Journal of the Operations Research
Society of Japan
Vol. 32, No. 1, March 1989



# ON THE RELAXATION TIME FOR SINGLE SERVER QUEUES

Masaaki Kijima
*Tokyo Institute of Technology*

(Received April 28, 1988; Revised October 13, 1988)

*Abstract* This paper gives a natural definition of the relaxation time for general single server queues. First, we describe a GI/GI/1 queue as a limit of GI/GPH/1 queues. Each GI/GPH/1 queue is transferred to some equivalent bulk arrival $GI^X/M/1$ queue, which is formulated by a spatially homogeneous Markov chain with reflecting barrior at zero. An upper bound, which is easy to calculate, of the relaxation time of the Markov chain is derived. It will be shown that the relaxation time of the GI/GI/1 queue, defined as a limit of the relaxation times of GI/GPH/1 queues, has a particularly simple upper bound. Some particular cases are finally treated, where the upper bound obtained is shown to be tight for M/M/1 and M/D/1 queues.

**Key words.** Single server queue, Relaxation time, Spatially homogeneous Markov chain

## 1 Introduction

A number of papers have devoted to obtaining stationary distributions of the number of customers, the waiting time, etc. in queuing systems, or approximations of them. These results are, however, meaningful only after the queueing process under consideration reaches to a statistical equilibrium. The transient analysis of the queueing process is in general desperate. Hence, from engineering points of view, the desire of having some measures indicating the time needed to wait for the use of stationary results arises. The term "relaxation time" is often used to represent such measures. In the preceding paper [6], we studied a measure of dependence and the relaxation time of finite state Markov chains. This paper discusses the relaxation time of general single server queues based on the convergence rate of spatially homogeneous Markov chains with reflecting barrior at zero.

Not many papers have treated the relaxation time of queueing systems, despite of the importance of the study in practice. Also, the existing literature seems restricted, to the author's best knowledge, to Poisson arrival queues. See, e.g. Keilson, Machihara and

Sumita [4] for M/G/1 queues, Keilson and Servi [5] for M/G/1 vacation queues, and Blanc [1] for Markovian Queues in tandem. The purpose of this paper is to define the relaxation time for general single server queues and to derive an upper bound of it, which is easy to calculate.

The main idea of how to define the relaxation time for a GI/GI/1 queue (denoted by $T_{REL}(GI/GI/1)$) is as follows. First we describe a GI/GI/1 queue as a limit of GI/GPH/1 queues (GPH indicates the service time distributions to be a generalized phase type, see Shanthikumar [10]). Each GI/GPH/1 queue is transferred to a bulk arrival $\mathrm{GI}^X/\mathrm{M}/1$ queue that behaves, in some sense, statistically as the GI/GPH/1 queue. The bulk arrival queue is represented as a spatially homogeneous discrete time Markov chain with reflecting barrior at zero. The relaxation time of the Markov chain is defined, in a natural manner, through the convergence rate of the transition probabilities. The relaxation time can be considered as $T_{REL}(GI/GPH/1)$. Then $T_{REL}(GI/GI/1)$ is defined as a limit of $T_{REL}(GI/GPH/1)$. Note that our relaxation time is the number of arrivals for which one should wait to use stationary results. When the expected interarrival time is $\tau$ (note that the arrival process is of renewal type), one may consider that, in average, the time of length $\tau T_{REL}(GI/GI/1)$ is needed for the queueing process to reach an equilibrium.

In the next section, we give a definition of $T_{REL}(GI/GI/1)$ along the line described above. An upper bound of $T_{REL}(GI/GI/1)$ is derived in Section 3. Section 4 treats some particular cases. Some remarks regarding the tightness of the upper bound and the relation to an existing result are also stated.

## 2 The Relaxation Time of GI/GI/1 Queues

Consider a GI/GI/1 queue having the interarrival time distribution function (DF) $A(x)$ and the service time DF $B(x)$. The Laplace-Stieltjes transform (LST) of a DF is denoted by the corresponding Greek letter of lower case (e.g. $\alpha(s) = \int_0^\infty e^{-st}dA(t)$). It is assumed throughout the paper that both $\alpha(s)$ and $\beta(s)$ have over-convergence beyond the imaginary axis and the queue is stable, i.e. $\alpha(s)$ and $\beta(s)$ exist in $Re(s) > \delta_1$ and $Re(s) > \delta_2$, respectively, for some $\delta_1$, $\delta_2 < 0$ and $\int_0^\infty t dA(t) > \int_0^\infty t dB(t)$.

Suppose first that $B(x)$ is GPH, i.e. $\beta(s) = \sum_{n=0}^\infty g_n(\frac{\mu}{s+\mu})^n$ where $\mu > 0$, $g_n \geq 0$ and $\sum_{n=0}^\infty g_n = 1$. Let $X$ be a random variable distributed by $(g_n)$. Consider then a bulk arrival queue $\mathrm{GI}^X/\mathrm{M}/1$, where the interarrival time DF is $A(x)$ and the service rate is $\mu$. Let $N_k$ be the number of customers just before the arrival of the $k$th bulk in the $\mathrm{GI}^X/\mathrm{M}/1$ queue. It is well known that $N_k$ forms a Markov chain in discrete time on the state space $\mathcal{N} = \{0, 1, 2, \cdots\}$. The Markov chain converges to a stationary distribution asymptotically at a geometric rate if the bulk queue is stable. That is, for some $0 < r < 1$,

$$\Pr[N_k \in A] - \pi(A) \sim O(r^k) \quad \text{as} \quad k \to \infty, \tag{2.1}$$

for any subset $A$ of $\mathcal{N}$. The relaxation time is usually defined, by using the convergence rate, as

$$T_{REL}(GI^X/M/1) = \frac{1}{1-r}. \tag{2.2}$$

Note that, since the states of the $\mathrm{GI}^X/\mathrm{M}/1$ queue are more refined than those of the GI/GPH/1, it should hold that $T_{REL}(GI^X/M/1) \geq T_{REL}(GI/GPH/1)$. On the other hand, by a particular choice of $A = \{0\}$ in (2.1), $\Pr[N_k = 0] - \pi(\{0\}) \sim O(r^k)$ as $k \to \infty$. But, $\Pr[N_k = 0]$ is equal to the probability that the $k$th arriving customer finds the GI/GPH/1 system empty. Thus, a natural definition of $T_{REL}(GI/GPH/1)$ may be to use $T_{REL}(GI^X/M/1)$ in (2.2).

The convergence rate $r$ in (2.1) coincides with the convergence norm of a substochastic matrix obtained from the transition probability matrix governing $N_k$. To see this, let

$$a_n = \int_0^\infty \frac{(\mu t)^n}{n!} e^{-\mu t} dA(t), \quad n \geq 0, \tag{2.3}$$

and let

$$p_n = \sum_{k=0}^{\infty} g_{n+k} a_k, \quad -\infty < n < \infty. \tag{2.4}$$

Thus, $a_n$ is the probability that the potential number of departures from the $\mathrm{GI}^X/\mathrm{M}/1$ queue during an interarrival time is $n$, and $p_n$ is the probability that the difference of the potential numbers of customers in the system at two successive arrivals of bulks is $n$. Denote $\bar{P}_n = \sum_{k=-\infty}^{n} p_k$ and define

$$\boldsymbol{P}_0 = \begin{pmatrix} \bar{P}_0 & p_1 & p_2 & \cdots \\ \bar{P}_{-1} & p_0 & p_1 & \cdots \\ \bar{P}_{-2} & p_{-1} & p_0 & \cdots \\ \cdots & \cdots & \cdots & \ddots \end{pmatrix}. \tag{2.5}$$

It is easy to see that $\boldsymbol{P}_0$ is the transition probability matrix governing $N_k$.

Let $\boldsymbol{P}$ be the submatrix of $\boldsymbol{P}_0$ obtained by deleting the first row and the first column, i.e.

$$\boldsymbol{P} = \begin{pmatrix} p_0 & p_1 & p_2 & \cdots \\ p_{-1} & p_0 & p_1 & \cdots \\ p_{-2} & p_{-1} & p_0 & \cdots \\ \cdots & \cdots & \cdots & \ddots \end{pmatrix}. \tag{2.6}$$

Define the matrix $\boldsymbol{T} = (t_{ij})$ as $t_{ij} = 1$ for $i \geq j$ and $t_{ij} = 0$ otherwise. The inverse matrix $\boldsymbol{T}^{-1} = (s_{ij})$ is given by $s_{ij} = 1$ for $i = j$, $s_{ij} = -1$ for $i = j+1$ and $s_{ij} = 0$ otherwise. It is not hard to see that

$$T^{-1}P_0^kT = (T^{-1}P_0T)^k = \begin{pmatrix} 1 & p_k^T \\ o & P^k \end{pmatrix}$$

for some vector $p_k$. Here $o = (0, 0, \cdots)^T$ and $T$ denotes the transpose. If $N_\infty = \lim_{k\to\infty} N_k$ exists, one sees that

$$(T^{-1}P_0T)^k \quad \to \quad \begin{pmatrix} 1 & p_\infty^T \\ o & O \end{pmatrix} \quad \text{as} \quad k \to \infty, \tag{2.7}$$

where $O$ is the zero matrix. This means that the convergence rate $r$ in (2.1) equals the rate of $P^k$ converging to $O$. This rate is given by the reciprocal of the convergence radius $R = \sup\{z : \sum_{n=0}^\infty z^n P^n < \infty\}$. Thus,

$$r = \frac{1}{R}; \quad R = \sup\{z : \sum_{n=0}^{\infty} z^n P^n < \infty\}, \tag{2.8}$$

as claimed. Note that, for a countable substochastic matrix, clearly $R \geq 1$. Further, if $P$ is finite then $r$ in (2.8) is the Perron-Frobenius eigenvalue of $P$. For this reason, the quantity $1/R$ is usually called the convergence norm of $P$ (see e.g. Seneta [9], Chapter 6).

Let $N_B$ be the number of arriving bulks in a busy period of the GI$^X$/M/1 and let $N$ be the number of customers served during a busy period of the GI/GPH/1. It is readily seen that $N = N_B$ in law. Let $f_n = \Pr[N = n]$, $n \geq 1$, and define $\phi(z) = \sum_{n=1}^\infty f_n z^{n-1}$. Starting with $u_1 = (p_1, p_2, \cdots)^T$, we generate $u_n$ successively by

$$u_{n+1}^T = u_n^T P = u_1^T P^n, \quad n \geq 1. \tag{2.9}$$

Then, $f_1 = \bar{P}_0$ and $f_{n+1} = u_n^T r$, $n \geq 1$, where $r = (\bar{P}_{-1}, \bar{P}_{-2}, \cdots)^T$. Denote $\mathbf{1} = (1, 1, \cdots)^T$ and let $e_n = u_n^T \mathbf{1}$. It is easily seen from (2.5), (2.6) and (2.9) that

$$e_n = u_1^T P^{n-1}\mathbf{1}; \quad e_{n+1} + f_{n+1} = e_n, \quad n \geq 1. \tag{2.10}$$

Thus, $f_{n+1} = e_n - e_{n+1}$, $n \geq 1$. It then follows that, after a little algebra,

$$\phi(z) = 1 - (1 - z)u_1^T \left(\sum_{n=0}^{\infty} z^n P^n\right) \mathbf{1}. \tag{2.11}$$

Therefore,

$$r = \frac{1}{R}; \quad R = \sup\{z : \phi(z) < \infty\}. \tag{2.12}$$

This is another interpretation of the convergence rate $r$ in (2.1).

Any distribution can be arbitrarily approximated by a GPH, in the sense that

$$\bar{H}(t) = \lim_{\mu\to\infty} \sum_{n=0}^{\infty} \bar{H}(n/\mu) e^{-\mu t} \frac{(\mu t)^n}{n!}$$

at continuity points of $\bar{H}(t) = 1 - H(t)$ (see e.g. Esary, Marshall and Proschan [3]). This means that

$$\beta(s) = \lim_{\mu\to\infty} \sum_{n=0}^{\infty} g_n(\mu) \left(\frac{\mu}{s+\mu}\right)^n \tag{2.13}$$

for each $Re(s) \geq 0$, where $g_n(\mu) = \bar{H}(\frac{n-1}{\mu}) - \bar{H}(\frac{n}{\mu})$, $n \geq 1$, see Shanthikumar [10]. It should be noted that the regularity structure of $\beta(s)$ obtained in (2.13) can be extended by the principle of analytical continuation [8]. Hence, in fact, $\beta(s)$ in (2.13) exists in $Re(s) > \delta_2$, $\delta_2 < 0$.

For each $\mu > 0$, the convergence rate $r(\mu)$ is obtained through either (2.8) or (2.12). Thus, by letting $r^* = \limsup_{\mu\to\infty} r(\mu)$, one may define the relaxation time of the GI/GI/1 queue as

$$T_{REL}(GI/GI/1) = \frac{1}{1-r^*}. \tag{2.14}$$

From mathematical points of view, it may be of interest to study the behavior of $r(\mu)$ as $\mu$ varies. However, as is easily realized, the convergence norm $r(\mu)$ of $\boldsymbol{P}(\mu)$ is hard to calculate. Hence, of desire from practice is to obtain upper bounds of $r(\mu)$ and $r^*$, which are easy to calculate. In the next section, we skip mathematical investigations of the behavior of $r(\mu)$ and focus on deriving such an upper bound.

## 3 An Upper Bound of the Relaxation Time

For a non-negative matix $\boldsymbol{Q}$, $\boldsymbol{x} \geq \mathbf{o}$ $(\neq \mathbf{o})$ is called a $\gamma$-subinvariant measure of $\boldsymbol{Q}$ for $\gamma > 0$ if

$$\gamma \boldsymbol{x}^T \geq \boldsymbol{x}^T \boldsymbol{Q} \tag{3.1}$$

(note the difference of the definition from the one in Seneta [9]). It is known that no $\gamma$-subinvariant measure can exist for $\gamma < r(\boldsymbol{Q})$ where $r(\boldsymbol{Q})$ is the convergence norm of $\boldsymbol{Q}$ (see Theorem 6.3 in p.205 of Seneta [9]). Thus, if one finds a $\gamma$-subinvariant measure of $\boldsymbol{P}$ in (2.6), the $\gamma$ must satisfy the relation $\gamma \geq r$, where $r$ is given in (2.8). $\gamma$ must be less than 1, otherwise the relaxation time in (2.2) becomes meaningless (note that 1 is an obvious upper bound of $r$).

For the GI/GPH/1 queue considered in Section 2, define the generating functions $P(z) = \sum_{n=-\infty}^{\infty} p_n z^n$, $A(z) = \sum_{n=0}^{\infty} a_n z^n$ and $G(z) = \sum_{n=0}^{\infty} g_n z^n$. It is easy to see that $P(z) = A(z^{-1})G(z)$. If the queue is stable, then $P(z)$ exists at least in $|z-1| < \delta$ for some $\delta > 0$ and $P'(1) < 0$ (the prime denotes the derivative). This together with the continuity of $P(z)$ implies that there exists $\sigma$ such that $0 < \sigma < 1$ and $P(\sigma^{-1}) < 1$. Let $\boldsymbol{x} = (1, \sigma, \sigma^2, \cdots)^T$. For this $\boldsymbol{x}$, the $(n+1)$th element of $\boldsymbol{x}^T\boldsymbol{P}$ is

$$p_n + p_{n-1}\sigma + \cdots + p_0\sigma^n + p_{-1}\sigma^{n+1} + \cdots \leq \sigma^n P(\sigma^{-1}), \tag{3.2}$$

for any $n \geq 0$. Thus, by choosing $\gamma = P(\sigma^{-1}) < 1$, one concludes that $(1-\sigma)\boldsymbol{x}$ is a $\gamma$-subinvariant measure of $\boldsymbol{P}$. Hence, letting

$$\gamma^* = \sup\{\gamma < 1 : P(\sigma^{-1}) \geq \gamma,\ \delta < \sigma < 1 \text{ for some } \delta > 0\}, \tag{3.3}$$

one has $T_{REL}(GI/GPH/1) \leq (1-\gamma^*)^{-1}$, see Figure 1.

To obtain an upper bound of $T_{REL}(GI/GI/1)$, we employ the limiting argument used in Section 2. Suppose $B(x)$ is approximated by a GPH. To identify the value of $\mu$, we write $A_\mu(z)$, $G_\mu(z)$, etc. Notice that $A_\mu(z) = \alpha(\mu - \mu z)$. Consider then the transform $s = \mu z - \mu$. Then, $z^{-1} = \frac{\mu}{s+\mu}$, from which

$$G_\mu(z^{-1}) = \sum_{n=0}^{\infty} g_n(\mu)\left(\frac{\mu}{s+\mu}\right)^n.$$

Hence, one has from (2.13) and the remark below the equation that, for each $\delta_2 < s < 0$,

$$\lim_{\mu\to\infty} A_\mu(z)G_\mu(z^{-1}) = \alpha(-s)\beta(s). \tag{3.4}$$

Therefore, combining (3.3) and (3.4), the next theorem follows.

**Theorem 1.** Let $\alpha(s)$ and $\beta(s)$ be the LSTs of the interarrival and the service time DFs, respectively, in a GI/GI/1 queue. Let (see Figure 2)

$$\gamma^* = \sup\{\gamma < 1 : \alpha(-s)\beta(s) \geq \gamma,\ \delta < s < 0 \text{ for some } \delta < 0\}. \tag{3.5}$$

Then, the relaxation time of the GI/GI/1 queue defined by (2.14) is bounded from above by

$$T_{REL}(GI/GI/1) \leq \frac{1}{1-\gamma^*}. \tag{3.6}$$

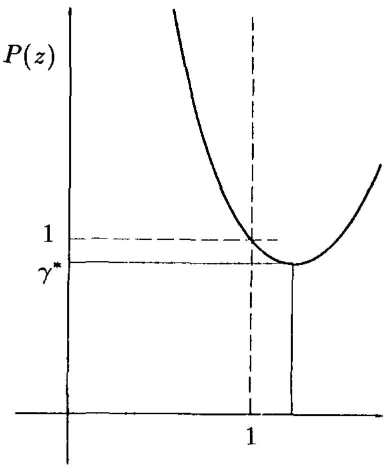


Figure 1

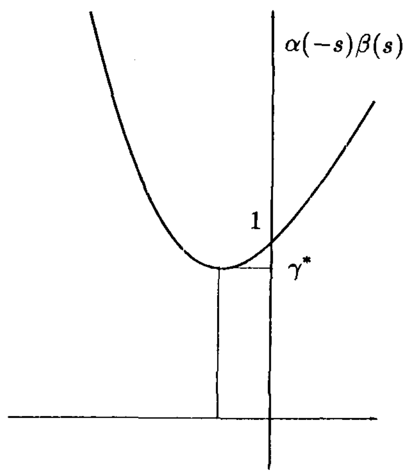


Figure 2

**Remark 1.** A lower bound of the convergence rate $r$ is easily obtained as follows (see Seneta [9] for details). Let $\boldsymbol{P}_{(n)}$ be the $n \times n$ north-west corner submatrix of $\boldsymbol{P}$. Since $\boldsymbol{P}_{(n)}$ is non-negative, it has the Perron-Frobenius eigenvalue. Denote it by $r_{(n)}$. It is known that $r_{(n)}$ is non-decreasing in $n$ and converges to $r$ from below. Thus, by choosing $n$ sufficiently large, one can approximate $r$ from below. However, the approximation error as well as the convergence speed is not clear. This together with the difficulty of finding the Perron-Frobenius eigrnvalue for large $n$ sometimes makes the approximation impractical.

## 4 Particular Cases

In this section, we consider some particular cases. The main result here is to show that the upper bound obtained in Theorem 1 is tight for M/M/1 and M/D/1 queues.

*4.1. M/M/1 Queues.* Let $\lambda$ and $\mu$ be the arrival and the service rates of an M/M/1 queue, respectively. Let $\rho = \lambda/\mu$ and define $g(s) = \alpha(-s)\beta(s)$. Solving $g'(s) = 0$, one easily finds $s = (\lambda - \mu)/2 < 0$ and

$$\gamma^* = \frac{4\rho}{(1+\rho)^2}; \quad T_{REL}(M/M/1) \le \left(\frac{1+\rho}{1-\rho}\right)^2. \tag{4.1}$$

To see that the upper bound in (4.1) is tight, we consider the distribution of the number served during a busy period. It is known (see e.g. [7]) that

$$f_n = \frac{1}{n}\binom{2n-2}{n-1}\rho^{n-1}(1+\rho)^{1-2n}, \quad n \ge 1. \tag{4.2}$$

By using Stirling's formula appropriately, one sees that

$$f_n \sim O\left(\left[\frac{4\rho}{(1+\rho)^2}\right]^n\right) \quad \text{as} \quad n \to \infty.$$

Thus, $\gamma^* = r$, see (2.12) for the definition of $r$. This means that our upper bound is tight for M/M/1 queues.

**Remark 2.** Let $N(t)$ denote the number of customers in the system at time $t$. Then $N(t)$ is a Markov chain on $\mathcal{N}$ governed by an infinitesimal generator of tri-diagonal form. It is known (see, e.g. [2], [11]) that the probability $\Pr[N(t) = j | N(0) = i]$ converges to $(1-\rho)\rho^j$ at an exponential rate $r_0$ for any $i$ and $j$. Moreover, $r_0 = \mu(1-\sqrt{\rho})^2$. In the literature, the relaxation time of M/M/1 queues is usually given as the reciprocal of the rate $r_0$. Recall at this point that our relaxation time is given in terms of the number of arriving customers. Thus, to compare our relaxation time with the ordinary one, it is plausible to consider $T_{REL}(M/M/1)/\lambda$. To be interesting,

$$\frac{T_{REL}(M/M/1)}{\lambda} > \frac{1}{\mu(1-\sqrt{\rho})^2}, \quad 0 < \rho < 1, \tag{4.3}$$

i.e. our relaxation time always exceeds. Note that the disagreement of the two relaxation times are due to the different definitions. Our definition for the M/M/1 case includes more steps, through which more uncertainty may come in.

*4.2. M/D/1 Queues.* Let $\lambda$ be the arrival rate and $\mu^{-1}$ be the service time of an M/D/1 queue, respectively. For this case, the same analysis as for the M/M/1 case leads to the conclusions

$$\gamma^* = \rho e^{1-\rho}; \quad T_{REL}(M/D/1) \leq \frac{1}{1-\rho e^{1-\rho}}. \tag{4.4}$$

The upper bound is tight, since appropriately using Stirling's formula bears

$$f_n = \frac{(n\rho)^{n-1}}{n!} e^{-n\rho} \sim O([\rho e^{1-\rho}]^n) \quad \text{as} \quad n \to \infty.$$

*4.3. GI/M/1 Queues.* For this case, $g_1 = 1$ so that (3.5) becomes

$$\gamma^* = \sup\{\gamma < 1 : \alpha(\mu - \mu z) \geq \gamma z, \delta < z < 1 \text{ for some } \delta > 0\}, \tag{4.5}$$

where $\mu$ is the service rate. The $\gamma^*$ can be easily obtained by drawing the line tangent to the curve $A(z) = \alpha(\mu - \mu z)$, $0 < z < 1$. Note that $A(z)$ is increasing and convex in $0 \leq z < \varepsilon$ for some $\varepsilon > 1$ and $A'(1) > 1$ if the queue is stable. Hence, the tangent line is unique (see Figure 3) as far as the queue is stable.

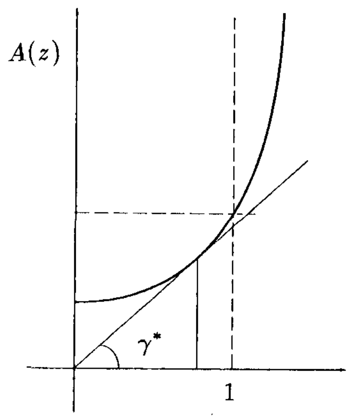


Figure 3

## References

[1] Blanc, J.P.C. (1985), "The Relaxation Time of Two Queueing Systems in Tandem," *Stoch. Models*, **1**, 1-16.

[2] Callaert, H. and Keilson, J. (1973), "On Exponential Ergodicity and Spectral Structure for Birth-death Processes II," *Stoch. Proc. Appl.*, **1**, 217-235.

[3] Esary, J.D., Marshall, A.W. and Proschan, F. (1973), "Shock Models and Wear Processes," *Ann. Prob.*, **1**, 627-649.

[4] Keilson, J., Machihara, H. and Sumita, U. (1984), "Spectral Structure of M/G/1 Systems – Asymptotic Behavior and Relaxation Time," Working Paper Series QM8414, Graduate School of Management, University of Rochester.

[5] Keilson, J. and Servi, L.D. (1987), "Dynamics of the M/G/1 Vacation Model," *Oper. Res.*, **35**, 575-582.

[6] Kijima, M. (1988), "Upper Bounds of a Measure of Dependence and the Relaxation Time for Finite State Markov Chains," to appear in *J. Oper. Res. Soc. Japan.*

[7] Kleinrock, L. (1976), *Queueing Systems*, Vol.1, Wiley, New York.

[8] Knopp, K. (1945), *Theory of Functions*, Part I, Dover Pub., New York.

[9] Seneta, E. (1981), *Non-negative Matrices and Markov Chains*, 2nd Ed. Springer-Verlag, New York.

[10] Shanthikumar, J.G. (1985), "Bilateral Phase-type Distributions," *Naval. Res. Log. Quart.*, **32**, 119-136.

[11] Van Doorn, E.A. (1985), "Conditions for Exponential Ergodicity and Bounds for the Decay Parameter of a Birth-death Process," *Adv. Appl. Prob.*, **17**, 514-530.

Masaaki KIJIMA
Department of Information Sciences
Tokyo Institute of Technology
Meguro-ku, Tokyo 152, Japan