Econ. J. of Hokkaido Univ.,
Vol. 33 (2004), pp. 37～52

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-

*This research was supported in part by the Grant-in-Aid for Scientific Research (No. 13480109) provided by the Japan Society for the Promotion of Science in 2001－2004.

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

$$\begin{aligned} E[J(t+\tau)-J(t)\,|J(t)=j]&=\beta(j,t)\tau+o(\tau),\\ E[\{J(t+\tau)-J(t)\}^2\,|J(t)=j]&=\alpha(j,t)\tau+o(\tau), \end{aligned}\tag{1}$$

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

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

$$\begin{aligned} b(x,t)&=\lim_{\tau\to 0}\frac{1}{\tau}E[X(t+\tau)-X(t)\,|\,X(t)=x],\\ a(x,t)&=\lim_{\tau\to 0}\frac{1}{\tau}E[\{X(t+\tau)-X(t)\}^2\,|\,X(t)=x]. \end{aligned}\tag{2}$$

Now, suppose that there exists a mapping $\theta : \mathbb{E} \to \mathbb{S}$ such that $\theta(i) \cap \theta(j) = \emptyset$ $(i, j \in \mathbb{E}, i \neq j)$ 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 $\mathbb{S}^\circ \times \mathbb{R}_+$, and if there exists a $x \in \theta(j)$ $(j \in \mathbb{E}^\circ)$ 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 $x \in \theta(j)$ $(j \in \mathbb{E}^\circ)$ 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 ($i.e.$, $\beta(j, t) \equiv \beta(t)$ and $\alpha(j, t) \equiv \alpha(t)$ for $t \geq 0$), the parameters b and a in the C model should be constant on $\mathbb{S}^\circ$ 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 :

Ⅰ. How do we obtain the infinitesimal parameters β and α (or b and a) ?
Ⅱ. How do we define the mapping θ ?
Ⅲ. How do we reflect boundary behavior of J into X ?

To illustrate possible solutions of the modeling issues Ⅰ－Ⅲ, 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_a^2 (c_s^2) be the squared coe.cient of variation (scv, *i.e.,* variance divided by the square of mean) of $F(G)$.

To solve the modeling issue I, we need to specify the structure of J more clearly. It is known that most queueing characteristic processes can be repre-

sented as

$$J(t)=J(0)+J_{+}(t)-J_{-}(t), \quad t \geq 0, \tag{3}$$

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 $\mathbb{E}^{\circ}$, 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 β and α.

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.,* $\beta(j, t) \equiv \beta$ and $\alpha(j, t) \equiv \alpha$ for all $(j, t) \in \mathbb{E}^{\circ} \times \mathbb{R}_{+}$. The infinitesimal parameters of such a homogeneous process J can be obtained by its asymptotic expectations

$$\beta=\lim_{t \to \infty} \frac{1}{t} \mathrm{E}\left[J(t) \mid F_t(\mathbb{E}^{\circ})\right], \tag{4}$$

$$\alpha=\lim_{t \to \infty} \frac{1}{t} \mathrm{Var}\left[J(t) \mid F_t(\mathbb{E}^{\circ})\right],$$

where $F_t(\mathbb{E}^{\circ})=\{J(u) \in \mathbb{E}^{\circ}$ for $u \in (0, t]\}$; see [55]. From (3) and (4), we immediately have

$$\beta=\lim_{t \to \infty} \frac{1}{t} \mathrm{E}[J_{+}(t) \mid F_t(\mathbb{E}^{\circ})]-\lim_{t \to \infty} \frac{1}{t} \mathrm{E}[J_{-}(t) \mid F_t(\mathbb{E}^{\circ})].$$

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

$$\alpha=\lim_{t \to \infty} \frac{1}{t} \mathrm{Var}[J_{+}(t) \mid F_t(\mathbb{E}^{\circ})]+\lim_{t \to \infty} \frac{1}{t} \mathrm{Var}[J_{-}(t) \mid F_t(\mathbb{E}^{\circ})].$$

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

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

$$\lim_{t\to\infty}\frac{1}{t}\mathrm{Var}[N_+(t)\mid \mathcal{F}_t(\mathbb{N}^\circ)]=\lim_{t\to\infty}\frac{1}{t}\mathrm{Var}[N_+(t)]=\lambda c_a^2.$$

Under the condition $\mathcal{F}_t(\mathbb{N}^\circ)$, 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

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

$$\lim_{t\to\infty}\frac{1}{t}\mathrm{Var}[N_-(t)\mid \mathcal{F}_t(\mathbb{N}^\circ)]=\mu c_s^2.$$

Consequently, the infinitesimal parameters of N are

$$\beta=\lambda-\mu\quad\text{and}\quad\alpha=\lambda c_a^2+\mu c_s^2. \tag{5}$$

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 $\mathcal{F}_t(\mathbb{R}_+^\circ)$, from which we obtain the infinitesimal parameters of V as

$$\beta=\rho-1\quad\text{and}\quad\alpha=\frac{\rho}{\mu}(c_a^2+c_s^2). \tag{6}$$

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

To regulate a diffusion process X in the interior $\mathbb{S}^\circ$, 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 probability mass at boundaries, which fails in modeling the empty/overflow probabilities 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 $p(x, t \mid y)$ be the pdf of $X(t)$ starting from $X(0)=y$, *i.e.* $p(x, t \mid y)dx = P\{x \leq X(t) < x+dx \mid X(0)=y\}$ for $x, y \in \mathbb{S}$ and $t \geq 0$. Then, it has been known that $p(x, t \mid y)$ satisfies the Kolmogorov forward equation

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

with the initial condition $p(x, 0 \mid y) = \delta(x-y)$ for $x, y \in \mathbb{S}$, where $\delta(\cdot)$ is the Dirac's delta function. Appropriate boundary conditions are added to (7) according to boundary behavior. Note that the PDE (7) should be modified for a process with elementary return boundaries. Solving the PDE (7) through certain analytical/numerical methods, we can obtain the pdf of X in a computationally 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. However, if $\mathbb{E}$ is discrete, we have to solve an additional modeling issue

Ⅳ. 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 Ⅳ, some heuristic solutions have been proposed, according to the mapping θ and the boundary condition. Consider, for example, the case where $\theta(j) = [j, j+1)(j \in \mathbb{N})$ 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

$$P\{N(t) = j \mid N(0) = i\} \approx \int_{\theta(j)} p(x, t \mid i)dx, \quad i, j \in \mathbb{N}, \quad t \geq 0.$$

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. 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

Abdou, E. (1983), "Diffusion approximation for preemptive-resume queueing system," in : *Proceedings of the 10 th International Teletraffic Congress,* Montreal, pp. 4. 4A. 5. 1-7.

Agrawal, S.C. (1985), *Metamodeling : A Study of Approximations in Queueing Models,* MIT Press, Cambridge.

Arató, M. (1975), "Diffusion approximation for multiprogrammed computer systems," *Computer & Mathematics with Applications*, 1, pp. 315-326.

Asgharzadeh, K. and Newell, G.F. (1978), "Optimal dispatching strategies for vehicles having exponentially distributed trip times," *Naval Research Logistics Quarterly,* 25, pp. 489-509.

Bhat, U.N. and Shalaby, M. (1979), "Approximation techniques in the solution of queueing problems," *Naval Research Logistics Quarterly*, 26, pp. 311-326.

Biswas, S.K. and Sunaga, T. (1980), "Diffusion approximation method for multiserver queueing system with balking," *Journal of the Operations Research Society of Japan,* 23, pp. 368-386.

Biswas, S.K. and Sunaga, T. (1981), "Diffusion approximation method for the solution of two-stage cyclic queueing problems," *Memoirs of the Faculty of Engineering, Kyushu University*, 41, pp. 17-32.

Browne, S. and Whitt, W. (1995), "Piecewise-linear diffusion processes," in: *Advances in Queueing : Theory, Methods and Open Problems,* J.H. Dshalalow (ed.), CRC Press, Boca Raton, FL, pp. 463-480.

Chiamsiri, S. and Leonard, M.S. (1981), "Diffusion approximation for bulk queues," *Management Science*, 27, pp. 1188-1199.

Choi, B.D. and Shin, Y.W. (1991), "Transient diffusion approximation for $M/G/m$ system," *Journal of the Operations Research Society of Japan*, 34, pp. 308-328.

Closs, F. (1973), "Packet arrival and buffer statistics in a packet switching node," in: *Proceedings of the 3rd Data Communications Symposium*, pp. 12-17.

Coffman, E.G.,Jr. and Reiman, M.I. (1984), "Diffusion approximations for computer/communication systems," in: *Mathematical Computer Performance and Reliability,* G. Iazeolla, P.J. Courtois and A. Hordijk (eds.), North-Holland, Amsterdam, pp. 33-53.

Crescenzo, A.D. and Nobile, A.G. (1995), "Diffusion approximation to a queueing system with time-dependent arrival and service rates," *Queueing Systems*, 19, pp. 41-62.

Duda, A. (1983), "Transient diffusion approximation for some queueing systems," *Performance Evaluation Review*, 12, pp. 118-128.

Duda, A. (1984), "Performance analysis of the checkpoint-rollback-recovery system via diffusion approximation," in: *Mathematical Computer Performance and Reliability,* G. Iazeolla, P.J. Courtois and A. Hordijk (eds.), North-Holland, Amsterdam, pp. 315-327.

Duda, A. (1986), "Diffusion approximations for time-dependent queueing systems," *IEEE Journal on Selected Areas in Communications,* SAC-4, pp. 905-918.

Falin, G.I. (1984), "Continuous approximation for a single server system with an arbitrary service time under repeated calls," *Engineering Cybernetics*, 22, pp. 66-71.

Filipiak, J. (1983), "Diffusion-equation model of slightly loaded $M/M/1$ queue," *Operations Research Letters*, 2, pp. 134-139.

Filipiak, J. (1985), "Control of arrivals to an exponential congestion system by diffusion," in: *Bridge Between Control Science and Technology,* Pergamon Press, Oxford, Vol. 3, pp. 1229-1234.

Filipiak, J. and Pach, A.R. (1985), "Selection of coefficients for a diffusion-equation model of multi-server queue," in: *Models of Computer System Performance,* E. Gelenbe (ed.), North-Holland, Amsterdam, pp. 135-146.

Filipiak, J. (1986), "Location of a boundary for a diffusion equation of a single server queue," *Performance Evaluation*, 6, pp. 205-218.

Fischer, M.J. (1977), "Analysis and design of loop service systems via a diffusion approximation," *Operations Research*, 25, pp. 269-278.

Fischer, M.J. (1977), "An approximation to queueing systems with interruptions," *Management Science*, 24, pp. 338-344.

Flores, C. (1990), "Diffusion approximations for computer communications networks," in: *Stochastic Analysis of Computer and Communication Systems,* H. Takagi (ed.), North-Holland, Amsterdam, pp. 173-213.

Foschini, G.J. and Salz, J. (1978), "Basic dynamic routing problem and diffusion," *IEEE Transactions on Communications,* COM-26, pp. 320-327.

Foschini, G.J. (1982), "Equilibria for diffusion models of pairs of communicating computers-symmetric case," *IEEE Transactions on Information Theory,* IT-28, pp. 273-284.

Gaver, D.P. (1968), "Diffusion approximations and models for certain congestion problems," *Journal of Applied Probability*, 5, pp. 607-623.

Gaver, D.P. (1971), "Analysis of remote terminal backlogs under heavy demand conditions," *Journal of the Association for Computing Machinery*, 18, pp. 405-415.

Gaver, D.P. and Shedler, G.S. (1973), "Approximate models for processor utilization in multiprogrammed computer systems," *SIAM Journal on Computing*, 2, pp. 183-192.

Gaver, D.P. and Shedler, G.S. (1973), "Processor utilization in multiprogramming systems via diffusion approximation," *Operations Research*, 21, pp. 569-576.

Gaver, D.P. and Lehoczky, J.P. (1976), "Gaussian approximations to service problems: A communication system example," *Journal of Applied Probability*, 13, pp. 768-780.

Gaver, D.P. and Lehoczky, J.P. (1977), "A diffusion approximation solution for a repairman problem with two types of failure," *Management Science*, 24, pp. 71-81.

Gaver, D.P. and Lehoczky, J.P. (1979), "A diffusion approximation model for a communication system allowing message interference," *IEEE Transactions on Communications,* COM-27, pp. 1190-1199.

Gaver, D.P. (1983), "Stochastic modeling : ideas and techniques," in: *Probability Theory and Computer Science,* G. Louchard and G. Latouche (eds.), Academic Press, London, pp. 3-49.

Gaver, D.P. and Jacobs, P.A. (1986), "Processor-shared time-sharing models in heavy trafffic," *SIAM Journal on Computing*, 15, pp. 1085-1100.

Gelenbe, E. (1975), "On approximate computer system models," *Journal of the Association for Computing Machinery*, 22, pp. 261-269.

Gelenbe, E. and Pujolle, G. (1976), "The behaviour of a single queue in a general queueing network," *Acta Informatica*, 7, pp. 123-136.

Gelenbe, E. and Pujolle, G. (1977), "A diffusion model for multiple class queueing networks," in : *Measuring, Modelling and Evaluating Computer Systems,* H. Beilner and E. Gelenbe (eds.), North-Holland, Amsterdam, pp. 189-199.

Gelenbe, E. (1979), "Probabilistic models of computer systems. Part II : diffusion approximations, waiting times and batch arrivals," *Acta Informatica*, 12, pp. 285-303.

Gelenbe, E. and Mitrani, I. (1980), *Analysis and Synthesis of Computer Systems,* Academic Press, London.

Gelenbe, E. and Pujolle, G. (1987), *Introduction to Queueing Networks,* John Wiley & Sons, New York.

Gelenbe, E., Mang, X. and Feng, Y. (1996), "A diffusion cell loss estimate for ATM with multi-class bursty traffic," in : *Performance Modeling and Evaluation of ATM Networks*, Vol. II, D. Kouvatsos (ed.), Chapman and Hall, London, pp. 233-248.

Gelenbe, E., Mang, X. and Önvural, R. (1996), "Diffusion based statistical call admission control in ATM," *Performance Evaluation*, 27, pp. 411-436.

Gelenbe, E., Mang, X. and Önvural, R. (1997), "Bandwidth allocation and call admission control in high-speed networks," *IEEE Communications Magazine*, 35, pp. 122-129.

Giorno, V., Nobile, A.G. and Ricciardi, L.M. (1986), "On some diffusion approximations to queueing systems," *Advances in Applied Probability*, 18, pp. 991-1014.

Giorno, V., Nobile, A.G. and Ricciardi, L.M. (1987), "On some time-non-homogeneous diffusion approximations to queueing systems," *Advances in Applied Probability*, 19, pp. 974-994.

Glynn, P.W. (1990), "Diffusion approximations," in: *Stochastic Models,* Handbooks in Operations Research and Management Science, Vol. 2, D.P. Heyman and M.J. Sobel (eds.), North-Holland, Amsterdam, pp. 145-198.

Gorov, G.V., Kogan, Ya.A. and Paradizov, N.V. (1985), "Jump diffusion approximation in single-server systems with interruption of service and variable rate of arrival of calls," *Automation and Remote Control*, 46, pp. 716-722.

Gross, D. and Harris, C.M. (1985), *Fundamentals of Queueing Theory*, 2nd ed., John Wiley & Sons, New York.

Halachimi, B. and Franta, W.R. (1977), "Diffusion approximate solution to the $G/G/k$ queueing system," *Computers & Operations Research*, 4, pp. 37-46.

Halachimi, B. (1978), "On approximating the conditional wait process of general multiserver queueing systems," *Computers & Operations Research*, 5, pp. 101-107.

Halachmi, B. and Franta, W.R. (1978), "Diffusion approximation to the multi-server queue," *Management Science*, 24, pp. 522-529.

Haryono and Sivazlian, B.D. (1985), "Analysis of the machine repair problem: A diffusion process approach," *Mathematics and Computers in Simulation*, 27, pp. 339-364.

Heyman, D.P. (1974), "An approximation for the busy period of the $M/G/1$ queue using a diffusion model," *Journal of Applied Probability*, 11, pp. 159-169.

Heyman, D.P. (1975), "A diffusion model approximation for the $GI/G/1$ queue in heavy traffic," *Bell system Technical Journal*, 54, pp. 1637-1646.

Heyman, D.P. (1977), "Optimal disposal policies for a single-item inventory system with returns," *Naval Research Logistics Quarterly*, 24, pp. 385-405.

Heyman, D.P. and Sobel, M.J. (1982), *Stochastic Models in Operations Research, Volume I : Stochastic Processes and Operating Characteristics,* McGraw-Hill, New York.

Hoshiai, T., Takahashi, H. and Akimaru, H. (1987), "Modified diffusion approximation of queueing systems," [in Japanese] *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* J70-B, pp. 279-287.

Jain, M. and Sharma, G.C. (1986), "Diffusion approximation for the $GI/G/r$ machine interference problem with spare machines," *Indian Journal of Technology*, 24, pp. 568-572.

Kimura, G. and Takahashi, Y. (1986), "Diffusion approximation for a token ring system with nonexhaustive service," *IEEE Journal on Selected Areas in Communications,* SAC-4, pp. 794-801.

Kimura, G. and Takahashi, Y. (1987), "An approximation for a token ring system with priority classes of messages," *Journal of Information Processing*, 10, pp. 86-91.

Kimura, G. and Takahashi, Y. (1987), "Traffic analysis for token ring systems with limited service," [in Japanese] *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* J70-B, pp. 327-336.

Kimura, T., Ohno, K. and Mine, H. (1979), "Diffusion approximation for $GI/G/1$ queueing systems with finite capacity : I — the first overflow time," *Journal of the Operations Research Society of Japan*, 22, pp. 41-68.

Kimura, T., Ohno, K. and Mine, H. (1979), "Diffusion approximation for $GI/G/1$ queueing systems with finite capacity : II — the stationary behaviour," *Journal of the Operations Research Society of Japan*, 22, pp. 301-320.

Kimura, T., Ohno, K. and Mine, H. (1980), "Approximate analysis of optimal operating policies for a $GI/G/1$ queueing system," *Memoirs of the Faculty of Engineering, Kyoto University*, 42, pp. 377-390.

Kimura, T. (1981), "Optimal control of an $M/G/1$ queuing system with removable server via diffusion approximation," *European Journal of Operational Research*, 8, pp. 390-398.

Kimura, T. (1983), "Diffusion approximation for an $M/G/m$ queue," *Operations Research*, 31, pp. 304-321.

Kimura, T. and Ohsone, T. (1984), "A diffusion approximation for an $M/G/m$ queue with group arrivals," *Management Science*, 30, pp. 381-388.

Kimura, T. (1985), "Diffusion approximation for a tandem queue with blocking," *Optimization,* 16, pp. 137-154.

Kimura, T. (1986), "Refining diffusion approximations for $GI/G/1$ queues: A tight discretization method," in: *Teletraffic Issues in an Advanced Information Society,* M. Akiyama (ed.), Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1, pp. 3. 1A. 2. pp. 1-7.

Kimura, T. (1987), "A unifying diffusion model for state-dependent queues," *Optimization,* 18, pp. 265-283.

Kimura, T. (1993), "A bibliography of research on heavy traffic limit theorems for queues," *Economic Journal of Hokkaido University*, 22, pp. 167-179.

Kimura, T. (1995), "An $M/M/s$-consistent diffusion model for the $GI/G/s$ queue," *Queueing Systems*, 19, pp. 377-397.

Kimura, T. (2001), "Refining diffusion models for state-dependent queues," in: *Operations Research/Management Science at Work: Applying Theory in the Area Pacific Region,* International Series in Operations Research & Management Science, E. Kozan and A. Ohuchi (eds.), pp. 395-404, Kluwer, Dordrecht.

Kimura, T. (2002), "Diffusion approximations for queues with Markovian bases," *Annals of Operations Research*, 113, pp. 27-40.

Kimura, T. (2003), "A consistent diffusion approximation for finite-capacity multi-server queues," *Mathematical & Computer Modelling*, 38, pp. 1313-1324.

Kleinrock, L. (1976), *Queueing Systems, Vol. II : Computer Applications,* John Wiley & Sons, New York.

Kobayashi, H. (1974), "Application of the diffusion approximation to queueing networks. I. equilibrium queue distributions," *Journal of the Association for Computing Machinery,* 21, pp. 316-328.

Kobayashi, H. (1974), "Application of the diffusion approximation to queueing networks. II. nonequilibrium distributions and applications to computer modeling," *Journal of the Association for Computing Machinery,* 21, pp. 459-469.

Kobayashi, H., Onozato, Y. and Huynh, D. (1977), "An approximate method for design and analysis of an ALOHA system," *IEEE Transactions on Communications,* COM-25, pp. 148-157.

Kobayashi, H. (1983), "Stochastic modeling: Queueing models," in: *Probability Theory and Computer Science,* G. Louchard and G. Latouche (eds.), Academic Press, London, pp. 51-121.

Kobayashi, H. and Ren, Q. (1993), "A diffusion approximation analysis of an ATM statistical multiplexer with multiple types of traffic. Part I : equilibrium state solutions," in: *Proceedings of IEEE International Conference on Communications,* Geneva, pp. 1047-1053.

Kobayashi, H. and Ren, Q. (1993), "Transient behavior of a statistical multiplexer for multiple types of traffic," in : *Proceedings of the Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management,* S. Osaki and D.N. Pra Murthy (eds.), World Scientific, Singapore, pp. 293-302.

Kobayashi, K. and Takahashi, Y. (1995), "Steady-state analysis of ATM multiplexer with variable input rate through diffusion approximation," *Performance Evaluation*, 23, pp. 145-162.

Kogan, Ya.A. and Litvin, V.G. (1979), "Piecewise diffusion approximations for queuing problems with heterogeneous arrivals and service," *Problems of Control and Information Theory*, 8, pp. 433-443.

Kogan, Ya.A., Litvin, V.G. and Lysyakov, P.K. (1987), "Approximate analysis of dynamic methods of data base reorganization in computer networks," *Automatic Control and Computer Sciences*, 21, pp. 47-51.

Korolyuk, V.S. and Vavrikovich, L.V. (1988), "Diffusion approximation of a redundant Markov system with repair," *Cybernetics*, 24, pp. 655-659.

Kruglikov, V.K. and Tarasov, V.N. (1983), "Analysis and calculation of queuing networks using the two-dimensional diffusion approximation," *Automation and Remote Control*, 44, pp. 1026-1034.

Kuribayashi, S. and Takahashi, Y. (1988), "Diffusion approximation for a multi-server nonpreemptive priority queue with batch poisson inputs and general service time," [in Japanese] *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* J72-B, pp. 121-128.

Lee, H.W., Yoon, S.H. and Lee, S.S. (1996), "A continuous approximation for batch arrival queues with threshold," *Computers & Operations Research*, 23, pp. 299-308.

Lehoczky, J.P. and Gaver, D.P. (1981), "Models for time-sharing computer systems with heterogeneous users," *Operations Research*, 29, pp. 550-566.

Lehoczky, J.P. and Gaver, D.P. (1981), "Diffusion approximations for the cooperative service of voice and data messages," *Journal of Applied Probability*, 18, pp. 660-671.

Lemoine, A.J. (1978), "Networks of queues — a survey of weak convergence results," *Management Science*, 24, pp. 1175-1193.

Mang, X. and Gelenbe, E. (1996), "Call admission control in ATM using the diffusion model," in: *Proceedings of IEEE GLOBECOM '96*, London, pp. 1700-1704.

Miyahara, H., Matsumoto, T. and Takashima, K. (1986), "Transient analysis of random access method via diffusion approximation," in: *Modelling Techniques and Tools for Performance Analysis '85*, N. Abu El Ata (ed.), North-Holland, Amsterdam, pp. 83-100.

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

ation, 3, pp. 123-129.

Naor, P. (1957), "Normal approximation to machine interference with many repairmen," *Journal of the Royal Statistical Society, Series* B, 19, pp. 334-341.

Newell, G.F. (1965), "Approximation methods for queues with applications to the fixed cycle traffic light," *SIAM Review*, 7, pp. 223-240.

Newell, G.F. (1968), "Queues with time-dependent arrival rates. I — the transition through saturation," *Journal of Applied Probability*, 5, pp. 436-451.

Newell, G.F. (1968), "Queues with time-dependent arrival rates. II — the maximum queue and the return to equilibrium," *Journal of Applied Probability*, 5, pp. 579-590.

Newell, G.F. (1968), "Queues with time-dependent arrival rates. III — a mild rush hour," *Journal of Applied Probability*, 5, pp. 591-606.

Newell, G.F. (1973), *Approximate Stochastic Behavior of n-Server Service Systems with Large n*, Lecture Notes in Economics and Mathematical Systems, 87, Springer-Verlag, Berlin.

Newell, G.F. (1979), *Approximate Behavior of Tandem Queues,* Lecture Notes in Economics and Mathematical Systems, 171, Springer-Verlag, Berlin.

Newell, G.F. (1982), *Applications of Queueing Theory*, 2nd ed., Chapman & Hall, London.

Newell, G.F. (1984), *The $M/M/\infty$ Service System with Ranked Servers in Heavy Traffic,* Lecture Notes in Economics and Mathematical Systems, 231, Springer-Verlag, Berlin.

Ohsone, T. (1984), "Diffusion approximation for a $GI/G/m$ queue with group arrivals," *Journal of the Operations Research Society of Japan*, 27, pp. 78-96.

Okamura, H., Dohi, T. and Osaki, S. (1996), "Optimal timing strategies for controlled $M/G/1$ queueing systems via diffusion approximation approach," in: *Proceedings of the 28th ISCIE International Symposium on Stochastic Systems Theory and Its Applications,* Kyoto, pp. 143-148.

Okamura, H., Dohi, T. and Osaki, S. (2000), "Optimal policies for a controlled queueing system with removable server under a random vacation circumstance," *Computers & Mathematics with Applications*, 39, pp. 215-227.

Ott, T.J. (1977), "The stable $M/G/1$ queue in heavy traffic and its covariance function," *Advances in Applied Probability*, 9, pp. 169-186.

Pronios, N.B. and Polydoros, A. (1988), "Approximate methods for analyzing slotted random-access spread-spectrum networks," in: *Proceedings of IEEE INFOCOM '88,* New Orleans, pp. 916-923.

Reiser, M. and Kobayashi, H. (1974), "Accuracy of the diffusion approximation for some queuing systems," *IBM Journal of Research and Development*, 18, pp. 110-124.

Ren, Q. and Kobayashi, H. (1993), "A diffusion approximation analysis of an ATM statistical multiplexer with multiple types of traffic. Part II: time-dependent solutions," in: *Proceedings of IEEE GLOBECOM '93*, Houston, pp. 772-777.

Sivazlian, B.D. and Wang, K.-H. (1989), "System characteristics and economic analysis of the $G/G/R$ machine repair problem with warm standbys using diffusion approximation," *Microelectronics and Reliability*, 29, pp. 829-848.

Sivazlian, B.D. and Wang, K.-H. (1990), "Diffusion approximation to the $G/G/R$ machine repair problem with warm standby spares," *Naval Research Logistics*, 37, pp. 753-772.

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

ing problems," *Journal of the Operations Research Society of Japan*, 21, pp. 29-44.

Sunaga, T., Biswas, S.K. and Nishida, N. (1982), "An approximation method using continuous models for queueing problems. II (multi-server finite queue)," *Journal of the Operations Research Society of Japan*, 25, pp. 113-128.

Takagi, H. and Kleinrock, L. (1984), "Diffusion process approximation for the queueing delay in contention packet broadcasting systems," in: *Performance of Computer-Communication Systems,* H. Rudin and W. Bux (eds.), Elsevier Science Publishers B.V. (North-Holland), Amsterdam, pp. 111-124.

Takahashi, H. and Akimaru, H. (1986), "Analysis of queueing systems via multidimensional elementary return process," in: *Teletraffic Issues in an Advanced Information Society,* M. Akiyama (ed.), Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1, pp. 3.3A.1.1-7.

Takahashi, H. and Akimaru, H. (1986), "Diffusion model for queues in randomly varying environment," *Transactions of the Institute of Electronics and Communication Engineers of Japan,* E69, pp. 13-20.

Takahashi, H. and Akimaru, H. (1986), "Diffusion approximation of first passage time and its application in queueing systems," [in Japanese] *Transactions of the Institute of Electronics and Communication Engineers of Japan,* J69-B, pp. 1046-1053.

Takahashi, H. and Wang, L.-S. (1989), "An approximate analysis of a queueing system with Markov-modulated arrivals," [in Japanese] *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* J72-B-I, pp. 823-830.

Takahashi, H. (1990), "Coupled processor: a second order continuous-state-space model," *Probability in the Engineering and Information Sciences*, 4, pp. 277-298.

Takahashi, Y. and Kimura, G. (1986), "Mean-delay approximation for a single-server priority queue with general low-priority arrival process," *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* E69, pp. 1173-1179.

Takahashi, Y. (1986), "Diffusion approximation for the single server system with batch arrivals of multi-class calls," [in Japanese] *Transactions of the Institute of Electronics and Communication Engineers of Japan,* J69-A, pp. 317-324.

Takahashi, Y. (1987), "Queueing analysis methods for mixed loss and delay systems: Exact and diffusion approximation results," *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* E70, pp. 1195-1202.

Takahashi, Y. (1989), "Mean-delay approximation for a single-server priority queue with batch arrivals of two classes," *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* E72, pp. 29-36.

Takahashi, Y. and Sumita, U. (1989), "Diffusion approximation for multi-server queue with Markov modulated Poisson process (MMPP) input," in: *Proceedings of TIMS XXIX,* Osaka, MD06.2.

Tien, T.M. and Jansen, U. (1984), "Remarks on the method of diffusion approximation and some further heuristic formulas," *Electronische Informationsverarbeitung und Kybernetik*, 20, pp. 382-394.

Varshney, K., Jain, M. and Sharma, G.C. (1987), "Diffusion approximation for $G/G/m$ queueing system with discouragement," *Journal of the Indian Statistical Association,* 25, pp. 91-96.

Wang, K.-H. and Sivazlian, B.D. (1990), "Comparative analysis for the $G/G/R$ machine repair problem" *Computers & Industrial Engineering*, 18, pp. 511-520.

Watanabe, M., Miyahara, H. and Hasegawa, T. (1978), "Diffusion approximation for data packet queues in integrated switched tandem network," *Transactions of the Institute of Electronics and Communication Engineers of Japan,* E61, pp. 898-899.

Whitt, W. (1974), "Heavy traffic limit theorems for queues: a survey," in: *Mathematical Methods in Queueing Theory,* Lecture Notes in Economics and Mathematical Systems, 98, Springer-Verlag, New York, pp. 307-350.

Whitt, W. (1982),"Refining diffusion approximations for queues," *Operations Research Letters*, 1, pp. 165-169.

Woodside, C.M., Pagurek, B. and Newell, G.F. (1980), "A diffusion approximation for correlation in queues," *Journal of Applied Probability*, 17, pp. 1033-1047.

Wu, J.-S. (1990), "Refining the diffusion-approximation for the $G/G/c$ queue," *Computers & Mathematics with Applications*, 20, pp. 31-36.

Wu, J.-S. and Wang, J.-Y. (1994), "Refining the diffusion-approximation for the $GI^{x}/G/c$ queue," *Journal of the Operational Research Society*, 45, pp. 431-439.

Yamashiro, M. and Yuasa, Y. (1986), "Simple queuing system in the case where a server joins the system when the queue length attains a certain length," *Bulletin of JSME,* 29, pp. 1952-1956.

Yamashiro, M. and Yuasa, Y. (1987), "Diffusion approximation of a queueing system with many heterogeneous servers," *Memoirs of Ashikaga Institute of Technology*, 13, pp. 17-24.

Yamashiro, M., Kaneda, T., Koyama, T. and Yuasa, Y. (1987), "Analysis of a system where an operative handles plural machine tools," *Memoirs of Ashikaga Institute of Technology*, 13, pp. 25-30.

Yanagiya, M. and Takahashi, H. (1989), "Diffusion approximation for queueing systems with nonexaustive cyclic service," [in Japanese] *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* J72-B-I, pp. 623-631.

Yao, D.D.W. and Leon-Garcia, A. (1982), "On the behaviour of the data queue in a voice/data integrated system," in : *Proceedings of the 20th Annual Allerton Conference on Communication, Control and Computing,* Monticello, pp. 977-984.

Yao, D.D.W. and Buzacott, J.A. (1985), "Queueing models for a flexible machining station. I. the diffusion approximation," *European Journal of Operational Research*, 19, pp. 233-240.

Yao, D.D.W. (1985), "Refining the diffusion approximation for the $M/G/m$ queue," *Operations Research*, 33, pp. 1266-1277.

Yokohira, T., Nishida, T. and Miyahara, H. (1986), "Analysis of dynamic behavior in ppersistent CSMA/CD using cusp catastrophe," *Computer Networks and ISDN Systems,* 12, pp. 277-289.