Modeling Fluctuations in the Quasi-static Approach Describing the Temporal Evolution of Retry Traffic

KOHEI WATABE
Graduate School of Information Science and Technology
Osaka University
Yamadaoka 1–5, Suita-shi, Osaka 565–0871
JAPAN
k-watabe@ist.osaka-u.ac.jp

MASAKI AIDA
Graduate School of System Design
Tokyo Metropolitan University
Asahigaoka 6–6, Hino-shi, Tokyo 191–0065
JAPAN
maida@sd.tmu.ac.jp

Abstract: In previous work, we introduced the quasi-static retry traffic model, which describes the behavior of retry traffic generated by users who are impatient when waiting for a response from the system. In other words, the model describes interactions between users and the system. This interaction can be described in a simple form if it is assumed that the system offers infinitely fast (ideal) processing. Moreover, we proposed a performance evaluation technique called the quasi-static approach that replicates the temporal evaluation of traffic in finite speed (real-world) systems. In the quasi-static approach, the difference between the behavior of the ideal system and that of the real-world system is expressed as stochastic fluctuation. In this paper, we model the fluctuation for exactly replicating the behavior of retry traffic caused by user impatience using the quasi-static approach, and show the validity of an evaluation of the quasi-static approach by comparing the results of the quasi-static approach and that of conventional Monte Carlo simulation, in M/M/1- and M/M/s-based systems with retry traffic.

Key–Words: Retry traffic, Quasi-static approach, Fluctuations, Traffic model, Queueing system, Langevin equation, Fokker-Plank equation

1 Introduction

One significant problem with the Internet is node failure due to congestion or overloading. A key factor behind overloading is retry traffic, where users repeatedly attempt to access links and in doing so generate duplicate service requests. To optimize resource allocation and construct stable systems, an evaluation method that accurately models retry traffic is therefore essential.

Retry traffic can be classified into the following two types:

- **Retry traffic due to request discard:** Caused by a shortage of system resources. Requests that exceed service capacity are discarded, regardless of whether the shortfall is temporary or chronic. Rejected users then reattempt service access.
- **Retry traffic due to impatience:** Human nature drives most users who have been kept waiting to issue duplicate service requests; the original request is not cancelled.

The M/G/s/s retrial queue model is a well-known attempt at using a queuing model to replicate retry traffic [1]. In that model, if all s servers are busy when a service request arrives at the system, the service request is discarded. Discarded requests reenter the system after a certain elapsed time that follows an exponential distribution. The conventional understanding is that the model describes communication services (including IP telephony) based on the Resource reSerVation Protocol (RSVP) [2]. In M/G/s/s, servers correspond to bandwidth, and a reattempt occurs if a service request arrives when all servers are busy, so there is no available bandwidth. Retry traffic in this model corresponds to the first of the two types mentioned above. That is, like most previous works [3, 4, 5, 6], the M/G/s/s retrial queue model does not consider retry traffic due to user impatience.

Taking IP telephony as an example, [7] modeled the behavior of retry traffic by considering not only request discards but also impatience; the result is called the *quasi-static retry traffic model*. Traffic behavior is determined by interaction between users and a system because users' decisions are affected by the state of a system. The model of [7] allows the interaction to be very simply described if the system can respond infinitely faster than the users.

To evaluate the behavior of traffic on systems that offer real-world speeds, [7] proposed the *quasi-static*

approach, in which the difference between the behavior of the infinitely high-speed system and that of the finite speed system is treated as stochastic fluctuation. Compared with the conventional Markov and Monte Carlo approaches, the quasi-static approach might be superior for estimating the probability of rare system outages on systems with finite but very high speed.

In this paper, we discuss how to model the fluctuation for exactly replicating the behavior of retry traffic caused by user impatience using the quasi-static approach. We demonstrate modeling of the fluctuation in M/M/1- and M/M/s-based systems with retry traffic, and confirm that the results of the quasi-static approach correspond to those of the slower conventional Monte Carlo simulation of queuing systems. Note that low-speed systems can be easily evaluated by conventional Monte Carlo simulation.

The rest of the paper is organized as follows. We summarize the quasi-static approach in Section 2. In Section 3, we compare the evaluation results of the quasi-static approach against those of Monte Carlo simulations of the input traffic under M/M/1- and M/M/s-based systems with retry traffic, and we introduce a fluctuation model for exactly replicating the behavior of retry traffic. The comparison demonstrates the validity of the quasi-static approach. We conclude the paper in Section 4.

2 Quasi-static Approach for IP Telephony System

2.1 Quasi-static Retry Traffic Model

Reference [7] introduced the *quasi-static retry traffic model* model to describe an IP telephony system with retry traffic due to both request discard and impatience. The model describes the interaction between users and a system with different timescales, since retry traffic is generated by user reattempts, which occur much more slowly than the responses of the system. In this subsection, we briefly explain the quasi-static retry traffic model introduced in [7].

Figure 1 shows a model of the IP telephony system investigated in [7]. The model is composed of a control plane and a data plane connected in series. It describes the behavior as related to call setup and data transmission processing. The control plane and the data plane are modeled by M/M/1 and M/G/s/s, respectively. Service requests first arrive at the M/M/1 queue and receive service from the M/M/1 server. Next, service requests receive service from one of the s servers in M/G/s/s. Subsequent service requests are discarded if all s servers are busy when they arrive at M/G/s/s. Service requests discarded on the data plane

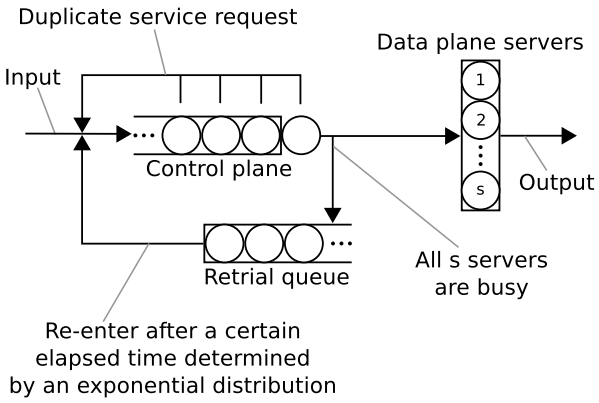

Figure 1: Model incorporating retry traffic from the control plane (M/M/1) and the data plane (M/G/s/s)

are stored in a retrial queue, and re-enter the system after an elapsed time determined by an exponential distribution. The volume of duplicate service requests is proportional to the number of users in the system of the control plane. Retry traffic from the control plane is the result of user psychology: increased service access delays trigger more reattempts. This retry traffic caused by user impatience characterizes the model.

We assume that reattempts due to impatience have extremely long timescales, as compared to the timescales of the transitions of the number in the system caused by service request arrival. By assigning a discrete time transition to the former and a continuous time transition to the latter, [7] constructed a traffic model that can express the difference in timescales.

For a certain constant $T\,(>0)$, where T denotes the timescale of user response, the traffic model is as follows:

- A change in the request arrival rate due to user reattempts occurs only when time $t = kT\,(k = 1, 2, \cdots)$.
- The arrival rate of retry traffic from the control plane in the time interval $(kT, (k+1)T]$ is proportional to the average number in the system in the time interval $((k-1)T, kT]$ (Fig. 2).
- Compared to the speed of user response, the system works at infinitely high speed, and the average number in the system in finite time interval $(kT, (k+1)T]$ is equal to the mean calculated from the stationary state probability.

Reference [7] used 1 s for T, based on a study of user response times [8].

The system attains the stationary state in the finite time interval $((k-1)T, kT]$ because the arrival rate

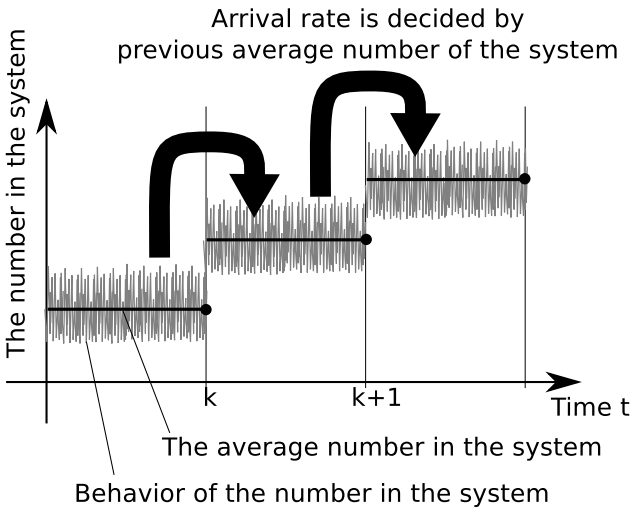

Figure 2: Relationship between a change in arrival rate and the number in the system

does not change in the interval, and we assume that the system can respond at infinitely high speed. Therefore, since the arrival rate λ_{k+1} of service requests in the time interval $(kT, (k+1)T]$ depends only on the arrival rate λ_k in the time interval $((k-1)T, kT]$, we get λ_{k+1} by the recurrence equation

$$\lambda_{k+1} = \lambda_0 + \lambda_k B(\lambda_k/\mu, s) + \varepsilon \frac{\lambda_k/\eta}{1-\lambda_k/\eta}\,, \qquad (1)$$

where λ_0 denotes the original arrival rate excluding retry traffic, η and μ denote service rates of the control and the data plane, respectively (i.e., the respective average service times are $1/\eta$ and $1/\mu$), and ε is a positive constant indicating the intensity of retry traffic generated from the control plane. $B(\rho, s)$ is the Erlang B formula and represents the discard rate on M/G/s/s, as follows:

$$B(\rho, s) = \frac{\rho^s/s!}{1+\rho+\rho^2/2!+\cdots+\rho^s/s!}\,.$$

The second and the third term on the right side of Eq. (1) represent retry traffic from the data and the control plane, respectively.

In this model, the rate of retry traffic from users in the control plane is proportional to the time average of the number in the system, and this means that the number of user reattempts is proportional to user time spent in the M/M/1 system on the control plane (the delay before service provision). According to the PASTA (Poisson Arrivals See Time Averages) property [9], in M/M/1 the event-average of the number in the system just before service request arrival equals the time average of the process of the number in the system. If the system works at infinitely high speed (namely, the limit for $\lambda_k \to \infty$, $\eta \to \infty$), the following equation holds:

$$\lim_{\lambda_k\to\infty} \frac{\lambda_k/\eta}{1-\lambda_k/\eta} = \lim_{\lambda_k\to\infty} \frac{1}{M(\lambda_k)} \sum_{i=1}^{M(\lambda_k)} Q_i^k \qquad \text{a.s.}, \qquad (2)$$

where $M(\lambda_k)$ and Q_i^k $(i = 1, 2, \cdots, M(\lambda_k))$ denote the number of arrivals in the time interval $((k-1)T, kT]$ and the number in the system just before the ith arrival in the time interval $((k-1)T, kT]$, respectively. $M(\lambda_k)$ is a random variable that follows a Poisson distribution with parameter $\lambda_k T$. Note that the left side of Eq. (2) corresponds to the third term on the right side of Eq. (1). Moreover, using Little's formula [10], we find

$$\lim_{\lambda_k\to\infty} \frac{1}{M(\lambda_k)} \sum_{i=1}^{M(\lambda_k)} Q_i^k = \lambda_k \lim_{\lambda_k\to\infty} \frac{1}{M(\lambda_k)} \sum_{i=1}^{M(\lambda_k)} W_i^k = \lim_{\lambda_k\to\infty} \sum_{i=1}^{M(\lambda_k)} W_i^k \qquad \text{a.s.},$$

where W_i^k $(i = 1, 2, \cdots, M(\lambda_k))$ denotes the time spent by the ith arrival in the system in the time interval $((k-1)T, kT]$, and the last equality follows from the following limit:

$$\lim_{\lambda_k\to\infty} \frac{M(\lambda_k)}{\lambda_k} = 1 \qquad \text{a.s.}$$

Therefore, if each user reattempts access in proportion to waiting time, the input traffic is determined by Eq. (1) on the limit $\lambda_k \to \infty$. Since we assume the system works at infinitely high speed, Eq. (2) holds, and we can get the transition of the arrival rate by determining λ_{k+1} in Eq. (1) from the original arrival rate λ_0. As a result, we can analyze the stability of the system [7, 11].

2.2 Quasi-static Approach

As mentioned above, Eq. (1) describes the behavior of a system with infinitely high response speed. Equation (1) may not describe real systems, since actual response speeds are finite. With real-world systems, since we cannot take the limit as in Eq. (2), we should add a fluctuation term $\phi(\lambda_k)$ as follows:

$$\frac{\lambda_k/\eta}{1-\lambda_k/\eta} + \phi(\lambda_k) = \frac{1}{M(\lambda_k)} \sum_{i=1}^{M(\lambda_k)} Q_i^k\,. \qquad (3)$$

Note that $\phi(\lambda_k)$ is a random variable whose mean is 0.

If we analyze the behavior of the left side of Eq. (3) using the conventional Markov approach, we must consider a Markov model with $M(\lambda_k)$-dimensional state space consisting of the past $M(\lambda_k)$ states $\{Q_i^k\}\,(i = 1, \cdots, M(\lambda_k))$. In the case of $M(\lambda_k) \gg 1$ (an extremely fast system, but not an infinitely fast one), the problem becomes intractable due to the excessive state space. When we use Monte Carlo simulation to analyze the behavior of a high-but-finite-speed system, the simulation must be quite long if we are interested in the probabilities of rare events (e.g. the probability of service failure is to be less than 10^{-6}). Therefore, it is difficult to analyze the behavior of high-speed systems by conventional approaches.

To solve the above problem, [7] proposed adding stochastic fluctuations to the behavior of a system with infinitely high response speed. This is called the quasi-static approach. The stochastic fluctuations mirror the difference between the behavior of finite speed systems and that of infinitely high-speed systems. Reference [7] defines $X(t)$ as the volume of input traffic, including retry traffic, at time t, and expresses the temporal evolution of $X(t)$ using the Langevin equation

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = F(X(t)) + \sqrt{D(X(t))}\,\xi(t)\,, \quad (4)$$

where $\xi(t)$ denotes white Gaussian noise with the following property:

$$\mathrm{E}[\xi(t)] = 0, \qquad \mathrm{E}[\xi(t)\xi(t')] = \delta(t - t')\,.$$

Note that, for expedience, we replace discrete time kT with continuous time t in Eq. (4). $F(X)$ and $D(X)$ are given as

$$F(X) = \lambda_0 + B\left(\frac{X}{tT}, s\right) + \varepsilon\frac{X/(\eta T)}{1 - X/(\eta T)} - \frac{X}{T}\,, \quad (5)$$

$$D(X) = \frac{X}{T} + c(X)\,, \quad (6)$$

where $c(x)$ is a simple step function representing the fluctuations in retry traffic from the data plane (see [7] for details). In Section 3, we discuss the derivation and the validity of $F(X)$ and $D(X)$ in detail for M/M/1- and M/M/s-based systems with retry traffic.

Moreover, we can eliminate X-dependence from the second (fluctuation) term on the right side of Eq. (4) by transforming the random variable using

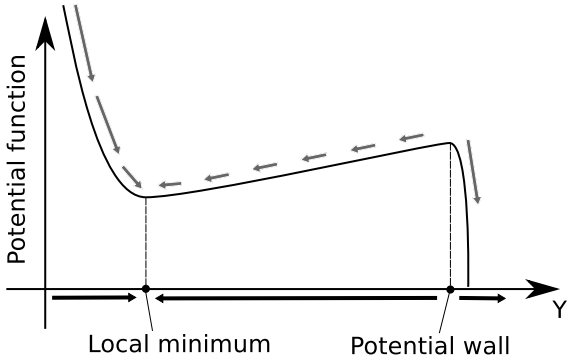

Figure 3: An example potential function

$Y = 2\sqrt{D(X)}$. In other words, we can treat the magnitude of fluctuations as a constant for any X. The result of the transformation is as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}Y(t) = G(Y(t)) + \xi(t),$$

$$G(y) = \frac{F(x) - 1/4}{\sqrt{x/T + c(x)}}, \qquad y = 2\sqrt{x/T + c(x)}\,.$$

We can investigate the behavior of $Y(t)$ using the potential function given by $-\int G(y)\mathrm{d}y$.

The potential function indicates the tendency of the temporal evolution of Y, but Y fluctuates by the effect of $\xi(t)$ and often moves against the potential function. On average, Y moves in the direction that lowers the potential function. Figure 3 shows an example potential function. In that example, Y tends to be distributed near the local minimum point. If, however, Y reaches the potential wall due to fluctuations $\xi(t)$, Y diverges (namely, the arrival rate diverges and the system suffers overloading).

It is well known that the Langevin equation Eq. (4) is equivalent to the Fokker-Planck equation [12] as shown by

$$\frac{\partial}{\partial t}p(y,t) = -\frac{\partial}{\partial y}G(y)p(y,t) + \frac{1}{2}\frac{\partial^2}{\partial y^2}p(y,t)\,, \quad (7)$$

where $p(y,t)$ denotes the probability density function (PDF) of $Y(t)$. Using Eq. (7), we can simulate the transition of the PDF of the volume of traffic, and can assess the probability of its divergence and so forth.

3 Comparing the Quasi-static Approach with the Conventional Approach

3.1 Verification in a Simple M/M/1 Model

We verify that the quasi-static approach, which adds random fluctuations to the behavior of infinitely high-speed systems, can appropriately describe the behavior of finite-speed systems. We start our verification using one of the simplest models: an M/M/1 system without retry traffic. The quasi-static approach can describe a system that contains no retry traffic, though it was proposed to analyze the behavior of retry traffic. The arrival rate of a simple M/M/1 system is constant, and it is well known that the volume of traffic follows a Poisson distribution.

First, we consider the Langevin equation corresponding to the M/M/1 system. Since the second and third terms on the right side of Eq. (5) correspond to retry traffic from the data and control plane, respectively, we find that

$$F(X) = \lambda_0 - \frac{X}{T}\,. \tag{8}$$

Similarly, since the second term on the right side of Eq. (6) represents the fluctuation magnitude of retry traffic from the data plane, we obtain

$$D(X) = \frac{X}{T}\,. \tag{9}$$

We therefore get the Langevin equation corresponding to a simple M/M/1 system by substituting Eqs. (8) and (9) for Eq. (4).

Moreover, the Fokker-Planck equation equivalent to this Langevin equation is derived as follows:

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}F(x)p(x,t) + \frac{1}{2}\frac{\partial^2}{\partial x^2}D(x)p(x,t)\,. \tag{10}$$

Note that we do not transform the random variable, because our aim is not to consider system stability with respect to the potential function, but rather to compare the distribution of X against a Poisson distribution.

Using Eq. (10), we compute the input traffic of the simple M/M/1 system, and in Fig. 4 the result of the stationary state is compared with the Poisson distribution that is the theoretical result. The parameters of the M/M/1 system are the arrival rate $\lambda_0 = 35$, the service rate $\eta = 50$, and the timescale $T = 1$ s. The horizontal axis represents the volume of traffic $X(t)$ that is the number of arrivals in the interval $(t-T, t]$.

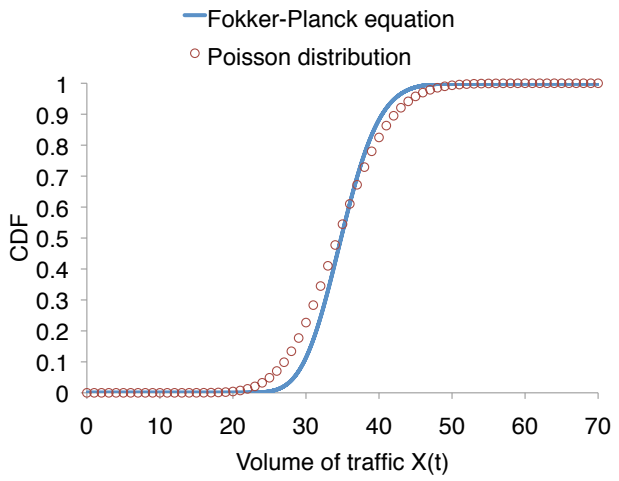

Figure 4: Poisson distribution and the distribution of $X(t)$ as computed by the Fokker-Planck equation with Eqs. (8) and (9)

Note that the figure displays the distributions as cumulative density functions (CDFs), not PDFs. According to the figure, the distribution does not correspond to a Poisson distribution, despite having already reached a stationary state. We must therefore reconsider Eqs. (8) and (9).

To solve this problem, we should exactly model the system as the Langevin equation. It is intuitive that $X(t)$ is defined as the actual number of arrivals in time interval $(t-T, t]$, since λ_k is the arrival rate in the time interval $((k-1)T, kT]$. If we define $\mathrm{d}X(t)$ as the change of $X(t)$ in a minute distance $\mathrm{d}t$, we find that

$$\begin{aligned}\mathrm{d}X(t) &= X(t+\mathrm{d}t) - X(t)\\ &= U(t,\mathrm{d}t) - U(t-T,\mathrm{d}t)\,,\end{aligned}$$

where $U(t,\mathrm{d}t)$ is the actual number of arrivals in the time interval $(t, t+\mathrm{d}t]$ (Fig. 5). The expectation and variance of random variable $U(t,\mathrm{d}t)$ are both $\lambda_0\mathrm{d}t$ because the future arrivals follow a Poisson arrival. Moreover, the conditional distribution of $U(t-T,\mathrm{d}t)$, given that $U(t-T,T) = X(t)$, obeys a binomial distribution $B(X(t), \mathrm{d}t/T)$ [13]. Therefore, the conditional expectation and variance of $U(t-T,\mathrm{d}t)$ is as follows:

$$\mathrm{E}[U(t-T,\mathrm{d}t)\,|\,U(t-T,T) = X(t)] = (X(t)/T)\mathrm{d}t,$$

$$\begin{aligned}&\mathrm{Var}[U(t-T,\mathrm{d}t)\,|\,U(t-T,T) = X(t)]\\ &\quad= (X(t)/T)\mathrm{d}t - (X(t)/T^2)(\mathrm{d}t)^2\\ &\quad\simeq (X(t)/T)\mathrm{d}t.\end{aligned}$$

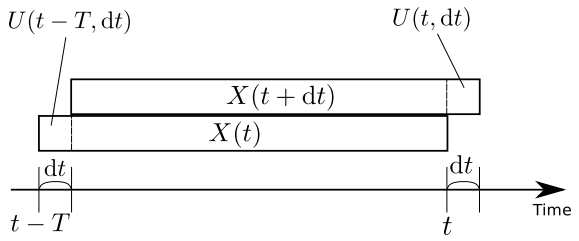

Figure 5: Transition of $X(t)$ that expresses actual input traffic in the past period T [s]

As a result, we can write $\mathrm{d}X(t)$ as

$$\mathrm{d}X(t) \simeq \lambda_0 \mathrm{d}t - \frac{X(t)}{T}\mathrm{d}t + \sqrt{\lambda_0 + \frac{X(t)}{T}}\, N(t)\sqrt{\mathrm{d}t}\,,$$

where $N(t)$ is a random variable that obeys a standard normal distribution and time series $N(t)$ are independent for different t. Now, we define $W(t)$ as a Wiener process. Since $N(t)\sqrt{\mathrm{d}t}$ is $\mathrm{d}W(t)$, we get the Langevin equation as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}X(t) &= \lambda_0 - \frac{X(t)}{T} + \sqrt{\lambda_0 + \frac{X(t)}{T}}\,\frac{\mathrm{d}W(t)}{\mathrm{d}t} \\ &= \lambda_0 - \frac{X(t)}{T} + \sqrt{\lambda_0 + \frac{X(t)}{T}}\,\xi(t)\,,\end{aligned}$$

and the appropriate fluctuation magnitude $D(X)$ is given by

$$D(X) = \lambda_0 + \frac{X(t)}{T}\,. \tag{11}$$

This is a necessary revision, because $X(t)$ is given as a function of time t, not as an amount that is defined in a time interval.

We recomputed the distribution of $X(t)$ using the above Fokker-Planck equation with exactly modeled fluctuation magnitude $D(X)$. Figure 6 shows the results; we can confirm that the behavior of input traffic for simple M/M/1 is described appropriately by the quasi-static approach. Figure 7 presents the potential function corresponding to the above experiment. Note that this potential function considers $X(t)$ (not $Y(t)$), and the fluctuations on each X are not constant. We can find the local minimum (stability point) at $X = 35.0$, and the volume of traffic X tends to be distributed around the stability point.

3.2 Verification in an M/M/1-based System with Retry Traffic

This subsection verifies the validity of the quasi-static approach for an M/M/1-based system with retry traffic. This model is significant because a control plane

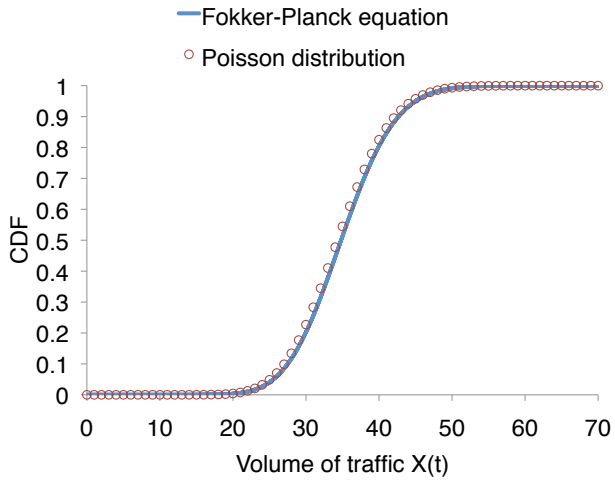

Figure 6: Poisson distribution and the distribution of $X(t)$, as computed by the Fokker-Planck equation with Eqs. (8) and (11)

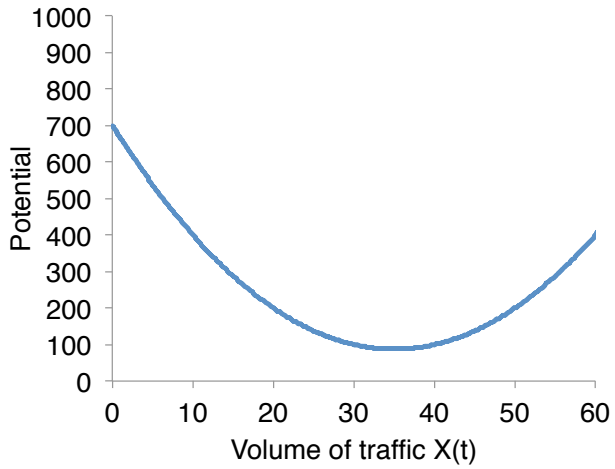

Figure 7: Potential function of the simple M/M/1

that generates retry traffic characterizes the model of an IP telephony system. In this subsection we treat M/M/1-FIFO-based systems, but the results can also apply to M/M/1-PS systems since a process of the number in the M/M/1-PS system is the same birth-death process as that of an M/M/1-FIFO system [14].

In the system of this section, the retry traffic rate at time t is proportional to the event-average of the number in the system at the time of request arrival in the time interval $(t-T, t]$, and it is added to the arrival rate at time t. Note that the traffic rate is changed not at discrete times as in the quasi-static retry traffic model, but rather using continuous time. Unfortunately, there is no analytical method to obtain the distribution of system traffic volume, unlike the case of

the simple M/M/1 system. We therefore compute the distribution of traffic volume (the number of arrivals including the retry traffic in an interval $(t-T,t]$) by Monte Carlo simulation, and compare with the results of the quasi-static approach.

As in the simple M/M/1 case, we investigate the Langevin equation corresponding to the M/M/1-based system with retry traffic. If we assume that the system works at infinitely high speed, the transition of the arrival rate in the quasi-static retry traffic model is given by

$$\lambda_{k+1} = \lambda_0 + \varepsilon \frac{\lambda_k/\eta}{1-\lambda_k/\eta}\,.$$

We rewrite this equation as difference $\Delta\lambda_k$,

$$\begin{aligned}\Delta\lambda_k &= \lambda_{k+1} - \lambda_k \\ &= \lambda_0 - \lambda_k + \varepsilon \frac{\lambda_k/\eta}{1-\lambda_k/\eta}\,.\end{aligned}$$

By a natural continuation to yield the Langevin equation, we have

$$\mathrm{d}\lambda(t) = \frac{1}{T}\left(\lambda_0 - \lambda(t) + \varepsilon \frac{\lambda(t)/\eta}{1-\lambda(t)/\eta}\right)\mathrm{d}t\,,$$

where $\lambda(t)$ denotes the arrival rate at time t. We substitute $X(t)/T$ (the actual number of arrivals per second) for $\lambda(t)$ to consider a finite speed system. As in the simple M/M/1 case, the change of $X(t)$, which is the number of arrivals in the past period T [s], is composed of the increment $U(t,\mathrm{d}t)$ and the decrement $-U(t-T,\mathrm{d}t)$. Their conditional expectation and variance are given by

$$\begin{aligned}&\mathrm{E}[U(t,\mathrm{d}t)\,|\,U(t-T,T)=X(t)] \\ &\qquad = \lambda_0 + \varepsilon\frac{X(t)/(\eta T)}{1-X(t)/(\eta T)}\mathrm{d}t\,,\end{aligned}$$

$$\begin{aligned}&\mathrm{Var}[U(t,\mathrm{d}t)\,|\,U(t-T,T)=X(t)] \\ &\qquad = \lambda_0 + \varepsilon\frac{X(t)/(\eta T)}{1-X(t)/(\eta T)}\mathrm{d}t\,,\end{aligned}$$

$$\mathrm{E}[-U(t-T,\mathrm{d}t)] = -\frac{X(t)}{T}\mathrm{d}t\,,$$

$$\mathrm{Var}[-U(t-T,\mathrm{d}t)] = \frac{X(t)}{T}\mathrm{d}t\,.$$

As a result, $F(X)$ and $D(X)$ of the Langevin equation that describes the temporal evolution of $X(t)$ are given by

$$F(X) = \lambda_0 - \frac{X(t)}{T} + \varepsilon\frac{X(t)/(\eta T)}{1-X(t)/(\eta T)}\,, \quad (12)$$

$$D(X) = \lambda_0 + \frac{X(t)}{T} + \varepsilon\frac{X(t)/(\eta T)}{1-X(t)/(\eta T)}\,. \quad (13)$$

The third term on the right side of Eq. (13) corresponds to the fluctuations of retry traffic. However, Eq. (6) does not contain this term since it can be neglected under the conditions assumed in [7]. By substituting Eqs. (12) and (13) for Eq. (10), we find the Fokker-Planck equation that can compute the behavior of the M/M/1-based system with retry traffic.

We used two methods to compute the CDF of input traffic including retry traffic when $t = 1$, 10, 20, and 30 s: Monte Carlo simulation and the Fokker-Planck equation for the quasi-static approach. Figure 8 shows the results. The parameters used are as follows: The original arrival rate λ_0 excluding retry traffic is 35, the service rate η is 50, the intensity of retry traffic ε is 0.1, and user timescale T is 1 s. We set the initial distribution $p(x,0)$ of the volume of traffic as a Poisson distribution with parameter $\lambda_0 T$. The figure confirms that the quasi-static approach yields results similar to those of the Monte Carlo simulation, though the model contains retry traffic. Figure 9 presents the potential function corresponding to the above experiment. Note that this potential function considers $X(t)$ (not $Y(t)$), and the fluctuations on each X are not constant. We can find the local minimum (stability point) and the wall of the potential at $X = 35.25$ and $X = 50$, respectively. The volume of traffic X tends to be distributed around the stability point, but X diverges once it crosses the wall of potential because it exceeds the capacity ηT that the system can process in a period T and the number in the system diverge. It is of prime interest to evaluate traffic divergence, which corresponds to system overload. In Fig. 8, the value of CDF at $X = 50$ falls with time, meaning that the diverge probability of traffic increases gradually. Note that we consider traffic to have diverged if $X(t)$ exceeds ηT once in the Monte Carlo simulation.

To verify the validity of the quasi-static approach, we check that its results correspond to those of Monte Carlo simulation with various parameters. First, we vary the original input traffic rate λ_0 as 30, 35, and 40. The parameters of the experiment are the same as in the above experiment, except for λ_0. Figure 10 shows the results of the probability of traffic divergence. The horizontal and vertical axes represent time and the probability of traffic divergence, respectively. The figure therefore indicates the CDF of the time that

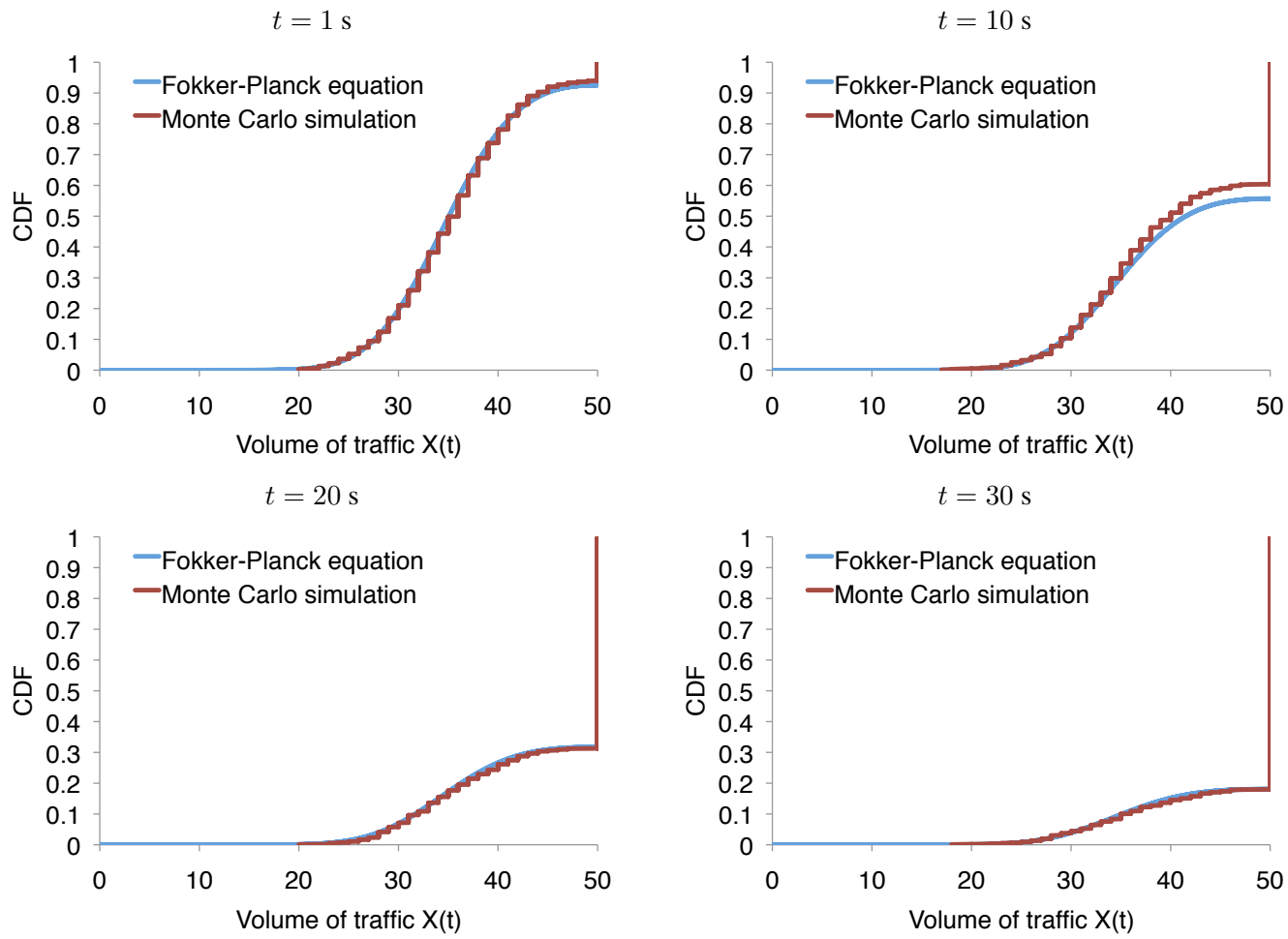

Figure 8: Distribution of input traffic at time t on the M/M/1-based system with retry traffic

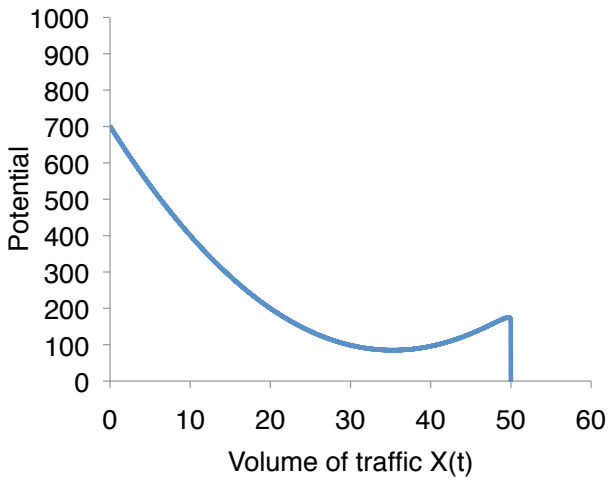

Figure 9: Potential function of the M/M/1-based system with retry traffic

the system will spend until the volume of traffic $X(t)$ in the interval $(t-T, t]$ first exceeds the capacity ηT that the system can process in a period T. The figure confirms that the divergence processes of the traffic calculated by the Fokker-Planck equation correspond to the results of Monte Carlo simulation for any λ_0. Similarly, we individually vary the user timescale T as 0.5, 1.0, and 1.5, and the intensity of retry traffic ε as 0.1 and 0.5. The other parameters of the experiment are the same as in the first experiment in this subsection. Figures 11 and 12 display the respective results and similar results are gained.

The above experiments confirm that the quasi-static approach can describe the behavior of retry traffic in an M/M/1-based system, like the conventional Monte Carlo simulation. We confirm that the probability of traffic divergence calculated by the quasi-static approach corresponds to the results of Monte Carlo simulation of the queuing system under slow system conditions. The target system that we want to evaluate using the quasi-static approach is a high-

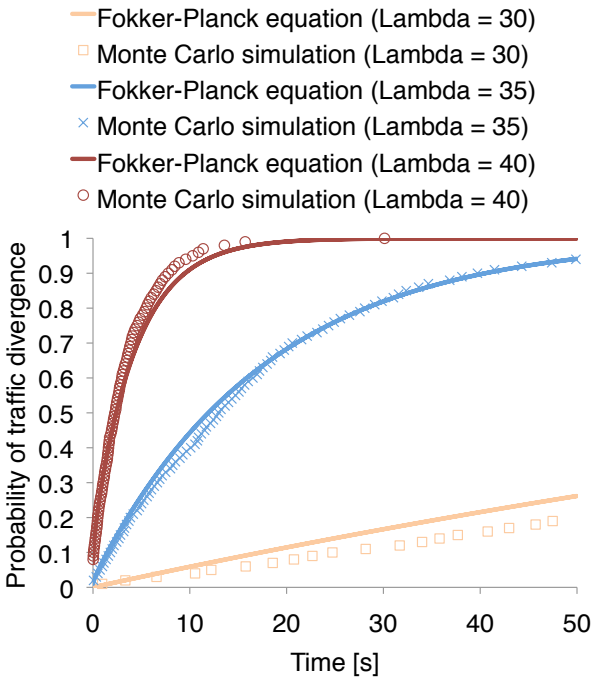

Figure 10: Probability of traffic divergence in an M/M/1-based system with retry traffic for various λ_0

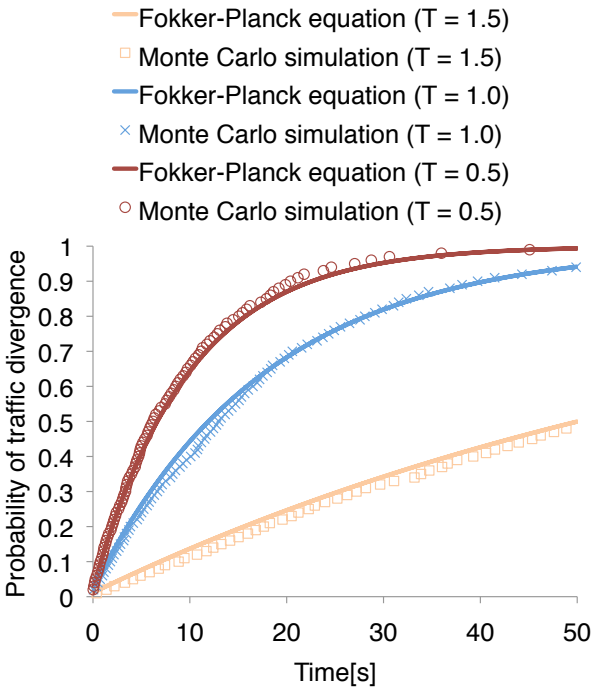

Figure 11: Probability of traffic divergence in an M/M/1-based system with retry traffic for various timescales T

speed system that is difficult to evaluate using conventional Monte Carlo simulation. We expect that the quasi-static approach can appropriately evaluate the behavior of retry traffic for a high-speed system, since the quasi-static approach can evaluate a low-speed system despite the difficulty of evaluating a large magnitude of fluctuations.

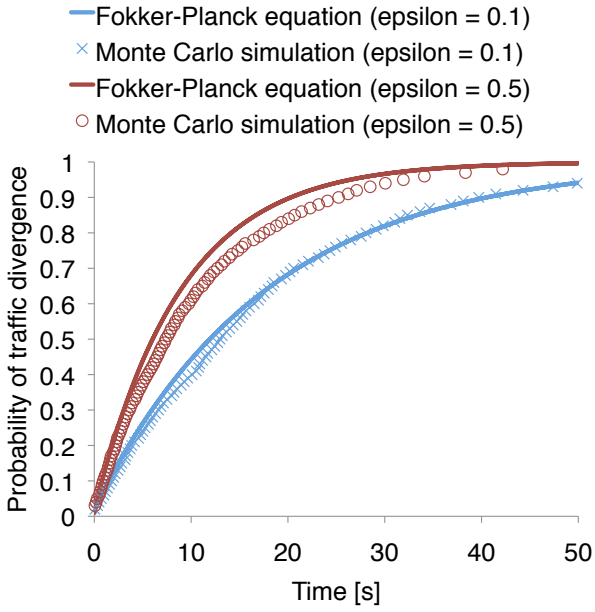

Figure 12: Probability of traffic divergence in an M/M/1-based system with retry traffic for various retry traffic intensities ε

3.3 Verification in an M/M/s-based System with Retry Traffic

Finally, we apply the quasi-static approach to an M/M/s-based system with retry traffic, and verify the validity of the quasi-static approach. We assume that retry traffic is generated at a rate proportional to the number in the system of M/M/s-based system, like the above-mentioned M/M/1-based system with retry traffic. The average number in the system of the M/M/s system $q(\rho, c)$ is given by

$$q(\rho, c) = \frac{\rho(c\rho)^c}{c!(1-\rho)^2} P_0 + c\rho,$$

$$P_0 = \frac{1-\rho}{(1-\rho)\sum_{i=0}^{c} \frac{(c\rho)^i}{i!} + \frac{(c\rho)^c}{c!}\rho},$$

where λ, η, ρ, and c denote arrival rate, service rate, utilization factor λ/η, and the number of servers, respectively [15]. By replacing the third terms, which describe the behavior of retry traffic, we can easily

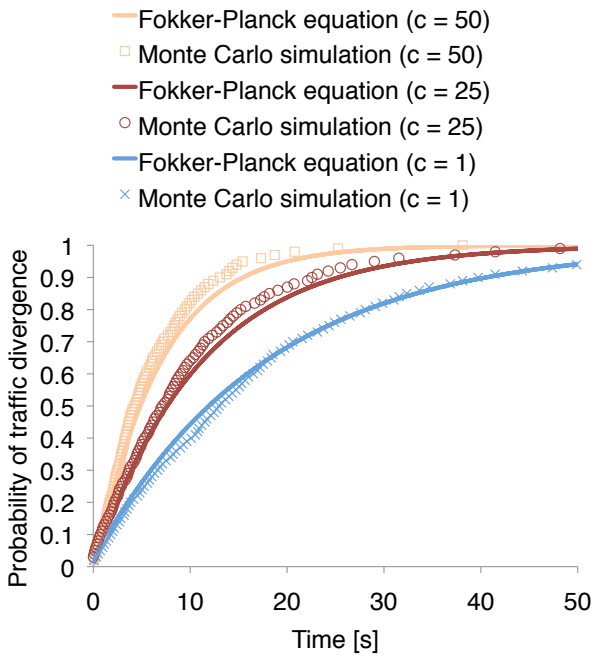

Figure 13: Probability of traffic divergence in an M/M/s-based system with retry traffic for various number of servers c

extend Eqs. (12) and (13) of the M/M/1 model to the M/M/s model as follows:

$$F(X) = \lambda_0 - \frac{X(t)}{T} + \varepsilon q\left(\frac{X(t)}{\eta T}, c\right), \qquad (14)$$

$$D(X) = \lambda_0 + \frac{X(t)}{T} + \varepsilon q\left(\frac{X(t)}{\eta T}, c\right). \qquad (15)$$

We calculate the probability of traffic divergence in an M/M/s-based system with retry traffic using the Fokker-Planck equation with Eqs. (14) and (15). Figure 13 compares of the results of the Fokker-Planck equation and the Monte Carlo simulation for the number of servers $c = 1$, 25, and 50. The figure confirms that the results of the Fokker-Planck equation correspond to those of Monte Carlo simulation.

The quasi-static approach evaluates the behavior of retry traffic due to user impatience in various queuing systems, though this paper focused on M/M/1- (or M/M/1-PS-) and M/M/s-based systems with retry traffic. We can obtain $F(X)$ and $D(X)$ of another system version by replacing the function q in Eqs. (14) and (15) as the average number in the system if we assume that the request arrival process is a Poisson process.

4 Conclusion

This paper modeled the fluctuation for exactly replicating the behavior of retry traffic due to user impatience using the quasi-static approach, and showed the validity of an evaluation of the quasi-static approach. We compared the distribution of the volume of traffic and the probability of traffic divergence calculated by the quasi-static approach and conventional Monte Carlo simulation, in M/M/1- and M/M/s-based systems with retry traffic under the condition of a slow system. We confirmed that the evaluation results of the quasi-static approach and Monte Carlo simulation correspond, and so expect that the quasi-static approach can appropriately evaluate the behavior of retry traffic in a high-speed system, because the quasi-static approach can evaluate a low-speed system despite the large magnitude of its fluctuations that make evaluation more difficult.

Acknowledgements: This work was supported by a JSPS Grant-in-Aid for JSPS Fellows, Grant Number 24 · 3184.

References:

[1] G. I. Falin and J. G. C. Templeton, *Retrial queues*. Chapman and Hall, 1997.

[2] B. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification," *RFC2205*, 1997.

[3] A. Gómez-Corral, "Stochastic Analysis of a Single Server Retrial Queue with General Retrial Times," *Naval Research Logistics*, vol. 46, no. 5, pp. 561–581, Aug. 1999.

[4] J. Wang and J. Cao, "Reliability Analysis of the Retrial Queue with Server Breakdowns and Repairs," *Queueing Systems*, vol. 38, no. 4, pp. 363–380, Aug. 2001.

[5] J. Wang and P. Zhang, "A Discrete-Time Retrial Queue with Negative Customers and Unreliable Server," *Computers & Industrial Engineering*, vol. 56, no. 4, pp. 1216–1222, May 2009.

[6] B. Krishna Kumar, G. Vijayalakshmi, A. Krishnamoorthy, and S. Sadiq Basha, "A Single Server Feedback Retrial Queue with Collisions," *Computers & Operations Research*, vol. 37, no. 7, pp. 1247–1255, Jul. 2010.

[7] M. Aida, C. Takano, M. Murata, and M. Imase, "A Proposal of Quasi-static Approach for Analyzing the Stability of IP Telephony Systems," in *International Conference on Networking (ICN 2008)*, Cancun, Mexico, Apr. 2008, pp. 363–370.

[8] J. Nielsen, "Usability Heuristics," in *Usability Engineering*. Academic Press, 1993, ch. 5, p. 135. [Online]. Available: http://www.useit.com/papers/responsetime.html

[9] R. W. Wolff, "Poisson Arrivals See Time Averages," *Operations Research*, vol. 30, no. 2, pp. 223–231, Mar. 1982.

[10] J. D. C. Little, "A Proof of the Queuing Formula: $L = \lambda W$," *Operations Research*, vol. 9, no. 3, pp. 383–387, Mar. 1961.

[11] M. Aida, C. Takano, M. Murata, and M. Imase, "A Study of Control Plane Stability with Retry Traffic : Comparison of Hard- and Soft-State Protocols," *IEICE Transactions on Communications*, vol. 91, no. 2, pp. 437–445, Feb. 2008.

[12] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*. Elsevier, 1992.

[13] N. L. Johnson, A. W. Kemp, and S. Kotz, "Poisson Distribution," in *Univariate Discrete Distributions*, 3rd ed. John Wiley & Sons, 2005, ch. 4, p. 166.

[14] E. G. Coffman, R. R. Muntz, and H. Trotter, "Waiting Time Distributions for Processor-Sharing Systems," *Journal of the ACM*, vol. 17, no. 1, pp. 123–130, Jan. 1970.

[15] M. Barbeau and E. Kranakis, "Medium Access Control," in *Principles of Ad Hoc Networking*. John Wiley & Sons, 2007, ch. 2, p. 42.