DOCID: 3927962

An Application of Queuing Theory to Computer Systems

BY JOSEPH BLUM

Unclassified

1. INTRODUCTION

1.1 *Importance and Current Status of Queuing Theory.*

Queuing theory has emerged as an established body of knowledge entirely within the twentieth century. Its initial impetus was derived from the study of congestion in telephone traffic, and its nourishment from the theory of stochastic processes. Since that time applications have been made in the fields of communications, transportation, warehousing and the operation of dams; computer systems are now being added to the growing list of applications. As the number of service systems increases and the problems of congestion become more significant queuing theory will be used more frequently as an analytic tool for improving the operation of current service systems and for designing the more powerful systems of the future.

The contributions made by numerous investigators swelled noticeably during the past two decades and these now constitute a large body of literature. Conferences on queuing theory are almost an annual occurrence. With the appearance of textbooks and university courses in queuing theory the field has been firmly established. The reader interested in early developments can find a brief historical resume in Saaty. [14, pp. 20–25]

1.2 *Limitations in the Use of Queuing Theory.*

Many service systems are inherently complex and the abstract models formulated from them contain networks of queues. A specific illustration which involves a multiprogramming computer system has been given in Blum. [3] These models generally resist mathematical analysis and in some instances computer simulation studies have provided a practical alternative. [2] However, even for relatively simple operational situations, the classical queuing models may fail to provide an adequate representation. This point can be illustrated with the following example of a computer system.

Suppose jobs arrive sequentially at a computer system in accordance with a Poisson arrival process, and the execution time of a job is a random variable. Jobs are executed in the order of arrival; if the computer is busy when a job arrives, the job is placed in a queue. Thus far it appears that the operation fits the classical single-server queue de-

Approved for release by NSA on 11-29-2011, Transparency Case# 63852

scribed rather precisely by Feller [7, p. 194] and shown diagrammatically in Fig. 1. In the terminology of queuing theory the computer is the "server" and the jobs are "customers." In this classical model the following rules hold:

(1) the server is in one of two allowable states—idle or serving a customer.

(2) the server is idle only when the queue is empty.

(3) an arriving customer who finds the server idle receives service with no delay.

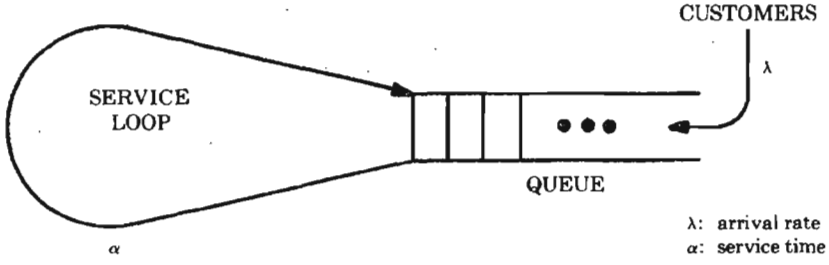

Fig. 1.—Structure of the Classical Single-Server Queue

To resume with our description of the computer system in operation—let us suppose that a job has just been completed. The computer remains idle while the operator unloads magnetic tapes from the completed job and then loads new tapes for the next job. The logical structure of the single-server queuing model can be restored with a simple device: let the server now represent the combined resources of computer and operator. While such devices can be useful in modelling systems, the discerning system designer will see that the emperor is naked in his postulational robes. Now consider the computer operation at the end of a busy period, when there is no job in the queue. The computer stands idle and the operator turns his attention to "other duties as assigned." However, in order to initiate service for a job the operator must be in attendance. Therefore, while the operator is absent, jobs which arrive must wait until he returns. Since it is hard to restore the basic structure of the classical single-server model, it would be better to develop a more suitable model to represent the computer service system described above. A model which will be useful for this system was actually introduced by Skinner [15] for other purposes and ingeniously extended in order to analyze a more complicated system involving two queues.

1.3 *A Modified Single-Server Model.*

The modified model proposed by Skinner is shown in Fig. 2. The following characteristics of the modified model show how it deviates from the rules given above for the classical model:

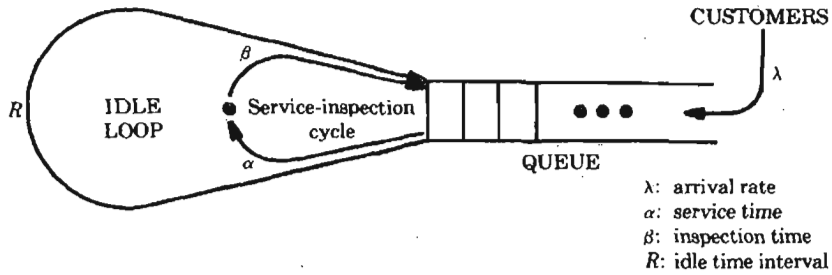

Fig. 2.—Structure of the Modified Single-Server Queue

(1) the server is in one of three allowable states—idle, serving a customer or inspecting the queue.

(2) the server may be idle even though the queue is occupied by customers.

(3) an arriving customer who finds the server idle has to wait for a period of time before service begins.

The customers arrive at the queue in accordance with a Poisson arrival process as in the classical model. The server's activities are governed by an algorithm whose structure is clearly exposed with the use of the following flow-chart (Fig. 3). The server, after completing service to a customer (block 2) inspects the queue (block 3). At the end of the inspection the server decides between two courses of action (block 1): (1) If the queue is empty the server becomes idle (block 4): (2) if the queue is occupied the server takes the first customer in the queue and completes another 'service-inspection' cycle (blocks 2 and 3). The server will repeat the service-inspection cycle until the queue is found to be empty. Note that no time is consumed in block 1. When an inspection reveals the queue to be empty the server becomes idle for a variable time R, after which the server inspects the queue in zero time (block 1) and either begins a service-inspection cycle or again becomes idle, depending on whether the queue is occupied or empty.

1.4 *Realizations for the Modified Model.*

Several examples are offered here as realizations of the modified model. These are suggested by the writer's experience with computers. Other realizations can probably be found among the service systems occurring in commerce and industry.

(1) The computer with operator in attendance.

This example was used above to indicate deficiencies in the classical single-server model. In this realization service time corresponds to the execution of a program. Inspection time is interpreted to be the time to tear down the old job plus the time to

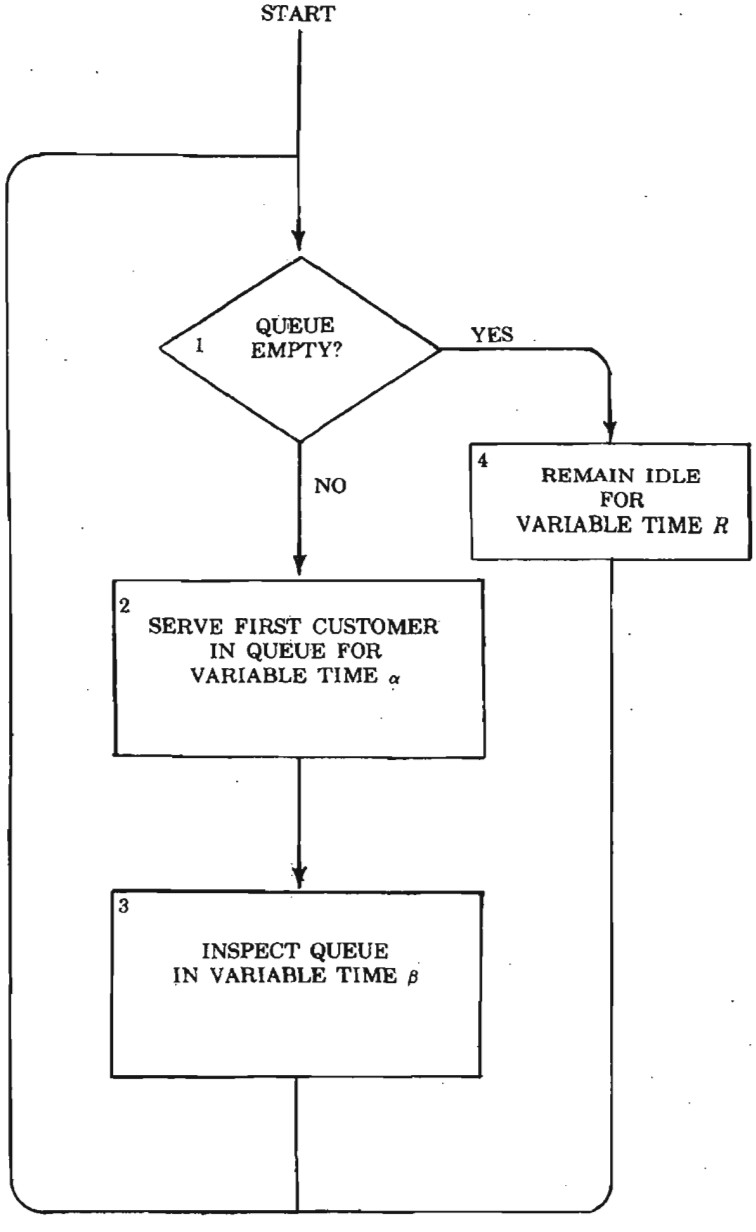

Fig. 3.—Flow-Chart of Server's Activities

set up the new job. Idle time corresponds to whatever the operator does when there are no jobs to run, e.g. a coffee break or other duties as assigned.

(2) The quick-service computer operation.

This mode of operation is often provided at many installations during the day shift. Only short jobs are accepted for this operation. However, it is assumed that an inexhaustible supply of long jobs is available for "background" work whenever the queue of short jobs is empty. Service time corresponds to the execution of a short job. Inspection time corresponds to the time for closing out the old job and setting up a new one. (For some installations this function may be performed by the operating system software.) Idle time corresponds to the time for executing a background job. It should be noted in this example that idle time is not synonymous with lost production. In the interest of providing quick and responsive service some installations accept a loss of production, i.e., the computer will stand idle when the queue becomes empty so that the period can terminate immediately upon the arrival of a job.

(3) The data-processing computer operation.

In this kind of operation large data-processing jobs are run on the computer. The existence of high-priority and low-priority jobs is postulated. Service time corresponds to the execution of a high-priority job. Inspection time corresponds to closing out the old job and setting up a new one. Assuming that low-priority jobs are always available, idle time corresponds to running a low-priority job when the queue of high-priority jobs is empty.

1.5 *Goals.*

The basic goal of this paper is to provide engineers, system analysts and others with an expository presentation containing results which are applicable to the study of certain computer service systems. A modified model of the single-server queue is presented which, for some applications, overcomes deficiencies in the classical single-server queuing model. For this modified model queuing theory will be applied to obtain results concerning the distributions of (1) queue length, (2) response time, (3) idle period, and (4) busy period. The paper attempts to expose the essential mathematical concepts and techniques which contribute to the analysis of the modified model, for they have been successfully applied to other queuing models. Section 2 presents sufficient mathematical material to provide background for the reader having a general familiarity with probability theory. Questions of mathematical rigor have been put aside in favor of emphasizing major concepts and techniques. This paper, furthermore, makes no pretense of providing a comprehensive account of queuing theory as applied to computer service systems.

2. MATHEMATICAL PRELIMINARIES

The reader is assumed to have a degree of mathematical maturity and to be acquainted with the basic concepts of random variables, probability functions, distribution functions, moments and expected values. A quite brief summary of these concepts is furnished herein, and further information can be found in Wilks [17].

2.1 *Random Variables.*

We shall have need to employ both discrete and continuous random variables. In the discrete case we shall be dealing exclusively with non-negative integer-valued random variables. The continuous random variables also will be non-negative.

2.2 *Probablity and Distribution Functions.*

Let G be a discrete random variable and $G_n = \text{Prob}\{G = n\}$; $n = 0, 1, \ldots$

Then G_n is the probability function of the random variable G.

Let X be a continuous random variable and $F(t) = \text{Prob}\{X \leq t\}$. Then $F(t)$ is the distribution function of the random variable X, and the probability that X lies in the interval $(t, t + dt)$ is given by $dF(t)$.

2.3 *Moments and Expected Values.*

Let G be a discrete random variable and g_n its probability function. If ϕ is a single-valued function, then $\phi(G)$, is also a random variable. The expected value of $\phi(G)$, denoted by $E[\phi(G)]$, is defined by

$$E[\phi(G)] = \sum_{k=0}^{\infty} g_k \, \phi(k). \tag{2.1}$$

The expected value of G^n, denoted by M_n, is called the moment of order n of the random variable G and is given by

$$M_n = \sum_{k=0}^{\infty} g_k \, k^n. \tag{2.2}$$

For notational convenience we shall use the symbol $\overline{G^n}$ to denote the moment of order n of the random variable G.

Let X be a continuous random variable and $F(t)$ be its distribution function. The expected value of $\phi(X)$, denoted by $E[\phi(X)]$, is given by

$$E[\phi(X)] = \int_0^{\infty} \phi(t) \, dF(t). \tag{2.3}$$

The expected value of X^n is called the moment of order n of the random variable X. Thus

$$M_n = E[X^n] = \int_0^{\infty} t^n \, dF(t). \tag{2.4}$$

For notational convenience the symbol $\overline{X^n}$ will be used to denote the moment of order n of the random variable X.

2.4 *Generating Functions.*

Let G be a discrete random variable and g_n its probability function. The generation function for g_n, denoted by $G(z)$, is defined by

$$G(z) = \sum_{k=0}^{\infty} g_k z^k. \tag{2.5}$$

It follows from equation (2.5) that

$$G(1) = 1, G'(1) = \overline{G}, \tag{2.6}$$

and $G^{(n)}(1)$ yields the nth factorial moment. The ordinary moments defined in section 2.3 are easily computed from the factorial moments (see Kendall [12, pp. 63–65]).

2.5 *The Poisson Arrival Process.*

We assume that t ranges over $[0, \infty)$ and denote by $t_1, t_2, \ldots t_n, \ldots$ the arrival instants of the customers to a single-server queue. The interarrival times, denoted by θ_n, are given by $\theta_n = t_{n+1} - t_n$ ($n = 0, 1 \ldots; t_0 = 0$) and are assumed to be mutually independent random variables with the same distribution function:

$$F(t) = 1 - e^{-\lambda t}. \tag{2.7}$$

Thus Prob $\{\theta_n \leq t\} = F(t)$; $n = 1, 2, \ldots$ The sequence $\{t_n$ is called Poisson arrival process with density λ and the reader may confirm that $E[\theta_n] = \frac{1}{\lambda}$. Since the mean interarrival time equals $\frac{1}{\lambda}$ it follows that λ equals the mean number of customers arriving per unit time.

Next, let $A_n(t)$ be the probability that n customers arrive during an interval of length t. It can be shown that

$$A_n(t) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}; n = 0, 1, \ldots \tag{2.8}$$

A derivation of this result and of the Poisson process from more basic assumptions can be found in Khintchine [13, pp. 11–15].

Finally, let us compute the value of $A(n; U)$, defined to be the probability that n customers arrive in the interval of length U, where U is a continuous random variable with distribution function $U(t)$. The probability that U is in the interval $(t, t + dt)$ is $dU(t)$ and the probability that n customers arrive during an interval of length t is $e^{-\lambda t} \frac{(\lambda t)^n}{n!}$. Thus the probability that n customers arrive in a random interval whose length is in $(t, t + tdt)$ is given by

$$e^{-\lambda t} \frac{(\lambda t)^n}{n!} dU(t).$$

To obtain $A(n; U)$ we integrate over all possible values of U and obtain

$$A(n; U) = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} \, dU(t). \tag{2.9}$$

2.6 *Laplace-Stieltjes Transforms.*

Let $F(t)$ be the distribution function for the continuous random variable X. The Laplace-Stieltjes transform of $F(t)$, denoted by $F^*(s)$, is defined by

$$F^*(s) = \int_0^\infty e^{-st} \, dF(t). \tag{2.10}$$

It follows from equation (2.10) that

$$F^*(0) = \int_0^\infty dF(t) = 1, \tag{2.11}$$

$$F^{*(k)}(0) = (-1)^k \int_0^\infty t^k \, dF(t) = (-1)^k \overline{x^k},$$

where $F^{*(k)}(0)$ denotes the kth derivative of $F^*(s)$ evaluated at $s = 0$. Further information on this transform can be found in Feller. [7, chapter 13]

2.7 *Convolutions and Their Transforms.*

Let X, Y be independent continuous random variables with distribution functions respectively $F(t)$, $G(t)$. The random variable Z defined by $Z = X + Y$ has the distribution function $H(t)$, where

$$H(t) = \int_0^\infty G(t - x) \, dF(x). \tag{2.12}$$

To justify equation (2.12) we observe that Prob $\{Z \leq t$ and X is in $(x, x + dx)\}$ is equal to Prob $\{Y \leq t - x$ and X is in $(x, x + dx)\} = G(t - x) \, dF(x)$.

The unconditional probability is obtained by integration over all possible values of X. Thus,

$$H(t) = \text{Prob}\{Z \leq t\} = \int_0^t G(t - x) \, dF(x).$$

The integral is called the convolution of the function $F(t)$ with the function $G(t)$. It can be shown that $H^*(s)$ satisfies the equation

$$H^*(s) = F^*(s). \tag{2.13}$$

This result can be extended for a sum of n independent random variables. Let $Y = X_1 + \ldots + X_n$, $F_i(t)$ be the distribution function of X_i. It can be shown that

$$H^*(s) = F_1^*(s) \, F_2^*(s) \ldots F_n^*(s). \tag{2.14}$$

The function $H(t)$ can be shown to be equal to the n-fold convolution of the functions $F_1(t), \ldots, F_n(t)$. For the special case where the X_i are distributed with the same function $F(t)$, we have

$$H^*(s) = [F^*(s)]^n. \tag{2.15}$$

Now let the random variable Z be defined by $Z = X_1 + \ldots + X_G$, where G (restricted to postive integers) is a random variable and every X_i is distributed like the random variable X with distribution function $F(t)$. Let g_n be the probability function for G, $G(z)$ be the generating function of G and let $H(t)$ be the distribution function for Z. Then Prob $\{Z \leq t \mid G = n\} = F_n(t)$, where $F_n(t)$ is the n-fold convolution of $F(t)$ with itself. It follows that

$$H(t) = \sum_{n=1}^{\infty} g_n F_n(t), H^*(s) = \sum_{n=1}^{\infty} g_n [F^*(s)]^n$$

and, therefore,

$$H^*(s) = G[F^*(s)]. \tag{2.16}$$

If we differentiate (2.16) and set $s = 0$ we obtain

$$\overline{Z} = \overline{G}\,\overline{X}, \tag{2.17}$$

a result which the reader might have anticipated on the basis of heuristic reasoning.

2.8 *Takács' Law of the Busy Period.*

We consider the classical single-server queue with a Poisson arrival process of density λ and a service time distribution function $B(t)$. A busy period is an interval of time during which the server is continuously engaged in giving service to customers. The busy period begins when a customer arrives to find the server idle and it terminates when a departing customer sees an empty queue. The idle period is defined as the interval of time between two consecutive busy periods. We denote by G the random variable representing the duration of a busy period and let $G(t)$ be its distribution function. It has been shown by Takács that the transform of $G(t)$ satisfies the functional equation

$$G^*(s) = B^*[s + \lambda - \lambda G^*(s)]. \tag{2.18}$$

This relation, frequently referred to as "Takács' law of the busy period", is discussed in Syski. [16, pp. 535–539] The following abbreviated argument, essentially due to Takács, indicates how equation (2.18) is obtained. Let the first customer served in a busy period engage the server for an interval I, called the "initial occupancy time". Suppose that n customers arrive during the interval I. Since the duration of the busy period does not depend on the order in which customers are served, let us place in an "auxiliary queue" only those customers which

arrive during the interval I. When the interval I terminates, the server takes his next customer from the auxiliary. During this service arriving customers are placed in the regular queue, and a busy period is started which terminates when the server finds no more customers in the regular queue. The server then returns to the auxiliary queue for the next customer and another busy period on the regular queue is initiated. This procedure continues until no more customers are left in the auxiliary queue and the whole busy period elapses with the termination of the busy period which is associated with the last customer in the auxilary queue. Thus the whole busy period has been split into an initial occupancy time I plus n busy periods. Let $G(n, x, y)\,dy$ be the probability that $G \leq x$, that n customers arrive during the interval I and that I is in $(y, y + dy)$. Then

$$G(n, x, y)\,dy = e^{-\lambda y}\frac{(\lambda y)^n}{n!}G_n(x - y)\,dB(y), \tag{2.19}$$

where $G_0(t) = 1$, $G_1(t) = G(t)$ and $G_n(t)$ is the n-fold convolution of $G(t)$ with itself for $n > 1$. Let $G(n, x)$ be the probability that $G \leq x$ and that n customers arrive during the initial occupancy time. Then $G(n, x)$ is obtained from (2.19) by integrating over the allowable range for y. Thus

$$G(n, x) = \int_0^\infty e^{-\lambda y}\frac{(\lambda y)^n}{n!}G_n(x - y)\,dB(y). \tag{2.20}$$

Finally, since

$$G(x) = \sum_{n=0}^{\infty} G(n, x),$$

we obtain

$$G(x) = \sum_{n=0}^{\infty}\int_0^\infty e^{-\lambda y}\frac{(\lambda y)^n}{n!}G_n(x - y)\,dB(y). \tag{2.21}$$

Let $G^*(n, s)$ be the transform of $G(n, x)$. Then using both the fact that $G(n, x)$ is a convolution and also the theorems relating to transforms of convolutions is section 2.7, we get

$$G^*(n, s) = [G^*(s)]^n\int_0^\infty e^{-sy}\,e^{-\lambda y}\frac{(\lambda n)^n}{n!}\,dB(y). \tag{2.22}$$

Since

$$G^*(s) = \sum_{n=0}^{\infty} G^*(n, s),$$

$$G^*(s) = \int_0^\infty e^{-y(s+\lambda)}\sum_{n=0}^{\infty}\frac{[\lambda y\,G^*(s)]^n}{n!}\,dB(y),$$

$$G^*(s) = \int_0^\infty e^{-y[s+\lambda-\lambda G^*(s)]}\,dB(y),$$

and equation (2.18) is an immediate consequence. Equation (2.18) is important because it provides an implicit solution for $G^*(s)$ and therefore, indirectly, a means for determining the distribution function for the busy period.

2.9 *Stochastic Processes and Markov Chains.*

A general stochastic process is a system $\{T, X_t, S\}$, where

(1) T denotes a set

(2) X_t denotes a random variable for each $t \in T$

(3) S denotes a set of values, called *states*, which the random variables X_t may take.

If $S = \{0, 1, 2, \ldots\}$ the process is called a discrete state stochastic process; if $T = \{0, 1, 2, \ldots\}$ the process is called a discrete parameter stochastic process.

A Markov process is a stochastic process with the property that, given the value of X_t, the values of X_s for $s > t$ do not depend on the values of X_u for $u \leq t$. A discrete state Markov process is called a Markov chain and it is customary to say that X_t is in state i if $X_t = i$. We shall be concerned in this paper with discrete parameter Markov chains. For such a Markov chain let $P_{ij}^{n,n+1}$ denote the probability defined by the equation

$$P_{ij}^{n,n+1} = \text{Prob}\{X_{n+1} = j \quad X_n = i\}; \; i, j, n = 0, 1, 2, \ldots \tag{2.23}$$

The probability $P_{ij}^{n,n+1}$ is called a one-step transition probability and in general depends on n. When these values are independent of n then the Markov chain is said to have stationary transition probabilities. We shall assume that the Markov chains used in this paper behave asymptotically (with increasing time) like Markov chains with stationary transition probabilities. We also shall assume that

$$\lim_{n \to \infty} \text{Prob}\{X_n = i\} = P_i; \; i = 0, 1, 2, \ldots \tag{2.24}$$

$$\sum_{i=0}^{\infty} P_i = 1.$$

The values P_i are called the steady-state or equilibrium probabilities for the Markov chain. Further information on stochastic processes can be found in Karlin. [9, chapters 1, 2]

3. MATHEMATICAL ANALYSIS OF THE MODIFIED MODEL

The reader is reminded that the modified model (Fig. 2) described in section 1.3 differs from the classical single-server queue; for the modified model:

(1) the server is in one of three allowable states—idle, serving a customer or inspecting the queue;

(2) the server may be idle even though the queue is occupied;

(3) an arriving customer who finds the server idle has to wait for a period of time before service begins.

The notation and definitions for the relevant parameters and variables are collected here for the reader's convenience.

λ – density parameter for the Poisson arrival process.

α – service time, a random variable governed by the distribution function $\alpha(t)$.

β – inspection time, a random variable governed by the distribution function $\beta(t)$.

S – a random variable defined by $S = \alpha + \beta$ and governed by the distribution function $S(t)$.

R – idle time, a random variable governed by the distribution function $R(t)$.

The random variables α, β and R are independent of the input Poisson process; R is independent of α and β; α and β may be taken as independent or not. The distribution functions are arbitrary except for those restrictions which are tacitly implied by the nature of the mathematical operations performed in the analysis presented here. This implies that a large number of distributions (including the majority of those found in common practice) are admissible.

3.1 *Queue Length Distribution.*

We shall use the method of the imbedded Markov chain for the analysis of queue lengths. This method was devised by D. G. Kendall and applied with success to the theory of queues. [9, 10] The essential strategy of this method (when confronted by a non-Markovian stochastic process) is to select a discrete set of points on the time axis and construct a discrete parameter stochastic process which is Markovian. In the Markov process we then determine the probability function for the random variable of interest and try to establish a relationship between this function and the corresponding probability function for the process in continuous time.

To begin, let $g_n(t)$ denote the probability that there are n jobs in the queue at time t. Within this continuous process we construct a Markov chain as follows. Select a set of points, t_i, where the t_i are defined as the instants when the server completes an inspection. From the description of the queuing model it should be noted that these points are the end points of two kinds of intervals (see Fig. 4):

(1) service-inspection intervals of length $S = \alpha + \beta$; and

(2) idle time intervals of length R.

It is appropriate to refer to these as R-intervals and S-intervals and to note in Fig. 4 how they relate to the t_i (called epochs of the process).

Let G_i denote the random variable representing the number of jobs in the queue at the ith epoch and let $g_{n,i} = \text{Prob}\{G_i = n\}$. Also let $A(n; R)$ and $A(n; S)$ denote probabilities defined by

$$A(n; R) = \text{Prob}\{n \text{ jobs arrive during a random } R\text{-interval}\}$$
$$A(n; S) = \text{Prob}\{n \text{ jobs arrive during a random } S\text{-interval}\}.$$

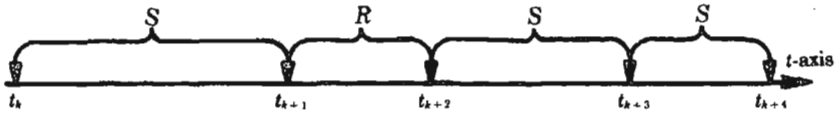

Fig. 4.—Epochs for the Markov chain g_n

It can be seen that the $g_{n,i}$ are the state probabilities for a Markov chain and that the state probabilities at the $(i + 1)^{th}$ epoch are computable from those at the ith epoch by

$$g_{n,i+1} = g_{0,i} A(n; R) + g_{1,i} A(n; S) + g_{2,i} A(n - 1; S) + \ldots + g_{n+1,i} A(0; S). \quad (3.1)$$

It follows from (3.1) that the steady-state probabilities, $g_n = \lim_{i \to \infty} g_{n,i}$, will satisfy

$$g_n = g_0 A(n; R) + \sum_{k=1}^{n+1} g_k A(n + 1 - k; S). \quad (3.2)$$

If we multiply equation (3.2) by z^n and sum with respect to n we obtain

$$G(z) = g_0 \sum_{n=0}^{\infty} A(n; R) z^n + \sum_{n=0}^{\infty} \sum_{k=1}^{n+1} g_k A(n + 1 - k; S) z^n,$$

where $G(z)$ is the generating function for the probability function g_n. Interchanging the order of summation in the double summation, we obtain

$$G(z) = g_0 \sum_{n=0}^{\infty} A(n; R) z^n + \frac{G(z) - g_0}{z} \sum_{n=0}^{\infty} A(n; S) z^n. \quad (3.3)$$

Using (2.9), $A(n; R) = \int_0^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} dR(t)$ and

$\sum_{n=0}^{\infty} A(n; R) z^n = \int_0^{\infty} e^{-t(\lambda - \lambda z)} dR(t) = R^*(\lambda - \lambda z)$. Similarly,

$\sum_{n=0}^{\infty} A(n; S) z^n = S^*(\lambda - \lambda z)$.

Substituting these results into (3.3) yields

$$G(z) = g_0 R^*(\lambda - \lambda z) + \frac{G(z) - g_0}{z} S^*(\lambda - \lambda z). \tag{3.4}$$

Differentiating this equation and setting $z = 1$ yields, after solving for g_0,

$$g_0 = \frac{1 - \lambda \bar{S}}{1 - \lambda \bar{S} + \lambda \bar{R}}. \tag{3.5}$$

We next construct a Markov chain h_n which is superior to the chain g_n in that its probability function agrees with the equilibrium probability function for the continuous process. For a discussion of this phenomenon see Cox and Miller. [4, pp. 268–269] The usefulness of the chain g_n resides in its affording a convenient transition to the evaluation of the h_n. To define the new chain we select a set of points τ_i on the time axis which correspond to instants when an inspection begins immediately following the departure of a job. The diagram in Fig. 5 shows the basic relationship between the t_i and τ_i. It can be seen that each S-interval is partitioned by a τ_i into a service interval (α-interval) and an inspection interval (β-interval). Let H be the random variable representing the number of jobs in the queue at a τ_i epoch and let h_n be its equilibrium probability function. Also let $A(n; \alpha)$ be the probability that n jobs arrive during a random α-interval. Then in a manner similar to that used to obtain (3.2) we can obtain

$$h_n = \frac{1}{1 - g_0} [g_1 A(n; \alpha) + g_2 A(n - 1; \alpha) + \ldots + g_{n+1} A(0; \alpha)]. \tag{3.6}$$

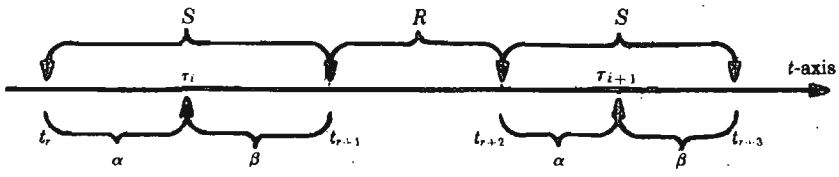

Fig. 5.—Epochs for the Markov chains g_n and h_n

The generating function for H is obtained from (3.6) and is given by

$$H(z) = \frac{1}{1 - g_0} \frac{G(z) - g_0}{z} \alpha^*(\lambda - \lambda z). \tag{3.7}$$

We can now eliminate $G(z)$ between equations (3.4) and (3.7) to get

$$H(z) = \frac{g_0}{1 - g_0} \frac{R^*(\lambda - \lambda z) - 1}{z - S^*(\lambda - \lambda s)} \alpha^*(\lambda - \lambda z). \tag{3.8}$$

If we differentiate (3.8) and let $z \to 1$ we arrive at the expected length of the queue:

$$\overline{H} = \lambda\{\overline{\alpha} + \frac{\overline{R^2}}{2\overline{R}} + \frac{\overline{S^2}}{2(1 - \lambda\overline{S})}. \tag{3.9}$$

Higher moments for the random variable H can be obtained from (3.8) through the use of higher derivatives of $H(z)$. From (3.9) we deduce that stability of the queue length distribution requires that

$$\lambda\overline{S} < 1 \text{ or } \lambda(\overline{\alpha} + \overline{\beta}) < 1.$$

3.2 *Response Time Distribution.*

We define the response time of a job to be the duration of the interval between the job's arrival and its departure. In the common parlance of the computer industry this is called the "turn around time" of the job. Let W denote the random variable representing the response time of a job. In order to obtain the equilibrium distribution of the response time we observe that the probability that a departing job j sees n jobs remaining in the queue equals the probability that n jobs arrive at the system between the time of j's arrival and j's departure. Let $A(n; W)$ denote the probability that n jobs arrive during a random response time interval. Using equation (2.9) we obtain

$$h_n = A(n; W) = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dW(t). \tag{3.10}$$

Applying the techniques described above, we get the generating function of H:

$$H(z) = \sum_{n=0}^{\infty} A(n; W)z^n = W^*(\lambda - \lambda z). \tag{3.11}$$

Differentiating this equation and setting $z = 1$ yields

$$\overline{H} = \lambda\overline{W}. \tag{3.12}$$

After substituting $\lambda\overline{W}$ for $\overline{H}$ in equation (3.9) and cancelling the factor λ we get

$$\overline{W} = \overline{\alpha} + \frac{\overline{R^2}}{2\overline{R}} + \frac{\overline{S^2}}{2(1 - \lambda\overline{S})}. \tag{3.13}$$

If we eliminate $H(z)$ between (3.8) and (3.11) we obtain

$$W^*(\lambda - \lambda z) = \frac{g_0}{1 - g_0} \frac{R^*(\lambda - \lambda z) - 1}{z - S^*(\lambda - \lambda z)} \alpha^*(\lambda - \lambda z). \tag{3.14}$$

The higher moments of W can be obtained from (3.14) since (see section 2.6):

$$\left[\frac{d}{dz^n} W^*(\lambda - \lambda z\right]_{z=1} = \lambda^n \overline{W^n}.$$

3.3 *Idle and Busy Periods Distributions.*

The logical rules governing the behavior of the queuing model require that an R-interval be repeated as long as the queue remains empty. Thus, the idle period will consist of a succession of R-intervals. Let p_n denote the probability that n customers arrive during a random R-interval. Then

$$p_0 = A(0; R) = \int_0^\infty e^{-\lambda t}\, dR(t), \text{ or}$$

$$p_0 = R^*(\lambda). \tag{3.15}$$

Let K denote the random variable representing the number of R-intervals contained in an idle period and let k_n be its probability function. Consider an idle period containing exactly n R-intervals. This event requires that no jobs arrive during the first $(n - 1)$ R-intervals and that one or more jobs arrive during the nth R-interval. It follows from this that

$$k_n = p_0^{n-1}(1 - p_0), \text{ and} \tag{3.16}$$

$$\overline{K} = \frac{1}{1 - p_0}. \tag{3.17}$$

Let I denote the random variable representing the duration of the idle period and $I(t)$ denote its distribution function. Let $I_n(t)$ denote the probability that the idle period does not exceed t and terminates after n R-intervals. We first consider $I_1(t)$. Let $0 < \tau < t$; the probability that the duration of R is in $(\tau, \tau + d\tau)$ and that one or more arrivals occur in $(0, \tau)$ is given by $(1 - e^{-\lambda\tau})\, dR(\tau)$.
Therefore,

$$I_1(t) = \int_0^t (1 - e^{-\lambda\tau})\, dR(\tau). \tag{3.18}$$

We next consider $I_2(t)$. The probability that the duration of the first R-interval is in $(\tau, \tau + d\tau)$, that no arrivals occur in $(0, \tau)$ and that the next R-interval terminates suitably in the interval (τ, t) is given by $e^{-\lambda\tau} I_1(t - \tau)\, dR(\tau)$. It follows that

$$I_2(t) = \int_0^t e^{-\lambda\tau} I_1(t - \tau)\, dR(\tau).$$

By induction we can establish

$$I_n(t) = \int_0^t e^{-\lambda\tau} I_{n-1}(t-\tau)\, dR(\tau). \tag{3.19}$$

Let $C[F(t), G(t)]$ denote the convolution of $F(t)$ with $G(t)$ and let $C^*[F, G]$ denote its Laplace-Stieltjes transform. From (3.19) we obtain

$$e^{\lambda\tau} I_n(t) = C[e^{\lambda\tau} I_{n-1}(t), R(t)]. \tag{3.20}$$

Applying transforms to equation (3.20) and using the fact that the transform of $e^{\lambda\tau} F(t)$ is given by $F^*(s-\lambda)$ (Feller [7, Chapter 13]), we get

$$I_n^*(s-\lambda) = C^*[e^{\lambda t} I_{n-1}(t), R(t)] = I_{n-1}^*(s-\lambda)\, R^*(s),$$

$$I_n^*(s) = I_{n-1}^*(s)\, R^*(s+\lambda). \tag{3.21}$$

With the use of the recursion formula (3.21) we can establish by induction that

$$I_n^*(s) = I_1^*(s)\, [R^*(s+\lambda)]^{n-1}. \tag{3.22}$$

Since

$$I(t) = \sum_{n=1}^{\infty} I_n(t)$$

we have

$$I^*(s) = \sum_{n=1}^{\infty} I_n^*(s),$$

with the use of (3.22) we obtain

$$I^*(s) = \frac{I_1^*(s)}{1 - R^*(s+\lambda)}. \tag{3.23}$$

By differentiating equation (3.23) repeatedly and setting $s = 0$ we can compute the moments of the random variable I. The first two moments are given by

$$\overline{I} = \frac{\overline{I_1} + p_1/\lambda}{1 - p_0}$$

$$\overline{I^2} = \frac{\overline{I_1^2} + 2p_2/\lambda^2 + \overline{I}(2p_1/\lambda)}{1 - p_0}. \tag{3.24}$$

We shall define the busy period to consist of the sequence of consecutive S-intervals which occurs between two idle periods. As in the classical model it corresponds to the interval during which the server gives uninterrupted service to jobs in the queue. However, unlike the classical model, the busy period is not initiated when a job arrives at an empty queue, but rather when the current R-interval terminates. Let B denote the random variable representing the duration of the busy period and $B(t)$ its distribution function. Also let $B(x\ n)$ denote the probability that $B \leq x$ given that n jobs arrive during the last R-interval of the preceding idle period. Finally, let $Q(t)$ denote the distribution function for the duration of a busy period whose initial occupancy time is S. Then $B(x\ n) = Q_n(x)$, where $Q_n(t)$ is the n-fold convolution of $Q(t)$ with itself. It follows from these definitions that

$$B(x) = \sum_{n=1}^{\infty} A(n; R)\, B(x\ n), \tag{3.25}$$

$$B^*(S) = \sum_{n=1}^{\infty} A(n; R)\, [Q^*(S)]^n. \tag{3.26}$$

The right member of (3.26) can be shown equal to

$$\int_0^{\infty} e^{-t[\lambda - \lambda Q^*(s)]}\, dR(t) - A(0; R),$$

Thus,

$$B^*(s) = R^*[\lambda - \lambda Q^*(s)] - R^*(\lambda), \tag{3.27}$$

where $Q^*(s)$ satisfies, by Takács' law of the busy period (see section 2.8),

$$Q^*(s) = S^*[s + \lambda - \lambda Q^*(s)]. \tag{3.28}$$

With the use of the last two equations we can compute, by differentiating and setting

$$s = 0, \overline{B} = \lambda \overline{Q} \overline{R} \text{ and } \overline{Q} = \frac{\overline{S}}{1 - \lambda \overline{S}}.$$

We therefore obtain the expected value of the busy period:

$$\overline{B} = \frac{\lambda \overline{R}\, \overline{S}}{1 - \lambda \overline{S}}. \tag{3.29}$$

With a second round of differentiation we can compute second moments and finally obtain the variance of the busy period:

$$\sigma_B^2 = \frac{\lambda \overline{R}\,\overline{S^2}}{(1 - \lambda\overline{S})^3} + \left(\frac{\lambda\overline{S}}{1 - \lambda\overline{S}}\right)^2 \sigma_R^2. \tag{3.30}$$

Let D denote the random variable representing the number of jobs executed in a busy period and d_n its probability function. Using equation (2.17) in section 2.7 we obtain $\overline{B} = \overline{D}\,\overline{S}$. Substituting for B as given in (3.29) we get

$$\overline{D} = \frac{\lambda\overline{R}}{1 - \lambda\overline{S}}. \tag{3.31}$$

The probability that n jobs are executed in an interval of length t is given by $S_n(t)$, where $S_n(t)$ is the n-fold convolution of $S(t)$ with itself. It follows that

$$d_n = \int_0^\infty S_n(t)\, dB(t). \tag{3.32}$$

For the special case when $S(t)$ is the negative exponential distribution, i.e.,

$$S(t) = 1 - e^{-\mu t}, \text{ then } \overline{S} = \frac{1}{\mu}, \text{ and } S_n(t) = e^{-\mu t}\,\frac{(\mu t)^n}{n!}.$$

The generating function of D, $D(z)$, can be calculated using the techniques described above to yield

$$D(z) = B^*(\mu - \mu z). \tag{3.33}$$

3.4 *Computational Aspects.*

In some cases the mean and variance of a distribution provide sufficient information with which to make a decision, but more frequently the system analyst would prefer to have a complete knowledge of the distribution function itself. From a theoretical point of view equations (3.8), (3.14), (3.23), (3.27), (3.28) uniquely determine the distributions of the random variables H, W, I, and B; however, it is often difficult (if not impossible) to obtain closed form expressions for these distributions. Under such difficulties it is necessary to rely on numerical techniques. The properties of $\alpha(t)$, $R(t)$, $S(t)$, and their transforms will influence the choice of numerical technique. The following are some of the techniques which may be useful:

(1) power series expansions;

(2) numerical integration;

(3) fitting distribution functions to numerical data (such as moments);

(4) numerical inversion of Laplace transforms.

The techniques in categories (1) and (2) are widely known. A treatment of item (3) can be found in Kendall and Stuart. [11, Chap. 6] A good discussion of item (4) is presented in Bellman, Kalaba and Lockett. [1, Chapter 2]

4. CONCLUSION

The engineer or systems analyst who has worked his way through this paper will have gained a degree of familiarity with the terminology and concepts of queuing theory and an understanding of its applicability to the study of computer service systems. It is hoped that some may have been stimulated to study the subject in greater depth. The writer will feel rewarded if this paper has served only to contribute a heightened awareness of queuing theory which may serve to improve communication between the system analyst and the applied mathematician.

The diligence with which the deficiencies of the classical single-server model have been pointed out should not be misunderstood. There are no intrinsic defects in the classical model as such; when applied to systems whose logical structure is compatible with that of the classical model, the classical model will yield useful results as it has done in the past. A general dictum (though not one which is easy to follow) is to avoid formulating abstract models which do not correspond logically with their counterpart physical systems. Nevertheless, it must be conceded that in practical situations an inexact solution is often preferable to no solution at all. System analysts, therefore, will be led to formulate deviant abstract models for the sake of gaining mathematical tractability. Such compromises should be examined critically by the applied mathematician with the aim of creating new mathematical techniques so that more precise models can be formulated and analyzed.

ACKNOWLEDGMENT

[redacted] has kindly reviewed the original draft of this paper.

STATUTORILY EXEMPT

REFERENCES

[1] R. Bellman, R. E. Kalaba and J. A. Lockett, "Numerical Inversion of the Laplace Transform," *American Elsevier*, 1966.

[2] J. Blum, "Modelling, Simulation and System Design," *3rd Congress on Information System Science and Technology*, Thompson Book Co., 1967.

[3] J. Blum, *Formulation of Problems for a Model with Multiple Queues and Feedback*, R45 Work Paper No. 67-7, 1967.

[4] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, John Wiley, 1965.

[5] A. Doig, "A Bibliography on the Theory of Queues," *Biometrika*, 44, pp. 490–514.

[6] W. P. Elderton, *Frequency Curves and Correlation*, 3rd edition, Cambridge Univ. Press, 1938.

[7] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, John Wiley, 1966.

[8] M. Fisz, *Probability Theory and Mathematical Statistics*, John Wiley, 1963.

[9] S. Karlin, *A First Course in Stochastic Processes*, Academic Press, 1966.

[10] D. G. Kendall, "Some Problems in the Theory of Queues," *J. Royal Statistical Soc.*, B13, 1951, pp. 151–185.

[11] D. G. Kendall, "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain," *Ann. Math. Statistics*, 24, 1954, 338–354.

[12] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. I, 2nd ed., C. Griffin and Co., 1963.

[13] A. Y. Khintchine, *Mathematical Methods in the Theory of Queues*, C. Griffin and Co., 1960.

[14] T. L. Saaty, *Elements of Queuing Theory*, McGraw-Hill, 1961.

[15] C. E. Skinner, "A Priority Queuing System with Server-Walking Time," *Operations Res.*, 15–2, 1967, 278–285.

[16] R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, 1960.

[17] S. S. Wilks, *Mathematical Statistics*, John Wiley, 1962.