1979

Measuring and Calculating Queue Length Distributions

Jeffrey P. Buzen

Peter J. Denning

Report Number:
79-317

MEASURING AND CALCULATING QUEUE LENGTH DISTRIBUTIONS*

Jeffrey P. Buzen
BGS Systems, Inc.
Box 128
Lincoln, MA 01773

and

Peter J. Denning
Senior Member, IEEE
Computer Science Department
Purdue University
W. Lafayette, IN 47907

November 1979
Revised December 1979
CSD-TR-317

This is a preprint of a paper to appear in IEEE Computer, February 1980.

*This work supported in part by NSF Grant MCS78-01729 at Purdue University.

MEASURING AND CALCULATING QUEUE LENGTH DISTRIBUTIONS

J. P. Buzen and P. J. Denning

Abstract: Queue length distributions can be described in three ways: as seen by an arriving job, by a completing job, and by an outside observer. We show: the arriver's distribution and the completer's distribution are the same if flow is balanced; the arriver's distribution is a renormalization of the outside observer's distribution if the arrival rate is independent of queue length (homogeneous); when arrivals are not exactly homogeneous, the error introduced by assuming it is can be quantified. The generalized "birth-death recursion" can be used to calculate an arbitrary queue's length distribution and a "prime queue recursion" can be used if the queue is part of a closed network. These results lead to mean value analysis, a simple method for computing mean response times and queue lengths in closed queueing networks. The basis for the entire discussion is operational analysis, whose simple assumptions help explain why queueing models are robust.

INTRODUCTION

Queueing network theory is the mathematical foundation for most analytic models of computer systems and communications networks. In the early 1970's, performance analysts began employing the equations of classical stochastic queueing network theory to calculate throughputs, queue lengths, and mean response times for real systems. Numerous experimental studies have shown that these models consistently estimate the real throughputs to within 5% and the real mean response times to within 25%. Queueing network models are robust.

The startling success of queueing network models caused a good deal of puzzlement among performance analysts. The derivations of the main results of stochastic queueing theory assume that the queueing network has time-invariant parameters, is in steady state, and has exponential distributions of service time at all FIFO (first in first out) devices. These assumptions are often seriously violated in practice. Yet, the models work.

The equations to stochastic queueing theory are relations among abstract quantities that cannot be observed directly. To simplify and solve the equations, the assumptions of time-invariant parameters, steady state, and exponential service distributions are introduced. To apply to solutions, the analyst substitutes measured or estimated values for the abstract parameters; he then compares the calculated results with other measured values. The important point is that the analyst replaces the abstract interpretation by an operational one whenever applying queueing equations.

This realization led to a new hypothesis for explaining the success of stochastic queueing theory: stochastic equations, when interpreted as relations among operational (measurable) quantities, have alternate derivations that depend on assumptions commonly satisfied in practice. This hypothesis is correct. In 1976 operational analysis was introduced as a framework for studying queueing systems during given periods [1,2]. The systems may be real or hypothetical, and the time periods may be past, present or future.

Operational variables represent quantities that can be measured by observing a system during any given interval, called the observation period. Operational analysis derives relations among operational

quantities and studies their consequences. Some relations, called laws, hold in every observation period [2,3]; others depend on assumptions that may be expressed in terms of more primitive operational quantities. The underlying principle is that all variables stand for observable values and all assumptions are experimentally verifiable. (It is not necessary, however, to actually observe the values or run the experiments for an operational analysis to make sense.)

Many of the informal, intuitive arguments used to motivate stochastic theorems become rigorous proofs in the formal context of operational analysis. Besides simplifying derivations, operational analysis extends stochastic theorems by demonstrating their validity in cases where the conventional stochastic assumptions do not appear justified. Operational analysis has led to new results about sensitivity factors and error bounds; these results are particularly valuable for prediction because the validity of operational (or stochastic) assumptions in future periods is uncertain.

In 1978 we published a tutorial on operational analysis of queueing network models [3]. We present here new topics not covered in that paper. We will study the three basic queueing distributions that can be measured at a device in the system: the queue length seen by an arriving job, by a completing job, and by an outside observer. We will study the relations among these distributions. For closed queueing network models, we will study the property that a job arriving at a given device sees the same queueing distribution as the outside observer would see with one less job (the arriver) in the system. This result is the basis of "mean value analysis", a new method introduced in 1978 by Reiser and Lavenberg [4] for numerically evaluating queueing network models.

THE THREE DISTRIBUTIONS AT A QUEUE

A single-resource queueing system comprises an input port, an output port, a queue, and a service facility. All arriving jobs (customers) enter by the input port and all completing jobs exit by the output port. The queue contains all jobs waiting for or receiving service. The service facility consists of one or more processing units.

Suppose that a single-resource queueing system is observed for an interval [0, T]. The state of the queue at time t, denoted n(t), is the number of jobs present. It varies from a minimum of 0 to a maximum of N during [0, T]. (A nonzero minimum on observed queue length changes the boundary condition but not the nature of the results.) A record of n(t) for $0 \le t \le T$ is called a behavior sequence of the system.

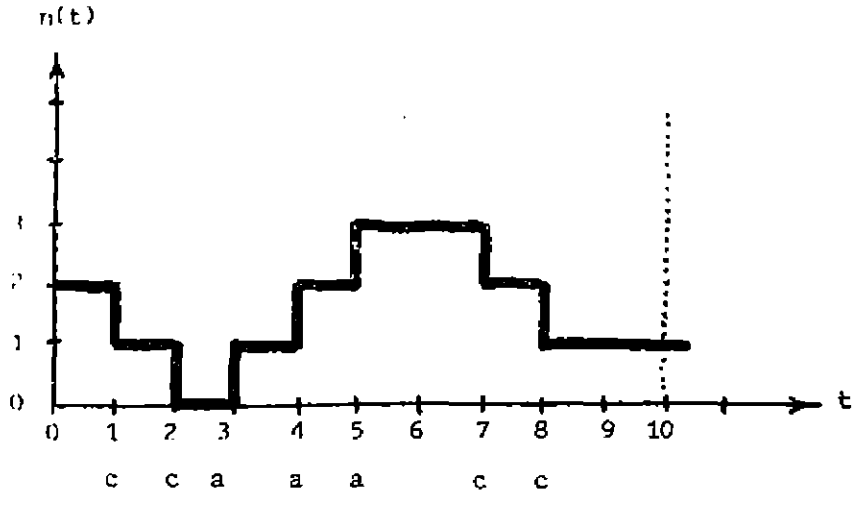
Figure 1(a) illustrates the behavior sequence of a queue for a 10-second observation (T = 10). There are 3 arrivals (marked by "a") and 4 completions (marked by "c").

The behavior sequence of Figure 1(a) satisfies the one-step assumption. This means that n(t) can change only in steps of plus or minus one, and that no arrival coincides with a completion. Each state transition from n to n+1 corresponds to an arrival; each state transition from n+1 to n corresponds to a completion. All behavior sequences in this paper are of this type.

There are three basic operational quantities for a given one-step behavior sequence:

A(n) - The number of arrivers who find n(t) = n
(i.e., who cause a state transition from n to n+1).

(a)



(b)

n	p(n)	$p_A(n)$	$p_C(n)$
0	1/10	1/3	1/4
1	4/10	1/3	2/4
2	3/10	1/3	1/4
3	2/10	-	-

FIGURE 1. Example showing that the three distributions may be different.

C(n) - The number of completions who leave when n(t) = n (i.e., who cause a state transition from n to n-1).

T(n) - The total time during which n(t) = n.

Note that A(n) counts entries at the input port, C(n) counts exits at the output port, and T(n) measures the holding times of states. Note that A(N) = 0 and C(0) = 0 since no arrival can occur while the queue is full and no completion can occur while the queue is empty. Grand totals are defined as follows:

$$A = A(0) + A(1) + \ldots + A(N-1)$$

$$C = C(1) + \ldots + C(N-1) + C(N)$$

$$T = T(0) + T(1) + \ldots + T(N-1) + T(N)$$

Box 1 lists various operational quantities that can be derived from these three basic quantities. These are the definitions from our 1978 paper [3]. The quantity S(n), which gives the mean time between completions for queue length n, is called the service function. The quantity Y(n), which gives the rate of arrivals for queue length n, is called the arrival function. The functions S(n) and Y(n) should not be confused with distributions of service times and arrival times, respectively; service and arrival distributions are not needed for the analysis of interest here. We distinguish the restricted arrival rate (Y) from the overall rate (Y_0). The restricted rate gives the arrival rate over just those intervals when arrivals are possible; arrivals are impossible when the queue is full (n(t) = N). Some of our results are exact only when expressed in terms of the restricted arrival rate.

Given the three basic quantities, we can define the three queueing distributions measurable at any single-resource queueing system:

$p(n)$ = Overall Distribution, the proportion of time n jobs are in the system,
= $T(n)/T$ for $n = 0,\ldots,N$

$p_A(n)$ = Arriver's Distribution, the fraction of arrivers who find n other jobs in the system,
= $A(n)/A$ for $n = 0,\ldots,N-1$

$p_C(n)$ = Completer's Distribution, the fraction of completions who leave behind n other jobs in the system
= $C(n+1)/C$ for $n = 0,\ldots,N-1$

Note that C(n+1), which counts state transitions from n+1 to n, also counts the number of completions who left behind n other jobs. Cooper refers to p(n) as the outside observer's distribution, to $p_A(n)$ as the arriving customer's distribution, and to $p_C(n)$ as the departing customer's distribution [5]. Figure 1(b), which illustrates these distributions for the behavior sequence of Figure 1(a), shows that, in general, the three distributions are different.

The overall distribution, p(n), is used primarily for calculating queueing statistics such as mean and variance of queue length. The arriver's distribution, $p_A(n)$, is used to calculate mean response times, which depend on the queue length experienced by the arriver. (Calculating response time distributions, however, is much harder [10].) The completer's distribution, $p_C(n)$, is seldum used because it is almost always identical to the arriver's distribution (this will be demonstrated below).

It is sometimes easier to derive one of the distributions and convert it to another. For example, the analysis of a Poisson-arrival, general-

service queue easily yields the distribution $p_C(n)$, while the analysis of the general-arrival, exponential-service queue easily yields $p_A(n)$. (See [5] or [6].) The subsequent sections study important relations among the distributions.

Box 2 is a summary of operational laws for these queueing quantities. Each law can be verified by substituting from the preceding definitions and reducing to an identity. These relations are called "laws" because they are valid for every possible behavior sequence [2,3]; logically, they are tautologies. Relations (6) and (7) of Box 2 have special importance. The utilization law, $U = XS$, states that the proportion of time a single-resource queueing system is busy (U) is the product of the mean time between completions and the output rate. Little's Law, $R = Q/X$, states that the mean response time per visit of a job to the queueing system is the ratio of the mean queue length to the output rate. Because it allows calculating response time from previously calculated values of mean queue and output rate, Little's law plays a central role in mean-value analysis, which is discussed later.

Flow Balanced Behavior Sequences

Flow Balance means that the overall arrival rate Y_0 is equal to the output rate X. It is equivalent to the condition that the total number of arrivals A is equal to the total numbers of completions C, and also to the condition that the initial state n(0) is equal to the final state n(T).

If the behavior sequence is flow-balanced and one-step, the number of transitions from state n to state n+1 must equal the number of transitions from state n+1 to state n:

$$A(n) = C(n+1) \qquad n = 0,\ldots,N-1.$$

But this implies that

$$p_A(n) = p_C(n) \qquad n = 0,\ldots,N-1.$$

Thus, flow balance and one-step behavior imply that the arriver's and completer's distributions are identical. In stochastic queueing theory, the steady-state arriving customer's distribution is identical to the steady-state departing customer's distribution; but the proof, which involves limits for infinite time, is more complex [6, pp 176 and 232].

If the given behavior sequence is one-step but not flow balanced, the difference between A(n) and C(n+1) cannot exceed 1. For this reason, the arriver's and completer's distributions seldom differ by much in practice. We will ignore the completers distribution hereafter.

Two important recursions hold in flow-balanced behavior sequences for the overall distribution and the arrivers distribution, respectively; [1,7]:

$$p(n) = S(n)\ Y(n-1)\ p(n-1) \qquad n = 1,\ldots,N$$

$$p_A(n) = S(n)\ Y(n)\ p_A(n-1) \qquad n = 1,\ldots,N-1$$

These are called operational birth-death recursions. They are easily proved by substituting the definitions of Box 1 and applying the flow balance condition A(n) = C(n+1). The birth-death recursions can be used to claculate values of a distribution from measurements or estimates of the service function and arrival rates. To calculate the overall distribution, for example, start with a positive value of p(0), say p(0)=1; iteratively compute p(1), p(2),..., p(N) from the birth-death recursion; and then normalize by dividing each p(n) by the sum p(0)+p(1)+...p(N).

The operational birth-death recursions produce the same formal distribution as the steady-state balance equations for a "birth-death queue" with state-dependent Poisson arrivals and state-dependent exponential service [5,6]. Note, however, that no Poisson or exponential assumptions were required to prove the operational-birth-death equations. These equations are valid for any one-step, flow-balanced behavior sequence, whether generated by a stochastic process or not. In other words, the birth-death recursions apply in cases where the conventional stochastic assumptions cannot be justified. This helps explain the robustness of these results.

The service function S(n) need not be related in an obvious way to the distribution of service requests at the queue; it can be different from the mean of the service requests even at a single-server queue [3,7,8,9,10]. An algorithm developed by Marie can be used in many cases to calculate the service function S(n) from the parameters of the distribution of service requests [8, p45].*

---

*Marie gives two algorithms. One calculates 1/S(n) for an M/G/1 queue with state-dependent arrival rate and service distribution of stage type [8, p45]. This algorithm is exact. It is valid also for a queue with operationally homogeneous arrivals and stage-type service. The second algorithm uses the first as a subroutine to calculate an approximate solution to an arbitrary queueing network [9]. Validation studies have shown that this algorithm does better than other known approximations in nonMarkovian networks with high coefficients of variation at the service stations [19].

Homogeneous Behavior Sequences

To apply the birth-death equations to an arbitrary system, one must measure or estimate the arrival function Y(n) for n = 0,...,N-1 and service function S(n) for n = 1,...,N, a total of 2N values. Now, the number of independent variables can be reduced significantly by making one or both of these assumptions:

$$Y(0) = Y(1) = \ldots = Y(N-1) = \text{constant}$$

$$S(1) = S(2) = \ldots = S(N) = \text{constant}$$

Equations (2) and (3) of Box 2 imply that the arrival constant must by Y, the restricted arrival rate. Likewise, Equation (4) of Box 2 implies that the service constant must be S, the overall mean time between completions.

The assumption that all Y(n) = Y is called homogeneous arrivals; it asserts that the arrival rate is independent of the queue size n. The assumption that all S(n) = S is called homogeneous services; it asserts that the mean time between completions is independent of n. These assumptions are examples of the general operational technique of simplifying problems by replacing a set of conditional values with a single, unconditional value corresponding to the average over the set. A further discussion of these assumptions may be found in our tutorial [3].

Homogeneity assumptions reduce the number of independent variables and thereby simplify both the algebraic form and the interpretation of the resulting equations. The simplified equations are of interest only when there is reason to believe that they will acceptably characterize actual performance. Long behavior sequences of real systems often do approximately satisfy these assumptions.

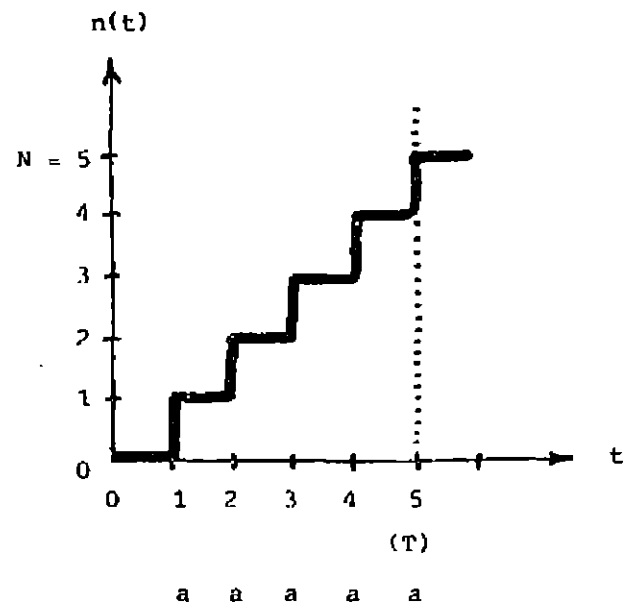
Since the aggregated ("macro") values of arrival rate (Y) and

mean time between completions (S) are usually easier to measure or estimate than the individual ("micro") values Y(n) and S(n), the homogeneity assumptions have another benefit: they enable the analyst to use independent variables that can be measured or estimated with high confidence. Even if the homogeneity assumptions are not satisfied precisely, it is often better to proceed as if they were, because approximate solutions based on a small set of stable variables are often more robust than exact solutions based on a large set of unstable variables.

Box 3 summarizes the consequences of homogeneous arrivals and services. (The proofs, which are straightforward applications of the laws in Box 2 and the birth-death equations, are elaborated in [11].) The main consequence of homogeneous arrivals is that the arriver's distribution is a renormalization of the overall distribution; the renormalization takes account of the arriver's distribution's being zero for a full queue (n(t) = N). This result does not require flow balance; to emphasize this point, Figure 2 presents a flow imbalanced behavior sequence with T = 5, N = 5, p(N) = 0, and $p_A = p$.

The main consequence of homogeneous services is the equation for mean response time, $R = S(Q_A+1)$. This equation says that the mean response time R is the same as the time that the set of Q_A+1 jobs in the queue just after an arrival would take to complete if each required exactly S seconds of service. This interpretation is valid even though the mean residual time of the job in progress at an arrival instant is not necessarily equal to S.

The main consequences of both types of homogeneity are the formulae for mean queue length and response time, which can be applied



$$p(n) = p_A(n) = 1/5$$

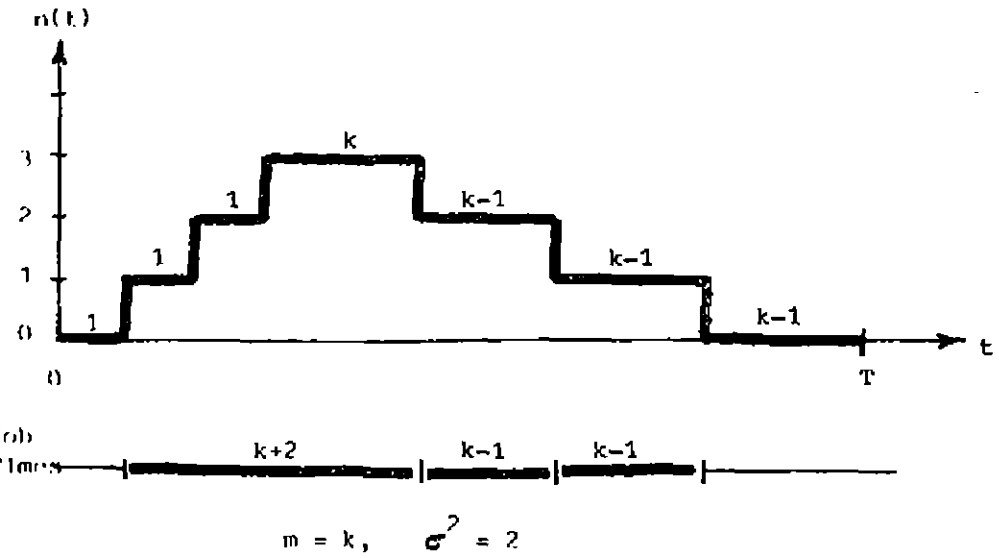
for $n = 0,\ldots,4$

FIGURE 2. A flow imbalanced behavior sequence for which $p = p_A$.

when just the values for utilization (U), the mean time between completions (S), and the proportion of time the queue is full (p(N)) are given.

The equations in Box 3 are operational counterparts of well-known stochastic results. They have the same form but different interpretations. For example, the result $p_A = p$ is familiar to those who have studied the unbounded Poisson-arrival, general-service queue in steady-state. Saaty [12, p 186] outlines a lengthy proof given by Khintchin [13], but Kleinrock's proof is shorter [6, p 118]. Cooper derives Equation (8) for the steady-state arriving customer's and outside observer's distributions for a Poisson-arrival, exponential-service queue with bounded waiting room of capacity N [5]. Equation (11) is the response-time formula usually encountered for queues with exponential service distributions [5,6]. Equation (12) is the mean queue usually derived for a Poisson-arrival, exponential-service queue with bounded waiting room of capacity N, and Equation (13) arises for the same queue with unbounded waiting room [5,6].

Because the equations in Box 3 have the same mathematical form as well-known stochastic results, it is legitimate to inquire whether the operational concepts of flow balance and homogeneity are equivalent to their stochastic counterparts, steady-state and Markovian assumptions. (Markovian assumptions include Poisson arrivals and exponential services.) They are not. The operational concepts are measurable properties of all behavior sequences whereas their stochastic counterparts are precisely observable only for infinite behavior sequences. Figures 3 and 4 are behaviors of a one-step, flow balanced system with homogeneous



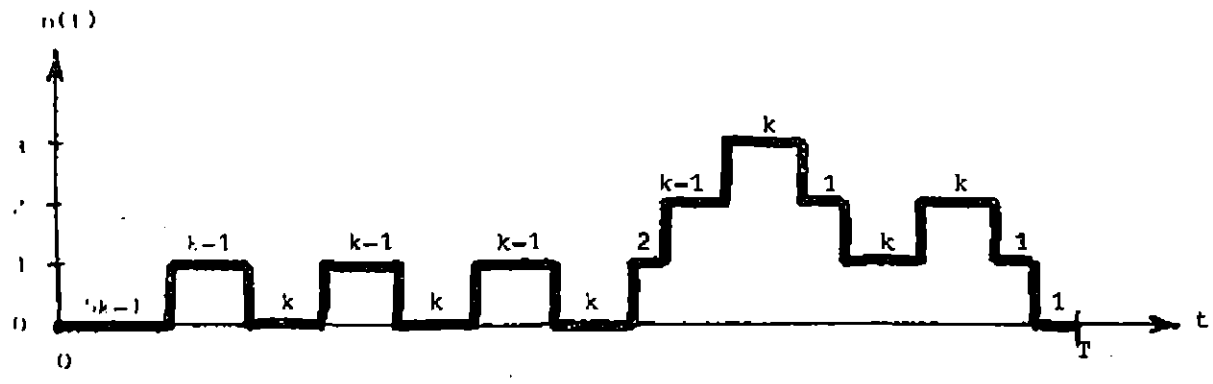
$m = k, \quad \sigma^2 = 2$

	n	A(n)	C(n)	T(n)	p(n)	$p_A(n)$	Y(n)	S(n)
	0	1	-	k	1/4	1/3	1/k	-
	1	1	1	k	1/4	1/3	1/k	k
	2	1	1	k	1/4	1/3	1/k	k
N =	3	-	1	k	1/4	-	-	k
TOTALS		3 (A)	3 (C)	4k (T)				

$$Y = \frac{A}{T-T(N)} = \frac{1}{k}$$

$$S = \frac{T-T(0)}{C} = k$$

FIGURE 3. A homogeneous, flow balanced behavior sequence of a single server queue.



$m = k \qquad \sigma^2 = (4k^2-2k+6)/7$

n	A(n)	C(n)	T(n)	p(n)	$p_A(n)$	Y(n)	S(n)
0	4	-	8k	8/15	4/7	1/2k	-
1	2	4	4k	4/15	2/7	1/2k	k
2	1	2	2k	2/15	1/7	1/2k	k
N - 1	-	1	k	1/15	-	-	k
TOTALS	7 (A)	7 (C)	15k (T)				

$$Y = \frac{A}{T-T(N)} = \frac{1}{2k}$$

$$S = \frac{T-T(0)}{C} = k$$

FIGURE 4. Homogeneous, flow balanced behavior sequence of a single server queue.

arrivals and services. In both cases the mean service time (m) is the same as the mean time between completions (S); the coefficient of variation of service time, which is the ratio of the standard deviation (σ) to the mean (m), differs significantly from 1, the value for an exponential distribution. These service times are not well modelled by exponential distributions. A similar statement holds for interarrival times.

We can extend either sequence by repeating the given pattern indefinitely. The resulting behavior sequence, being periodic, has no steady-state limit; and, being deterministic, it will fail any statistical test for goodness-of-fit to exponential service times (or Poisson arrivals) at any given level of confidence. In other words, the extended replication of Figures 3 and 4 are non-steady-state, nonMarkovian behavior sequences that satisfy the operational conditions of flow balance and homogencity. Whereas a homogeneous behavior sequence need not be Markovian, a sufficiently long Markovian behavior sequence will be homogeneous.

The extended replications of the examples of Figures 3 and 4 are contrived to make the point that the operational assumptions of flow balance and homogeneity are weaker than their stochastic counterparts. It is easy to imagine a deterministic system with the prescribed behavior; given the knowledge that the system is deterministic, a reasonable observer could not explain the extended replications by postulating a Markovian stochastic process. Nevertheless, the formulae originating in Markovian queueing theory are valid in this case because they can be derived under operational assumptions that are satisfied. This fact helps explain the robustness of classical queueing formulae.

Analysis of Sensitivity to Assumptions

In operational analysis it is possible to study the error introduced by simplifying assumptions such as homogeneity. Box 4 illustrates two analyses of the error in estimates of the arriver's distribution caused by an assumption of homogeneous arrivals. The relative error in an estimate of a value $p_A(n)$ is the same as the relative error in the homogeneity assumption, and the weighted relative error in the whole distribution is proportional to the weighted relative error in the homogeneous arrival assumption. Similar analyses are possible for errors caused by flow balance [14] or by homogeneous services.

Sensitivity analyses of this type measure the error introduced by the assumptions of an analysis. Such analyses have no stochastic counterparts. Some stochastic analyses do use the Ergodic Theorem and the Law of Large Numbers to study the error between an estimate of a stochastic parameter and the postulated true value, or between a steady-state probability and a corresponding proportion of time. But these analyses focus only on the error caused by observing a stochastic process during a finite interval, not on the error caused by possible violations of Markovian assumptions. In contrast, operational analysis deals explicitly with finite intervals; it is a tool for studying the errors caused by possible violations of flow balance or homogeneity.

QUEUEING NETWORKS

The preceding discussion deals with single-resource queueing systems. Many real systems are networks of such queues. Models of closed queueing networks are the basis of almost all successful analyses of computer systems. Closed queueing networks can be applied when the "system" includes both the computing devices and the users who submit jobs [3]; they can be applied to a subsystem of fixed capacity operating under a backlog [3]; and they can be applied to communications networks with a fixed number of packets circulating between a given pair of communicants [15]. Although they require exponential service distributions at FIFO queues, closed network models deal with multiple job classes and general service distributions at processor sharing servers, delay servers, and LIFO (last in first out) servers [3,14,15,18].

The discussion of previous sections applies to any queue, whether it is part of a network or not. When a queue belongs to a network, its arrival function is determined jointly by the output (service) functions of the other queues that feed it. For this reason it is helpful to derive additional results that relate queueing distributions to all the service functions throughout a network.

The main parameters of a closed queueing network are the service functions, $S_i(n)$, defined as in Box 1, and the visit ratios V_i [3]. A visit ratio V_i is the average number of times a job visits device i during its sojourn in the system. A system has an input port and an output port; in a "closed network," a new job is admitted (from an external backlog) as soon as a job leaves the system. We let $X_0(N)$ denote the system throughput for network load N, that is, the job flow from the output port to the input port. The throughput at device i is

given by the forced flow law [3]:

$$X_i(N) = V_i \, X_0(N)$$

The overall mean response time can be calculated from the mean response times per visit to the devices, according to the response time law [3]:

$$R_0(N) = V_1 R_1(N) + \ldots + V_K R_K(N)$$

(K is the number of devices in the system.)

Box 5 summarizes the principal results for the queueing distributions in closed queueing networks. The Prime Queue Recursion relates the overall distribution at the next smaller load, using the product of the device's throughput and service function as the scale factor. It is similar to the birth-death recursion in its iteration over queue length, but different in its iteration over network load. The Prime Queue Recursion holds for any flow-balanced behavior sequence that satisfies network-homogeneity. Network homogeneity means that the job flow rate between any pair of devices depends only on the queue length at the source of the flow [3,16]. Like arrival and service homogeneity for a single queue, network-homogeneity is not a Markovian assumption; Sevcik and Klawe have presented examples of network-homogeneous behavior sequences that do not satisfy Markovian assumptions [17]. Sufficiently long Markovian behavior sequences are network-homogeneous. Network-homogeneity does not imply homogeneous arrivals or homogeneous services at any device.

Box 5 suggests that the Prime Queue Recursion can be used to calculate a queue distribution. There is, however, some question about the numerical stability of this algorithm. The original algorithms,

based on computing normalizing constants, appear to be more robust [20,25].

Box 5 also points out the important relation between the arriver's distribution and overall distribution at a device in a flow balanced, network-homogeneous behavior sequence of a closed queueing network. Each arriver sees the same distribution as the outside observer with one job (the arriver) removed. In other words, each arriver acts as an outside observer of the device when it arrives. This result is the basis of mean value analysis, a technique for calculating queueing network statistics.

Mean Value Analysis

Mean value analysis is a new technique for computing mean response times, throughputs, and queue lengths at devices in closed queueing networks [4, 15, 23, 24]. We will present the method in its simplest form -- for networks with a single class of jobs and with homogeneous services.

Mean value analysis uses Little's Law, the Forced Flow Law, and the arriver's-distribution theorem (Equation 16 of Box 5) to calculate iteratively mean values of response time, throughput, and queue length for successively larger values of the network load N. The iteration stops when the desired load is reached. Three basic equations are used at each stage of the iteration. Under the assumption of homogeneous services, Equation 11 of Box 3 tells us that the mean response time per visit to device i must be, for network load N, $R_i(N) = S_i(Q_{Ai}(N) + 1)$. Now, the arriver's-distribution theorem (Equation 16 of Box 5) tells us that the mean queue length seen by an arriver is the same as the overall mean queue length for one less job in the network: $Q_{Ai}(N) = Q_i(N-1)$. Thus we have

$$R_i(N) = S_i(Q_i(N-1) + 1). \tag{A}$$

This is the first basic equation of mean value analysis.

The response time law tells us that the mean response time across the input/output ports of the system, $R_0(N)$, is the sum, over the devices, of the product of the mean response time per visit and the mean number of visits. When applied to the entire system, Little's Law tells us that the mean response time, $R_0(N)$, is the product of the number in the system, $Q_0(N)$, and the throughput of the system, $X_0(N)$. Solving for the system throughput, we have $X_0(N) = Q_0(N)/R_0(N)$. But because the system

is closed $Q_0(N) = N$. Thus

$$(B) \qquad X_0(N) = N \Big/ \sum_{i=1}^{K} V_i R_i(N).$$

This is the second basic equation of mean value analysis.

The forced flow law states that the throughput at device i, $X_i(N)$, is $V_i X_0(N)$. Little's law tells us that the mean queue length at device i, $Q_i(N)$, is the product of the throughput, $X_i(N)$, and the mean response time per visit, $R_i(N)$. Thus

$$(C) \qquad Q_i(N) = V_i \; X_0(N) \; R_i(N).$$

This is the third basic equation of mean value analysis.

Box 6 illustrates a calculation for a simple three-device network. The iteration begins with load N = 0 and all mean queue lengths zero. It then applies Equations (A)-(C) until N reaches the desired load.

The values of $R_i(N)$, $X_0(N)$, and $Q_i(N)$ could also be computed by an algorithm based on normalizing constants [25]. When $R_i(N)$ and $Q_i(N)$ must be computed for every server in the network, the normalizing-constant method and mean-value method require approximately the same number of arithmetic operations and the same amount of storage. There are, however, many applications in which the throughput $X_0(N)$ is the only quantity sought; in this case, the normalizing-constant method requires approximately half as much storage as the mean-value method. It remains an open question which of the two methods is numerically the more robust for calculating queue-length distributions. Because

mean-value analysis is new, it is too early to present a comprehensive analysis of its strengths and weaknesses relative to normalizing-constant analysis.

Box 6 also illustrates a simple, but accurate, approximation suggested by P. Schweitzer, Y. Bard, and J. Zahorjan. The approximation finds the mean response times, queue lengths, and system throughput without enumerating these values for smaller loads, but it does involve other types of iterations. It is based on the assumption that changing the load from N-1 to N scales up the mean queue lengths by a factor of $N/(N-1)$.

Mean value analysis can be extended in several ways. Equation (A), the formula for mean response time, is valid at a device with FIFO, LIFO, or processor-sharing scheduling [4,18,19,20,23]. If device i is a "delay server", which has at least as many internal processors as there are jobs in the network, each job will experience the same delay S_i (its own service time) irrespective of the queue length; in this case we replace Equation (A) by

$$R_i(N) = S_i \qquad \text{(all } N\text{)} .$$

In general, we can substitute a response time formula for each particular type of device [24].

Another extension removes the homogeneous-service assumption. The response time per visit to a general, load-dependent device i is, from Little's Law,

$$R_i(N) = Q_i(N)/X_i(N) = \sum_{i=1}^{N} n\, p_i(n,N)/X_i(N) .$$

Applying the prime queue recursion (Box 5), this reduces to

$$(A') \qquad R_i(N) = \sum_{i=1}^{N} n\, S_i(n)\, p_i(n-1, N-1) \ .$$

For the general server, we use Equation A' to compute mean response time (instead of Equation A). To do this, however, we must compute the entire queue distribution for each network load, which implies that Equation C must be replaced with the evaluation of the overall queueing distribution (see Box 5). Because the computations are considerably more complex for a load-dependent device, most implementations are hybrid: they require the user to specify whether each device is load-dependent, then they apply the more complex algorithm only at those devices where it is needed [18,19,20,25].

Another extension is for multiclass systems. Each workload consists of a fixed number of jobs having common visit ratios and service functions. The visit ratio for this case, V_{ij}, is the mean number of visits a type j job makes to device i, and the service function, $S_{ij}(n_{ij})$, is the mean time between completions of type j jobs at device i, given that n_{ij} jobs of type j are present. The algorithms then calculate the quantities $R_{ij}(N)$, $Q_{ij}(N)$, and $X_{0j}(N)$ for each job class j at each network load N. The details of these algorithms are given by Reiser and Lavenberg [4, 15, 23] and by Bruell and Balbo [18, 19, 20].

Mean value analysis can also be extended to approximate solutions for closed queueing networks that are not network-homogeneous. For example, Bard suggests extensions to handle priority scheduling in multiclass networks [24]. Such extensions may ultimately prove to be

the most practical contribution of mean value analysis. Even when the system is network-homogeneous, mean value analysis suggests approximations that work well (especially for heavy loads) without calculating all the intermediate results. Box 6 shows an example. Bard [24] and Reiser [15] present others.

CONCLUSIONS

The derivations of this paper illustrate the power of operational analysis. Although most results were already known as stochastic theorems, the operational proofs are significantly simpler. These proofs also demonstrate that the theorems are valid in many practical cases where stochastic assumptions cannot be justified. This helps explain the robustness of queueing network results in practice. Operational analysis both extends and simplifies stochastic analysis.

This paper has also initiated study of operational bounds on the errors that can arise if homogeneity is only satisfied approximately. Such bounds are important for prediction, because the future validity of all assumptions is uncertain. Since it is difficult to quantify the concept of "approximate validity" for stochastic assumptions such as ergodicity or Poisson arrivals, operational analysis has an advantage over stochastic modeling in this regard.

The discussions of the prime queue recursion, mean value analysis, and approximation methods illustrate that important new results continue to be discovered in the theory of queueing networks. Because of its simplicity and intuitive appeal, mean value analysis is a particularly promising tool for investigating approximate solutions of queueing network models.

Achnowledgement

We are grateful to John Spragins and his referees for their thoughtful and constructive criticisms of a rough version of the manuscript; to John Zahorjan and Y. Bard for pointing out the approximation used in Box 6; and to Martin Reiser for his inspiration. The work was supported in part by NSF Grant MCS78-01729 at Purdue University.

BIBLIOGRAPHY

1. Buzen J.P. "Operational analysis: The key to the new generation of performance prediction tools", Proceedings COMPCON 76- IEEE Computer Society Conference Washington, D.C., September 1976, 166-171.

2. Buzen, J.P. "Fundamental operational laws of computer system performance". Acta Informatica, 7, 2(1976), 167-182.

3. Denning, P.J. and Buzen, J.P., "The operational analysis of queueing network models," Computing Surveys 10, 3 (September 1978), 225-261.

4. Reiser, M. and Lavenberg, S.S., "Mean value analysis of closed multichain queueing networks," IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, Report RC 7023 (March 1978).

5. Cooper, R.B., Introduction to Queueing Theory, New York: Macmillan (1972).

6. Kleinrock, L., Queueing Systems, Vol. 1, John Wiley and Sons, New York (1975).

7. Buzen, J.P. "Operational analysis: An alternative to stochastic modeling", In Ferrari, D., (ed.), Performance of Computer Installations, North-Holland, 1978, 175-194.

8. Marie, R. "Modélisation par réseaux de'attente", Ph.D Thesis, Universite de Rennes, (November 1978).

9. R.A. Marie, "An approximate method for general queueing networks," to appear in IEEE Trans Softw. Engr (Fall 1979).

10. Lazowska, E.D., "Characterizing service time and response time distributions in queueing network models of computer systems," Ph.D. Thesis, U. Toronto, Dept. Computer Sciences, TR-CSRG-85 (October 1978).

11. J.P. Buzen and P.J. Denning, "Operational Treatment of queue distributions and mean value analysis, "Technical Rept. CSD-TR-309, Computer Sciences Dept., Purdue University, W. Lafayette, IN 47907 (August 1979).

12. Saaty, T.L., Elements of Queueing Theory, McGraw-Hill, New York (1961).

13. Khintchin, A., Mathematical Methods in the Theory of Queueing, Charles Griffin and Co., London (1960).

14. J. Buzen, P.J. Denning, D.B. Rubin, L.S. Wright, "Operational Analysis of Markov Chains, "Technical Rept. 79-1, BGS Systems, Inc., Box 128, Lincoln MA 01773 (January 1979).

15. M. Reiser, "A queueing network analysis of computer communication networks with window flow control," IEEE Trans. Communications COM-27, 8 (August 1979), 1199-1209.

16. Denning P.J., and Buzen, J.P., "Operational analysis of qucueing networks," Proc. 3rd. Int'l Symp. on Computer Performance Modeling Measurement, and Evaluation (H. Beilner and E. Gelenbe, Eds.), North-Holland Publishing Co., Amsterdam, The Netherlands (1977).

17. Sevcik, K. and Klawe, M., "Operational analysis versus stochastic modeling of computer systems." Proc. Symp. on the Interface, U. Waterloo, Ontario, Canada (May 1979).

18. S.C. Bruell, "On single and multiple job class queueing network models of computer systems," Ph.D thesis, Computer Sciences Dept., Purdue Univ., West Lafayette, IN 47907 (December 1978).

19. G. Balbo, "Approximate Solutions of Queueing Network Models of Computer Systems," Ph.D Thesis, Computer Sciences Dept., Purdue University, W. Lafayette, IN 47907 (September 1979).

20. S.C. Bruell and G. Balbo, Computational Algorithms for Single and Multiple Class Queueing Network Models, Elsevier/North-Holland Publishing Co. (New York), Series on Programming and Operating Systems (1980).

21. Lavenberg, S.S. and Reiser, M., "Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers", IBM, T.J. Watson Research Center, Report RC 7592 (April 1979).

22. Sevcik, K. and Mitrani, I., "The distribution of queueing network states at input and output instants," Research Report No. 307, IRIA, Rocquencout, France (1978); to appear in J. ACM.

23. Reiser, M. "Mean Value Analysis of Queueing Networks, a New Look at an Old Problem", Proc. 4th. Int'l Symposium on Modeling and Performance Evaluation of Computer Systems, IFIP WG 7.3, North-Holland Publishing Co., Amsterdam, The Netherlands (February 1979).

24. Y. Bard, "Some extensions to multiclass queueing network analysis," Proc. 4th. Int'l Symposium on Modeling and Performance Evaluation of Computer Systems, IFIP WG 7.3 (February 1979). North-Holland Publishing Co., Amsterdam, The Netherlands.

25. Buzen, J.P. "Computational algorithms for closed queueing networks with exponential servers", Commun. ACM. 16,9 (September 1973), pp. 527-531.

Box 1 - Operational Quantities

$S(n)$ = mean time between completions when $n(t) = n$

= $T(n)/C(n)$ (defined only if $C(n) > 0$)

B = total busy time

= $T(1) + T(2) + \ldots T(N)$

S = overall mean time between completions

= B/C

U = utilization

= B/T

X = output rate

= C/T

W = job-seconds of accumulated waiting time

R = $\sum_{n=1}^{N} n\,T(n)$

Q = mean queue length

= W/T

R = mean response time per completed job

= W/C

Y = arrival rate when $n(t) = n$

= $A(n)/T(n)$ (defined only if $T(n) > 0$)

Y_0 = overall arrival rate

= A/T

Y = restricted arrival rate

= $A/(T-T(N))$ (defined only if $T(N) < T$)

Box 2 - Operational Laws

(1) $p_A(n) = p(n)(Y(n)/Y_0)$ (if Y(n) defined)

(2) $Y/Y_0 = 1/(1-p(N))$ (if T(N)<T)

(3) $Y_0 = \sum_{n=0}^{N-1} p(n)\ Y(n)$ (for defined Y(n))

(4) $S = \sum_{n=1}^{N} p_C(n-1)\ S(n)$ (for defined S(n))

(5) $X = \sum_{n=1}^{N} p(n)/S(n)$ (for defined S(n))

(6) $U = SX = 1-p(0)$ [Utilization Law]

(7) $R = Q/X$ [Little's Law]

Box 3 -- Consequences of Homogeneity and Flow Balance [11]

Arrivals to a single-server queue are homogeneous if the arrival function is a constant equal to the restricted arrival rate, that is, if $Y(n) = Y$ for $n = 0,\ldots,N-1$. Services are homogeneous if the service function is a constant equal to the overall mean time between completions; that is, if $S(n) = S$ for $n = 1,\ldots,N$.

If arrivals are homogeneous, the arrivers distribution is directly proportional to the overall distribution:

$$(8) \qquad p_A(n) = \frac{p(n)}{1 - p(N)} , \qquad n = 0,\ldots,N-1,$$

where $p(N)$ is the proportion of time the queue is at its maximum value. This implies that the mean queue length seen by an arriver (Q_A) is related to the overall mean queue length (Q) by

$$(9) \qquad Q_A = \frac{Q - Np(N)}{1 - p(N)} .$$

If $p(N)$ is small, Q_A is approximately Q.

If services are homogeneous, the mean queue length seen by an arriver is related to the overall mean queue length and the utilization by

$$(10) \qquad Q_A = Q/U - 1.$$

Box 3, continued

and the mean response time is

$$R = S(Q_A+1). \tag{11}$$

If arrivals and services are both homogeneous, the overall mean queue length is

$$Q = \frac{U}{1 - U - p(N)} \left(1 - (N+1)p(N)\right). \tag{12}$$

The mean response time is calculated from this by applying Little's Law (BOX 2), $R = Q/X$. If $p(N)$ is small, the overall mean queue length is approximately

$$Q \simeq \frac{U}{1-U}. \tag{13}$$

In many real applications, however, Equation (12) gives better results if an estimate or measurement of $p(N)$ is available.

Box 4 -- Example of an Operational Sensitivity Analysis

Suppose that arrivals are approximately homogeneous. The deviation from exact homogeneity can be expressed so:

$$\left| \frac{Y - Y(n)}{Y(n)} \right| < \varepsilon \qquad \text{for } n = 0,\ldots,N-1$$

Let $\hat{p}_A$ denote the estimate of the arrival distribution obtained by using the homogeneous arrival assumption. The relative error caused by the homogeneous arrival assumption is

$$e_A(n) = \left| \frac{\hat{p}_A(n) - p_A(n)}{p_A(n)} \right| .$$

By Equation (1) of Box 2, each term $p_A(n)$ is equal to $p(n)Y(n)/Y_0$. If arrivals are assumed to be homogeneous, then $\hat{p}_A(n)$ is equal to $p(n)Y/Y_0$. Substituting,

$$e_A(n) = \left| \frac{Y - Y(n)}{Y(n)} \right| < \varepsilon .$$

In words, the relative error in the approximation to $p_A(n)$ is the same as the relative error in the approximation to $Y(n)$.

Relative errors can be deceptive: large errors can sometimes be tolerated if they are infrequent. It is sometimes useful to employ another measure of the deviation from exact homogeneity:

$$\sum_{n=0}^{N-1} \left| \frac{Y - Y(n)}{Y} \right| \frac{p(n)}{1-p(N)} < \varepsilon .$$

The weighted relative error caused by the homogeneous arrival assumption is then

$$e_A = \sum_{n=0}^{N-1} p_A(n)\, e_A(n).$$

Substituting $\left| (Y-Y(n))/Y(n) \right|$ for $e_A(n)$ and $p(n)Y(n)/Y_0$ for $p_A(n)$,

$$e_A = \sum_{n=0}^{N-1} \frac{p(n)Y(n)}{Y_0} \left| \frac{Y - Y(n)}{Y(n)} \right|$$

Cancelling Y(n), multiplying and dividing by Y, and setting Y/Y_0 = 1((1-p(N)) [Equation (2) of Box 2], this reduces to

$$e_A < \varepsilon.$$

Box 5 -- The Prime Queue Recursion

Consider a closed queueing network containing N jobs and K devices. For a given behavior sequence, the overall distribution $p_i(n,N)$ at device i denotes the proportion of time that n jobs are enqueued at device i (n = 0,...N). If the behavior sequence is flow balanced and network-homogeneous (see text), the overall distribution satisfies

$$p_i(n,N) = X_i(N)\ S_i(n)p_i(n-1,N-1) \tag{14}$$

where $X_i(N)$ is the throughput at device i (for network load N) and $S_i(n)$ is the service function for device i. We call this the prime queue recursion because it can be used to derive formulae for many other performance quantities of queueing networks [18, 19, 20]. It was first proved by Reiser and Lavenberg [14].

The prime queue recursion suggests an algorithm for computing a queue distribution given throughputs for various network loads N. Begin with $p_i(0,0) = 1$. Having calculated $p_i(n-1, N-1)$ for $n = 1,...N$, apply Equation (14) to calculate $p_i(n,N)$; then choose $p_i(0,N)$ by subtracting the utilization $U_i(N) = p_i(1,N)+...+p_i(N,N)$ from 1.0. This is repeated for successive N until the desired load is reached. If this algorithm is applied at a bottleneck in the system, it will eventually get to a load N at which $U_i(N)$ is approximately 1.0 whereupon the subtraction to calculate $p_i(0,N)$ may generate a round-off error that propagates to larger values of N. To help control the error, we check that $p_i(0,N) \geq 0$ after the subtraction $1.0-U_i(N)$; if the subtraction yields a negative

Box 5, continued

result, set $p_i(0,N) = 0$. Even with this fix, however, the algorithm seems to generate significant errors for larger N; the normalizing-constant algorithms [20,25] appear to be more numerically stable.

For a given behavior sequence, let $p_{Ai}(n,N)$ denote the arrivers distribution at device i; that is $p_{Ai}(n,N)$ is the fraction of arrivers who find n in the queue at device i (given network load N). If the behavior sequence is flow balanced,

$$p_{Ai}(n,N) = \frac{p_i(n+1,N)}{X_i(N)\ S_i(n+1)} \quad ; \tag{15}$$

this can be proved by substituting the definitions of overall distribution, throughput, and service function and using the flow balance condition to reduce the right side to the left [11]. On applying the prime queue recursion to $p_i(n+1,N)$, this equation reduces to

$$p_{Ai}(n,N) = p_i(n,\ N-1) \quad . \tag{16}$$

Equation (16) was noted informally by Reiser and Lavenberg [4] and later proved by them [21]; it is a special case of a stochastic theorem proved by Sevcik and Mitrani [22]. The details leading to Equation (16) show that is is true operationally [11]. Its interpretation is that the arrivers see the same distribution as the outside observer with one job (the arriver) removed form the network.

Box 6 -- Examples of Mean Value Calculations

Consider a system with K = 3 and these visit ratios and mean service times:

$V_1 = 1$ $V_2 = 2$ $V_3 = 3$

$S_1 = 2$ $S_2 = 1$ $S_3 = 1$ (seconds)

The results of the mean-value calculation up through N = 4 are shown in the table

N	$R_1(N)$	$R_2(N)$	$R_3(N)$	$X_0(N)$	$Q_1(N)$	$Q_2(N)$	$Q_3(N)$
0	-	-	-	-	0.000	0.000	0.000
1	2.000	1.000	1.000	.143	.286	.286	.429
2	2.571	1.286	1.429	.212	.545	.545	.909
3	3.091	1.545	1.909	.252	.779	.779	1.443
4	3.557	1.779	2.443	.277	.985	.985	2.030

We will illustrate the calculation of the mean queue length $Q_3(4)$. We first calculate mean response times $R_1(4)$, $R_2(4)$, and $R_3(4)$ using Equation A. For example,

$$R_2(4) = S_2(Q_2(3) + 1)$$
$$= (1)(.779 + 1)$$
$$= 1.779 \text{ seconds}$$

Then we calculate the overall mean response time:

Box 6, continued (2)

$$R_0(4) = V_1R_1(4) + V_2R_2(4) + V_3R_3(4)$$
$$= (1)(3.557) + (2)(1.779) + (3)(2.443)$$
$$= 14.444 \text{ seconds.}$$

Equation B gives the system throughput:

$$X_0(4) = 4/R_0(4) = 4/14.444 = .277 \text{ jobs/sec.}$$

Finally, Equation C gives the mean queue at device 3:

$$Q_3(4) = V_3X_0(4)R_3(4)$$
$$= (.277)(3)(2.443)$$
$$= 2.030 \text{ jobs.}$$

If we are willing to sacrifice some accuracy in the solution, we can obtain the values R_i, Q_i, and X_0 for a given value of N without enumerating all these quantities for loads smaller than N. A reasonable approximation is that $Q_i(N-1)$ is $(N-1)/N$ as large as $Q_i(N)$. (This approximation was suggested to us by J. Zahorjan, P. Schweitzer and Y. Bard.) Therefore, we seek values of R_i, Q_i, and X_0 satisfying

$$R_i = S_i(Q_i\frac{N-1}{N} + 1) \qquad i = 1,\ldots,K$$
$$X_0 = N/\Sigma V_iR_i$$
$$Q_i = V_i\ X_0\ R_i \qquad i = 1,\ldots,K$$

Box 6, continued (3)

We can obtain a solution iteratively: start with a guess of all the mean queue lengths Q_i (summing to N), and then compute a new guess by evaluating these equations; repeat this until successive guesses for the Q_i do not differ by much. (Termination in guaranteed [24].) This iteration is asymptotically correct for large N [24]. The table below presents the values obtained by applying this approximation for N = 4; the errors between the approximate and the true values (from the previous table) are quite good.

	$R_1(N)$	$R_2(N)$	$R_3(N)$	$X_0(N)$	$Q_1(N)$	$Q_2(N)$	$Q_3(N)$
Exact (N=4)	3.557	1.779	2.443	.277	.985	.985	2.030
Approx (N=4)	3.393	1.697	2.606	.274	.929	.929	2.141
% Error	-4.6	-4.6	+6.7	-1.1	-5.7	-5.7	+5.5