1975

Some Distribution-Free Properties of Throughput and Response Time

Peter J. Denning

Kevin C. Kahn

Report Number:
75-159

SOME DISTRIBUTION-FREE PROPERTIES OF THROUGHPUT AND RESPONSE TIME*

Peter J. Denning
Kevin C. Kahn

Department of Computer Sciences
Purdue University
W. Lafayette, IN 47907
317-494-8566

Abstract: A system model characterized by networks of service stations is used to study trhoughput and response time as a function of system load, and to characterize the effects of saturation in a system. The results are expressed in terms of only three sets of parameters: the matrix of interstation transition frequencies, the light load service rate of each station, and the service capacity of each station. No assumptions about service time distributions are made. The results hold for arbitrary systems, as long as Little's formula ($\bar{n} = \lambda W$) can be applied approximately to the system and stations in it.

Key Words: Queueing systems, queueing networks, throughput, response time, saturation.

*Work reported herein was supported in part by NSF Grant GJ-41289.

NOTE: It was discovered after this paper was completed that many of the results herein had been developed independently by R. Muntz & J. Wong, "Asymptotic properties of closed queueing network models," in Proc. 8th Princeton Conference on Information Science and Systems, March 1974.

1. INTRODUCTION

Networks of queues have attracted considerable attention as models of computer system performance.* These models have displayed a sometimes uncanny ability to predict utilizations and response times for practical systems, even when these systems fall short of meeting the usual model assumption of exponential service times at the network's service stations (Buz71a, LaS72, Moo71, Mun75, Sch67). It is easy to wonder how crucial the exponential assumptions really are, and how much one can learn without them. A careful study of the literature shows that a great deal can be deduced about the behavior of a system with minimal knowledge of the system. These deductions can be reached without knowledge of the service time distributions at network stations and without appeal to queueing theory.

Our purpose in this paper is to give a unified presentation of some "distribution-free" properties of general networks of service stations. These properties characterize light load and heavy load asymptotes for the throughput and response time of systems as a function of the load on them. They specify the monotonicity and convexity of these functions. They also characterize intermediate loads, at which the error between the bounds and true values is greatest. The data required to parameterize the results in a given system is easily obtained. To the extent that the data reveal system equilibrium during the measurement interval, the results show how to predict waiting times and throughputs for the given system under heavy and light loads, irrespective of the load under which the data were collected.

*The most recent advances have been reported by Baskett et al. for general open or closed networks with one or more classes of customers (BCM75). Applications of these models to multiprogrammed memory management are reported recently by Brandwajn et al. (Bra74, BBG74), by Denning and Graham (DeG75), and by Muntz (Mun75). Of special interest are computationally efficient procedures for evaluating such measures as utilizations, throughputs, and response times (Buz71a, Buz74, CHW75, Mun75).

These results extend and unify results reported by Kleinrock (Kle68), Scherr and Lassettre (Sch67, LaS72), Moore (Moo71), and Chang and Lavenburg (ChL72); they incorporate results reported by Buzen (Buz71a, Buz71b) and by Baskett and Muntz (BaM73).

2. A SYSTEM MODEL

Consider some service system S comprising a network of M stations numbered 1,...,M (Figure 1). A job consists of some number of tasks, each being a demand on some particular station. The tasks of a given job must be performed sequentially; for this reason a job is said to "visit" station i whenever one of its tasks is performed there. A job circulates around the network, visiting stations until all its tasks are completed. Each service station i has a load dependent service rate $b_i(k)$ giving the number of tasks completed per unit time, given that k tasks are present at that station; in other words, $1/b_i(k)$ is the mean time between departures from station i when its load is k. We assume: a) the light load service rate a_i is $b_i(1)$; b) the capacity A_i is the limit of $b_i(k)$ for large k, and is finite for $i = 1,...,M$; and c) the service rate function $b_i(k)$ is nondecreasing and concave down -- i.e., successive increments of load at station i produce increases in service rate in nonincreasing increments. Assumption (c), which holds for most practical systems, is made here to avoid anomalies in system throughput functions which might, for example, result from some station's

switching in extra capacity when its instantaneous load exceeds some threshold. By a load independent station we mean a station i for which $b_i(k) = a_i$ for all $k \geq 1$; for such a station, $a_i = A_i$.

To study the effect of load on the system S we can imagine driving it with a finite source as shown in Figure 2. The number of jobs in the extended system (source and original system S) is held fixed at N. When n jobs are in the system S, the source is independently submitting each of the remaining N-n jobs at rate a_0; that is, the mean time between job submissions by the source is $1/(N-n)a_0$. Under these assumptions we can regard the source as an M^1st station (station 0) in a closed system; its light load service rate is a_0 and its capacity infinite.

Associated with the closed system of Figure 2 is a transition matrix $Q = [q_{ij}]$ giving the relative frequencies of interstation transitions over an observation interval. In other words, over an observation interval containing K task completions, we observe $q_{ij}K$ jobs proceed from station i to station j. Conservation of work requires that the row sums of Q are 1, that is, $q_{i0}+q_{i1}+\ldots+q_{iM} = 1$ for each i. We assume that the matrix Q is determined only by the job's demands and is independent of the number of jobs in the system. We assume also that the system is not decomposable into disjoint subsystems -- i.e., for any i and j it is possible for a job to proceed from station i to station j in one or more transitions.

Associated with station i is a work rate function r_i; it depends on the other work rate functions, the network topology as expressed by the transition matrix Q, and the load N on the system. When we want to make explicit the dependence on load, we shall write $r_i = r_i(N)$. In particular, the system throughput is the same as the source's work rate, r_0. The

interpretation of work rate is that, over an observation interval of length T, we observe $r_i T$ tasks complete at station i. If station i is a load independent station, we may write $r_i = a_i u_i$, where u_i is that station's utilization (over an observation period of length T we observe station i busy for $u_i T$ units of time).

Associated with the system S is the response time function W(N), which is the mean holding time of a job in S, measured from the time it is submitted by the source until the time it returns to the source from S. The system cycle time is the mean time for a job to pass once through the source and the system; it is $W(N)+1/a_0$.

In the above, the light load service rates a_i, the capacities A_i, and the transition matrix Q are the independent parameters. They are easily established: a_i and A_i from service station specifications, Q from measurements on the system. The work rate functions r_i and the response time function W(N) are the derived functions. We shall characterize certain properties of these functions in terms of the independent parameters. Most of the analysis to follow assumes that the system does not generate any backlogs of work, so that work conservation principles and Little's formula (see Appendix) can be applied throughout the network; as shown in the Appendix, this amounts to stipulating that the results can be used with confidence over any observation interval which is long compared to the maximum holding time observed in the system or source during the observation period.

3. INVARIANT PROPERTIES OF WORK RATE FUNCTIONS

3.1. Relative Work Rates

The work rate function r_i must satisfy $0<r_i\leq A_i$ at every load $N>0$.* Equations expressing conservation of work flow can be derived by assuming that, over an observation interval, the same number of tasks enter a station as leave it. Since the flow rate from station j to i is $r_j q_{ji}$, the work rates must satisfy the conservation law

$$r_i = \sum_{j=0}^{M} r_j q_{ji} , \qquad i=0,1,\ldots,M. \tag{3.1}$$

Note that eq. 3.1 holds for every load N. Defining a vector $\underline{r}=(r_0,r_1,\ldots,r_M)$, we can rewrite equations 3.1 in the compact vector form

$$\underline{r} = \underline{r}Q. \tag{3.2}$$

Unfortunately, eqs. 3.2 cannot be solved for a unique vector $\underline{r}$, because there are M+1 unknowns and at most M linearly independent equations; to see this, note that

$$\sum_{i=0}^{M} r_i = \sum_{i=0}^{M} \sum_{j=0}^{M} r_j q_{ji} = \sum_{j=0}^{M} r_j \sum_{i=0}^{M} q_{ji} = \sum_{j=0}^{M} r_j ,$$

an identity resulting from the row sums of Q being 1. The assumption that each station is reachable from any other guarantees that M of the equations 3.2 are independent. Therefore, we can obtain a solution of 3.2 in terms of one of the unknowns, say r_0; but we cannot determine the value of r_0 without additional assumptions.

---

*In fact, if f_i is the fraction of W(1) a long job requires at station i, then $f_i a_i = r_i(1)$, and $f_i a_i \leq r_i \leq A_i$.

Now, if $\underline{r}$ is a solution of 3.2, so is $\underline{R} = \underline{r}/r_0$. In other words, the relative work rates

$$R_i = r_i/r_0, \qquad i=0,1,\ldots,M, \tag{3.3}$$

can be determined uniquely. They are, in fact, obtained by solving the equations

$$\underline{R} = \underline{R}Q \tag{3.4}$$

subject to the constraint $R_0=1$. Because $\underline{R}$ depends only on Q, the relative work rates are unique, independent of load and station service distribution times. Put another way, the work rate functions stand in fixed ratios independent of load:

$$\frac{r_i(N)}{r_j(N)} = \frac{R_i}{R_j}, \qquad N>0. \tag{3.5}$$

That R_i is the number of tasks completed by station i between two job submissions to the system, suggests that R_i can be interpreted as the mean number of tasks v_i generated by a job for station i (that is, R_i is the mean number of visits made by a job to station i). This can be seen more clearly from the following argument. By our assumptions, the mean time to cycle once through the system and the source must be $W(N)+1/a_0$. Since the source work rate is r_0, Little's formula tells that $N = (W(N)+1/a_0)r_0$. Since $R_i = r_i/r_0$, we have $N = (W(N)+1/a_0)(r_i/R_i)$, or

$$R_i = \frac{(W(N)+1/a_0)r_i}{N}.$$

The numerator of this expression denotes the total expected number of tasks

completed by station i during one system cycle time. Since the load is N, 1/N of this number must be attributable to one job. Hence $R_i = v_i$.

That the work rate functions stand in fixed ratios (cf 3.5) implies that the work rate functions must be nondecreasing in system load N:

$$r_i(N+1) \geq r_i(N), \qquad N>0. \tag{3.6}$$

For if an increase in load were to decrease the work rate at any station, the work rate at every station would decrease proportionately — which is patently impossible under an increase in load.

3.2. Incremental Work Rates

Define the incremental work rate at load N to be

$$\Delta_i(N) = r_i(N+1) - r_i(N), \qquad N\geq 0, \tag{3.7}$$

and the vector $\underline{\Delta}(N) = (\Delta_0(N),\ldots,\Delta_M(N))$. Now, observe from eqs. 3.2 that

$$r_i(N)+\Delta_i(N) = r_i(N+1) = \sum_{j=0}^{M} r_j(N+1)q_{ji} = \sum_{j=0}^{M} r_j(N)q_{ji} + \sum_{j=0}^{M} \Delta_j(N)q_{ji} \tag{3.8}$$

Since $\underline{r} = \underline{r}Q$, equality can hold in 3.8 if and only if

$$\underline{\Delta}(N) = \underline{\Delta}(N)Q. \tag{3.9}$$

By analogy with the properties of $\underline{r} = \underline{r}Q$, we see that the work rate increments must stand in the same fixed ratios as the work rates:

$$\frac{\Delta_i(N)}{\Delta_j(N)} = \frac{R_i}{R_j}, \qquad N\geq 0. \tag{3.10}$$

This property has the important consequence that the work rate increments must be nonincreasing in system load N:

(3.11) $$\Delta_i(N+1) \leq \Delta_i(N), \qquad N \geq 0, \qquad i=0,1,\ldots,M.$$

For the failure of 3.11 at any one station would imply its failure at all stations, which is inconsistent with our earlier assumption that every station's service rate function is concave down. Equation 3.11 implies that the work rate functions $r_i(N)$ are concave down.

3.3. Light Load Work Rates

From Little's formula, the system throughput for load N=1 must satisfy

(3.12) $$r_0(1)(W(1)+1/a_0) = 1$$

where $W(1)+1/a_0$ is the mean time the one job requires to complete one cycle through the system and source. If the job requires an average number v_i tasks at station i, the total time it spends there is expected to be v_i/a_i; using $R_i=v_i$,

(3.13) $$W(1) = \sum_{i=1}^{M} \frac{R_i}{a_i} .$$

Noting that $1/a_0 = R_0/a_0$, we obtain from 3.12 that

(3.14) $$r_0(1) = \frac{1}{\sum_{i=0}^{M} \frac{R_i}{a_i}} .$$

Thus the light load system throughput can be obtained without knowledge of service time distributions. The work rate of any other station can be obtained from the relation $r_i(1) = R_i r_0(1)$.

3.4. Saturation Work Rates

Station i is considered to saturate when its work rate approaches its capacity; that is, when

$$\lim_{N\to\infty} r_i(N) = A_i \,. \tag{3.15}$$

In general, it will not be possible for all stations to saturate simultaneously. It is easy to see that

$$\frac{R_i}{A_i} > \frac{R_j}{A_j} \tag{3.16}$$

implies

$$\frac{r_i(N)}{A_i} > \frac{r_j(N)}{A_j} \,, \qquad N>0. \tag{3.17}$$

Thus it is obvious that if $r_i(N)\to A_i$ then

$$r_j(N) \to A_i \frac{R_j}{R_i} < A_j \,, \tag{3.18}$$

where the inequality follows from 3.16.

These observations lead to a simple characterization of saturation. Let s denote the index of any station for which

$$\frac{R_s}{A_s} \geq \frac{R_i}{A_i} \,, \qquad i=0,1,\ldots,M, \quad s\neq 0. \tag{3.19}$$

Then station s is the only one guaranteed to saturate:

$$\begin{aligned} r_s(N) &\to A_s \\ r_i(N) &\to A_s \frac{R_i}{R_s} \,, \quad i=0,1,\ldots,M. \end{aligned} \tag{3.20}$$

(Note that $s\neq 0$ is legitimate, since A_0 is infinite and R_0/A_0 is 0.)

We have shown so far that the relative work rates satisfy the vector equation $\underline{R} = \underline{R}Q$ with $R_0=1$; a similar statement holds for the incremental work rates. The system throughput function $r_0(N)$ is a nondecreasing, concave down function with limit values

$$r_0(N) = \frac{1}{\sum_{i=0}^{M} \frac{R_i}{a_i}} \qquad r_0(N) \to \frac{A_s}{R_s}$$

where s satisfies 3.19. These properties are summarized in Figure 3.

3.5. Relation to Markov Chains

The matrix Q is a stochastic matrix (it has row sums 1). It can be regarded as defining a Markov chain describing the task transition behavior of a job: whenever a job enters state i of the chain, it generates a task for station i of the system; exiting state 0 correspondes to a job's initiating its first task, entering state 0 to a job's completing its final task.

Solving the equation $\underline{p} = \underline{p}Q$ with constraint $p_0+\ldots+p_M = 1$ is equivalent to finding the equilibrium probability vector of the chain. Since the holding times in the various states of the chain are different (they depend on the station service times), p_i cannot be interpreted as the probability of finding a given job at station i. However, for load N=1, the mean holding time $1/a_i$ weighted by the probability p_i is proportional to the fraction of time f_i the job spends at station i:

$$f_i = \frac{p_i/a_i}{\sum_{j=0}^{M} p_j/a_j} \quad . \tag{3.21}$$

Since the relative work rate vector $\underline{R}$ is also a solution of $\underline{R} = \underline{R}Q$, normalizing it will produce the vector p; since $R_0=1$ we must have

(3.22) $$1/p_0 = \sum_{i=0}^{M} R_i \; ; \qquad p_i = p_0 R_i \text{ for } i=1,\ldots,M.$$

Written compactly, the relation between $\underline{R}$ and $\underline{p}$ is simply $\underline{R} = \underline{p}/p_0$.

If a Markov chain has been observed for a large number K of transitions, Kp_i of them are expected to be entries to state i. The mean number of transitions between entries to state 0 must be $K/(Kp_0) = 1/p_0$, and the mean number of entries to state i between visits to state 0 is $Kp_i/Kp_0 = p_i/p_0$. Therefore we expect v_i, the mean number of visits to state i per job, to satisfy $v_i = p_i/p_0$. From 3.22, this implies that $1/p_0$ is the mean number of tasks generated by the job.

4. INVARIANT PROPERTIES OF RESPONSE TIME FUNCTIONS

4.1. Limiting Values of W(N).

Suppose that $\bar{n}_i$ is the mean number of tasks at station i. Obviously $\bar{n}_0+\bar{n}_1+\ldots+\bar{n}_M = N$. Let $\bar{n} = N-\bar{n}_0$ denote the mean number of jobs in the system. From Little's Formula,

(4.1) $$W(N) = \frac{\bar{n}}{r_0(N)} \; .$$

Letting $T_i(N)$ denote the mean holding time at station i, Little's formula gives also

(4.2) $$\bar{n}_i = T_i(N)\, r_i(N)$$

where in particular $\bar{n}_0 = r_0(N)/a_0$. Using this and 4.1, we obtain this alternative expression for W(N):

(4.3) $$W(N) = \frac{N}{r_0(N)} - 1/a_0 \; .$$

(Note that 4.3 also states that $W(N)+1/a_0 = N/r_0(N)$, which is Little's formula applied to a full cycle through the system and source.) Employing 4.2,

$$(4.4) \qquad W(N) = \sum_{i=1}^{M} \frac{\bar{n}_i}{r_0(N)} = \sum_{i=1}^{M} R_i T_i(N) = \sum_{i=0}^{M} R_i T_i(N) - 1/a_0 .$$

The last equality was obtained by adding and subtracting $R_0 T_0(N) = 1/a_0$. By noting that the resulting sum i=0 to M is the full cycle time $N/r_0(N)$, we see the equivalence to 4.3.

Obviously, W(1) of 3.13 is a lower bound on W(N). A lower bound asymptote for large N can be found using 3.20 for the saturating station s:

$$(4.5) \qquad W(N) \approx \frac{N}{r_0(N)} - 1/a_0 = \frac{NR_s}{r_s(N)} - 1/a_0 \geq \frac{NR_s}{A_s} - 1/a_0 .$$

In 4.5, equality is a good approximation for large N.

A proof that $W(N+1) \geq W(N)$ is obtained with the help of Figure 4. Let $W'(N) = W(N)+1/a_0$. The slope of the line from the origin to point A is $r_0(N)/N = 1/W'(N)$. That $r_0(N)$ is concave down implies point A is higher than $r_0(N+1)$. Now, if $W'(N+1) < W'(N)$, we would have

$$\frac{N+1}{r_0(N+1)} = W'(N+1) < W'(N)$$

or

$$\frac{N+1}{W'(N)} < r_0(N+1)$$

but then

$$\frac{N+1}{W'(N)} = \frac{N}{W'(N)} + \frac{1}{W'(N)} = r_0(N) + \frac{1}{W'(N)} < r_0(N+1)$$

which contradicts the concave downness of $r_0(N)$. Accordingly, we must assume $W'(N+1) \geq W'(N)$ and, therefore, $W(N+1) \geq W(N)$.

The properties above are suggested in Figure 5, showing that W(N) is a nondecreasing function originating at the known value W(1) and approaching an asymptote that depends linearly on the load and parameters of the saturating service station. (This suggests that, as load increases, the additional jobs tend to queue up at the saturated station, which then dominates the network. See below.)

We conjecture that, but have not found a satisfactory proof, W(N) is concave up.

4.2. The Saturation Point

The intersection of the two asymptotes in Figure 5 occurs at load N^*, called the saturation point, found by setting $W(1) = NR_s/A_s - 1/a_0$:

$$N^* = \frac{A_s}{R_s} W(1) + \frac{A_s}{a_0 R_s} \quad . \tag{4.6}$$

The intersection of the heavy load asymptote with the horizontal axis occurs at load N_0^*, found by setting $NR_s/A_s = 1/a_0$:

$$N_0^* = \frac{A_s}{a_0 R_s} \quad . \tag{4.7}$$

The importance of the point N^* is that the true value of W(N) deviates from the asymptotes by the maximum amount there; that is, N^* is the value maximizing the uncertainty

$$W(N) - \min\left\{ W(1),\ NR_s/A_s - 1/a_0 \right\} . \tag{4.8}$$

This is a simple consequence of the fact that the difference $W(N)-W(1)$ increases in N and the difference $W(N)-(NR_s/A_s-1/a_0)$ decreases in N.

The points N_0^* and N^* have important interpretations in terms of queueing in the network at the onset of saturation. Suppose we want to find a load L^* such that $N>L^*$ implies that queueing must exist somewhere in the system. At loads $N\leq L^*$, therefore, we may hypothesize for the moment that no queueing occurs anywhere -- i.e., a job's holding time at any station i is $1/a_i$ and in the system $W(1)$. Let $\bar{n}_i$ denote the expected number of tasks at station i, and note that these assumptions and Little's formula imply $r_i(N)/a_i = \bar{n}_i$. Then,

$$(4.9) \qquad N = \sum_{i=0}^{M} \bar{n}_i = \sum_{i=0}^{M} \frac{r_i(N)}{a_i} = \frac{r_s(N)}{R_s} \sum_{i=0}^{M} \frac{R_i}{a_i} = \frac{r_s(N)}{R_s}(W(1)+1/a_0)$$

where s is a station that saturates, and 3.13 was used to reduce the sum. Since $r_s(N) \leq A_s$, we obtain the desired bound on N,

$$(4.10) \qquad N \leq \frac{A_s}{R_s}(W(1) + 1/a_0) = L^* \ .$$

Comparing with 4.6, we see that $L^*=N^*$. Hence N^* represents the maximum load beyond which queueing is certain to occur in the system.

Noting that the mean number of jobs still in the source is $\bar{n}_0 = r_0(N)/a_0 = r_s(N)/a_0R_s$ and that $r_s(N) \leq A_s$, we have

$$(4.11) \qquad \bar{n}_0 \leq \frac{A_s}{a_0R_s} = N_0^* \ .$$

The interpretation of N_0^* is the maximum number of jobs in the source under all loads, being achieved in saturation. It follows that $N^*-N_0^*$ is the average load on the system S beyond which queueing must occur.

A bound on the largest attainable value of N^* can be obtained from 4.9, recalling that $R_s/A_s \geq R_i/A_i$ and $r_s(N) \leq A_s$:

$$(4.12) \qquad N = \frac{r_s(N)}{R_s} \sum_{i=1}^{M} \frac{R_i}{A_i} \frac{A_i}{a_i} + \frac{r_s(N)}{R_s a_0} \leq \sum_{i=1}^{M} \frac{A_i}{a_i} + N_0^* \ .$$

Since A_i/a_i can be interpreted as the effective number of servers in station i, 4.12 implies that the best possible value of N^* is N_0^* plus the total effective number of servers in the system S; this value will be achieved only in a balanced system, one in which the ratios R_i/A_i are all equal.

4.3. Queueing in Saturation

We suggested earlier that the linear asymptote of W(N) for large N suggests that all additional jobs are queueing only at the saturating station. Consider a saturated system, in which $i \neq s$ implies $r_i(N) \cong A_s R_i / R_s < A_i$; were an increase in load to produce further queueing at station i (an increase in $\bar{n}_i$), the fact of unused capacity at station i (viz., $A_i - r_i(N)$) would imply an increase in $r_i(N)$ — a contradiction. Therefore the mean queue length at each nonsaturating station reaches some maximal value as load increases, implying that the extra jobs are queueing at the saturating station.

As suggested in Figure 6, the mean total number of jobs in saturation among stations $i \neq s$ is a constant $\bar{k}$ independent of N in saturation. This means that the asymptotic waiting time can be written as

$$(4.13) \qquad W(N) = R_s T_s + \sum_{\substack{i=1 \\ i \neq s}}^{M} R_i T_i = R_s \frac{N - \bar{k}}{A_s} + \sum_{\substack{i=0 \\ i \neq s}}^{M} R_i T_i - 1/a_0$$

where we used Little's formula at the saturating station to deduce that $T_s A_s = N - \bar{k}$. (This formula for W(N) is given by Baskett and Muntz [BaM73] paraphrasing Moore [Moo71].) This formula, however, does not add significantly

to our knowledge about W(N). Using Little's formula, $\bar{n}_i = r_i T_i = A_s R_i T_i / R_s$ in saturation; thus

$$\bar{k} = \sum_{\substack{i=0 \\ i \neq s}}^{M} \bar{n}_i = \frac{A_s}{R_s} \sum_{\substack{i=0 \\ i \neq s}}^{M} R_i T_i \quad . \tag{4.14}$$

Substituting into 4.13,

$$W(N) = \frac{R_s}{A_s} N - \frac{R_s}{A_s} \frac{A_s}{R_s} \sum_{\substack{i=0 \\ i \neq s}}^{M} R_i T_i + \sum_{\substack{i=0 \\ i \neq s}}^{M} R_i T_i - 1/a_0 = \frac{R_s}{A_s} N - 1/a_0 \ , \tag{4.15}$$

which is identical to our earlier formula for the asymptote of W(N).

With only the assumptions of this paper, but without assumptions about the service distributions of the stations, it is not possible to specify $\bar{k}$.

Using the property that, for $\bar{n}$ in the system and $N-\bar{n}$ in the source, $r_0(N) = (N-\bar{n})a_0$, we find that

$$\bar{n} = N - \frac{r_0(N)}{a_0} \quad . \tag{4.16}$$

From this and earlier results, it is clear that for N=1,

$$\bar{n} = \frac{W(1)}{W(1) + 1/a_0}$$

and

$$\bar{n} \geq N - \frac{A_s}{R_s a_0} = N - N_0^*$$

and that $\bar{n}$ is concave up and nondecreasing in N. These properties are displayed in Figure 7.

4.4. Generalizations

Consider the generalization of Figure 8, showing system S_0 (a generalization of the source) driving system S. If the saturating station s is in S_0, the intersystem flow $r_0(N)$ will approach a constant and the response

time of S will approach a constant. If the saturating station s is in S, the response time of S will follow a curve as shown in Figure 9, under these assumptions:

1. The matrix Q used to solve $\underline{R} = \underline{R}Q$ is the same as before, with state 0 representing the entire system S_0.
2. W(1) is the same as before (eq. 3.13).
3. W_0 is the response of S_0 in saturation, in which case the mean number of jobs in S_0 is N_0^*.

5. EXAMPLE

Figures 10 and 11 show models which have been used for time sharing applications [Bra74, Buz71a, Buz71b, LaS72, Mun75, Sch67], Figure 10 being the classical "machine repairman" model. The parameters of both have been chosen to correspond to typical situations. For the simple system (Figure 10), the work rate equation is trivially $r_0=r_1$, for which the relative work rates are

$$R_0 = 1, \quad R_1 = 1.$$

For the network system, the work rate equations are

$$\begin{aligned} r_0 &= r_1 q_1 \\ r_1 &= r_0 + r_2 + r_3 \\ r_2 &= r_1 q_2 \\ r_3 &= r_1 q_3 \end{aligned}$$

for which the relative work rates are

$$R_0 = 1, \quad R_1 = 1/q_1, \quad R_2 = q_2/q_1, \quad R_3 = q_3/q_1.$$

For the parameter values shown in the figures, the following are the formulas pertaining to throughput and response time:

	Simple System	Network System
$\underline{R}$	(1,1)	(1,10,5.45,3.55)
station s	CPU	DISK
$r_0(1) = 1/\sum R_i/a_i$	1/31	1/31
max r_0: A_s/R_s	1.00	1.34
$W(1) = \sum R_i/a_i - 1/a_0$	1.00	1.00
max W: NR_s/A_s	N - 30	0.745N - 30
saturation point N^*	31	41.6
source load bound N_0^*	30	40.3

We chose the parameters so that W(1) is the same in both systems — i.e., each job places the same demands on both. Using the methods of [Buz74] we computed the throughput and response time curves and plotted them in Figures 12 and 13, respectively. In these cases, the curves W(N) are within 10% of their asymptotes when N is about 25% in excess of N^*. Of course the larger percentage errors between W(N) and W(1) for $N<N^*$ are not necessarily serious, since the response time function assumes tolerable values in that range. The heavy load asymptotes of the two systems are of course different because the network system has more inherent processing capacity and saturates less rapidly.

6. SUMMARY

We have shown that it is possible to glean a considerable amount of information only from data on interstation transition frequencies, light load service rates, and service capacities. This information allows a characterization of asymptotes for throughput (work rate) and response time functions, with the least error between the true curves and the asymptotes occurring at light and heavy loads. A characterization of the saturation point permits estimating the load at which the error between the asymptotes and functions is maximum.

The analysis exploits work conservation properties and Little's formula for systems in equilibrium. The work conservation principles led to the conclusion that the relative work rates are invariant under load and changes in service time distributions. Little's formula allowed us to obtain relations between waiting times and work rates.

The primary interest of these results lies in their holding for a large class of network systems independently of service time distributions and queueing effects. They permit a designer or system evaluator to determine the limits of a system, without having to drive it to its extremes or to take extensive measurements.

APPENDIX - Little's formula as an Approximation

Consider a service system which, over a long period of measurement (0,T), is observed to contain a mean number of jobs $\bar{n}$, to complete them with an average holding time in the system W, and to have a throughput rate of r jobs per unit time. Suppose that n_m and w_m are, respectively, the maximum number of jobs in the system at one time and the maximum holding time of any job in the system, during the measurement period. Then $Wr = \bar{n}$ is a good approximation whenever $n_m w_m/T$ is small compared to $\bar{n}$.

Index the jobs i=1,2,... in order of arrival to the system; let t_{ia} and t_{id} denote respectively the arrival and departure times of job i, and $w_i = t_{id}-t_{ia}$ denote its holding time in the system. Define the "job presence function" H(i,t) to be 1 if $t_{ia} \leq t \leq t_{id}$, and 0 otherwise. Let n(t) denote the number of jobs in the system at time t>0, and observe that

$$n(t) \doteq \sum_{i=1}^{\infty} H(i,t) .$$

Over the observation interval (0,T), the mean number in the system is estimated as

$$\bar{n} = \frac{1}{T}\int_0^T n(t)\,dt = \frac{1}{T}\sum_{i=1}^{\infty}\int_0^T H(i,t)\,dt$$

Let A denote the number of arrivals in the observation interval. Then the righthand expression can be written

$$\frac{1}{T}\sum_{i=1}^{A} w_i - \varepsilon , \qquad \varepsilon>0,$$

where the error ε cannot exceed $n_m w_m/T$ (for $t_{ia}>T$ the integral is 0; for $t_{id}\leq T$ it is w_i; and otherwise it is $T-t_{ia} \leq w_i \leq w_m$ and there are at most n_m such jobs in the system). Thus

$$\bar{n} = \left(\frac{A}{T}\right)\left(\frac{A}{T}\sum_{i=1}^{A} w_i\right) - \varepsilon .$$

By noting that the left parenthesized term is the estimate of the arrival rate r, and the right is the estimate of mean holding time W, we have $\bar{n} = rW - \varepsilon$. Since $\varepsilon < n_m w_m / T$ is by assumption small compared to $\bar{n}$, the equation $\bar{n} = rW$ is a good approximation.

REFERENCES

[Bam73] Baskett, F., and Muntz, R., "Networks of queues." Proc. 7th Princeton Conf. on Info. Sci. and Syst., Dept. Elec. Engrg., Princeton Univ. (March 1973).

[BBG74] Brandwajn, A., Buzen, J., Gelenbe, E., and Potier, D., "A model of performance for virtual memory systems." Proc. ACM SIGMETRICS Symposium (October 1974), 9.

[BCM75] Baskett, F., Chandy, K., Muntz, R., and Palacios, F., "Open, Closed, and mixed networks of queues with different classes of customers." J. ACM 22, 2 (April 1975), 248-260.

[Bra74] Brandwajn, A., "A model of a time sharing virtual memory system solved using equivalence and decomposition methods." Acta Informatica 4 (1974), 11-47.

[Buz71a] Buzen, J., "Queueing network models of multiprogramming." Ph.D. thesis, Div. Engrg. and Appl. Sci, Harvard Univ. (1971).

[Buz71b] Buzen, J., "Analysis of system bottlenecks using a queueing network model." Proc. ACM SIGOPS Workshop on Syst. Perf. Eval. (April 1971), 82-103.

[Buz74] Buzen, J., "Computational algorithms for closed queueing networks with exponential servers." Comm. ACM 16, 9 (September 1973), 527-531.

[ChL72] Chang, A., and Lavenburg, S., "Work rates in closed queueing networks with general independent servers." IBM T. J. Watson Research Report RJ-898 (March 1972).

[CHW75] Chandy, K., Herzog, U., and Woo, L., "Parametric analysis of queueing networks." IBM J. of R. and D. 19, 1 (January 1975), 36-42.

[DeG75] Denning, P., and Graham, G. S., "Multiprogrammed memory management." Proc. IEEE on Interactive Computer Systems (June 1975).

[Kle68] Kleinrock, L., "Certain analytic results for time shared processors." Proc. IFIP Congress 1968, 838-845.

[LaS72] Lassettre, E., and Scherr, A., "Modelling the performance of the OS/360 time sharing option (TSO)." In Statistical Computer Performance Evaluation (W. Freiberger, Ed.), Academic Press (1972), 57-72.

[Moo71] Moore, C., "Network models for large scale time sharing systems." TR-71-1, Dept. Indust. Engrg., Univ. of Michigan, Ann Arbor (April 1971).

[Mun75] Muntz, R., "Analytic modeling of interactive systems." Proc. IEEE on Interactive Computer Systems (June 1975).

[Sch67] Scherr, A., An Analysis of Time Shared Computer Systems. M.I.T. Press (1967).

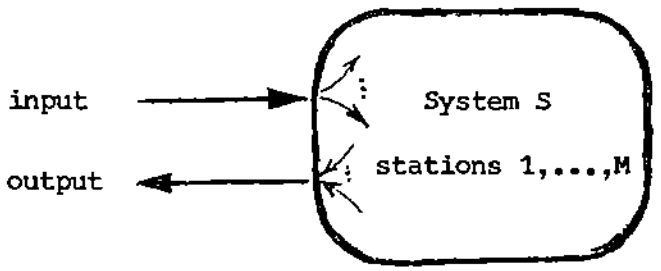


Figure 1. A system.

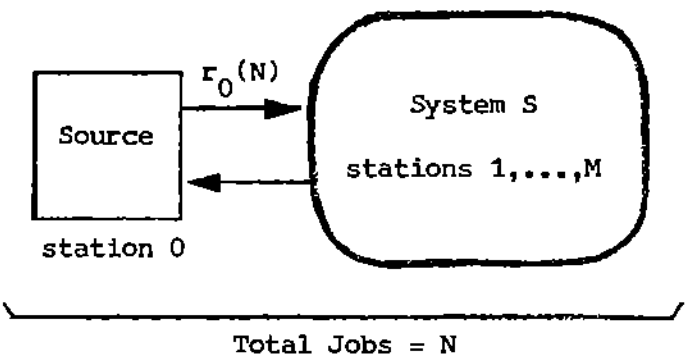


Figure 2. Driving system with a source.

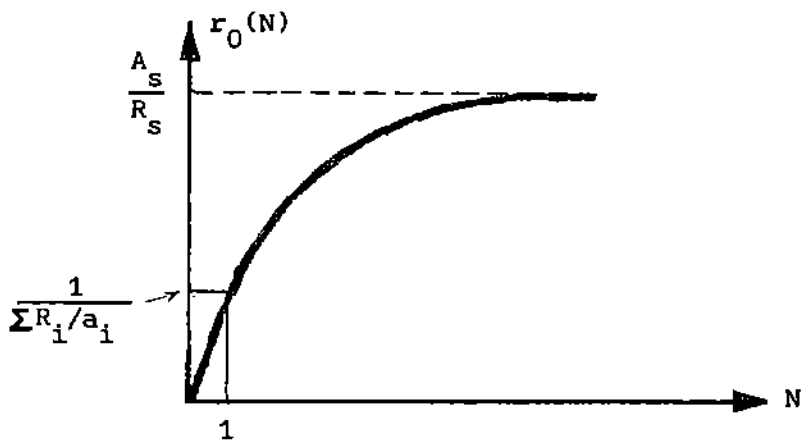


Figure 3. System throughput function.

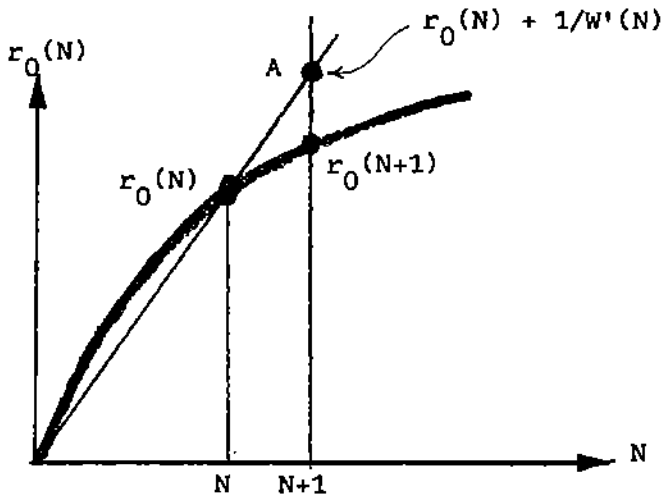


Figure 4. Showing W'(N) increasing.

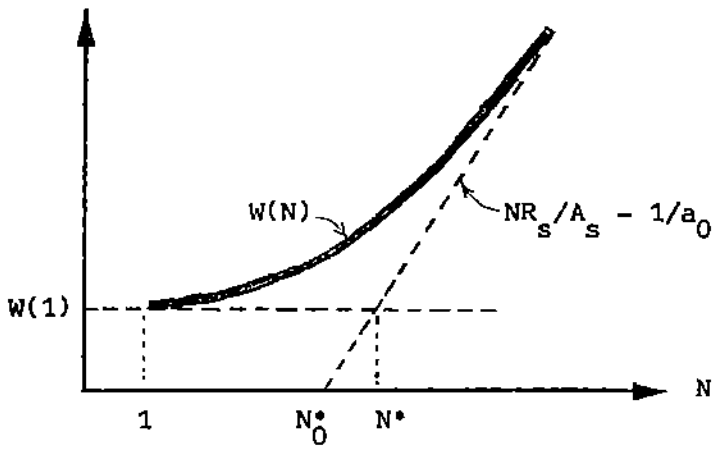


Figure 5. Response time curve.

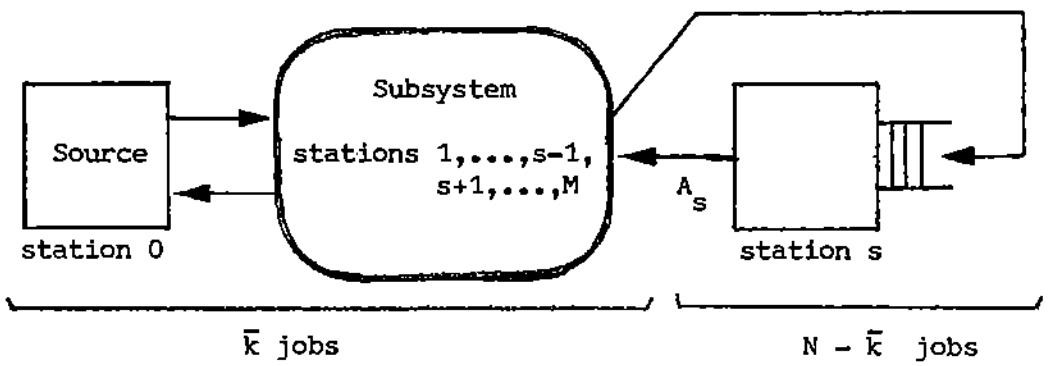


Figure 6. Queueing at saturated station.

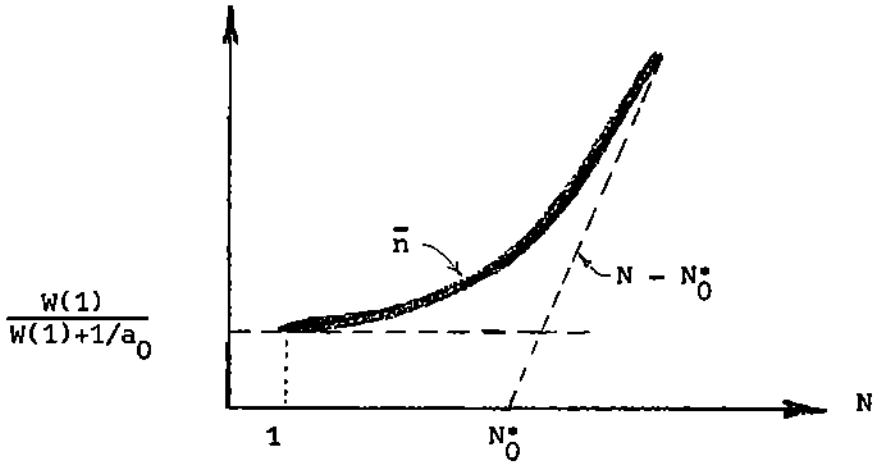


Figure 7. Mean number in system.

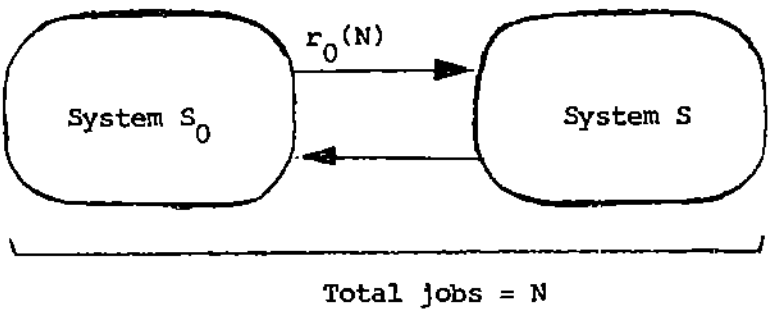


Figure 8. Generalized system.

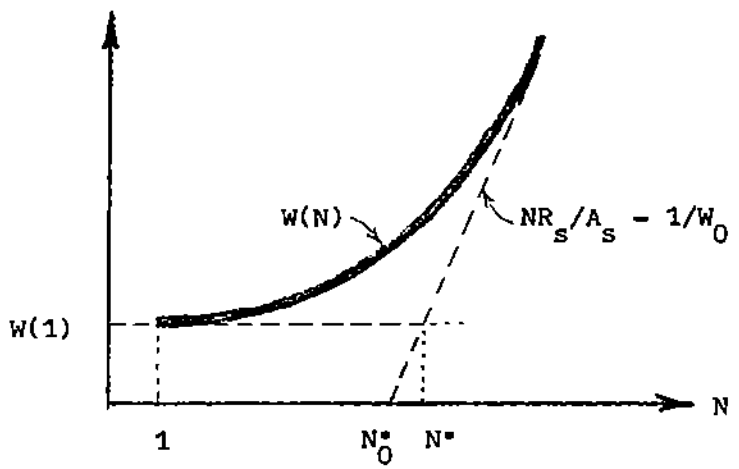
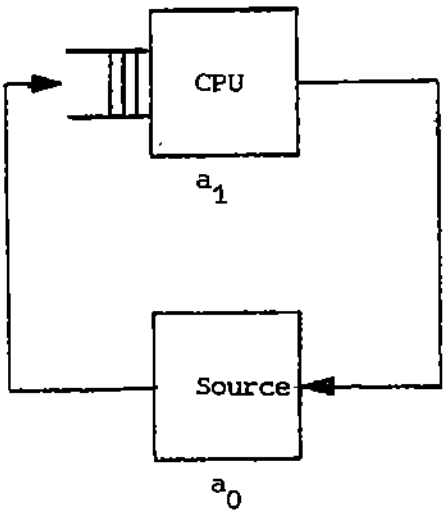


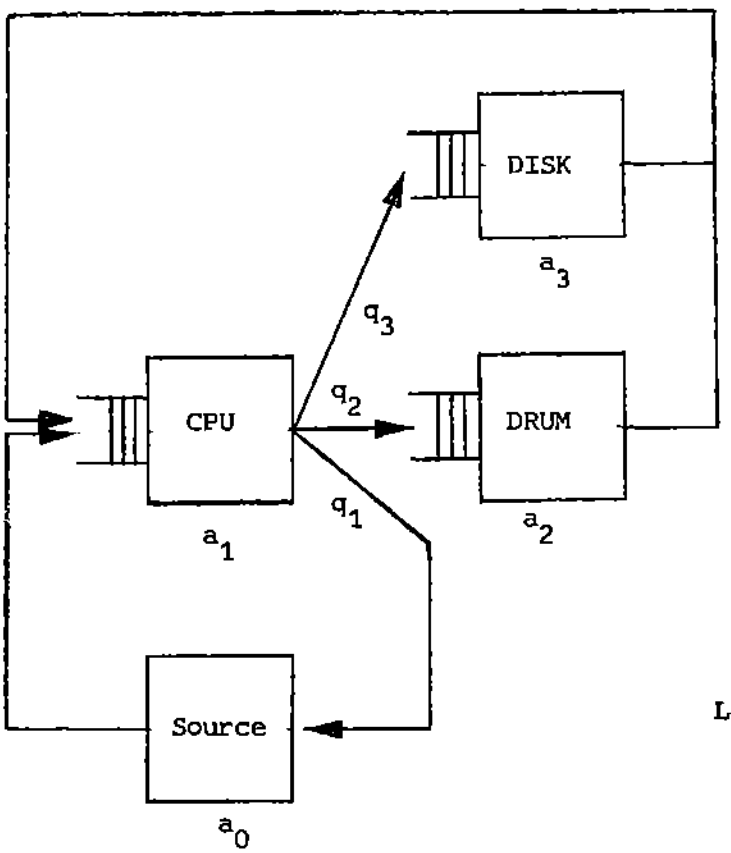
Figure 9. Response time of S with generalized source.



$1/a_0$ = 30 sec.

$1/a_1$ = 1 sec.

Figure 10. Simple System.



$1/a_0$ = 30 sec.

$1/a_1$ = 20 ms.

$1/a_2$ = 10 ms.

$1/a_3$ = 210 ms.

q_1 = 0.1

q_2 = 0.545

q_3 = 0.355

Load independent ($a_i=A_i$)

Figure 11. Network system.

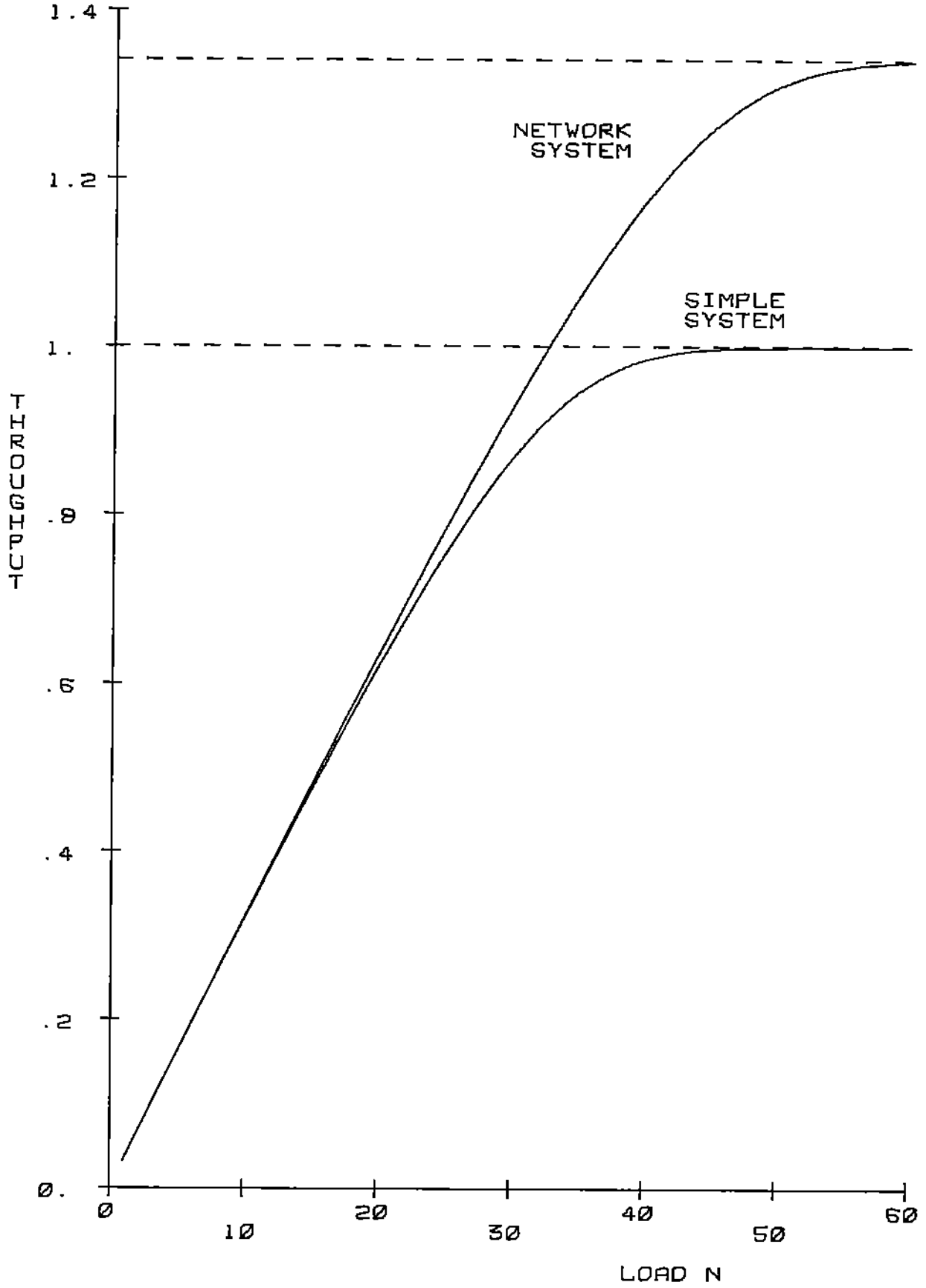


Figure 12. Throughput curves.

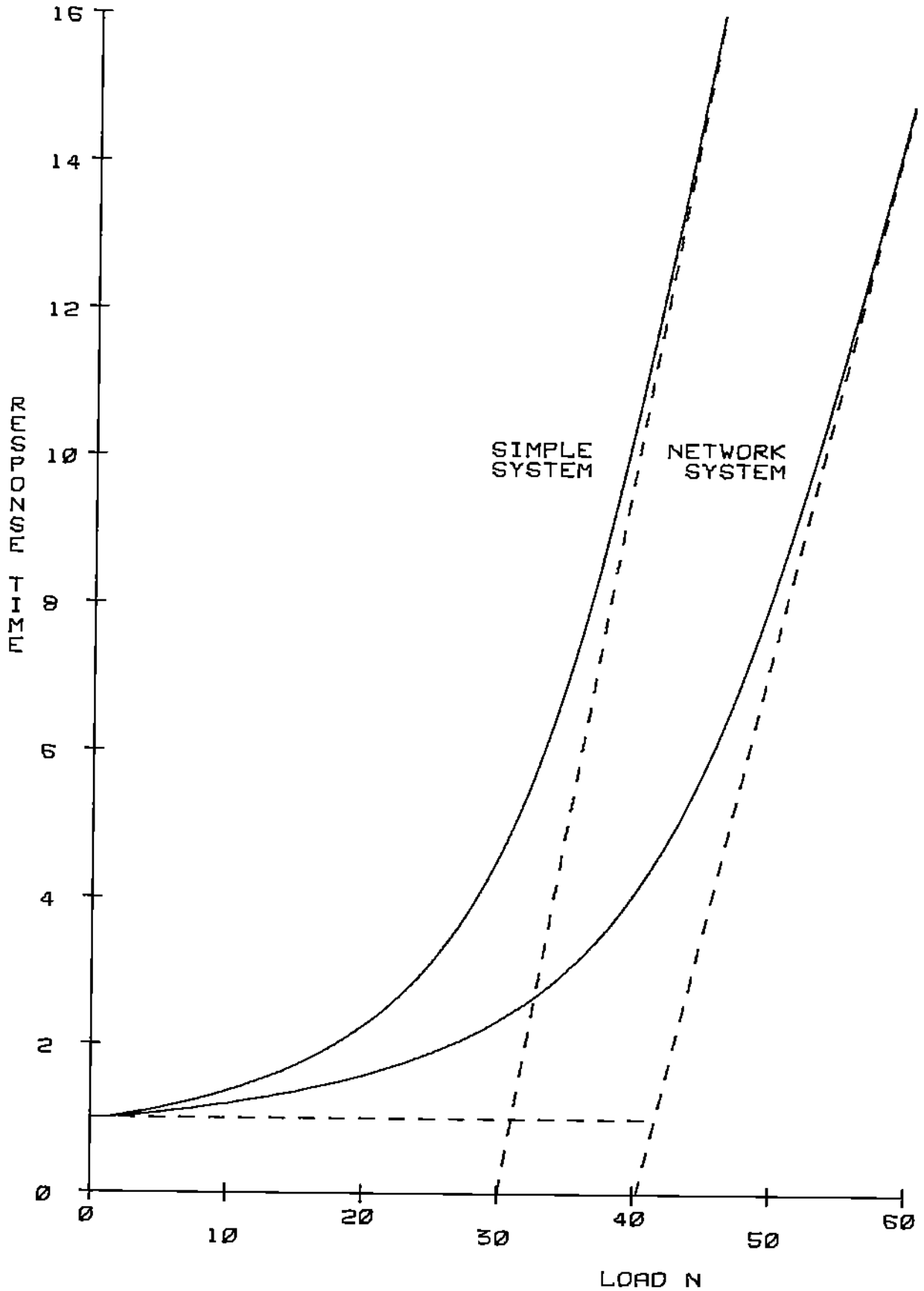


Figure 13. Response time curves.