Diffusion Models for Computer/Communication Systems*

Toshikazu Kimura

Diffusion approximations for queues contain two conceptually different kinds of approximations, i.e., diffusion limits justified by heavy traffic limit theorems for unstable queues and diffusion models as continuous approximations for stable queues. The purpose of this paper is to provide (i) some basic concepts and issues in diffusion modeling, and (ii) a bibliographical guide to diffusion models for queues that are typically found in computer/communication systems.

JEL Classification Numbers: C44, C63
Key Words: Diffusion Models, Queues

1. Introduction

Performance evaluation of computer/communication systems has been often addressed through the analysis of queues therein. Exact solutions are available only under restrictive assumptions that scarcely fit in with the reality, making approximate solutions a practical necessity. Approximation methods for queues have thus been of practical interests. One of widely used approximations for queues is a diffusion approximation, where a queueing characteristic process (e.g., queue length, waiting time, work load, etc.) is approximated by a diffusion (Brownian motion) process. A number of textbooks of the theory of queues have dealt with diffusion approximation as a standard approximation method [2, 40, 41, 49, 57, 77, 104]; see also [5, 34, 39, 81].

It has been known that diffusion approximations for unstable queues can be often justified by heavy traffic limit theorems (HTLTs): When a queueing characteristic process is appropriately scaled and translated, HTLTs show that the translated process in an unstable queue converges weakly to a Brownian motion process; see [12, 24, 47, 72, 93, 132] for surveys as well as bibliographical guides. However, this does not necessarily imply that the Brownian motion process still gives an accurate approximation when the queue is stable, because HTLTs provide us no information on the process behavior in stable situations. Hence, we must clearly distinguish two kinds of “diffusion approximations”, i.e., diffusion limits justified by HTLTs for unsta-
ble queues and diffusion models as continuous approximations for stable queues. Such a clear distinction has not sufficiently been made in the literature. This paper is a first review of diffusion models for queues that are typically found in computer/communication systems. While I do not claim that this review is comprehensive, I hope that it will serve as a useful reference for researchers and engineers who are interested in diffusion modeling.

Here is how this paper is organized: In Section 2, we introduce some basic concepts and issues in diffusion modeling. In Section 3, we review previously established diffusion models for single-server queues, state-dependent queues, queueing networks and specific queues in computer/communication systems.

2. Fundamentals in Diffusion Modeling

2.1 What is a diffusion model?

Let $J = \{J(t) : t \geq 0\}$ be a queueing characteristic process being approximated and let $E$ denote its state space. In this section we focus on the one-dimensional case $E \subseteq \mathbb{N} = \{0, 1, 2, \ldots\}$ if $E$ is discrete or $E \subseteq \mathbb{R} = [0, \infty)$ if $E$ is continuous, because the dimension of $J$ is not essential for the fundamentals in diffusion modeling. Let $E^\circ$ denote the interior of $E$, where $E^\circ$ for the discrete case is defined by $E^\circ = \{j \in E ; \{j \pm 1\} \subseteq E\}$. To deal with discontinuity in sample paths of $J$, it is sufficient to assume that $J \in D_E[0, \infty)$. In addition, assume that there exist functions $\beta : E^\circ \times \mathbb{R}_+ \rightarrow \mathbb{R}$ and $\alpha : E^\circ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that

$$E[J(t + \tau) - J(t) | J(t) = j] = \beta(j, t)\tau + o(\tau),$$

$$E[(J(t + \tau) - J(t))^2 | J(t) = j] = \alpha(j, t)\tau + o(\tau),$$

for $\tau > 0$ and $(j, t) \in E^\circ \times \mathbb{R}_+$. These functions are termed the infinitesimal parameters of the process $J$. In particular, $\beta$ is called the infinitesimal mean and $\alpha$ the infinitesimal variance. For convenience, we assume that $\beta$ and $\alpha$ are continuous with respect to $t$.

As a continuous approximation of $J$, we introduce a diffusion process $X = \{X(t) : t \geq 0\} \in C_s [0, \infty)$, where $S \subseteq \mathbb{R}$ is an appropriate state space of $X$. As with $E$, let $S^\circ$ denote the interior of $S$. For $(x, t) \in S^\circ \times \mathbb{R}_+$, we define the infinitesimal parameters of $X$ by

$$b(x, t) = \lim_{\tau \to 0} \frac{1}{\tau} E[X(t + \tau) - X(t) | X(t) = x],$$

$$\alpha(x, t) = \lim_{\tau \to 0} \frac{1}{\tau} E[(X(t + \tau) - X(t))^2 | X(t) = x].$$
Now, suppose that there exists a mapping \( \theta : \mathbb{E} \rightarrow \mathbb{S} \) such that \( \theta(i) \cap \theta(j) = \emptyset \) for all \( (i, j \in \mathbb{E}, i \neq j) \) and \( \bigcup_{j \in \mathbb{E}} \theta(j) = \mathbb{S} \), i.e., \( \theta(j) ; j \in \mathbb{E} \) is a partition of \( \mathbb{S} \).

**Definition 1** \( X \) is called a **continuously parameterized diffusion model** of \( J \), if \( b \) and \( a \) are continuous functions on \( \mathbb{S} \) \( \times \mathbb{R}_{+} \), and if there exists a \( x \in \theta(j) \) such that \( b(x, t) = \beta(j, t) \) and \( a(x, t) = \alpha(j, t) \) for all \( t \geq 0 \).

**Definition 2** \( X \) is called a **piecewise-constantly parameterized diffusion model** of \( J \), if \( b(x, t) = \beta(j, t) \) and \( a(x, t) = \alpha(j, t) \) for all \( x \in \theta(j) \) such that \( j \in \mathbb{E}^{\ast} \) and all \( t \geq 0 \).

Concisely, we call the diffusion models defined above **C model** and **PC model** in order; cf. [8]. If \( J \) is a state-independent process \( (i.e., \beta(j, t) = \beta(t) \) and \( \alpha(j, t) = \alpha(t) \) for \( t \geq 0) \), the parameters \( b \) and \( a \) in the C model should be constant on \( \mathbb{S} \) for a fixed \( t \), so that both models are equivalent. Both models are also equivalent when \( \mathbb{E} = \mathbb{S} \) and \( \theta(j) = \{ j \} \) for all \( j \in \mathbb{E} \), which can be found, for example, in the work load process of a single server queue.

### 2.2 From discrete to continuous

Because Definitions 1 and 2 are no more than general frameworks of diffusion models, there are various modeling issues to be solved for obtaining \( X \). The most elementary modeling issues are:

1. How do we obtain the infinitesimal parameters \( \beta \) and \( \alpha \) (or \( b \) and \( a \))?
2. How do we define the mapping \( \theta \)?
3. How do we reflect boundary behavior of \( J \) into \( X \)?

To illustrate possible solutions of the modeling issues \( 1 - 3 \), we use the standard \( GI/GI/1 \) queueing system specified as follows: Customers are arriving at the system according to a renewal process. That is, if we let \( u_n \) \( (n \geq 1) \) be the interarrival time between \( (n-1) \)-st and \( n \)-th arriving customers, then \( \{u_n; n \geq 1\} \) is a sequence of nonnegative iid random variables. Customers are served in order of arrivals. Let \( v_n \) \( (n \geq 1) \) be the service time of the \( n \)-th customer and assume that \( \{v_n; n \geq 1\} \) is a sequence of non-negative iid random variables being independent of \( \{u_n; n \geq 1\} \). In addition, let \( F(G) \) denote the interarrival-time (service-time) cdf with mean \( \lambda^{-1}(\mu^{-1}) \); let \( \rho = \lambda / \mu \) be the traffic intensity; and let \( c_r^2 \) (\( c_v^2 \)) be the squared coefficient of variation \( \langle \text{scv}, i.e., \text{variance divided by the square of mean} \rangle \) of \( F(G) \).

To solve the modeling issue 1, we need to specify the structure of \( J \) more clearly. It is known that most queueing characteristic processes can be repre-
sented as
\[ J(t) = J(0) + J_+(t) - J_-(t), \quad t \geq 0, \]  \hspace{1cm} (3)

where \( J_+(t) \) (\( J_-(t) \)) denotes a cumulative amount of inputs into (outputs from) a queue in the time interval \((0, t]\). The input and output processes \( J_\pm \) are not only correlated but also strongly affected by boundary behavior at the upper (if exists) and lower boundaries, respectively. However, if \( J \) is conditioned on the interior \( \mathbb{E}^o \), we can see that they behave like mutually independent processes and free from boundary effects. This conditioning will enable us to obtain explicit forms of \( \beta \) and \( \alpha \).

As target processes in the GI/GI/1 queue, we consider (i) \( N = \{N(t); t \geq 0\} \): the number of customers (either waiting or being served) in the system, and (ii) \( V = \{V(t); t \geq 0\} \): the unfinished work load process. Clearly, \( \mathbb{E} = \mathbb{N} \) for \( J = N \) and \( \mathbb{E} = \mathbb{R}_+ \) for \( J = V \). Due to the time- and space-homogeneity of system parameters in the GI/GI/1 queue, the infinitesimal parameters of both processes are constant, i.e., \( \beta(j,t) = \beta \) and \( \alpha(j,t) = \alpha \) for all \( (j,t) \in \mathbb{E}^o \times \mathbb{R}_+ \). The infinitesimal parameters of such a homogeneous process \( J \) can be obtained by its asymptotic expectations

\[ \beta = \lim_{t \to \infty} \frac{1}{t} \mathbb{E}[J(t) \mid \mathcal{F}_t(\mathbb{E}^o)], \] \hspace{1cm} (4)

\[ \alpha = \lim_{t \to \infty} \frac{1}{t} \text{Var}[J(t) \mid \mathcal{F}_t(\mathbb{E}^o)], \]

where \( \mathcal{F}_t(\mathbb{E}^o) = \{J(u) \in \mathbb{E}^o \text{ for } u \in (0, t]\} \); see [55]. From (3) and (4), we immediately have

\[ \beta = \lim_{t \to \infty} \frac{1}{t} \mathbb{E}[J_+(t) \mid \mathcal{F}_t(\mathbb{E}^o)] - \lim_{t \to \infty} \frac{1}{t} \mathbb{E}[J_-(t) \mid \mathcal{F}_t(\mathbb{E}^o)]. \]

Since the processes \( J_\pm \) in (3) can be regarded as being independent under the condition \( \mathcal{F}_t(\mathbb{E}^o) \), the infinitesimal variance is given by

\[ \alpha = \lim_{t \to \infty} \frac{1}{t} \text{Var}[J_+(t) \mid \mathcal{F}_t(\mathbb{E}^o)] + \lim_{t \to \infty} \frac{1}{t} \text{Var}[J_-(t) \mid \mathcal{F}_t(\mathbb{E}^o)]. \]

For the process \( N \), the input process \( N_+ \) can be given by

\[ N_+(t) = \max\{n \mid u_1 + \cdots + u_n < t\}, \quad t > 0; \quad N_+(0) = 0. \]

Hence, we obtain

Toshikazu Kimura
\[ \lim_{t \to \infty} \frac{1}{t} \mathbb{E}[N_+(t) \mid \mathcal{F}_t(\mathbb{N}^+)] = \lim_{t \to \infty} \frac{1}{t} \mathbb{E}[N_-(t)] = \lambda, \]

\[ \lim_{t \to \infty} \frac{1}{t} \text{Var}[N_+(t) \mid \mathcal{F}_t(\mathbb{N}^+)] = \lim_{t \to \infty} \frac{1}{t} \text{Var}[N_-(t)] = \lambda c^2. \]

Under the condition \( \mathcal{F}_t(\mathbb{N}^+) \), the output process \( N_- \) can be identified with

\[ N_-(t) = \max \{ n \mid v_1 + \cdots + v_n < t \}, \quad t > 0; \quad N_-(0) = 0, \]

and hence

\[ \lim_{t \to \infty} \frac{1}{t} \mathbb{E}[N_-(t) \mid \mathcal{F}_t(\mathbb{N}^+)] = \mu, \]

\[ \lim_{t \to \infty} \frac{1}{t} \text{Var}[N_-(t) \mid \mathcal{F}_t(\mathbb{N}^+)] = \mu c^2. \]

Consequently, the infinitesimal parameters of \( N \) are

\[ \beta = \lambda - \mu \quad \text{and} \quad \alpha = \lambda c^2 + \mu c^2. \quad (5) \]

See [29] for a modification of (5).

On the other hand, for the process \( V \), the input and output processes \( V_\pm \) are given by

\[ V_+(t) = \sum_{n=1}^{N_+(t)} v_n, \quad t > 0; \quad V_+(0) = 0, \]

and \( V_-(t) = t \) (a.s.) for \( t \geq 0 \) under the condition \( \mathcal{F}_t(\mathbb{R}^+) \), from which we obtain the infinitesimal parameters of \( V \) as

\[ \beta = \rho - 1 \quad \text{and} \quad \alpha = \frac{\rho}{\mu}(c^2 + c^2). \quad (6) \]

For the process \( N \) in the GI/GI/1 queue, we have a few different solutions of the modeling issue, e.g., \( \theta(j) = [j - 0.5, j + 0.5] (j \in \mathbb{N}) \) with \( S = [0, \infty) \), \( \theta(j) = [j, j + 0.5] (j \in \mathbb{N}) \) with \( S = [0, \infty) \), \( \theta(0) = \{0\} \) and \( \theta(j) = (j - 1, j] (j \in \mathbb{N}^+) \) with \( S = [0, \infty) \), and so on. Of course, for the process \( V \), we have the only solution \( \theta(j) = \{j\} \) with \( S = \mathbb{R}_+ \).

To regulate a diffusion process \( X \) in the interior \( S^+ \), we need an impenetrable barrier at the lower boundary of state space of \( N \) or \( V \). The simplest one is a reflecting boundary described by the HTLT for \( \rho = 1 \); see, e.g., [132]. An alternative solution is a sticky (elementary return) boundary where the process is absorbed for a random time interval. After this time has elapsed, the process jumps into an interior point determined by a probability distribution, and
then the process restarts from that point. Each of these boundaries has its own
defects: An apparent defect of the reflecting boundary is the lack of probabil-
ity mass at boundaries, which fails in modeling the empty/overflow probabili-
ties explicitly. On the other hand, it is pointed out in [12] that a primary defect
of the elementary return boundary is the lack of continuous paths, which fails
in establishing return processes as limits of stable queues being modeled. The
latter indication is, however, mainly based on diffusion limits and that is not a
defect from the standpoint of diffusion modeling.

2.3 From continuous to discrete
The major advantage of using diffusion models is a connection to the
study of certain partial differential equations (PDEs). The connection to PDEs
makes diffusion models computationally and analytically attractive. Let
\( p(x, t | y) \) be the pdf of \( X(t) \) starting from \( X(0) = y \), i.e. \( p(x, t | y)dx = P\{x \leq X(t) < x + dx \mid X(0) = y\} \) for \( x, y \in \mathbb{S} \) and \( t \geq 0 \). Then, it has been known that
\( p(x, t | y) \) satisfies the Kolmogorov forward equation

\[
\frac{\partial p}{\partial t} = \frac{1}{2} \frac{\partial^2}{\partial x^2} \{a(x, t)p(x, t | y)\} - \frac{\partial}{\partial x} \{b(x, t)p(x, t | y)\} \tag{7}
\]

with the initial condition \( p(x, 0 | y) = \delta(x - y) \) for \( x, y \in \mathbb{S} \), where \( \delta(\cdot) \) is the Di-
rac’s delta function. Appropriate boundary conditions are added to (7) accord-
ing to boundary behavior. Note that the PDE (7) should be modified for a
process with elementary return boundaries. Solving the PDE (7) through cer-
tain analytical/numerical methods, we can obtain the pdf of \( X \) in a computa-
tionally efficient way. If \( \mathbb{E} \) is continuous and \( \mathbb{E} = \mathbb{S} \) just as for the work load
case, the pdf can be immediately considered as an approximate pdf of \( J \). How-
ever, if \( \mathbb{E} \) is discrete, we have to solve an additional modeling issue

\( N \). How do we discretize the pdf of \( X \) to obtain an approximate distribution of
\( J \)?

While it seems from the nature of the issue that there is no exact solution of
the issue \( N \), some heuristic solutions have been proposed, according to the
mapping \( \theta \) and the boundary condition. Consider, for example, the case where
\( \theta(j) = [j, j + 1) \in \mathbb{N} \) as well as the reflecting boundary at the origin are used in
a diffusion model for the process \( N \) in the GI/GI/1 queue. Then, a natural
discretization method is defined by

\[
P\{N(t) = j \mid N(0) = i\} \approx \int_{\theta(j)}^{\theta(j+1)} p(x, t | i)dx, \quad i, j \in \mathbb{N}, \quad t \geq 0.
\]
As an index of judging whether or not \( X \) is a “good” approximation for \( J \), we introduce the notion of consistency. We mean by consistency that approximations by a diffusion model agree with available exact results for particular cases. More definitely, we define

**Definition 3** A diffusion model \( X \) for \( J \) is said to be exactly consistent with a particular queueing system, if the approximate distribution of \( J \) agrees with the exact distribution.

**Definition 4** A diffusion model \( X \) for \( J \) is said to be first-order consistent with a particular queueing system, if the approximate mean of \( J \) agrees with the exact mean.

To obtain a consistent diffusion model in the sense of Definitions 3 and 4, we need certain heuristic ideas rather than conventional analyses. That is,

\[ V. \text{ How do we realize wide consistency of } X? \]

should be addressed as another important modeling issue.

3. A Bibliographical Guide to Diffusion Models for Queues

3.1 Single-server queues

The \( GI/GI/1 \) queue and its variants are basic to computer/communication systems. By the use of the diffusion model described in Section 2.2, the standard \( GI/GI/1 \) queue in steady state is analyzed in \([18, 21, 27, 55, 70, 111, 133]\), while its transient analysis in \([14, 16, 45]\) and spectral analysis in \([109, 134]\). Diffusion models are proposed for some variants of the \( GI/GI/1 \) queue, including a bulk queue \([9, 90]\), a queue with repeated calls \([17]\) and the \( GI/GI/1/N \) queue \([29, 30, 36, 64, 70, 75]\); see also \([98]\) for a transportation application. A first passage time in a diffusion process is applied to analyze the probability distributions of busy periods in the \( M/G/1 \) queue \([54]\) and first overflow times in the \( GI/GI/1/N \) queue \([63]\). Diffusion processes with time-dependent infinitesimal parameters are analyzed in \([13, 46, 99, 100, 101]\) as continuous approximation models of non-stationary \( GI/GI/1 \) queues.

3.2 State-dependent queues

Diffusion models for state-dependent queues are just analogous to birth-death queueing models. Some PC models for general state-dependent queues are proposed in \([71, 75, 85, 128]\). Multi-server queues form a typical subclass of state-dependent queues, for which various diffusion models are analyzed in \([20, 50, 51, 52, 67, 73, 115, 138, 143]\) (the \( GI/GI/s \) queue), \([68, 106, 136]\) (the
$GI^x/GI/s$ queue) and $[50, 74, 76, 116, 142]$ (the $GI/GI/s/N$ queue) for their steady-state analysis. Transient behavior of the $M/G/s$ queue is analyzed in [10] by using a PC model [67]. Another typical subclass of state-dependent queues is the $GI/GI/s/\cdot/K$ queue, for which a number of C models are developed [7, 32, 53, 59, 97, 113, 114, 130, 139]. See also [4, 6, 129, 137] for other state-dependent queues and [102, 105] for queues with infinite number of servers.

### 3.3 Queueing networks

A multi-dimensional extension of the process $J$ in Section 2 can be considered as a natural diffusion model for a network of queues. For general queueing networks, see [16, 37, 38, 78, 79, 111, 118]. See also [25, 26, 69, 88, 91, 103, 122] for detailed analyses of two-node queueing networks.

### 3.4 Specific queues in computer/communication systems

In addition to traditional queues stated above, we can observe numerous queueing phenomena in computer/communication systems, according to particular assumptions on the arrival/service processes. Priority queues [1, 48, 61, 123, 126] and queues with service interruptions [4, 22, 23, 48, 60, 61, 96] may be rather classical models of computer systems; see also [3, 35, 91]. In the context of queueing synthesis, diffusion models are proposed for reliability problems [15, 86, 87] and control problems [4, 19, 56, 65, 66, 85, 107, 108]. In communication systems, there exist various diffusion models developed for random access schemes [80, 95, 110, 117, 144], voice/data integrated schemes [92, 141] and bursty inputs in ATM [11, 42, 43, 44, 82, 83, 84, 85, 94, 112, 119, 127]; see also [31, 33, 125, 131] for other communication systems.

Extensive recent statistical analyses of packet traffic in modern high-speed networks have proved that actual packet traffic is more complex or bursty than voice traffic and it has statistical self-similarity or fractal characteristics. This implies that traditional teletraffic models based extensively on Poisson or Markovian assumptions may become inadequate for communication systems with self-similar nature, for which only few theoretical results exist to date. As a direction of future research, diffusion models are expected to deal with self-similarity. A possible approach may be to extend the diffusion models for queues in random environments.

Professor, Hokkaido University
References


Nain, P. (1983), "Queueing systems with service interruptions: an approximation model," Performance Evalu-


Sunaga, T., Kondo, E. and Biswas, S.K. (1978), "An approximation method using continuous models for queue-


