Computer Architectures and Networks, E. Gelenbe / R. Mahl (Eds.)

BOUNDS FOR THE WAITING TIME IN QUEUEING SYSTEMS

Hisashi KOBAYASHI

Computer Science Department
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT: The present paper discusses bounds on the waiting time distributions in queueing systems. First, we review closed form solutions of the waiting time distribution for M/M/m, GI/M/m, M/G/1, and GI/G/1 queues. Then, we derive exponential bounds for the waiting time distribution (in both transient and equilibrium states) of a GI/G/1 queue. Our result is an extension of Kingman's bound and is based on Kolmogorov's inequality for submartingale. In the final section we give a new treatment of the heavy traffic approximation, in which close relationships will be found between the heavy traffic theory for GI/G/1, Lindley's theory, and the exponential bound derived above.

I. INTRODUCTION

In recent years we have been observing an increasing interest in approximation and bounding techniques for queueing theory. The diffusion approximation method (Newell, 1971; Gaver and Shedler, 1973; Kobayashi, 1974a), for example, is an attempt to find reasonably accurate solutions for server utilization and queue size distribution for queueing systems with general interarrival and service time distributions. Another fruitful approach will be to establish upper and lower bounds on such variables as waiting time and queue size, which are valid for a wide range of input and service mechanisms (Kleinrock, 1969). The present paper will focus on bounds for the waiting time.

In Section 2, we review closed form solutions for the equilibrium state waiting time distribution in M/M/m, GI/M/m, M/G/1, and GI/G/1 queues. In Section 3, we derive an exponential bound on the complementary distribution of waiting time $W_n^c(t) = \Pr\{w_n \geq t\}$, where w_n is the waiting time of the n th job in a busy period in a GI/G/1 queueing system. The result is an extension of an earlier work by Kingman (1964, 1970) who obtained a bound on the equilibrium state distribution. Section 3 discusses Kingman's heavy traffic approximation (1962a, 1965). Our treatment, however, is different from his argument and will shed light on the relation between Lindley's theory and the exponential bound discussed in the earlier sections.

II. WAITING TIME DISTRIBUTIONS - A REVIEW:

Consider a simple queueing system in which the number of servers may be greater than one but a single queue is shared. Furthermore, we assume the FCFS (first-come, first-served) discipline. Let us number the sequence of jobs in the order of arrival and define the waiting time of the n th job, w_n, to be the time from its entering the system to the instant at which its service begins, whereas its response time, r_n, is the time from entering the system to the instant at which its service is completed. Thus, the response time equals the waiting time plus its service time s_n.

Let us denote the steady-state distribution of the random variable w by

$$W(t) = \Pr\{w \leq t\} \quad , \tag{2.1}$$

and the complementary distribution of waiting time by

$$W^c(t) = 1 - W(t) = \Pr\{w > t\}. \tag{2.2}$$

In the sequel, we adopt the conventional short-hand notations due to Kendall (1951) to describe the type of queueing situation. The following are known results on the waiting and response time distributions (see for example Syski, 1960; Cooper, 1972).

2.1 M/M/m System: (Poisson arrival with rate λ; exponentially distributed service time with mean $1/\mu$; m servers):

$$W^c(t) = W^c(0) \cdot e^{-m\mu(1-\rho)t} \quad , \tag{2.3}$$

$$W^c(0) = \Pr\{w > 0\} = \frac{P(m)}{1-\rho} = \frac{1}{1-\rho} \cdot \frac{\frac{a^m}{m!}}{1+a+\frac{a^2}{2!}+\ldots+\frac{a^{m-1}}{(m-1)!}+\frac{a^m}{m!}\cdot\frac{1}{1-\rho}} \tag{2.4}$$

where

$$a = \frac{\lambda}{\mu} \, , \quad \rho = \frac{a}{m} = \frac{\lambda}{m\mu} \tag{2.5}$$

and P(j) is the probability of queue size (including a job in service if any) being j:

$$P(j) = \frac{\rho^j}{j!} P(0), \quad j < m$$

$$P(j) = P(m)\, \rho^{m-j}, \quad j \geq m \, . \tag{2.6}$$

The constant P(0) is determined from the normalizing condition.

The mean waiting time $\bar{w}$ is found to be

$$\bar{w} = \frac{W^c(0)}{m\mu(1-\rho)} = \frac{P(m)}{m\mu(1-\rho)^2} \quad . \tag{2.7}$$

By convolving W(t) and the service time distribution we obtain the response time distribution of the M/M/m system

$$R(t) = \frac{m(1-\rho) - W(0)}{m(1-\rho)-1} \{1-e^{-\mu t}\} - \frac{W^c(0)}{m(1-\rho)-1} \{1 - e^{-m\mu(1-\rho)t}\}, \quad t \geq 0. \tag{2.8}$$

An M/M/1 system is obtained as a special case in which m = 1, hence $a = W^c(0) = \rho$, leading to

$$W^c(t) = \rho e^{-(1-\rho)\mu t} \tag{2.9}$$

and

$$R(t) = 1-e^{-\mu(1-\rho)t} \ . \tag{2.10}$$

2.2 GI/M/m: (General independent interarrival with distribution A(t); exponentially distributed service time with mean $1/\mu$; m servers):

$$W^c(t) = W^c(0)\ e^{-m\mu(1-\beta)t} \tag{2.11}$$

$$W^c(0) = \Pr\{w > 0\} = \frac{Q(m)}{1-\beta} \ . \tag{2.12}$$

Here β is the unique root of the characteristic equation

$$\beta = \int_0^\infty e^{-m\mu(1-\beta)t}\, dA(t) = \tilde{A}[m\mu(1-\beta)] \tag{2.13}$$

where $\tilde{A}(\cdot)$ is the Laplace-Stieltjes transform of A(t). Q(j) is the distribution of the imbedded Markov chain associated with the GI/M/m queueing process, and represents the probability distribution of the number of jobs in the system just prior to the moment a job arrives in the system. If the traffic intensity ρ is less than unity, the limiting distribution exists and is a geometric series except for modifications of its first m-1 terms, the common ratio being β. In particular for j = m,

$$Q(m) = \frac{1}{\frac{1}{1-\beta} + \sum_{i=1}^{m} \gamma_i} \tag{2.14}$$

where

$$\gamma_i = \frac{\binom{m}{i}}{c_i(1-\phi_i)} \cdot \frac{m(1-\phi_i)-i}{m(1-\beta)-i} \tag{2.15}$$

$$c_0 = 1, \qquad c_i = \prod_{j=1}^{i} \frac{\phi_j}{1-\phi_j} \tag{2.16}$$

and

$$\phi_i = \tilde{A}(i\mu) \ . \tag{2.17}$$

The mean waiting time is

$$\bar{W} = \frac{Q(m)}{m\mu(1-\beta)^2} \ . \tag{2.18}$$

There is a striking resemblance of these expressions to the case M/M/m, the difference being that ρ is replaced by β.

In the case of a single server, GI/M/1, the distribution Q(j) is a geometric one,

$$Q(j) = (1-\beta)\beta^j, \quad j = 0, 1, 2, \ldots \tag{2.19}$$

whereas the waiting time distribution is exponential

$$W^c(t) = \beta e^{-\mu(1-\beta)t} \tag{2.20}$$

which again resembles formally the solution of the case M/M/1. The response time distribution for GI/M/m and GI/M/1 are given by exactly the same formulas (2.8) and (2.9) except for ρ being replaced by β.

2.3. M/G/1: (Poisson arrival with rate λ; general service time distribution B(t) with mean $1/\mu$; a single server):

Pollaczek-Khinchine Formula:

$$\tilde{W}(s) = \frac{s(1-\rho)}{s-\lambda[1-\tilde{B}(s)]} \tag{2.21}$$

where $\tilde{W}(s)$ and $\tilde{B}(s)$ are the Laplace-Stieltjes transforms of W(t) and B(t), respectively. Equation (2.21) can be expanded as

$$\tilde{W}(s) = (1-\rho) \sum_{j=0}^{\infty} \rho^j \left[\frac{\mu\{1-\tilde{B}(s)\}}{s}\right]^j \tag{2.22}$$

which yields

Benes's formula:

$$W(t) = (1-\rho) \sum_{j=0}^{\infty} \rho^j \hat{B}^{*j}(t) \tag{2.23}$$

where $\hat{B}^{*j}(t)$ is the j-fold convolution with itself of the distribution $\hat{B}(t)$ defined by

$$\hat{B}(t) = \mu \int_0^t [1-B(s)]ds = \frac{\int_0^t [1-B(s)]ds}{\int_0^\infty [1-B(s)]ds} . \tag{2.24}$$

2.3 GI/G/1: (General independent interarrival distribution A(t)/general service time distribution B(x)/single server):

Define the following distribution function

$$F(x) = \int_0^\infty B(x+y)\, dA(y), \tag{2.25}$$

Then, W(t) is given as the solution of

Lindley's Integral Equation:

$$W(t) \begin{cases} = \int_{-\infty}^{t} W(t-x)\, dF(x), & t \geq 0 \\ = 0 , & t < 0 . \end{cases} \tag{2.26}$$

We assume that the probability density associated with the interarrival time drops off at least as fast as an exponential for very large interarrival times and let θ be a real number greater than zero such that

$$\lim_{t\to\infty} \frac{dA(t)/dt}{e^{-\theta t}} < \infty \tag{2.27}$$

which in turn implies that

$$\lim_{x\to-\infty} \frac{F(x)}{e^{\theta x}} < \infty \quad . \tag{2.28}$$

Let $\hat{F}(s)$ be the Laplace-Stieltjes transform of F(x) and find the factorization of $\hat{F}(s)-1$

$$\hat{F}(s) - 1 = \frac{\psi_+(s)}{\psi_-(s)} \tag{2.29}$$

such that for $Re\{s\}>0$ $\psi_+(s)$ is an analytic function of s and it contains no zeros in this half plane, and for $Re\{s\}<\theta$, $\psi_-(s)$ is an analytic function of s and contains no zeros in this half plane. Furthermore, these functions should satisfy

$$\lim_{|s|\to\infty} \psi_+(s) = s \quad \text{for } Re\{s\} > 0 \tag{2.30}$$

$$\lim_{|s|\to\infty} \psi_-(s) = -s \text{ for } Re\{s\} < \theta \quad . \tag{2.31}$$

The Laplace-Stieltjes transform of W(t) is then given by

$$\hat{W}(s) = \frac{s\,W(0)}{\psi_+(s)} \tag{2.32}$$

where

$$W(0) = \lim_{s\to 0} \frac{\psi_+(s)}{s} = \frac{1-\rho}{\lambda}\ \psi_-(0) \ . \tag{2.33}$$

III. EXPONENTIAL BOUNDS ON WAITING TIME DISTRIBUTIONS

In this section we will derive a tight upper bound on $W_n^c(t) = Pr\{w_n > t\}$, where w_n is the waiting time of the n th job in a busy cycle. First we start from the background mathematics.

3.1 Kolmogorov's Inequality and Exponential Bounds on Waiting Time Distribution

Chebyshev's inequality is among the most frequently used inequalities in probability theory. The inequality in most general form is stated as follows (Gallager, 1968).

Chebyshev's Inequality

Let y be a nonnegative random variable with its finite mean $\bar{y}$. Then for any $\delta > 0$,

$$P\{y \geq \delta\} \leq \frac{\bar{y}}{\delta} \ . \tag{3.1}$$

Kolmogorov generalized Chebyshev's inequality to martingales and submartingales (or semi-martingale)

Definition of Martingale:

A random variable sequence $\{y_n\}$ is called a martingale if $E[|y_n|] < \infty$ for all n and if

$$E[y_n \mid y_1 y_2, \ldots, y_{n-1}] = y_{n-1}. \tag{3.2}$$

Definition of Submartingale:

A random variable sequence $\{y_n\}$ is called a submartingale (or semimartingale) if $E[|y_n|] < \infty$ for all n and if

$$E[y_n \mid y_1, y_2, \ldots, y_{n-1}] \geq y_{n-1} \tag{3.3}$$

for all n.

We now present Kolmogorov's inequality for arbitrary submartingales.

Kolmogorov's Inequality for Submartingales:

Let $\{y_n\}$ be a sequence of variables for which eq. (3.3) holds and $y_n \geq 0$ for all n. Then for any $\delta > 0$

$$P\{\max[y_1, \ldots, y_n] \geq \delta\} \leq \frac{\bar{y}_n}{\delta} \quad . \tag{3.4}$$

Proof: See Feller (1966: pp. 235-236).

3.2 An Upperbound on $W_n^c(t)$ in a GI/G/1 System

We now consider a GI/G/1 queue, in which jobs $\{J_n\}$ arrive and are served in the order of their arrivals. As in section 2, we denote the waiting time and service time of J_n by w_n and s_n, respectively, and let the interval between the arrivals of J_{n+1} and J_n be denoted by t_n. We now define a random variable x_n by

$$x_n = s_n - t_n \ . \tag{3.5}$$

Note that x_n's are i.i.d. random variables and have the common distribution given by (2.25), i.e.,

$$P\{x_n \leq u\} = F(u). \tag{3.6}$$

Then the sequence $\{w_n\}$ of waiting times is a sequence of random variables defined recursively by

$$w_0 = 0 \tag{3.7a}$$

$$w_{n+1} = \max[0, w_n + x_n] \tag{3.7b}$$

where we assume that job J_0 arrives at epoch 0 at a free server and so his waiting time is $w_0 = 0$. By solving (3.7) recursively we have

$$w_n = \max[0, x_{n-1}, x_{n-1}+x_{n-2}, \ldots, x_{n-1}+x_{n-2}+ \ldots +x_0] \ . \tag{3.8}$$

For given n, we define a sequence $\{y_j\}$ by

$$y_0 = 1$$

$$y_j = e^{\theta(x_{n-1}+x_{n-2}+ \cdots + x_{n-j})}, \quad 1 \leq j \leq n, \tag{3.9}$$

where θ is a real-valued parameter to be determined below. If $\theta > 0$, then eqs. (3.8) and (3.9) imply that

$$e^{\theta w_n} = \max[y_0, y_1, \ldots, y_n] \tag{3.10}$$

We define the moment generating function $f(\theta)$ of the i.i.d. random variables x by

$$f(\theta) \triangleq E[e^{\theta x}] \tag{3.11}$$

where θ is a real variable. Note that $f(\theta)$ is defined over an interval I_θ in which $f(\theta)$ is bounded. This domain I_θ includes the origin $\theta = 0$. It is not difficult to show that the function $f(\theta)$ is a convex U function. Furthermore, $f(0) = 1$ and $f'(0) = E[x] < 0$. Let $\theta > 0$ be any value in I_θ that satisfies

$$f(\theta) \leq 1 \tag{3.12}$$

then

$$\begin{aligned} &E[y_n \mid y_1, y_2, \ldots, y_{n-1}] \\ &= E[e^{\theta(x_{n-1}+x_{n-2}+\ldots+x_0)} \mid e^{\theta x_{n-1}}, e^{\theta(x_{n-1}+x_{n-2})}, \ldots, e^{\theta(x_{n-1}+x_{n-2}+\ldots+x_1)}] \\ &= E[e^{\theta x_0}]\, e^{\theta(x_{n-1}+x_{n-2}+\ldots+x_1)} = f(\theta)\, y_{n-1} \leq y_{n-1} \,. \end{aligned} \tag{3.13}$$

Therefore, the sequence $\{y_n\}$ is a submartingale. Note here that $f(\theta)$ is equal to the Laplace-Stieltjes transform of F(t) evaluated at $s = -\theta$, since

$$f(\theta) = \int_{-\infty}^{\infty} e^{\theta x}\, dF(x) = \tilde{F}(-\theta) = \tilde{A}(\theta)\, \tilde{B}(-\theta) \,. \tag{3.14}$$

Then by applying the Kolmogorov's inequality to this submartingale, we obtain

$$\begin{aligned} W_n^c(t) = P\{w_n \geq t\} &= P\{e^{\theta w_n} \geq e^{\theta t}\} \\ &= P\{\max[y_o, y_1, \ldots y_n] \geq e^{\theta t}\} \\ &\leq \frac{E[y_n]}{e^{\theta t}} = e^{-\theta t + n g(\theta)} \end{aligned} \tag{3.15}$$

where $g(\theta)$ is called the semi-invariant moment generating function and is defined by

$$g(\theta) \triangleq \ln f(\theta) \tag{3.16}$$

Then for a given n and t, the tightest bound is attained by finding

$$\min \{g(\theta) - \frac{t}{n} \cdot \theta\} \tag{3.17}$$

with the constraint (3.12) or equivalently

$$g(\theta) \leq 0 \,. \tag{3.18}$$

If it were not for the constraint (3.18), the minimization of (3.17) would always be achieved by choosing θ^* such that

$$g'(\theta^*) = \frac{f'(\theta^*)}{f(\theta^*)} = \frac{t}{n} \,. \tag{3.19}$$

Therefore if the solution θ^* of (3.19) is larger than θ_o which is the unique positive root (in the domain I_θ) of the following equation:

$$g(\theta) = 0, \tag{3.20}$$

or equivalently

$$f(\theta) = 1 , \tag{3.21}$$

then we have

$$W_n^c(t) \leq e^{-\theta^* t + n\, g(\theta^*)} = e^{-n[\theta^* g'(\theta^*) - g(\theta^*)]} \tag{3.22}$$

If $\theta^* < \theta_o$, then

$$W_n^c(t) \leq e^{-\theta_o t} . \tag{3.23}$$

It is to be noted that the tighter bound (3.22) is valid for small n and/or large t.

By letting $n \to \infty$ in (3.23), we obtain an upper bound of the tail of the waiting time distribution at equilibrium state:

$$W^c(t) \leq e^{-\theta_o t} . \tag{3.24}$$

This inequality for equilibrium distribution has been obtained by Kingman (1964, 1970). It is important to notice that the zero of $\psi_+(s)$ of (2.29) that is closest to the origin s = 0 is given by

$$s_o = -\theta_o . \tag{3.25}$$

Therefore, the upperbound (3.23) is best possible in the sense that there is no $\theta_o' > \theta_o$ for which

$$W_c(t) = O(e^{-\theta_o' t}) . \tag{3.26}$$

In fact Kingman (1970) has derived the following lower bound

$$W_c(t) \geq a\, e^{-\theta_o t} \tag{3.27}$$

where

$$a = \inf_{t>0} \frac{\int_t^\infty dF(x)}{\int_t^\infty e^{\theta_o (x-t)} d\, F(x)} \tag{3.28}$$

where F(x) is the distribution of (3.6). It is also possible to show that

$$a \geq \inf_{t>0} \frac{\int_t^\infty dB(s)}{\int_t^\infty e^{\theta_o (s-t)}\, dB(s)} \tag{3.29}$$

where B(s) is the distribution of the service time s_n.

In (Kobayashi, 1974b) it is shown how the above results can be applied to various queueing systems, and the bounds are compared with the corresponding exact solutions via Lindley's theory. Because of the space limitation these results are not reproduced here.

IV. THE HEAVY TRAFFIC APPROXIMATION AND UPPER BOUNDS FOR THE WAITING TIME

In this section we discuss the behavior of the system GI/G/1 in the "heavy traffic" situation, where the traffic intensity ρ is just below its critical value $\rho=1$. The central result in heavy traffic theory is that the distributions of the waiting time and similar other variables of interest are insensitive to the nature of the input and service processed. Here we derive the heavy traffic approximation in a different manner from the argument given in (Kingman, 1962a, 1965), whereby we can show more clearly relationships between the results of sections 2 and 3.

Before we consider a heavy traffic situation, let us review the formulae given in sections 2.1 and 2.2. There we saw that an exponential form of $W^C(t)$ holds exactly in the M/M/m and GI/M/m queues. In a GI/G/1 queue, when the Laplace-Stieltjes transform $\tilde{F}(s)$ is a rational function of s, W(t) can be expressed as a sum of negative exponential functions, since $\tilde{F}(s)$ has in general more than one zero in the half plane $\mathrm{Re}\{s\} < 0$.

Let $s_o < 0$ be such zero that is closest to s = 0, i.e.,

$$\tilde{F}(s_o) = 1 \ . \tag{4.1}$$

It is not difficult to show that s_o is real, and it is clear (see eq. (3.24) that

$$s_o = -\theta_o \ . \tag{4.2}$$

Then for sufficiently large $t > 0$, the exponential term due to the zero $s_o = -\theta_o$ predominates, and thus we have the following asymptotic expression

$$W^C(t) \simeq C_o \, e^{-\theta_o t} \tag{4.3}$$

where C_o is determined by applying the partial fraction method to eq. (2.32) and is given by

$$C_o = -W(0) \lim_{s \to s_o} \frac{s-s_o}{\psi_-(s_o)} = -\frac{\psi_+'(0)}{\psi_+'(s_o)} \ . \tag{4.4}$$

We now consider the following Taylor series expansion of $f(\theta)$ defined by (3.12)

$$f(\theta) = E[e^{\theta x}] = 1 + \theta \, E[x] + \frac{\theta^2}{2} E[x^2] + O(\theta^3) \tag{4.5}$$

Therefore, the characteristic equation of (3.21)

$$f(\theta) - 1 = \theta \cdot \{E[x] + \frac{\theta}{2} E[x^2] + O(\theta^2)\} = 0 \tag{4.6}$$

has a zero $\theta=0$ and additional zero θ_o near $\theta=0$, which is given by

$$\theta_o \simeq -\frac{2E[x]}{E[x^2]} > 0 \ . \tag{4.7}$$

By the definition of random variable x (eq. (3.5))

$$E[x] = \bar{s} - \bar{t} = -\bar{t}(1-\rho) \simeq 0 \tag{4.8}$$

and

$$E[x^2] = \bar{x}^2 + \mathrm{var}[x]$$

$$\simeq \mathrm{var}[x] = \sigma_s^2 + \sigma_t^2 \tag{4.9}$$

under a heavy traffic condition. Thus

$$\theta_o \simeq -\frac{2\bar{x}}{\mathrm{var}[x]} . \tag{4.10}$$

As we have discussed above, the exponential term which contains θ_o predominates in the waiting time distribution

$$W^c(t) \simeq C_o e^{-\theta_o t} . \tag{4.11}$$

The coefficient C_o of (4.4) is close to unity under the heavy traffic condition, and we can obtain, using (3.14) and (4.6), the following factorization

$$\tilde{F}(s) - 1 = -s\{E[x] - \frac{s}{2}E[x^2] + O(s^2)\} \simeq \frac{E[x^2]}{2} s(s+\theta_o) \tag{4.12}$$

near the origin, i.e. around s=0. Therefore, $\psi_+(s)$ defined by (2.29) will be well approximated by

$$\psi_+(s) \simeq s(s+\theta_o)\phi_+(s) \tag{4.13}$$

near the origin where $\phi_+(s)$ contains no zeros or poles in $\mathrm{Re}\{s\} > 0$ and remains relatively unchanged near s=0 and $s = -\theta_o$. Since $\psi_+(s)$ is a parabola in this region, it follows that

$$C_o = \frac{\psi_+'(0)}{\psi_+'(-\theta_o)} \simeq 1 . \tag{4.14}$$

Therefore, we obtain the following essential result of heavy traffic theory

$$W^c(t) \simeq e^{-\frac{2\bar{t}(1-\rho)}{\sigma_t^2 + \sigma_s^2} t} \tag{4.15}$$

Thus, the bound (3.23) is quite tight upper bounds in the heavy traffic situation i.e., for $\rho \simeq 1$.

The results discussed above apply only to GI/G/1. Kingman (1965), however, has made a conjecture that when the queue GI/G/m is in a situation of heavy traffic, the equilibrium waiting time distribution is approximately negative exponential

$$W^c(t) \simeq e^{-\frac{2(m\bar{t} - \bar{s})}{m\sigma_t^2 + \sigma_s^2/m}} \tag{4.16}$$

which, to the author's knowledge, remains yet to be proved.

V. CONCLUDING REMARKS

Sometimes we may be content with the mean waiting time rather than the distribution. Bounds for the mean waiting time can be easily derived from the exponential bounds for the waiting time distribution discussed above. But tighter bounds for the mean waiting time have been found via different approaches by Kingman (1962b, 1970), Marshall (1968) and Brummelle (1971, 1973). For a review of this subject, see also (Kobayashi, 1974b).

In this paper the exponential tail of the waiting time distribution was found to be a characteristic common to a wide class of distributions of interarrival and service times. This property seems, however, critically dependent on the queue service discipline. Chow (1974) has recently shown that the response time distribution of M/M/1 queue with processor sharing discipline is given by the following hyperexponential distribution with infinite stages

$$R(t) = (1-\rho) \sum_{j=0}^{\infty} \rho^j \left(1-e^{-\frac{\mu t}{j+1}}\right)$$

for which there is no $\theta > 0$ such that $R(t) \leq e^{-\theta t}$. Bounds for the waiting and response time distributions under non-FCFS queue disciplines are research subjects yet to be explored.

REFERENCES

Brummell, S. L., 1971, Some Inequalities for Parallel Server Queues Operations Research, 19, 402-413.

Brummell, S. L., 1973, Bounds on the Wait in a GI/M/m Queue, Management Science, 19, 773-777.

Chow, W-M., 1974, The Response Time Distribution of M/M/1 Queue with Processor Sharing Discipline, to appear as an IBM Research Report.

Cooper, R. B., 1972, Introduction to Queueing Theory, (MacMillan Co.) 203-248.

Feller, F., 1966, An Introduction to Probability Theory and Its Applications, vol. 2, (John Wiley & Sons, Inc.) 149-150 and 234-238.

Gallager, R. G., 1968, Information Theory and Reliable Communication, (John Wiley & Sons, Inc.), 126-127.

Kendall, D. G., 1951, Some Problems in the Theory of Queues, J. Roy. Statist. Soc. Ser. B13, 151-185.

Gaver, D. P. and G. S. Shedler, 1973, Processor Utilization in Multiprogramming Systems Via Diffusion Approximation, Operations Research, 21, 569-576.

Kingman, J. F. C., 1962a, On Queues in Heavy Traffic, J. Roy Statist. Soc. B, 24, 383-392.

Kingman, J. F. C., 1962b, Some Inequalities for the Queue GI/G/1, Biometrika, 49, 315-324.

Kingman, J. F. C., 1964, A Martingale Inequality in the Theory of Queues, Proc. Comb. Phil. Soc., 59, 359-361.

Kingman, J. F. C., 1965, The Heavy Traffic Approximation in the Theory of Queues, in Proceedings of the Symposium on Congestion Theory (Univ. of North Carolina Press, Chapel Hill), 137-169.

Kingman, J. F. C., 1970, Inequalities in the Theory of Queues, Journal of Royal Statistical Society, Series B32, 102-110.

Kleinrock, L., 1969, Queueing Systems: Theory and Applications, Lecture note of Engineering 220, University of California, Los Angeles.

Kobayashi, H., 1974a, Application of the Diffusion Approximation to Queueing Networks, Parts I and II, Journal of ACM, April and July 1974. Also IBM Research Reports, RC 3943 and RC 4054, (1972).

Kobayashi, H., 1974b, Bounds for the Waiting Time in Queueing Systems, IBM Research Report, RC 4718.

Marshall, K. T., 1968, Some Inequalities in Queueing, Operat. Research, 16, 651-665.

Newell, G. F., 1971, Applications of Queueing Theory (Chapman and Hall Ltd.).

Syski, R., 1960, Introduction to Congestion Theory in Telephone Systems, (Oliver & Boyd, Ltd.), 285-287.