#217

INFORMATION PROCESSING 71 – NORTH-HOLLAND PUBLISHING COMPANY (1972)

TIGHT BOUNDS ON THE AVERAGE RESPONSE TIME FOR TIME-SHARED COMPUTER SYSTEMS*

Leonard KLEINROCK, Richard R. MUNTZ and Jiunn HSU**

Computer Science Department, University of California, Los Angeles, California, USA

In this paper, some fundamental properties are established which apply to the average response time functions for all time-shared computer systems. The first property is one of monotonicity. The second is a conservation law which provides insight into the trade-offs available as one varies the response time function by changing the scheduling algorithm.

The main thrust of the paper is to establish tight upper and lower bounds on the average response time. All these equilibrium results are good for Poisson arrivals, arbitrary service time distribution and arbitrary (but work-conserving) scheduling algorithms which can take advantage only of arrival time and attained service time. Examples of these properties are given for a number of service-time distributions and scheduling algorithms.

1. INTRODUCTION

We are in the midst of a veritable explosion regarding the number of published papers which give analytical results for computer systems! This seems especially true in the modeling and analysis of time-shared computer systems [1].

It is fair to say that the recognition of probabilistic models as the appropriate method for studying these systems was that which permitted the breakthrough in analysis. In particular, the use of queueing theory has been most profitable in this analytic work.

As a result of this flood of results, each applying to a slightly different set of assumptions, it is natural that we should seek some order in this embarrassment of riches. For example, do there exist any invariants in behavior? Can we bound the possible range of performance, regardless of structure? What constitutes feasible performance profiles for these systems? These, and many more, are reasonable inquiries to make amidst the confusion of results.

In this paper we adopt the point of view that such questions are important and must be answered. Our focus is on a class of models for time-shared computer systems. For these systems we are able to state a monotonicity property, a conservation law, and tight upper and lower bounds on the system performance as measured by average response time.

It is worthwhile mentioning that numerous papers have recently been published which address themselves to bounds, inequalities and approximate solutions to general queueing systems. Among these are Marshall [2, 3], Kingman [4], Iglehart [5], Daley and Moran [6], and Gaver [7] to mention a few.

2. THE CLASS OF SYSTEMS

Our objective is to create some order among many of the results available in the analysis of time-shared computer systems. Let us consider the class of systems described below.

We adopt the well-known [8] feedback queueing model for time-shared systems shown in fig. 1.

In this model it is assumed that the central processing unit (CPU) is the only resource being accessed. Jobs arrive according to a Poisson process with an average arrival rate λ jobs/sec. They each bring a de-

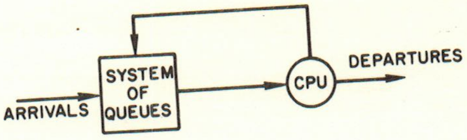

Fig. 1. General feedback queueing model.

* This work was supported by the Advanced Research projects Agency of the Department of Defense (DAHC-15-69-C-0285).

** This author (J.H.) wishes to acknowledge his gratitude to the International Business Machines Corporation for the granting of an IBM Fellowship

[1] See, for example, the recent survey by McKinney [1].

mand for service by the CPU in an amount equal to t seconds, where these demands are chosen independently from the service time distribution $B(t)$.

$$B(t) = P\,[\text{service time} \leqslant t \text{ seconds}] \tag{2.1}$$

We define the usual moments of service time as[2)]

$$\overline{t^n} = E\,[t^n] = \int_0^\infty t^n \mathrm{d}B(t) \ . \tag{2.2}$$

We further define the utilization factor[3)]

$$\rho = \lambda \overline{t} \ . \tag{2.3}$$

Upon arrival, a job enters the systems of queues where he waits for a "turn" at service. When, finally, his turn comes up, he is provided a quantum of service equal to q seconds. If he requires less than (or equal to) q seconds, he departs upon completion; if not, returns to the system of queues having been partially served, in which case we say that he has an *attained service* of q seconds. Eventually, he will be permitted a second quantum, etc., finally leaving when his total attained service equals his required service time. We assume that no overhead (in time) is incurred in transferring customers in and out of service (i.e., no loss or swap-time); it is possible to account for swap-time [9] in these models, but we do not pursue that matter here.

The decision rule which chooses the next customer to receive a quantum is referred to as the *scheduling algorithm.* We assume that the scheduling algorithm makes use only of λ, $B(t)$, a job's arrival time and a job's attained service.

In this paper, we consider a very useful special case of the above model in which we permit the quantum q to approach zero. This limit is known as the processor-sharing model [10] for time-shared systems. In this case, our model in fig. 1 becomes that of fig. 2 in which more than one customer (say n) may be sharing the processor simultaneously; in such a case each customer receives service at a rate of $1/n$ seconds of service/second.

Response time is the interval measured from when a customer arrives demanding service until he departs fully serviced. For a customer requiring t seconds of service, the average response time is denoted.[4)]

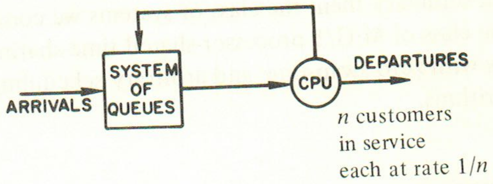

Fig. 2. Feedback queueing model for processor sharing.

$$T(t) = \text{average response time for customer requiring } t \text{ seconds of service} \tag{2.4}$$

This quantity is usually taken as the measure of performance for time-shared systems for good reason. In particular, it is usually desired that short jobs (small t) be given preferential treatment over long jobs; this discriminatory performance is easily seen through the function $T(t)$.

A function closely related to the average response time $T(t)$, is the average *wasted* or *waiting* time $W(t)$ defined as

$$W(t) = T(t) - t \ . \tag{2.5}$$

Furthermore, we consider a third related function, $W(t)/t$ which may be interpreted as the *penalty rate* to jobs requiring t seconds of service since it gives the ratio of the cost in time ($W(t)$ which must be paid per second of useful service time t).

It is convenient to introduce some additional notation at this point. Let us define

$$\overline{t^n_x} = \int_0^\infty t^n \mathrm{d}B(t) + x^n\,[1-B(x)] \tag{2.6}$$

which is the nth moment of the service time distribution if service times are truncated at x seconds. Also let

$$\rho_x = \lambda \overline{t}_x \tag{2.7}$$

and

$$\overline{W}_x = \lambda \overline{t^2_x}/2(1-\rho_x) \ . \tag{2.8}$$

Note that $\overline{t^n_\infty} = \overline{t^n}$, $\rho_\infty = \rho$ and that $\overline{W}_\infty$ is the expected

[2)] where E denotes the expectation operator.

[3)] The systems we consider are assumed to be in equilibrium, which requires $\rho < 1$.

[4)] Since we have $\rho < 1$, we consider steady-state results only, an example of which is $T(t)$.

One might naturally inquire as to whether these curves are confined to any particular region in the $(W(t), t)$ plane. The answer is definitely yes,[7] and we develop these and other constraints in the next section.

4. RESULTS

In this section we present results concerning the response functions $(W(t))$ which are feasible when the scheduling discipline is based only on attained service times and elapsed waiting times of jobs. In section 4.1 below we describe several fundamental characteristics of $W(t)$ and, in particular, we give a conservation relationship which the response function must satisfy. In sections 4.2 and 4.3, tight lower and upper bounds are derived for response functions in the sense that for any $W(t)$, $W_l(t) \leqslant W(t) \leqslant W_u(t)$.

4.1. *A monotonicity property and a conservation law for W(t)*

We are considering scheduling disciplines in which each job is characterized by : (1) its attained service time, t_s, and (2) its elapsed waiting time, t_w. Therefore, the state of the system is the number of jobs in the system and t_s and t_w for each job. A particular scheduling discipline may effectively ignore one or both of these parameters, but this information is assumed to be available for each job. Because scheduling decisions are made only on the basis of these two parameters, the following statement is self-evident. The history of a job requiring $t_1 \geqslant t$ seconds of service from the time of its arrival at the system until it has received t seconds of service is independent of the exact value of t_1. A direct consequence of this fact is that $W(t)$ is a nondecreasing function or equivalently

$$W'(t) \equiv \frac{\mathrm{d}W(t)}{\mathrm{d}t} \geqslant 0 \ . \tag{4.1}$$

In deriving $W_l(t)$ and $W_u(t)$ we shall need another result which is given below. From [8] we have that

$$n(t) = \lambda[1-B(t)]\,[W'(t)+1] \ , \tag{4.2}$$

where $n(t)$ is the density of jobs in the system with t seconds of attained service time. We define the "work" in the system at the time t as the additional time required to empty the system if no new arrivals are permitted entry; this is also referred to as the "unfinished work" and as the "virtual waiting time." The mean work $\overline{W}$ in the system can be expressed as

$$\overline{W} = \int_{0^-}^{\infty} n(t) E\,[\text{remaining service time for a job with attained service time of } t]\,\mathrm{d}t$$

or

$$\overline{W} = \int_{0^-}^{\infty} n(t) \int_{t}^{\infty} (\tau-t)\,\frac{\mathrm{d}B(\tau)}{1-B(t)}\,\mathrm{d}t \ .$$

Substituting from (4.2)

$$\overline{W} = \lambda \int_{0^-}^{\infty} (\overline{W}'(t)+1) \int_{0}^{\infty} (\tau-t)\mathrm{d}B(\tau)\mathrm{d}t \ .$$

By changing the order of integration

$$\overline{W} = \lambda \int_{0}^{\infty} \left[\int_{0^-}^{\tau} (W'(t)+1)(\tau-t)\mathrm{d}t\right] \mathrm{d}B(\tau) \ . \tag{4.3}$$

Integrating the inner integral by parts,

$$\int_{0^-}^{\tau} (W'(t)+1)(\tau-t)\mathrm{d}t$$

$$= (\tau-t)(W(t)+t)\Big|_{0^-}^{\tau} + \int_{0^-}^{\tau} [W(t)+t]\,\mathrm{d}t$$

$$= \int_{0^-}^{\tau} [W(t)+t]\,\mathrm{d}t \ .$$

Substituting into eq. (4.3),

$$\overline{W} = \lambda \int_{0}^{\infty} \int_{0^-}^{\tau} [W(t)+t]\,\mathrm{d}t\,\mathrm{d}B(\tau).$$

Again changing the order of integration,

$$\overline{W} = \lambda \int_{0^-}^{\infty} [W(t)+t] \int_{t}^{\infty} \mathrm{d}B(\tau)\mathrm{d}t$$

$$= \lambda \int_{0}^{\infty} [W(t)+t]\,[1-B(t)]\,\mathrm{d}t \ .$$

But in general, we have that

$$\int_{0}^{\infty} t[1-B(t)]\,\mathrm{d}t = \tfrac{1}{2}\overline{t^2} \ .$$

[7] In fact, if the reader looks at this figure and squints his eyes, he can almost guess the shape of such bounds.

$$Pr\,[\text{service time} = kq] = p_k \quad k = 1, 2, 3, \ldots$$

where q is the time quantum discussed in section 2. Therefore, the only possible service time requirements are multiples of q. We shall also assume that arrivals may take place only during the instant before the end of a quantum and that the processor is assigned to a job for a quantum at a time. The probability that an arrival takes place at the end of a quantum is λq so that the mean arrival rate is λ. It should be clear that any continuous service time distribution can be approximated arbitrarily closely by a discrete time distribution by letting q approach 0. Also, these restrictions on the service discipline and arrival mechanism are effectively eliminated when $q \to 0$. In this discrete time model our goal is to maximize $W(kq)$.

We claim that the following scheduling rule is necessary and sufficient to maximize $W(kq)$: no allocation of a kth quantum is made to any job where there is some other job in the system waiting for its jth quantum where $j \neq k$. We note in passing that many scheduling disciplines will satisfy this rule.

We relabel the time axis so that $t = 0$ at an arbitrary point in some idle period. The times at which some job is allocated a kth quantum we call "critical times." Let c_i be the time that the ith critical time occurs. We wish to maximize $\bar{c}_l$ (the average of c_l) for some fixed l, and we will show that to accomplish this it is necessary and sufficient to satisfy the condition that at the lth critical time no job is waiting for a jth quantum where $j \neq k$. Certainly this condition is necessary since if a proposed scheduling discipline did not have this property then c_l can easily be increased when the condition is not satisfied as follows: follow the proposed schedule until the point where the lth critical time would occur and then assign a quantum to a job waiting for its jth ($\neq k$) quantum.

Since we have already shown necessity, to prove the sufficiency of the condition for maximizing $\bar{c}_l$, we need only show that any schedule satisfying the condition yields the same value for $\bar{c}_l$. Let A be any scheduling algorithm which satisfies the rule that at the lth critical time no job is waiting for jth quantum where $j \neq k$. Let a_l be the time at which the lth job arrives which will require at least kq seconds of service. The state of the system at a_l will, in general, depend on the algorithm A. In particular, the number of critical times that have occurred prior to a_l (let this be s) is a function of A. Let $E_{\mathrm{A}}\,[c_l{-}a_l|$ state of system at $a_l]$ be the expected value of $c_l{-}a_l$ under algorithm A conditioned on the state of the system at a_l. The state of the system is given by the number of jobs in the system, the attained service time of each job in the system and s, the number of critical times that have occurred. Thus, we have

$$\begin{aligned}
&E_{\mathrm{A}}\,[c_l{-}a_l|\,\text{state of system at } a_l]\\
&\quad= E_{\mathrm{A}}\,[\text{remaining work in system not requiring a } k\text{th quantum}\,|\,\text{state of system at } a_l]\\
&\quad+ (l{-}s{-}1)E\,[\text{remaining service time for job with } (k{-}1)q \text{ seconds of attained service}]\\
&\quad+ (k{-}1)q\\
&\quad+ \lambda\bar{t}_{(k-1)q}E_{\mathrm{A}}\,[c_l{-}a_l|\,\text{state of the system at } a_l]\;.
\end{aligned} \tag{4.8}$$

But the sum of the first two terms on the right-hand side of this equation is equal to the expected amount of work in the system at a_l given the state at a_l. Thus

$$\begin{aligned}
&E_{\mathrm{A}}\,[c_l{-}a_l|\,\text{state of system at } a_l]\\
&\quad= E_{\mathrm{A}}\,[\text{work in system at } a_l|\,\text{state at } a_l]\\
&\quad+ (k{-}1)q\\
&\quad+ \lambda\bar{t}_{(k-1)q}E_{\mathrm{A}}\,[c_l{-}a_l|\,\text{state of system at } a_l]\;.
\end{aligned}$$

Removing the condition on the state of the system at a_l we have

$$\begin{aligned}
E_{\mathrm{A}}\,[c_l{-}a_l] &= E_{\mathrm{A}}\,[\text{work in the system at } a_l]\\
&+ (k{-}1)q + \lambda\bar{t}_{(k-1)q}E_{\mathrm{A}}\,[c_l{-}a_l]
\end{aligned}$$

or

$$E_{\mathrm{A}}\,[c_l{-}a_l] = \frac{E_{\mathrm{A}}\,[\text{work in system at } a_l] + (k{-}1)q}{1{-}\lambda\bar{t}_{(k-1)q}}\;.$$

But $E_{\mathrm{A}}\,[$work in system at $a_l]$ is not a function of the particular scheduling algorithm and therefore $E_{\mathrm{A}}\,[c_l{-}a_l]$ does not depend on A. Since $E\,[c_l] = E\,[c_l{-}a_l] + E\,[a_l]$ and the right-hand side is independent of A, $E\,[c_l]$ is independent of A. Note that the form of eq. (4.8) depended on A having the property that at c_l there are no jobs in the system waiting for a jth quantum where $j \neq k$. We have now shown that this condition is necessary and sufficient to maximize $E\,[c_l]$ $(= \bar{c}_l)$.

We now show that the general scheduling rule to maximize $W(kq)$ is the same rule which maximizes $\bar{c}_l$ applied for all l. We have

$$W(kq) = \lim_{n\to\infty} \left(\sum_{l=1}^{n} \bar{c}_l - \sum_{l=1}^{n} \bar{a}_l\right)\Big/ n\;. \tag{4.9}$$

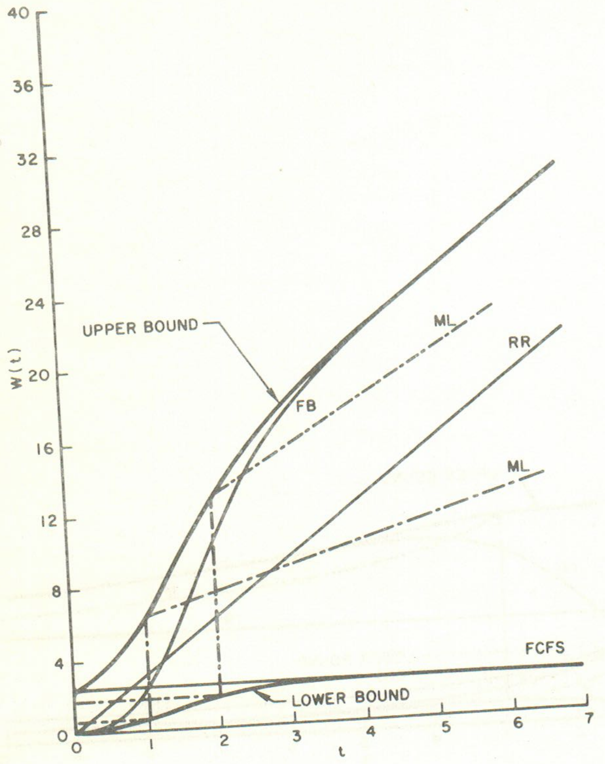

Fig. 6. Bounds on response for $M/E_2/1$, $t = 1.0$, $\lambda = 0.75$, $\rho = 0.75$.

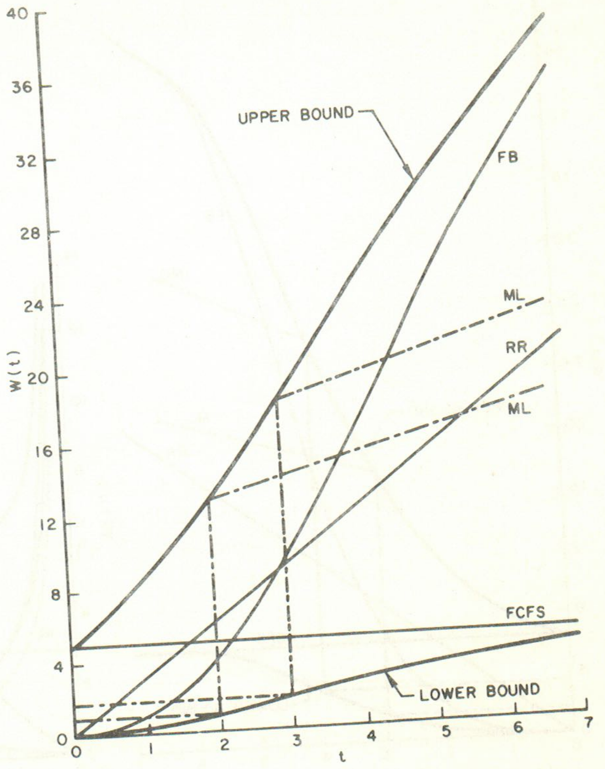

Fig. 7. Bounds on response for $M/H_2/1$, $\bar{t} = 1.0$, $\lambda = 0.75$, $\rho = 0.75$.

infinity; conversely, the most discriminating scheduling algorithm (FB) touches the lower bound at $t = 0$ and forms the asymptote for the upper bound as t approaches infinity. The above-mentioned behavior of the upper and lower bounds applies not only for the M/M/1 system, but also holds true for any M/G/1 system in general, although the rate of convergence for the bounds to their respective limits varies for different service distributions.

For the second example we choose the system $M/E_2/1$. In this system we have

$$\frac{dB(x)}{dx} = (2\mu)^2 x e^{-2\mu x} \quad , x \geqslant 0 \qquad (5.1)$$

with mean service time equal to $1/\mu$; the second moment of this distribution is $3/2\mu^2$. Because the second moment is smaller than that of the exponential distribution (whose value is $2/\mu^2$), the bounds are tighter in this example than the M/M/1 case, just as one would expect. Fig. 6 shows the behavior of this system with $\mu = 1$ and $\lambda = 0.75$. It is obvious from the figure that for $t > 5/\mu$, the upper and lower bounds have essentially reached their asymptotic form.

In the third example we show the bounds for the $M/H_2/1$ system, where H_2 stands for hyperexponential service distribution with

$$\frac{dB(x)}{dx} = 0.5\mu_1 e^{-\mu_1 x} + 0.5\mu_2 e^{-\mu_2 x} \quad , \ x \geqslant 0 \ . \qquad (5.2)$$

We choose $\mu_1 = 5\mu$, $\mu_2 = (5/9)\mu$, resulting in a mean service time of $1/\mu$. The second moment of this distribution is $82/25\mu^2$. Fig. 7 shows the behavior of the $M/H_2/1$ system with $\mu = 1$ and $\lambda = 0.75$. The upper and lower bounds approach their respective limits at a slower rate than either M/M/1 or $M/E_2/1$ because of the larger second moment.

For our last example we choose the system M/U/1 where U stands for uniform service distribution. For this particular example we have

$$\frac{dB(x)}{dx} = \begin{cases} 0.25 & 2 \leqslant x \leqslant 6 \\ 0 & \text{otherwise} \end{cases} \qquad (5.3)$$

and $\lambda = 0.1875$, $\bar{t} = 4.0$, $\rho = 0.75$. Fig. 8 shows the behavior of this system. Notice that when $t \geqslant 6$, the

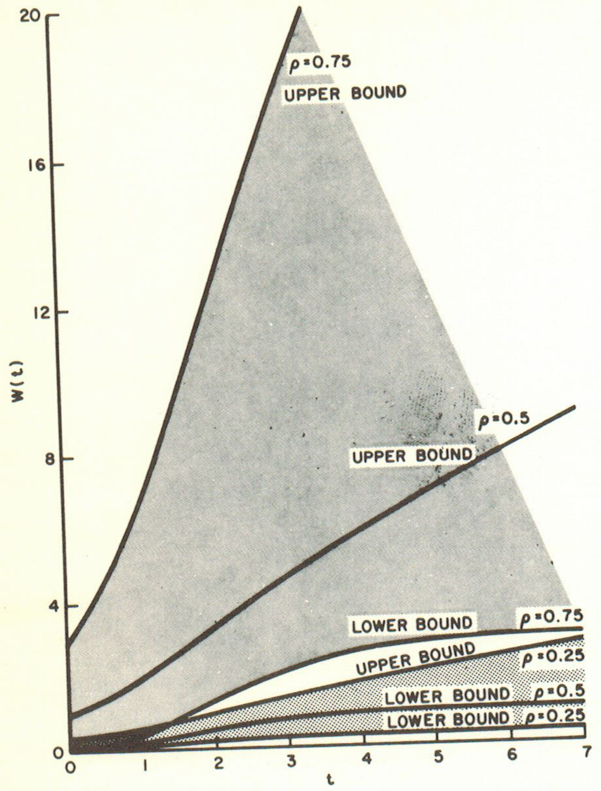

Fig. 10. Variation of bounds for M/M/1 with ρ = 0.25, 0.50, 0.75.

by the curves given in section 5. We note here that although the results were expressed for processor-shared systems, the same type of results apply to the case $q > 0$.

We might observe some additional properties which follow from our results. First we see that any $W(t)$ may touch the lower bound at most once (except over the semi-infinite interval $t_1 \leqslant t$ when $B(t_1) = 1$); the same may be said for the upper bound.

Secondly, we find that we are able to respond to the following kind of specification. Suppose that a designer requests that all jobs of duration $t \leqslant t^*$ should have an average wasted time $W(t) \leqslant W^*$. Then if $W^* \geqslant W_l(t^*)$, it is possible to guarantee at least this behavior (for example, by an ML system where the first level is FCFS out to t^*). Such a specification seems to us to be quite natural. The next obvious need is to specify the bounds on $W(t)$ which exist for $t > t^*$.

Lastly, we pose the more general question which, at the time of this writing remains unsolved, namely, what are the necessary and sufficient conditions for a given response function to be feasible? This paper has presented some important necessary conditions.

REFERENCES

[1] J.M. McKinney, A survey of analytical time-sharing models, Computing Surveys, Vol. 1, No. 2 (June 1969) 105-116.

[2] K.T. Marshall, Some inequalities in queueing, Operations Research, Vol. 16, No. 3 (May-June 1968) 651-665.

[3] K.T. Marshall, Bounds for some generalizations of the GI/G/1 queue, Operations Research, Vol. 16, No. 4 (July-August 1968) 841-848.

[4] J.F.C. Kingman, Some inequalities for the GI/G/1 queue, Biometrika, Vol. 49, 315-324.

[5] D.L. Iglehart, Diffusion approximations in applied probability, Math. of the Decision Sciences (part 2), G.B. Dantzig and A.F. Veinott, Jr., eds., Amer. Math. Soc., Providence, R.I., (1968).

[6] D.J. Daley and P.A.P. Moran, Two-sided inequalities for waiting time and queue size distributions in GI/G/1, Theory of Probability, Vol. XIII, No. 2 (1968) 338–341.

[7] D.P. Gaver, Diffusion approximations and models for central congestion problems, J. of Applied Prob., Vol. 5 (1968) 607-623.

[8] L. Kleinrock and E. Coffman, Distribution of attained service in time-shared systems, J. of Computers and Systems Science, Vol. 3 (October 1967) 287–298.

[9] L. Kleinrock, Swap time considerations in time-shared systems, IEEE Trans. on Computers, (June 1970) 534-540.

[10] L. Kleinrock, Time-shared systems: a theoretical treatment, J. Assoc. Computing Machinery, Vol. 14, No. 2 (April 1967) 242-261.

[11] D.R. Cox and W.L. Smith, Queues, (Methuen, 1961).

[12] L. Kleinrock, A continuum of time-sharing scheduling algorithms, Proc. 1970 SJCC, Atlantic City, (May 1970) 453-458.

[13] L.E. Schrage, The queue M/G/1 with feedback to lower priority queues, Management Science, Vol. 13, No. 7 (1967).

[14] L. Kleinrock and R.R. Muntz, Multilevel processor-sharing queueing models for time-shared systems, Proc. of the 6th ITTC, Munich, Germany, (September 1970) 341/1-341/8.