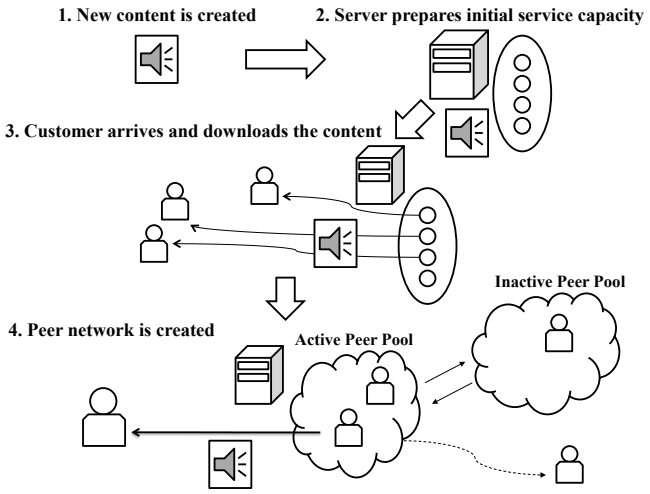


Figure 1: System illustration

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 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 manages one queue for waiting customers and allocates a new customer in 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 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

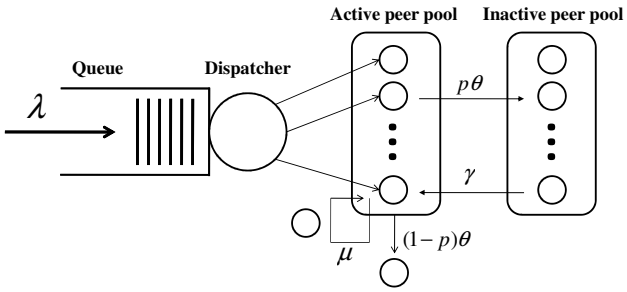


Figure 2: Simplified system model

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 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 paper, 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 paper. 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 paper, 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.

2.2 Mathematical model

Let $X(t) = (x(t), y(t), z(t))^{\mathrm{T}}$ 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 one queue, which is managed by the company. Therefore, this process can be characterized as a $M/M/y(t)$ 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 [17] and Yang and de Veciana [21] by comparing real trace data from a BitTorrent network.

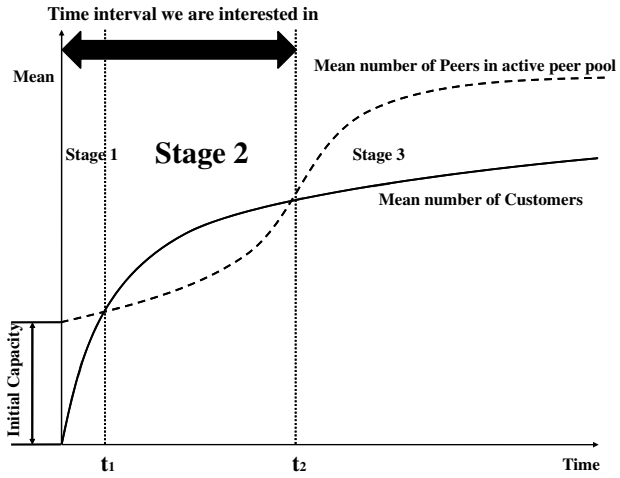


Figure 3: Typical evolution of peer networks on average

2.3 Objective

Figure 3 illustrates a typical evolution of peer networks. From Figure 3, 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 (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) accurately 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 [12] and Mandelbaum et al. [14] for our problem. After developing the results, we will show the inadequacy of these approximations. Fluid and diffusion approximations are used by several previous studies (Mandelbaum et al. [13], Mandelbaum et al. [14], Whitt [19], Whitt [20], Qiu and Srikant [17]). The first step of this approach is to define a sequence of stochastic processes and to obtain the fluid model by taking limit of the sequence. Fluid model takes the role of the expected value for each time point. The second step is to obtain a diffusion model 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 $X(t) = (x(t), y(t), z(t))^{\mathrm{T}}$ 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. $X(0) = (0, C, 0)^{\mathrm{T}}$. Then, for our model, the sample path can be constructed using the following integral equation:

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

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's up, peer's leaving, and peer's down respectively. To apply fluid and diffusion approximations to equation (2), consider a sequence of stochastic processes $\{X_n(t)\}_{t\geq 0}$ so that $X_n(t)$ is the solution to the following integral equation:

$$
\begin{aligned}
X_n(t) &= \begin{pmatrix} x_n(t) \\ y_n(t) \\ z_n(t) \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ C \\ 0 \end{pmatrix} + \frac{1}{n}\Big\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} A_1\Big(n\int_0^t \lambda ds\Big) + \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} A_2\Big(n\int_0^t \mu\min(x_n(s), y_n(s))ds\Big) \\
&\quad + \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} A_3\Big(n\int_0^t p\theta y_n(s)ds\Big) + \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} A_4\Big(n\int_0^t (1-p)\theta y_n(s)ds\Big) \\
&\quad + \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} A_5\Big(n\int_0^t \gamma z_n(s)ds\Big)\Big\}. \qquad (3)
\end{aligned}
$$

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

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

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

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

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

$$
X_n(t) = \begin{pmatrix} 0 \\ C \\ 0 \end{pmatrix} + \sum_{i=1}^{5} \frac{1}{n} l_i A_i\Big(n\int_0^t f_i\big(X_n(s)\big)ds\Big),
$$

where $l_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $l_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$, $l_3 = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$, $l_4 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}$, and $l_5 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$.
Then, it is easy to verify that the $f_i(\cdot)$'s are Lipschitz and there exist ϵ_i's such that $|f_i(X)| \le \epsilon_i(1+|X|)$. Since $\sum |l_i|^2\epsilon_i < \infty$, by Theorem 2.1 and 2.2 in Kurtz [12], $\lim_{n\to\infty} X_n = \bar{X}(t)$ 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 its graph is identical to Figure 3. 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 fluid approximation results;

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

Notice t_1 and t_2, depicted in Figure 3, for further clarification. The switching times t_1 and t_2 can be obtained directly by solving the integral 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 $D_n(t)$ be the scaled centered process; i.e. $D_n(t) = \sqrt{n}\big(X_n(t) - \bar{X}(t)\big)$ and assume measures zero at t_1 and t_2. Then, we can define the diffusion approximation model as*

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

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

$$\begin{aligned} K_1 &= \begin{pmatrix} -\mu & 0 & 0 \\ \mu & -\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix}, \\ K_2 &= \begin{pmatrix} 0 & -\mu & 0 \\ 0 & \mu-\theta & \gamma \\ 0 & p\theta & -\gamma \end{pmatrix}, \text{ and} \\ L(t) &= \begin{pmatrix} \sqrt{\lambda} & -\sqrt{\mu \min(\bar{x}(t), \bar{y}(t))} & 0 & 0 & 0 \\ 0 & \sqrt{\mu \min(\bar{x}(t), \bar{y}(t))} & -\sqrt{p\theta\bar{y}(t)} & -\sqrt{(1-p)\theta\bar{y}(t)} & \sqrt{\gamma\bar{z}(t)} \\ 0 & 0 & \sqrt{p\theta\bar{y}(t)} & 0 & \sqrt{\gamma\bar{z}(t)} \end{pmatrix}. \end{aligned}$$

Then, $D(t)$ is the solution of the following integral equation:
for $0 \leq t < t_1$,

$$D(t) = \int_0^t L(s) \cdot \begin{pmatrix} dB_1(s) \\ dB_2(s) \\ dB_3(s) \\ dB_4(s) \\ dB_5(s) \end{pmatrix} + \int_0^t K_1 \cdot D(s) ds, \tag{5}$$

for $t_1 \leq t < t_2$,

$$D(t) = \begin{pmatrix} d_1(t_1) \\ d_2(t_1) \\ d_3(t_1) \end{pmatrix} + \int_{t_1}^t L(s) \cdot \begin{pmatrix} dB_1(s) \\ dB_2(s) \\ dB_3(s) \\ dB_4(s) \\ dB_5(s) \end{pmatrix} + \int_{t_1}^t K_2 \cdot D(s) ds, \tag{6}$$

and for $t \geq t_2$,

$$D(t) = \begin{pmatrix} d_1(t_2) \\ d_2(t_2) \\ d_3(t_2) \end{pmatrix} + \int_{t_2}^t L(s) \cdot \begin{pmatrix} dB_1(s) \\ dB_2(s) \\ dB_3(s) \\ dB_4(s) \\ dB_5(s) \end{pmatrix} + \int_0^t K_1 \cdot D(s) ds, \tag{7}$$

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 X, l_i's and $f_i(\cdot)$'s as in the proof of Theorem 1, define $F(X)$ as follows:

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

Then, by Kurtz [12], 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 $0 \leq t < t_1$, (5) is straightforward. However, according to Kurtz [12], the drift matrix of (5) and (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. [14] considers. Then, we can obtain (6) for $t_1 \leq t < t_2$ and (7) for $t \geq t_2$. □

Note that the diffusion model in (5), (6), and (7) turns out to be a Gaussian process and is closely related to the fluid model ($\bar{X}(t)$). 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 $D(t)$ in the following way.

Theorem 3 (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 2, $m(t)$ is the solution to the following differential equation: for $0 \le t < t_1$ or $t \ge t_2$,*

$$\frac{d}{dt}m(t) = K_1 \cdot m(t), \tag{8}$$

and for $t_1 \le t < t_2$,

$$\frac{d}{dt}m(t) = K_2 \cdot m(t). \tag{9}$$

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

$$\frac{d}{dt}\Sigma(t) = K_1 \cdot \Sigma(t) + \Sigma(t) \cdot K_1^T + L(t) \cdot L(t)^T, \tag{10}$$

and for $t_1 \le t < t_2$,

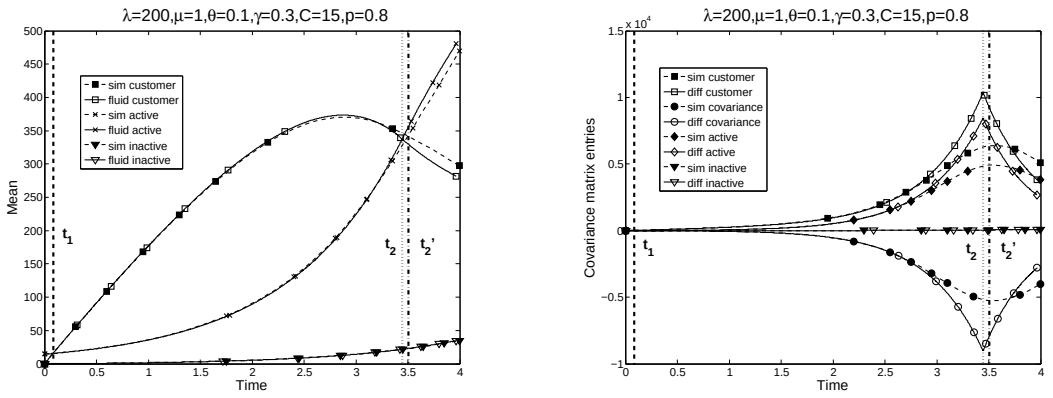
$$\frac{d}{dt}\Sigma(t) = K_2 \cdot \Sigma(t) + \Sigma(t) \cdot K_2^T + L(t) \cdot L(t)^T. \tag{11}$$

Proof. For $0 \le t < t_1$, we know that $E\big(D(0)\big) = 0 < \infty$ since $D(0) = 0$. Then, by Theorem 8.2.6 in Arnold [2], $m(t)$ and $D(t)$ satisfy (8) and (10). From (8), we also have $E\big(D(t_1)\big) < \infty$. Therefore, we can also apply Theorem 8.2.6 in Arnold [2] and obtain (9) and (11). Since $E\big(D(t_2)\big) < \infty$, we obtain (8) and (10) for $t \ge t_2$. □

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 $D(t)$, we use $\bar{X}(t) + D(t)$ as an approximation of $X(t)$ (i.e. $X(t) \approx \bar{X}(t) + D(t)$). By Theorem 3, we obtain $E[D(t)] = m(t) = 0$ for all $t \ge 0$ since $m(0) = E[D(0)] = E[\lim_{n\to\infty} \sqrt{n}(X_n(0) - \bar{X}(0))] = E[\lim_{n\to\infty} \sqrt{n}(x_0 - x_0)] = 0$. Therefore,

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

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 ($C = 15$).



(a) Mean number of customers and peers

(b) Covariance matrix entries

Figure 4: Standard fluid and diffusion approximations

Note that Figure 4 (a) is for $\bar{X}(t)$ and Figure 4 (b) for $\Sigma(t)$. The simulation result is obtained by averaging 5,000 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 a big difference, especially in covariance matrix entries around t_2. We find two significant problems in the fluid and diffusion models from Figure 4. Let t_2' denote the switching time between stages 2 and 3 in the simulation result. Then,

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 significant difference from the simulation result around t_2. These spikes come from the sudden change of the drift matrix from K_2 and K_1 at time t_2 in Theorem 2 and this switching is caused by the non-differentiability of the $\min\left(\bar{x}(t), \bar{y}(t)\right)$ 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 Theorem 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 [17] 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 $D(\infty)$ be the scaled centered process $D(t)$ defined in Theorem 2 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 (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)-(14), 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 2. *Notice that the steady state number of customers, active peers, and inactive peers via equations (12)-(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 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:

$$\begin{aligned} \frac{d}{dt}\bar{x}(t) &= \lambda - \mu \min(\bar{x}(t), \bar{y}(t)), && (15) \\ \frac{d}{dt}\bar{y}(t) &= \mu \min(\bar{x}(t), \bar{y}(t)) - \theta\bar{y}(t) + \gamma\bar{z}(t), \text{ and} && (16) \\ \frac{d}{dt}\bar{z}(t) &= p\theta\bar{y}(t) - \gamma\bar{z}(t). \end{aligned}$$

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

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

$$X(t) = x_0 + \sum_l l A_l\bigg(\int_0^t f_l\big(X(s)\big)ds\bigg), \tag{17}$$

where $l \in Z^d$, $x_0 = X(0)$ which is constant, as described in Section 3, A_l's are independent Poisson processes, and f_l's are non-negative and satisfy the conditions defined in Kurtz [12]. Then, $E[X(t)]$ is the solution to the following equation:

$$E\big[X(t)\big] = x_0 + \sum_l l \int_0^t E\Big[f_l\big(X(s)\big)\Big]ds. \tag{18}$$

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

$$
\begin{aligned}
E\big[X(t)\big] &= E\bigg[x_0 + \sum_l l A_l\Big(\int_0^t f_l\big(X(s)\big)ds\Big)\bigg] \\
&= x_0 + \sum_l lE\bigg[A_l\Big(\int_0^t f_l\big(X(s)\big)ds\Big)\bigg] \\
&= x_0 + \sum_l lE\bigg[\int_0^t f_l\big(X(s)\big)ds\bigg] \text{ due to Poisson process's expected value} \\
&= x_0 + \sum_l l\int_0^t E\Big[f_l\big(X(s)\big)\Big]ds \text{ by the conditions in Kurtz [12] and Fubini theorem.}
\end{aligned}
$$

□

Corollary 1. *If $f_l(X)$'s are constant or a linear combination of the components of X, then,*

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

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

Proof. Using the linearity of expectation, the integral equation (18) is the same as that for the fluid model. □

Remark 3. *In many situations, f_l's are constant or linear combinations of components of 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 (18), we encounter a fundamental problem. To obtain $E\big[\min(x(t), y(t))\big]$, 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 $X(t)$ is approximated by the Gaussian process, and mean and variance were obtained from $\bar{X}(t)$ and $D(t)$ respectively. In addition, from previous research studies such as Mandelbaum and Pats [15] and Mandelbaum et al. [14], we notice that empirical densities of original processes are well matched with Gaussian density in several applications, even if rate functions are non–differentiable. Therefore, it could be reasonable to use a Gaussian density function to calculate $E\big[\min(x(t), y(t))\big]$. Then, we can rewrite (18) as a differential equation form to fit our model as follows:

$$\frac{d\bar{x}(t)}{dt} = \lambda - \mu\big\{q(t)\bar{x}(t) + (1-q(t))\bar{y}(t) - \sigma^2(t)\phi\big(0, \bar{x}(t) - \bar{y}(t), \sigma(t)\big)\big\} \quad (19)$$

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

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

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

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

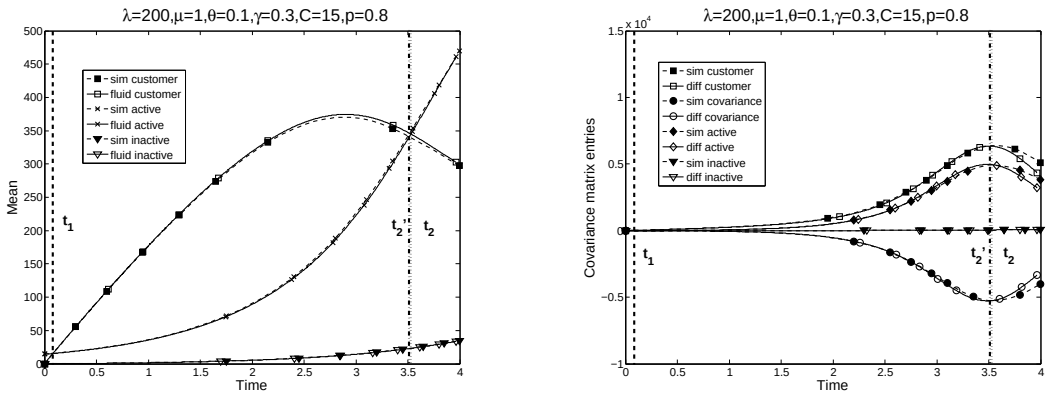
If we use the adjusted fluid model obtained from equations (19)-(21), we can eliminate the non-differentiability of rate functions and obtain a new drift matrix $K(t)$ and a diffusion coefficient matrix $L(t)$ as follows:

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

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

where $\alpha(t) = q(t)\bar{x}(t) + (1-q(t))\bar{y}(t) - \sigma^2\phi\big(0, \bar{x}(t) - \bar{y}(t), \sigma\big)$.
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 $K(t)$ and $L(t)$. With the newly obtained $K(t)$ and $L(t)$, we have an additional differential equation from Theorem 3.

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



(a) Mean number of customers and peers

(b) Covariance matrix entries

Figure 5: Adjusted fluid and diffusion approximations with adjustment

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

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

Figure 5 shows the results from the adjusted fluid and diffusion models with same parameters in Figure 4. From Figure 5, we see that the fluid model 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 model, however, shows great accuracy regardless of the values of parameters.

Remark 5. *We consider the constant rates for arrival, service, peer's up and down times. However, the fluid and diffusion models can extend to time-varying rates by substituting λ, μ, θ, and γ with $\lambda(t)$, $\mu(t)$, $\theta(t)$, and $\gamma(t)$ since Theorem 1-3 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 through Sections 3 and 4. We show more 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.

Table 1: Comparison between standard and adjusted models

	Standard model	Adjusted model
Rate functions	$f_i(\cdot,\cdot)$'s	$E[f_i(\cdot,\cdot)]$'s
Fluid model	obtained independently	obtained simultaneously
Diffusion model	obtained using fluid model	
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

5.1 Comparison between the standard and adjusted models

Table 1 summarizes key characteristics of the standard and adjusted models. Before looking at the numerical examples, one can see the difference between two models. We provide two examples to demonstrate that the adjusted models outperform the standard models 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 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.

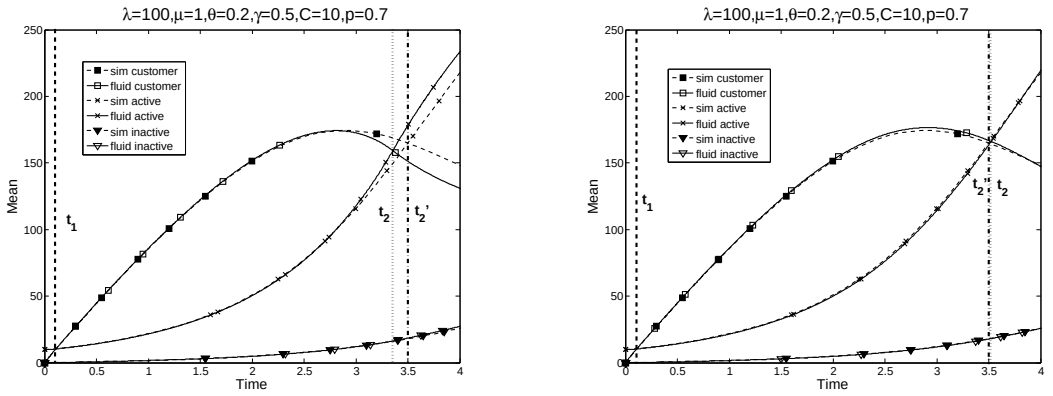
However, parameter values are selected arbitrarily based on the above conditions. We conducted 5,000 simulation runs for each example and compared the simulation results with the results of the standard and adjusted models to see how accurate each model is. Figures 6 and 7 illustrate the comparison of 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 reasonably well estimate t_2, especially as the arrival rate becomes higher, which is desirable for the 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

Table 2: 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

examples, we can verify that the adjusted models are more suitable for the 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



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

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

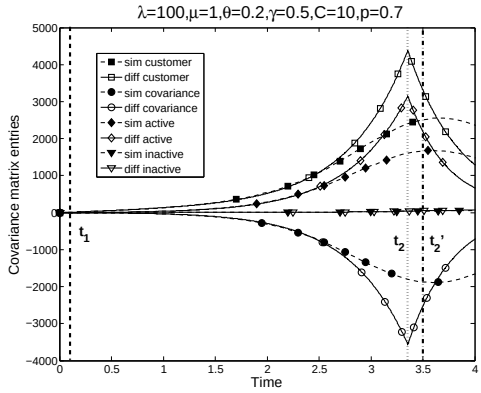
Figure 6: Comparison of mean numbers between standard and adjusted models in Example 1

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:

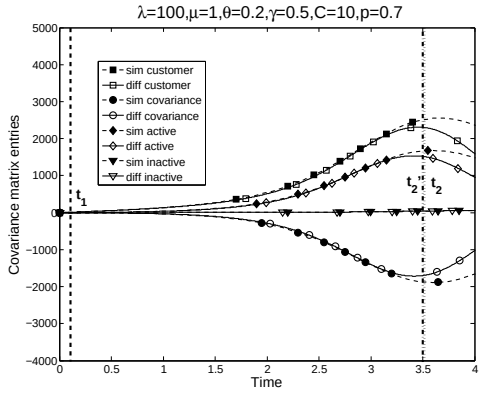
- The arrival rate implies the demand for the content. When operating a peer network, preparing a burst of the demand is crucial. Therefore, it is important to see when to reach stage 3 and how many peers (customers also) reside in the system at the end of stage 2, according to the arrival rates.

- 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 10 and 11 show the changes of t_2 and $E[X(t_2)]$ over λ and p respectively. As seen in Figure 10, 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 $E[x(t_2)](= E[y(t_2)])$ decrease according to p whereas $E[z(t_2)]$ increases. This implies that

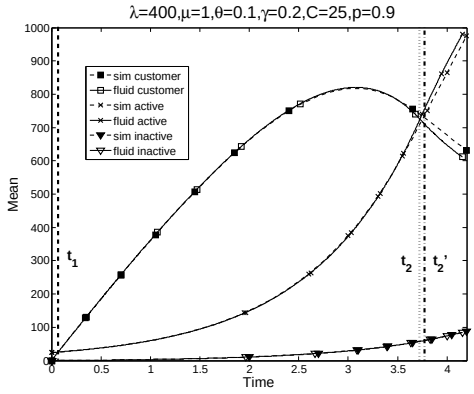


(a) Covariance matrix entries of standard model

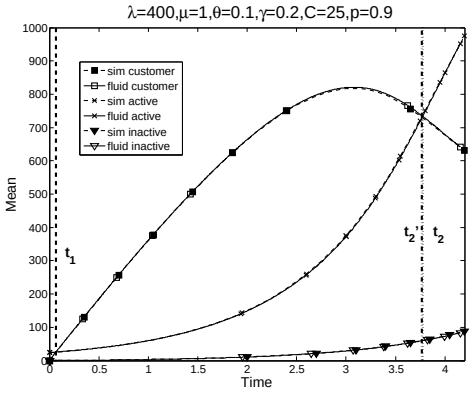


(b) Covariance matrix entries of adjusted model

Figure 7: Comparison of covariance matrices between standard and adjusted models in Example 1

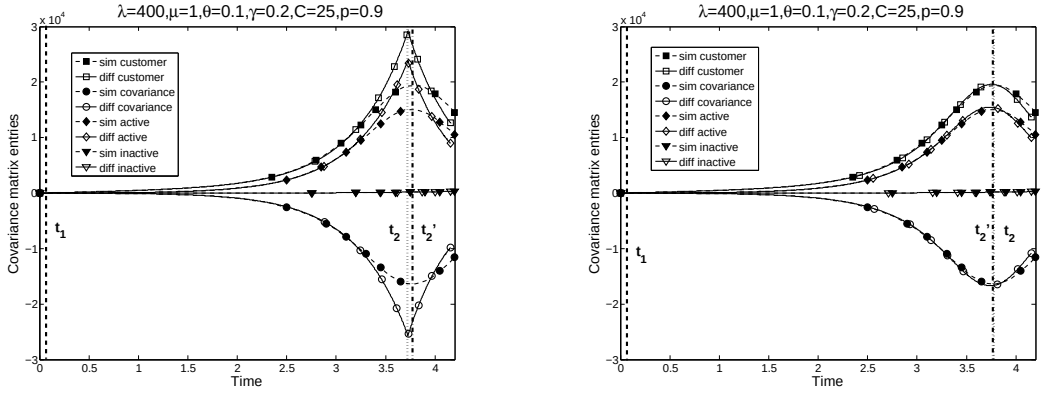


(a) Mean number of customers and peers of standard model



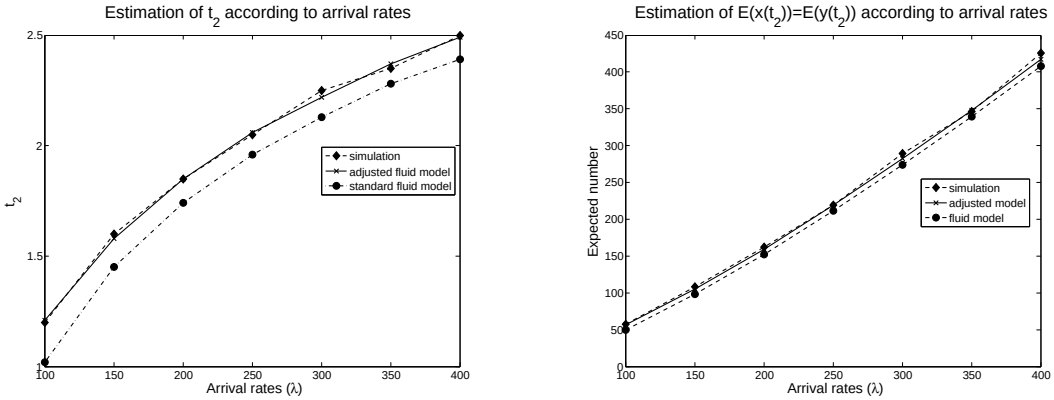
(b) Mean number of customers and peers of adjusted model

Figure 8: Comparison of mean numbers between standard and adjusted models in Example 2



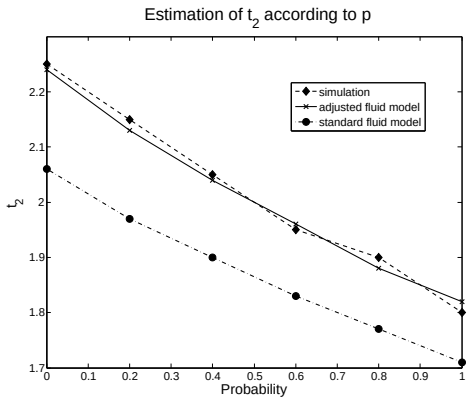
(a) Covariance matrix entries of standard model (b) Covariance matrix entries of adjusted model

Figure 9: Comparison of covariance matrices between standard and adjusted models in Example 2

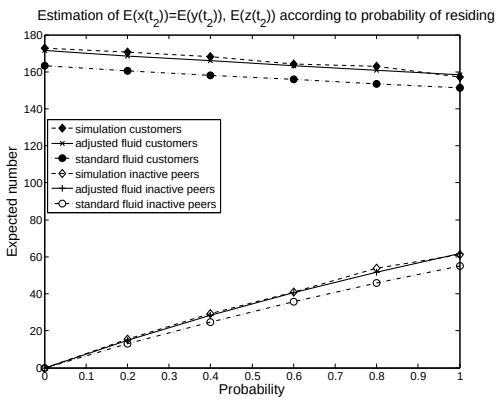


(a) Estimation of t_2 according to λ (b) Estimation of $E(x(t_2))$ according to λ

Figure 10: Estimation of t_2 and $E(x(t_2))$ according to λ



(a) Estimation of t_2 according to p

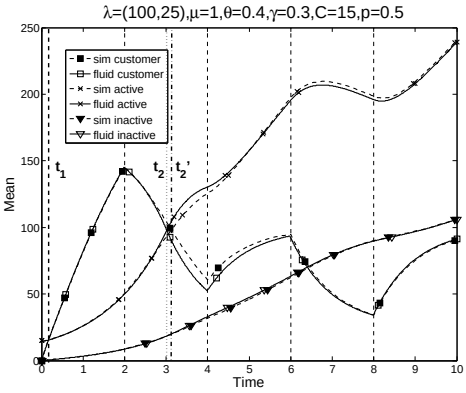


(b) Estimation of $E(x(t_2))$ and $E(z(t_2))$ according to p

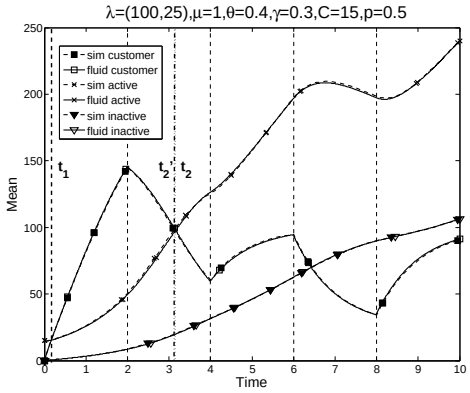
Figure 11: Estimation of $E(X(t_2))$ according to p

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 model provides more accurate t_2 and $E[X(t_2)]$ than the standard fluid model.

5.2 Time-varying rate functions



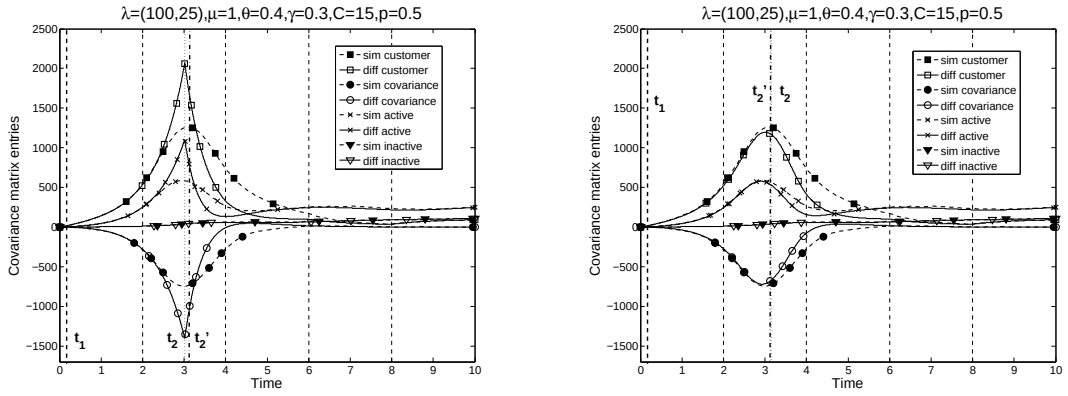
(a) Standard model with time-varying arrival rate



(b) Adjusted model with time-varying arrival rate

Figure 12: Mean number of customers and peers with time-varying arrival rate

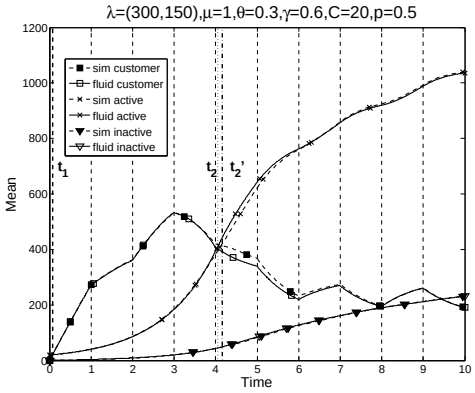
In Remark 5, we mentioned that fluid and diffusion approximations can be extended to time varying rate functions; i.e. arrival rate is $\lambda(t)$, service rate is $\mu(t)$, and peer's up and down times are $1/\theta(t)$ and $1/\gamma(t)$ on average, respectively. In this section, we show two numerical examples in that the arrival rate changes over time (μ, θ, and γ are held constant over time only for illustration purposes).



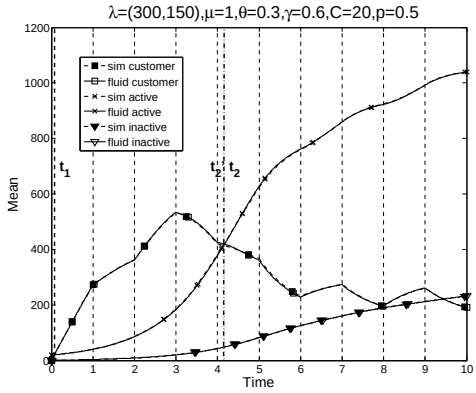
(a) Standard model with time-varying arrival rate (b) Adjusted model with time-varying arrival rate

Figure 13: Covariance matrix entries with time-varying arrival rate

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 two time units. We apply both the standard and adjusted models and compare them with simulation results. As seen in Figure 12, the adjusted model gives quite accurate results in all ranges of time intervals, whereas the standard model shows some error around $t \in [1.2, 2.5]$ and gives accurate results after $t > 3$. 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 $t = 1.5$ and after that it remains in $\bar{x}(t)$. Therefore, we can explain the reason for this phenomenon by Theorem 4 and Lemma 1, similar to the case of the 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 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 14 and 15, we observe results similar to the first example. The standard fluid model shows inaccuracy around $t \in [3.7, 6]$ while 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 think that our adjusted fluid and diffusion models work great during the time interval we are interested in, i.e. $0 \leq t \leq t_2$.

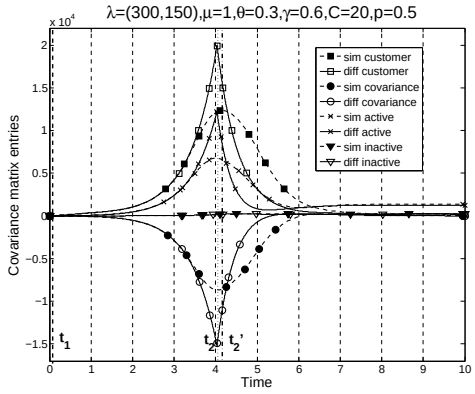


(a) Standard model with time-varying arrival rate

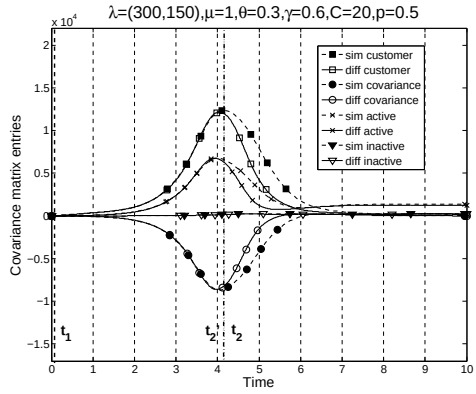


(b) Adjusted model with time-varying arrival rate

Figure 14: Mean number of customers and peers with time-varying arrival rate



(a) Standard model with time-varying arrival rate



(b) Adjusted model with time-varying arrival rate

Figure 15: Covariance matrix entries with time-varying arrival rate

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 Figure 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 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 customers 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.

There are several extensions to this paper that can be considered in the future. Firstly, we assume that the $X(t)$ 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 for obtaining the asymptotic distribution of $X(t)$ are required. Empirically, we observe that the distribution of $X(t)$ shows the extreme value type distribution near the switching time. Secondly, we assume that all the times follow exponential distributions. In some situations, this assumption is not realistic. Therefore, one could consider relaxing this assumption in the model formulation in the future.

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 the NSF grant CMMI-0946935.

References

[1] 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*, pages 1538–1549.

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

[3] Bassamboo, A., Kumar, S., and Randhawa, R. S. (2009). Dynamics of New Product Introduction in Closed Rental Systems. *To appear in Operations Research*.

[4] Bassamboo, A. and Randhawa, R. S. (2009). Optimal control in a Netflix-like closed rental system. *Submitted for Publication*.

[5] Billingsley, P. (1999). *Convergence of Probability Measures*. A John Wiley & Sons, Inc., Publication.

[6] 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*, pages 86–95.

[7] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence*. A John Wiley & Sons, Inc., Publication, 1 edition.

[8] 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.

[9] 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*.

[10] 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. *Proceedings of the ACM SOSP*.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] Massey, W. A. (2002). The Analysis of Queues with Time-Varying Rates for Telecommunication Models. *Telecommunication Systems*, 21(2-4):173–204.

[17] 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, pages 367–378.

[18] Whitt, W. (2002). *Stochastic Process Limits*. Springer, 1 edition.

[19] Whitt, W. (2004). Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments. *Management Science*, 50(10):1449–1461.

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

[21] Yang, X. and de Veciana, G. (2004). Service Capacity of Peer to Peer Networks. In *Proceedings of the IEEE INFOCOM*, volume 4, pages 2242–2252.