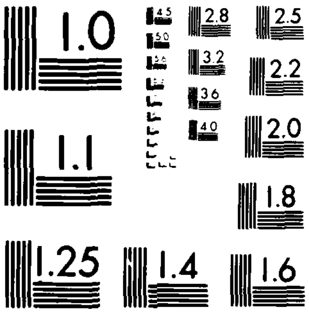
AD-A083 827 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 9/2
ON HOMOGENEITY AND ON-LINE-OFF-LINE BEHAVIOR IN M/G/1 QUEUEING --ETC(U)
FEB 80 R M BRYANT DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-2045 NL

1 OF 1
AD
A083827

END
DATE
FILMED
6-80
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

P

MRC Technical Summary Report #2045

ON HOMOGENEITY AND ON-LINE=OFF-LINE BEHAVIOR IN M/G/1 QUEUEING SYSTEMS

Raymond M. Bryant

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

February 1980

(Received December 7, 1979)

(14

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

ON HOMOGENEITY AND ON-LINE=OFF-LINE BEHAVIOR IN M/G/1 QUEUEING SYSTEMS

Raymond M. Bryant

Technical Summary Report #2045
February 1980

ABSTRACT

Operational analysis replaces certain classical queueing theory assumptions with the conditions of "homogeneous service times" and "on-line=off-line behavior." In the general case, it has been conjectured that these conditions hold as $t \to \infty$ only if the service times are exponentially distributed. In this paper, we show that this is correct for stable M/G/1 queueing systems. We also state dual results for inter-arrival times in G/M/1. Finally, we consider the relationship between the operational quantities S(n) and the mean service time in M/G/1. This relationship is shown to depend on the form of the service time distribution. It follows that using operational analysis to predict the performance of an M/G/1 queueing system will be most successful when the service time is exponential. Simulation evidence is presented which supports this claim.

AMOS(MOS) Subject Classification 68C15

Key Words: operational analysis, queueing theory, computer system modeling

Work Unit No. 5 - Operations Research

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

SIGNIFICANCE AND EXPLANATION

Queueing theory models of computer systems nave been extremely successful in spite of the numerous mathematical assumptions required to make the queueing analysis tractable. These assumptions are often unverifiable and sometimes obviously incorrect. Operational analysis attempts to explain this success by developing a theory of computer system modeling which does not depend on the classical assumptions.

For example, the assumption of exponential service times is replaced by the condition of "homogeneous service times"; this means that the average job interdeparture time does not depend on the number of jobs in system. The "on-line=off-line behavior" condition asserts that the average job interdeparture time does not depend on the system arrival rate. Finally, the operational analysts maintain that queueing system behavior can be predicted based only on measured data and without making any distributional assumptions.

In this paper we show that a M/G/1 queueing system has homogeneous service times and satisfies on-line=off-line behavior as t-->∞ if and only if the service times are exponential. Thus operational analysis is in a certain sense equivalent to Markovian queueing theory. Additionally, it is shown that a particular prediction problem of operational analysis is unsolvable for M/G/1 queues unless the service distribution is specified.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

ON HOMOGENEITY AND ON-LINE=OFF-LINE BEHAVIOR IN M/G/1 QUEUEING SYSTEMS

Raymond M. Bryant

1. INTRODUCTION

Operational Analysis [3; 4; 5; 6; 7] is a non-classical approach to analysis of queueing systems in which the system parameters $\lambda(n)$ and $\mu(n)$ are replaced by observed quantities $I(n)$ and $S(n)$ respectively. Assumptions about arrival and service time distributions are replaced by conditions on $S(n)$ and $I(n)$. Two of the key conditions are "homogeneous service times" which states that $S(n)$ is constant in n and "on-line=off-line behavior" which states that the $S(n)$ can be estimated by observing the system under a constant load.

Whenever a new idea like this appears, it is natural to explore its relation to the existing theory. This paper examines the relationship between operational and classical concepts by considering the limiting values (as t-->∞) of $I(n)$ and $S(n)$ for the sample paths of an M/G/1 queueing system. The primary results are that on-line=off-line behavior and homogeneous service times occur in M/G/1 if and only if the service times are exponentially distributed. (More precisely, it is shown that in any stable, non-exponential M/G/1 queueing system, any set of sample paths which have these properties must be a set of probability zero.) Dual results for the G/M/1 queue are stated.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

It is also shown that open, feed-forward networks of single-server queues with Poisson external arrivals can have product form solutions valid across a range of arrival rates if and only if all the service times are exponential. Finally, exact values for S(n) in M/G/1 queueing systems are derived and their dependence on the mean service time is depicted for several standard service time distributions. This discussion implies that the usual way of using the S(n) values to predict the behavior of a system is correct only when the service times are exponential. As an example, operational and stochastic methods are used to attempt to predict the performance of a simulation of some M/G/1 queueing systems. The operational method is most successful in the exponential service time cases.

In Section 2 we describe the notation of the paper and give definitions of "homogeneous service times" and "on-line=off-line behavior." Section 3 discusses what it means for an M/G/1 queueing system to have these operational properties; this section also contains the main results of the paper. These results are used to provide a method of calculating S(n) for arbitrary service times in an M/G/1 queue. Graphs of these values versus mean service time are then given in Section 4. This section concludes with an empirical comparison of the accuracy of operational and stochastic methods for predicting the mean number of jobs in an M/G/1 queueing system when the mean service time is halved.

2. NOTATION AND DEFINITIONS

Throughout this paper, whenever we are considering an M/G/1 queueing system, we will assume that the system is stable, is load independent, has arrival rate λ and service distribution B(t). We let $\bar{x}$ denote the mean service time and $\mu=1 / \bar{x}$. We will let ρ denote the system utilization and p(n) denote the stationary probability of finding n customers in system. For a G/M/1 queueing system, we let A(t) denote the inter-arrival time distribution, $\bar{a}$ denote the mean inter-arrival time, and μ denote the system service rate.

We will use a superscript * to indicate the Laplace-Stieljes transform; for example, $B^*(s)$ is the transform of B(t) and is defined as:

$$B^*(s) = \int_{-\infty}^{+\infty} e^{-st} \, dB(t)$$

Where necessary to distinguish real numbers from real valued random variables, we will use an underline to indicate the random variable.

For any particular realization of an M/G/1 queueing system, we define the sample path $\omega(t)$ as the right continuous function defined for $t \geq 0$ which gives the number of jobs in system versus time. We assume that $\omega(0) = 0$ for all sample paths, and that ω is a sample point in some probability space Ω.

We will use the term "behavior sequence" to refer to a finite length sample path resulting from the observation of a physical system.

For the purposes of this paper, we wish to make a formal distinction between a "queueing system" and an arbitrary "system" in which some queueing happens to take place. The distinction we wish to make is that a queueing system is a stochastic process characterized by inter-arrival and service times which are independent and identically distributed random variables, the inter-arrival and service times themselves being independent, service in order of arrival, and no idle server being allowed if any customers are waiting. Thus when we refer to an M/G/1 queueing system, we are really referring in a shorthand way to a particular probability space which could in principle be formally specified, but in most cases is not.

Many real systems allow queues to form and are not representable by queueing systems. Inter-arrival and service times may not be independent, service times may not be representable by random variables but instead may be cyclic or deterministically formed; there are numerous reasons that a system is not well modeled by a queueing system. The key distinction is that the characteristics of a real system are contained in a finite set of observed behavior sequences. If the system is sufficiently simple, then one can determine the set of all possible behavior sequences, but for most interesting systems this is not the case. With this distinction in mind one sees that operational analysis is primarily the study of sets of behavior sequences and their properties, while queueing theory is the study of sets of sample paths upon which a probability counterparts of conditional arrival and service rates in classical queueing theory. In fact, for any behavior sequence such that the total number of jobs in system is the same at times 0 and t, and for which arrivals and departures only occur one at a time, we have the following exact relationship among P(n,t), S(n,t) and I(n,t) [4]:

$$(2.1) \qquad P(n,t) = G \prod_{i=1}^{n} \frac{S(n,t)}{I(n-1,t)}$$

where G is a normalizing constant. We will refer to this equation as the "generalized birth-death formula."

We will primarily be interested in the asymptotic values of S(n,t) and I(n,t) as t-->∞. We will indicate this limiting value (assuming it exists) by dropping the parameter t. Thus:

$$S(n) = \lim_{t \to \infty} S(n,t).$$

It is clear that in order for I(n) and S(n) to be defined that the underlying behavior sequence must be defined for all values of t. We will refer to S(n) and I(n) as the (asymptotic) service and arrival functions, respectively. We adopt the convention that the adjective "asymptotic," when applied to the definitions of this section, implies that the quantities S(n) and I(n) have been substituted for S(n,t) and I(n,t) in the associated definition.

Finally, we wish to define certain operational terms so that they can be conveniently referred to in the sequel:

<u>Definition</u> <u>2.1</u>: A behavior sequence is said to have

measure has been defined.

We now give some definitions from operational analysis. Most of this material is contained in [4], however we prefer a notation more similar to that of [7]. To emphasize the fact that these quantities depend on values observed from a particular behavior sequence and during a finite time interval [0,t), we will modify the notation of [7] to explicitly include the parameter t.

We begin by defining the "basic operational measures" of a system during [0,t):

A(n,t) is the number of customers who arrive in [0,t) to find exactly n customers already in system.

C(n,t) is the number of customers who left the system during [0,t) when there were exactly n customers in system.

T(n,t) is the amount of time during [0,t) when there were exactly n customers in system.

Given these quantities, we then may define the following "operational performance measures." (We follow the convention of [7] and leave undefined any quantity with a zero denominator.):

S(n,t) =T(n,t)/C(n,t) is the mean service time between job departures during [0,t) given n jobs in system.

I(n,t) =T(n,t)/A(n,t) is the mean inter-arrival time during [0,t) given n jobs in system.

P(n,t) =T(n,t)/t is the proportion of time there were n jobs in system during [0,t).

We note that I(n,t) and S(n,t) are the operational homogeneous arrival times (HAT) during [0,t) if I(n,t) is constant in n.

Definition 2.2: A behavior sequence is said to have homogeneous service times (HST) during [0,t) if S(n,t) is constant in n.

Now S(n,t) is calculated by observing a system as it interacts with its environment (i. e. "on-line"). Thus S(n,t) can be referred to as the "on-line" service function. For the purpose of performance prediction, it is necessary to estimate values of S(n,t) from system service requirements, and independently of interactions with the system's environment [7]. In the terminology of operational analysis, one would estimate S(n,t) by observing an "off-line" experiment where the system was subjected to a constant load of n jobs. *This can be done* by placing n jobs in the system and then causing an an arrival to occur every time a job departs. Let $S_o(n)$ denote the mean service time between job departures during an off-line experiment with n jobs in system. When interpreted as a function of n, $S_o(n)$ is the off-linc service function. If the on-line and off-line service functions are the same then the behavior sequence satisfies on-line=off-line behavior [6]:

Definition 2.3: A behavior sequence is said to satisfy on-line=off-line behavior (on=off-line) during [0,t) if $S(n,t)=S_o(n)$ for all values of n for which S(n,t) is defined.[1]

[1] In [6], "on-line=off-line behavior" is referred to as "homogeneity."

Implicit in this definition is the fact that a system satisfies on=off-line with respect to a particular, but unspecified, set of off-line experiments. The off-line experiments may be based on a model of the system rather than on measurement of an actual system. For example, if the service distribution is known, one might use an analytic or simulation model to estimate $S_o(n)$.

We point out that in operational analysis, the HST condition is the counterpart of an exponential service time assumption in classical queueing analysis [7]. However it is easy to construct finite behavior sequences for single server queueing systems which have HST and/or satisfy on=off-line [5]. By replicating such sequences, one can create deterministic behavior sequences of arbitrary length for which S(n) can be defined, S(n) is constant in n, and $S(n)=\bar{x}$. The existence of such behavior sequences neither demonstrates nor denies the existence of non-exponential queueing systems which have HST. All that such examples demonstrate is that there exist behavior sequences which have these operational properties.

Oftentimes sets of such behavior sequences can occur with non-zero probability as the [0,t) portion of a sample path of a queueing system, but when extended by letting $t \to \infty$, the probability measure of these sets must tend toward zero. The only exceptions are sets of behavior sequences of queueing systems which themselves have HST or satisfy on=off-line (e. g. stable D/D/1 systems). Thus to explore the relationship between HST, on=off-line and M/G/1 queues, the definitions we have presented must be extended so that they apply to the ensemble of sample paths which we implicitly think of when we consider a queueing system. This is done in the next section.

3. OPERATIONAL ANALYSIS AND M/G/1 QUEUEING SYSTEMS

The operational performance measures defined in the last section are calculated from a particular behavior sequence during a particular time interval [0,t). In the context of an M/G/1 queueing system, we would say that they have been defined for a particular sample path, ω_0. Thus, we have defined what it means to say that "ω_0 has HST during [0,t)" but we have yet to define what it means to say that "an M/G/1 queueing system has HST." It is the purpose of this section to define the latter phrase in what we believe is a natural way and to explore the consequences of such a definition.

For any sample path ω in Ω, let $A(n,t,\omega)$, $C(n,t,\omega)$, $S(n,t,\omega)$, and $I(n,t,\omega)$ be the values of $A(n,t)$, $C(n,t)$, etc. associated with ω during [0,t). Let $\underline{A}(n,t)$, $\underline{C}(n,t)$, etc. denote the random variables thus defined on Ω. Also, let $\underline{A}(n)$, $\underline{C}(n)$, etc. denote the limits of the random variables $\underline{A}(n,t)$, $\underline{C}(n,t)$, etc. as t-->∞. Let $E_a(n)$ be the event that an arrival occurs to find n jobs already in system, and let $E_d(n)$ be the event that a departure occurs when there were n jobs in system. Finally, if E is a recurrent event, let m(E) denote the mean recurrence time of the event. Then we note that for any stable M/G/1 queueing system:

(1) With probability one, $\underline{T}(n,t)/t \rightarrow p(n)$ as $t \rightarrow \infty$.

(2) Since the embedded Markov Chain defined at departure instants is irreducible and positive recurrent, it follows that $m(E_d(n)) < \infty$, for all $n > 0$. Furthermore, since the probability of two or more arrivals in $[t,t+h)$ is $o(h)$, it follows that $0 < m(E_d(n))$.

(3) The recurrence times of $E_d(n)$ are independent random variables. Therefore, by an elementary result of renewal theory [12, p. 36]:

$$\lim_{t \rightarrow \infty} \underline{C}(n,t)/t = 1/m(E_d(n))$$

with probability one.

(4) $\underline{S}(n) = \lim_{t \rightarrow \infty} (\underline{T}(n,t)/t)/(\underline{C}(n,t)/t).$

We have thus shown:

Theorem 3.1: The limiting random variables $\underline{S}(n)$ are constant with probability one and $\underline{S}(n) = p(n)\ m(E_d(n))$. □

To get a similar statement for $\underline{I}(n)$, we need the following Lemma, which we will find useful later in this section:

Lemma 3.2: In any stable M/G/1 queueing system,

$$\lim_{t \rightarrow \infty} \frac{\underline{A}(n-1,t)}{t} = \lim_{t \rightarrow \infty} \frac{\underline{C}(n,t)}{t},$$

with probability one, for all $n \geq 1$.

Proof: Let $\{t_i\}$ be the starting instances of the busy cycles of the queue. Clearly $t_i \rightarrow \infty$ as $i \rightarrow \infty$ and $t_i < \infty$ for all i, both statements with probability one. Similarly,

$\underline{A}(n-1,t_i)=\underline{C}(n,t_i)$ with probability one, since the number of up-crossings of level n-1 must be the same as the number of down-crossings level n at the start of each busy cycle. (Note that the arrival at time t_i is not counted in $\underline{A}(0,t_i)$ since $\underline{A}(0,t_i)$ is the number of arrivals which found the system empty during $[0,t_i)$.) Finally, we note that $A(n-1,t,\omega) \geq C(n,t,\omega) \geq A(n-1,t,\omega)-1$ for all t and all sample paths ω. Thus with probability one

$$\lim_{t\to\infty} \frac{\underline{A}(n-1,t)}{t} = \lim_{t\to\infty} \frac{\underline{C}(n,t)}{t} . \quad \square$$

Therefore, $E_a(n)$ is a recurrent event whenever $E_d(n+1)$ is, and we have

Theorem 3.3: The limiting random variables $\underline{I}(n)$ are constant with probability one and $\underline{I}(n)=p(n)\ m(E_a(n))$. $\square$

Since $\underline{I}(n)$ and $\underline{S}(n)$ are almost everywhere constant, we will drop the distinction between these random variables and their values.

With these facts in mind, is seems natural to suggest the following definitions:

Definition 3.4: An M/G/1 queueing system will be said to have stochastic homogeneous arrival times (S-HAT) if and only if I(n) is constant in n.

Definition 3.5: An M/G/1 queueing system will be said to have stochastic homogeneous service times (S-HST) if and only if S(n) is constant in n.

Since we are only considering load independent M/G/1 queueing systems, it is clear that the off-line service function is given by $S_o(n) = \bar{x}$. We therefore have:

Definition 3.6: An M/G/1 queueing system will be said to satisfy stochastic on-line=off-line behavior (stochastic on=off-line) if and only if $S(n) = \bar{x}$ for all n.

We have added the adjective "stochastic" to these definitions to distinguish the properties of the queueing system from the properties of individual sample paths and not because we believe these concepts to be fundamentally different from the definitions of Section 2. For example it is clear that if an M/G/1 system has S-HST then almost every sample path has asymptotic HST, while if an M/G/1 system does not have S-HST then almost no sample path has HST. Thus an M/G/1 queueing system either does or does not have S-HST; one need not say that an M/G/1 queueing system has S-HST with probability one. Similar comments apply to queueing systems which satisfy stochastic on=off-line.

Now we wish to determine what types of M/G/1 queueing systems satisfy these definitions. We begin with a basic Lemma:

Lemma 3.7: In any stable M/G/1 queueing system:

$$(3.1)\quad S(1) = \frac{1}{\lambda}\left[\frac{1}{B^*(\lambda)} - 1\right].$$

Proof: $E_d(1)$ occurs if and only if the system becomes idle. Thus $m(E_d(1))$ is the mean busy cycle length. Now the Laplace transform of the busy period distribution, $G^*(s)$, is known to satisfy the functional equation:

$$G^*(s) = B^*[s + \lambda - \lambda G^*(s)]$$

(see, for example, [9, p. 212]). From this equation it is easy to determine the mean busy period length, and upon adding the mean idle time we obtain the mean busy cycle length:

$$1 / \lambda + \bar{x} / (1 - \rho).$$

Also, we know $p(1) = Q'(\emptyset)$, where $Q(z)$ is the generating function of $p(n)$. Thus $p(1)$ can be found from the Pollaczek-Khinchin transform equation (see, for example, [9, p. 194]):

$$(3.2)\quad Q(z) = B^*(\lambda-\lambda z) \frac{(1 - \rho)\ (1-z)}{B^*(\lambda-\lambda z)-z}.$$

Calculating $p(1)$ from equation (3.2) and using Theorem 3.1 shows that $S(1)$ has the indicated form. □

We observe that stochastic on=off-line implies that the $S(n)$'s cannot depend on λ. This observation is the basis for the following theorem.

<u>Theorem 3.8</u>: Suppose $B^*(s)$ is analytic for $\emptyset < \text{Re}(s) < \mu$. Then the M/G/1 queueing system satisfies stochastic on=off-line if and only if $B(t)$ is exponential.

<u>Proof</u>: (i) If $B(t)$ is exponential then the result is straightforward.

(ii) Suppose that the system satisfies stochastic on=off-line. Then, in particular, $S(1)$ does not depend on λ. Solving equation (3.1) for $B^*(\lambda)$ we get

$$B^{*}(\lambda) = \frac{1}{1 + \lambda\, S(1)}, \quad 0 < \lambda < \mu.$$

But we have thus determined $B^{*}(s)$ on a set with limit point, and hence determined $B^{*}(s)$ throughout its region of analyticity. Therefore $B(t) = 1 - \exp(t/S(1))$. □

Intuitively, this Theorem says that if an M/G/1 queueing system is decomposable in the sense that the off-line service functions can be used to estimate the on-line service functions for a variety of different arrival rates, then the service distribution is exponential. The Theorem does not state that there do not exist particular values of λ for particular service distributions $B(t)$ such that $S(n) = \bar{x}$. However, such values of λ must be isolated in the sense that they cannot have a limit point, or otherwise $B(t)$ is exponential.

If $B(t)$ is non-exponential, the dependence of $S(1)$ on λ can be quite pronounced. In Figure 3.1 we have plotted $S(1)$ versus λ for some typical service distributions. (All distributions have $\bar{x}=1.0$.) The horizontal line at $S(1)=1.0$ represents the $S(1)$ versus λ curve for exponential service times. We see that those distributions with coefficients of variation less than one have $S(1)$ versus λ curves which lie above the exponential case; and that those distributions with coefficients of variation greater than one have $S(1)$ versus λ curves which lie below the exponential case.

We have shown that if the $S(n)$'s are constant in n and do

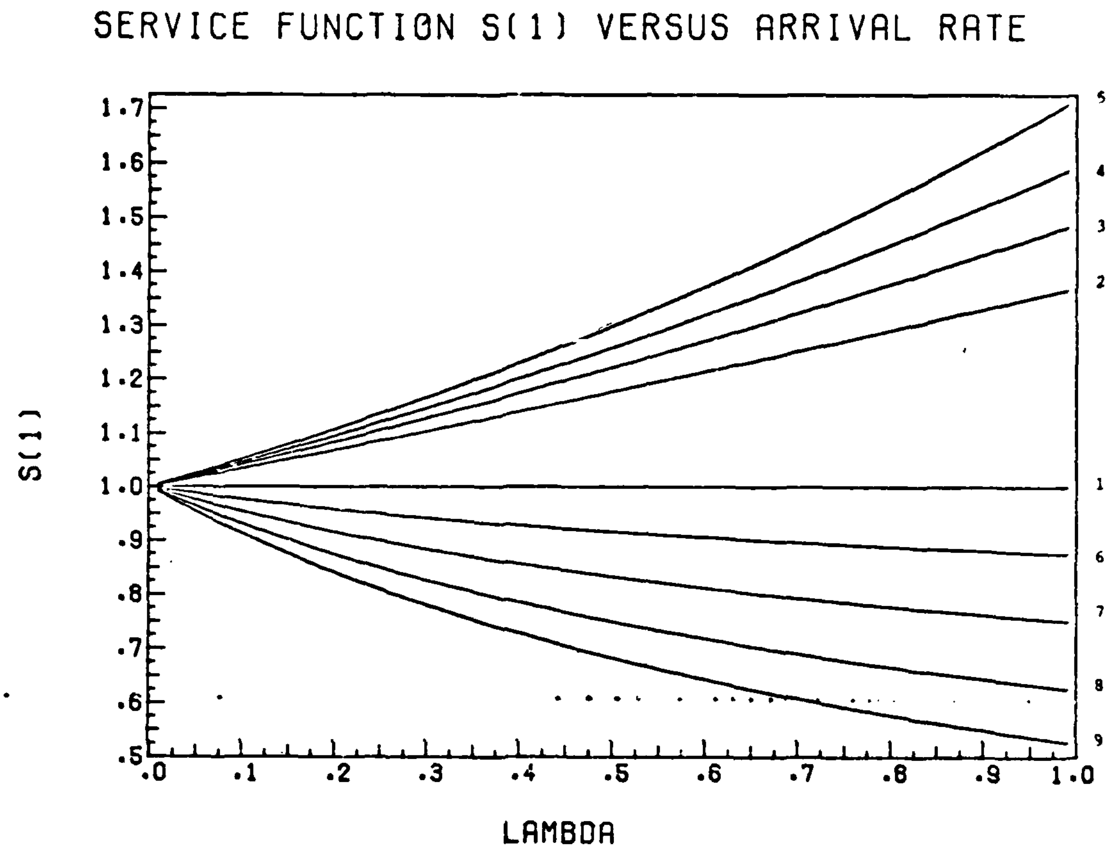


Figure 3.1

Curve Number	Distribution
1	Exponential
2	Erlang-3
3	Erlang-5
4	Erlang-10
5	Constant
6	Hyperexponential, α=0.5, CV^2=1.5
7	Hyperexponential, α=0.5, CV^2=2.0
8	Hyperexponential, α=0.5, CV^2=2.5
9	Hyperexponential, α=0.5, CV^2=2.9

not depend on λ, then the service time must be exponential. But is it possible that the S(n)'s are constant in n, but all of them depend on λ in the same way? If so, we would still have S-HST but not stochastic on=off-line. To study this situation, we begin with:

Lemma 3.9: In M/G/1, $I(n)=\lambda^{-1}$ for all n.

Proof: Because the arrival process is Poisson, $m(E_a(n)) = p(n)\lambda$. The result then follows from Theorem 3.3. □

We now prove the stochastic analogue of equation (2.1).

Lemma 3.10: In M/G/1, $p(n) = p(n-1)\, S(n) / I(n-1)$, $n \geq 1$.

Proof: By Lemma 3.2 we know that $m(E_a(n-1)) = m(E_d(n))$. Using this fact, combining the formulas for S(n) and I(n) from Theorems 3.1 and 3.3, and solving for p(n) gives the desired result. □

We can now show that S-HST is essentially equivalent to stochastic on=off-line:

Theorem 3.11: An M/G/1 queueing system has S-HST if and only if the service distribution is exponential.

Proof: (i) As before, if the service times are exponential the result is straightforward.

(ii) Let S denote the common value of S(n). From the Lemmas it follows that $p(n) = \lambda S\, p(n-1)$, $n \geq 1$. We know $p(\emptyset) = 1 - \rho$ in any M/G/1 queue. Thus

$$p(n) = (1 - \rho)\,(\lambda S)^n,\ n \geq \emptyset.$$

But $\sum_n p(n) = 1$ implies that

$$\frac{1}{1 - \lambda S} = \frac{1}{1 - \rho}.$$

Hence $\rho = \lambda\, S$ or $S = 1 / \mu$. Thus S cannot depend on λ.

To finish the proof without the explicit analyticity condition of Theorem 3.8, we let P(z) be the generating function of p(n). Equating P(z) and Q(z) from equation (3.2) gives us:

$$\frac{(1 - \rho)}{(1 - \rho\, z)} = B^{*}(\lambda - \lambda z)\, \frac{(1 - \rho)\ (1 - z)}{B^{*}(\lambda - \lambda z) - z}.$$

Solving for $B^{*}(\lambda - \lambda z)$ and substituting $s = \lambda - \lambda z$ yields:

$$B^{*}(s) = \frac{1}{1 + S\, s}. \ \Box$$

Given this Theorem, how does one account for tne existence of sample paths for single server queueing systems which have asymptotic HST, but which do not appear to have exponential service times (such as are presented in [5])? The answer is that the Theorem implies that any set of such sample paths must be assigned probability zero in any stable M/G/1 queueing system except perhaps when B(t) is exponential. Even in that case, unless the set of sample paths exhibits the rest of the properties required (e. g. independence of inter-arrival and service times), the set will still be assigned probability zero.

We can summarize the results of our discussion in the following way:

<u>Corollary 3.12</u>: Suppose that in an M/G/1 queueing system,

p(n) satisfies the product form

$$p(n) = G \prod_{i=1}^{N} S(n) \;/\; I(n-1)$$

where G is a normalizing constant and p(Ø)=G. Then provided that B(t) is well-behaved, the following conditions are equivalent:

(i) The S(n) do not depend on the I(n).

(ii) The S(n) are constant in n.

(iii) Almost every sample path of the system has asymptotic HST.

(iv) Almost every sample path of the system satisfies asymptotic on=off-line with respect to the off-line service function $S_o(n) = \bar{x}$.

(v) B(t) is the exponential distribution. □

Before turning to the case of G/M/1, we wish to state a limited result for queueing networks. We recall that a "feed-forward" network of queues is a network in which a customer is allowed to visit a server at most one time [1Ø]:

Theorem 3.13: Consider any open, stable, feed-forward network of single-server queues, and assume that external arrivals to the network are Poisson with overall intensity λ and independent of the network state. Let λ_i denote the arrival rate at the i^{th} node calculated from the routing probability matrix as if the service times were all exponential. Then the network stationary probability distribution p(n) is of the product form

(3.3) $$p(\underline{n}) = G \prod_{i=1}^{N} \prod_{j=1}^{n_i} \lambda_i S_{ij}$$

where the S_{ij} do not depend on λ if and only if all of the service times are exponential.

Proof: (i) If the service times are all exponential, the result is due to Jackson [8].

(ii) Pick any queue with only external arrivals. By the form of equation (3.3) the marginal stationary distribution of this queue must be of the form:

$$p(n_i) = G_i \prod_{j=1}^{n_i} \lambda_i S_{ij}$$

where the S_{ij} do not depend on λ_i. For this queue, Lemmas 3.9 and 3.10 imply that $S(j) = S_{ij}$, so that the service function for this queue does not depend on λ. Hence, by Theorem 3.8, the service time at this queue is exponential. By Burke's Theorem [2], we then know that the departure process from this queue is Poisson. Repeating this argument at all queues with only external arrivals shows that all queues in the network have Poisson arrivals. Hence all queues in the network have Poisson arrivals and a marginal distribution of the form indicated above. Therefore all of the service times must be exponential. □

We note that exactly as in the case of single server systems, it is easy to construct behavior sequences for networks of queues such that the behavior sequences satisfy a type of product form but for which the service times do not appear to be exponential [7; 13]. Once again, the existence of such behavior sequences does not confirm or deny the existence of stochastic processes which satisfy product form and which are defined by non-exponential networks of queueing systems. However if we restrict our attention to the types of networks discussed in the Theorem, we see that any stochastic process generated by a network of non-exponential queueing systems must assign probability zero to any such set of sample paths.

We state without proof some dual results for G/M/1 queueing systems:

Theorem 3.14: If $A^*(s)$ is analytic in $re(s) > 1 / \bar{a}$, and $I(n)$ does not depend on μ, then $A(t)$ is the exponential distribution. □

Theorem 3.15: A G/M/1 queueing system has S-HAT (see Definition 3.4) if and only if the inter-arrival time distribution is exponential. □

4. USING S(n) TO PREDICT BEHAVIOR IN M/G/1

Let us consider the following performance prediction problem:

> Random arrivals from a very large population are served one at a time, in FCFS order. The system appears to be stable, but still is very heavily loaded. What would

> the mean number of jobs in system be if a server twice as fast as the current one were installed? Observed values of I(n,t) and S(n,t) are available.

To solve this problem using operational analysis, one would adjust the S(n,t) values to represent the service function for the faster server and leave the I(n,t) values unchanged. From the generalized birth-death formula (equation 2.1), one could then estimate new values of P(n,t) and hence new values of the mean number of jobs in system.

The difficult part of the problem is estimating the new values for S(n,t). One obvious estimate is to let the new values be one-half of the old values. However, this is not correct for all service distributions. Instead the exact way that the S(n,t) depend on $\bar{x}$ varies with the service time distribution. To illustrate this dependence, we will consider the relationship between S(n) and $\bar{x}$ in an M/G/1 queueing system. (We note that Balbo and Denning [1] have considered the relationship of S(n) and the coefficient of variation of the service time of a particular server in a network of queues. Their analysis was based on simulation data.)

To evaluate S(n) we recall Lemma 3.10, which relates S(n), I(n), and p(n). Since in M/G/1, $I(n)=\lambda^{-1}$, it follows that the p(n) determine the S(n). This observation (originally due to Buzen [4]), allows the calculation of S(n) from the service time distribution. To do so we can use a power series expansion of the P-K transform formula (equation 3.2) to calculate p(n). While this process is algebraically involved, suitable tools exist to assist in the calculation and evaluation of the derivatives of $Q(z)$.[2] Thus for a particular service distribution, one can determine $p(n)$ and hence $S(n)$ as a function of the distribution parameters, at least for a few values of n.

For values of $n > 10$, the expressions for $p(n)$ become too complex to handle. Even if such large expressions could be generated, round off error would probably make any evaluated results meaningless. As we shall see below, values of $S(n)$ for $n > 5$ are usually not needed.

We believe it is unlikely that significantly simpler formulas for $S(n)$ and arbitrary $B(t)$ will ever be found. It is clear that if simple closed form expressions for $S(n)$ were known, then simple closed form expressions for $p(n)$ could easily be constructed. Since no known formulas for the latter exist, it seems unlikely that any will be found for $S(n)$. However, [4] discusses how to derive simple recursive formulas for calculating $S(n)$ when $B(t)$ is a Coxian type distribution. We now return to our discussion of the relationship between $S(n)$ and $\bar{x}$.

We have already given a formula for $S(1)$ in M/G/1, and we begin our discussion by considering a graph of $S(1)$ versus $\bar{x}$ for some typical service distributions. (See Figure 4.1.) Throughout this discussion, we are considering a stable M/G/1 queueing system with $\lambda=1.0$. In this figure we have included the $S(1)$ versus $\bar{x}$ curve for exponential service times as a

[2] The calculations in this paper used the FORMAL system [11], a FORMAC like system developed at the University of Maryland.

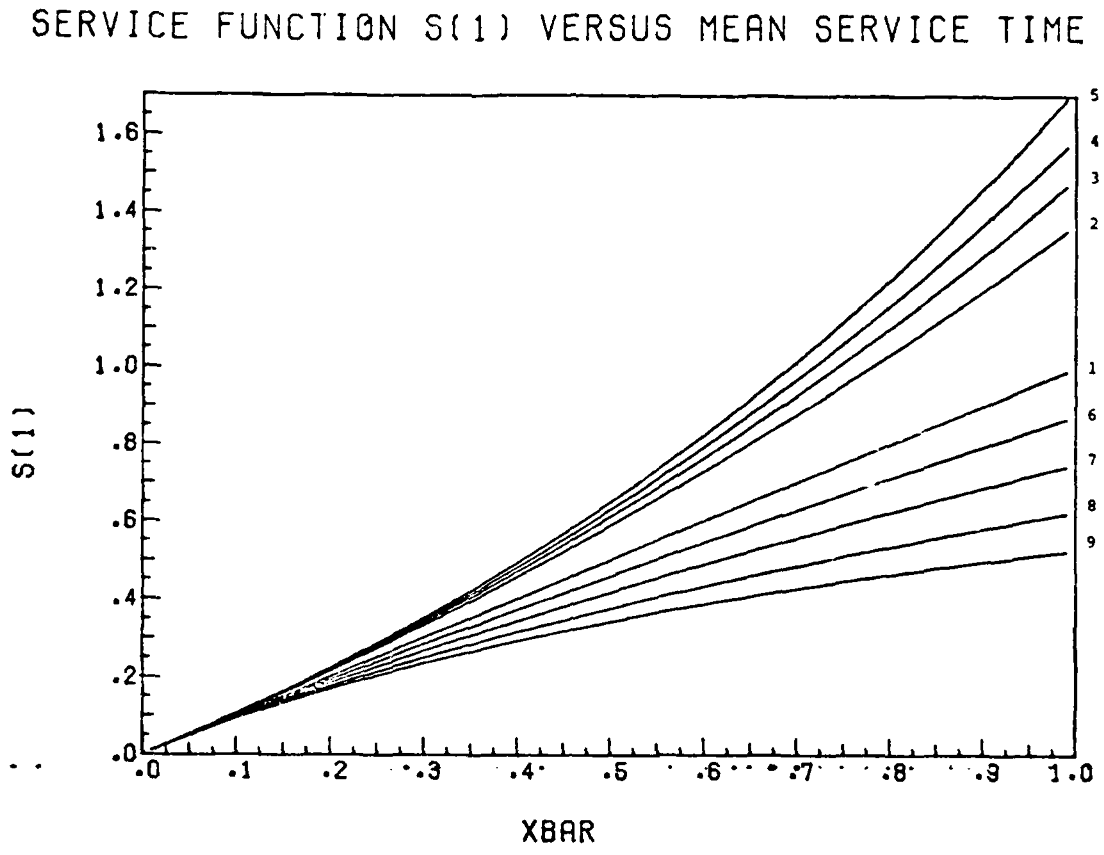


Figure 4.1

Curve Number	Distribution
1	Exponential
2	Erlang-3
3	Erlang-5
4	Erlang-10
5	Constant
6	Hyperexponential, α=0.5, CV_2^2=1.5
7	Hyperexponential, α=0.5, CV_2^2=2.0
8	Hyperexponential, α=0.5, CV_2^2=2.5
9	Hyperexponential, α=0.5, CV^2=2.9

comparison. We see that of the service distributions we have considered, the distributions with squared coefficients of variation (CV^2) greater than 1 have S(1) versus $\bar{x}$ curves which lie below the exponential case, and those distributions with $CV^2 < 1$ have curves which lie above the exponential case. Thus to estimate new values of S(1,t) in our performance prediction problem, we should more than halve the observed S(1,t) values when the service distribution has $CV^2 > 1$, and less than halve the observed S(1,t) values when $CV^2 < 1$.

Figures 4.2, 4.3, and 4.4 give S(n) versus $\bar{x}$ graphs for higher values of n. In each graph the S(n) values for a specific distribution have been plotted. (As before, we have included the S(n) versus $\bar{x}$ curve for the exponential case as a reference.) The first observation about these graphs is that S(n) rapidly approaches a limiting value as n increases. Apparently, the tail of the distribution of number in system is approximately geometric for large values of n. A second observation is that the limiting S(n) versus $\bar{x}$ curve always lies on the other side of the exponential case curve from the S(1) curve. Thus we cannot extend the statement of the last paragraph to higher values of n. Exactly how to estimate these values of S(n,t) depends on the current value of $\bar{x}$ and the other parameters of the service distribution. Third, from these examples it appears that if $CV^2 < 1$, then the S(n) versus $\bar{x}$ curve is convex upward; if $CV^2 > 1$ then the curve is convex downward.

In brief, the performance problem we have posed does not appear to be solvable without making additional assumptions about

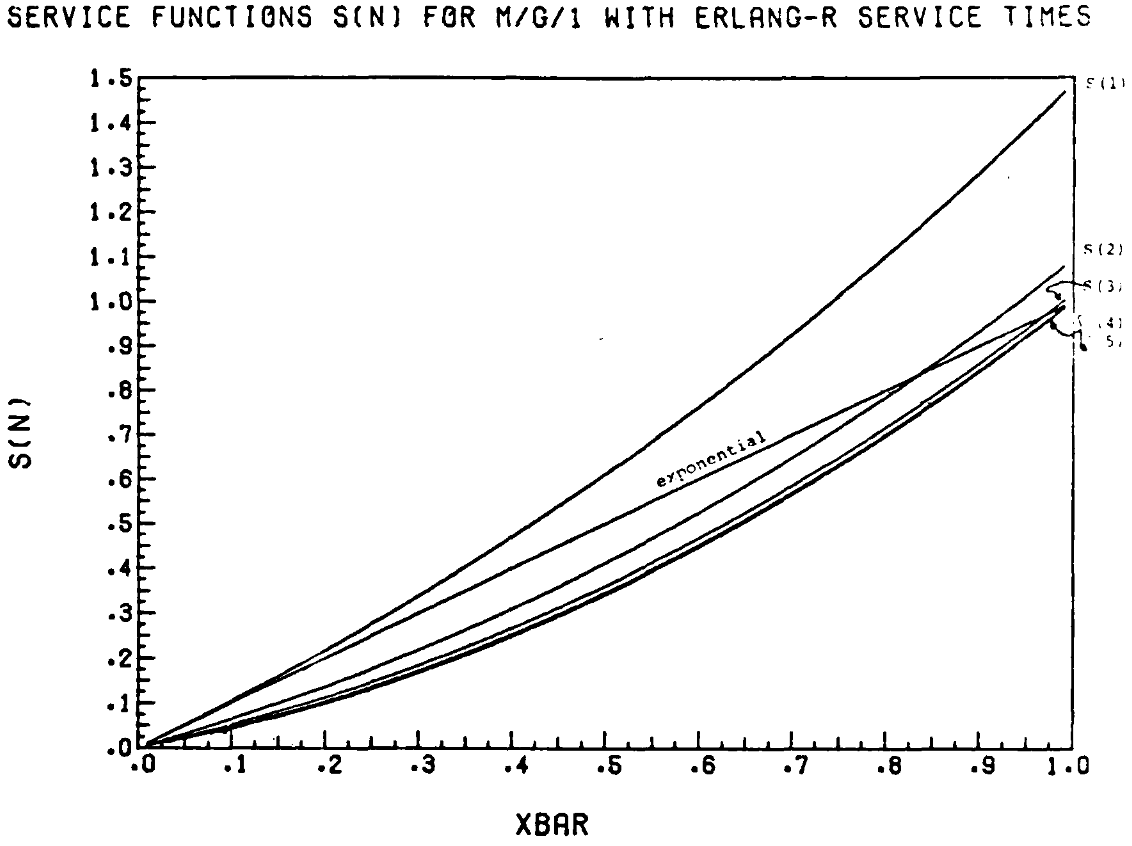


Figure 4.2

Values of service functions S(1), . . . ,S(5) versus mean service time for Erlang-5 distribution.

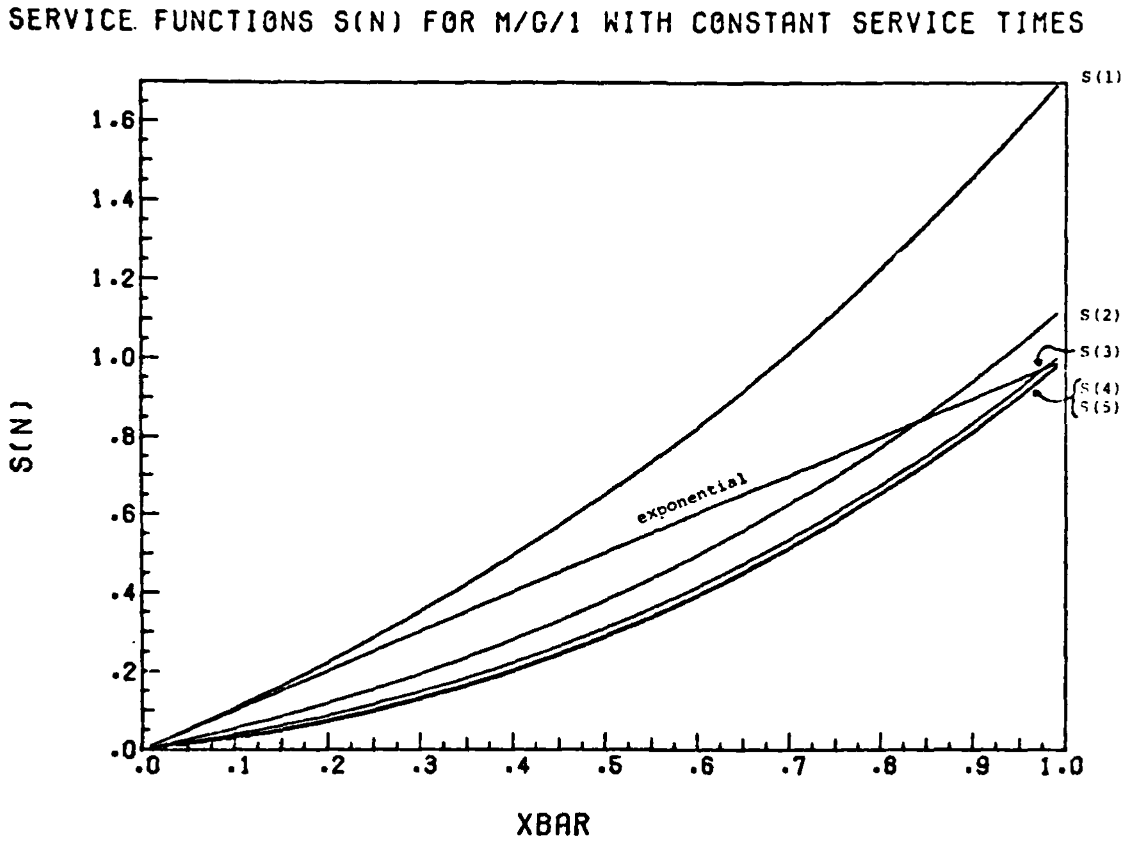


Figure 4.3

Values of service functions S(1),. . .,S(5) versus mean service time for constant service times

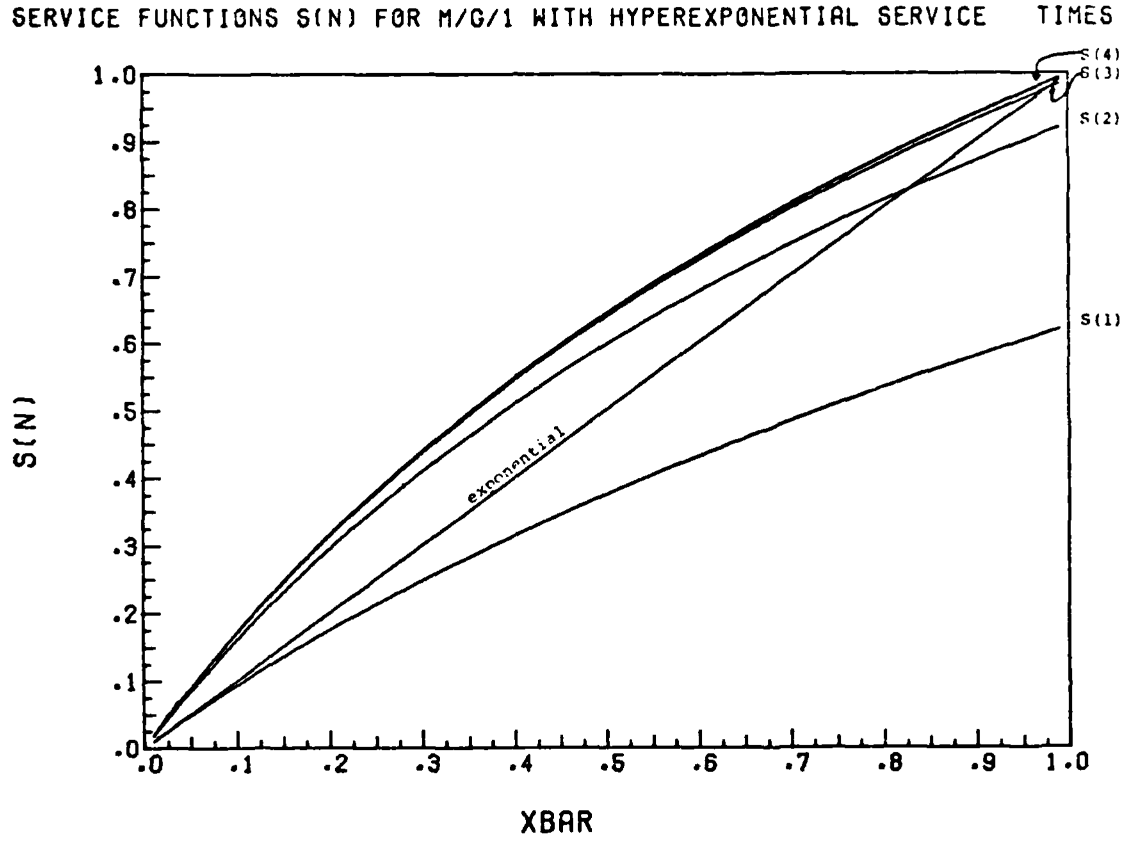


Figure 4.4

Values of service functions S(1), . . .,S(5) versus mean service time for two stage hyperexponential distribution, α=0.5, CV^2=2.5.

the distribution of service time. To verify this statement, we have used both operational and stochastic methods to attempt to predict the performance of a simulation of an M/G/1 queueing system when a server twice as fast as the original server is installed. In this way, we can study in a controlled environment the problems of using measured data to predict the performance of a system.

The operational predictions were based on values of I(n,t) and S(n,t) measured during a baseline simulation. Then the S(n,t) values were halved, and the resulting S(n,t) values (together with the old I(n,t) values) were substituted into the generalized birth-death formula. The resulting values of P(n,t) were used to calculate the mean number of jobs in system, $\bar{n}$, given that the service times were halved. The stochastic predictions were based on observing the mean and variance of the service time as well as the arrival rate during the baseline interval. These numbers were then modified to reflect a server which was twice as fast, and the modified numbers were inserted in the Pollaczek-Khinchin mean value formula [9]:

$$\bar{n} = \rho + \frac{\lambda^2 [\bar{x}^2 + \sigma^2]}{2 (1 - \rho)}.$$

The prediction case simulations were then performed and the observed values of $\bar{n}$ were recorded. The prediction case simulations were arranged so that (1) the same random number streams as the corresponding baseline case were used, (2) the same number of arrivals and departures occurred as in the corresponding baseline case, and (3) each service time random variable was one-half of the corresponding service time during the baseline case. In all of the simulations the system started and ended empty. By using the same random number streams in the baseline and prediction cases, we have attempted to isolate prediction errors in the prediction methods themselves, and without extraneous errors being introduced through randomization.

Tables 4.5, 4.6, and 4.7 summarize the results of these simulations. Each table gives the results for nine pairs of baseline and prediction case simulations. In the first three pairs of each table, the simulations were run until 100 arrivals and departures had occurred; in the second set of three pairs there were 1000 arrivals and departures; in the third set of three pairs there were 10000 arrivals and departures. Within each set of three pairs, the only differences among the simulation runs were the random number seeds. In all cases $\lambda=1.0$, the mean service time in the baseline cases was 0.8, and the mean service time in the prediction cases was 0.4.

As a comparison to the operational approach, the baseline values for λ, $\bar{x}$, and σ were also substituted into the P-K mean value formula. These values are displayed under the heading "M/G/1 Fit." Since the simulations started and ended with the system empty, it is a characteristic of the generalized birth-death formula that the baseline P(n,t) values would be identically correct when the S(n,t) and I(n,t) values from the baseline case are substituted into the equation. Thus there is no need to do a corresponding operational calculation, since the

Case	Arrival Count	Base-Line $\bar{n}$	M/G/1 Fit	OA Pred.	M/G/1 Pred.	Pred. Case $\bar{n}$	OA Error	M/G/1 Error
1	100	2.66	6.15	1.28	0.66	0.64	+100%	+3%
2	100	2.47	2.36	0.71	0.53	0.58	+22%	-9%
3	100	1.79	2.91	0.86	0.56	0.51	+69%	+10%
4	1000	1.97	2.42	0.81	0.53	0.53	+53%	≅0%
5	1000	2.35	2.43	0.77	0.54	0.54	+43%	≅0%
6	1000	2.35	2.53	0.79	0.54	0.54	+44%	+2%
7	10000	2.25	2.31	0.75	0.53	0.53	+44%	+2%
8	10000	2.37	2.56	0.80	0.54	0.53	+48%	≅0%
9	10000	2.29	2.48	0.79	0.54	0.53	+49%	+2%

Table 4.5

Operational and Stochastic Predictions for M/G/1
with Constant Service Times

Case	Arrival Count	Base-Line $\bar{n}$	M/G/1 Fit	OA Pred.	M/G/1 Pred.	Pred. Case $\bar{n}$	OA Error	M/G/1 Error
1	100	5.28	10.80	0.79	1.03	0.99	-20%	+4%
2	100	4.20	5.64	0.48	0.86	0.91	-47%	-5%
3	100	3.91	7.65	0.71	0.91	0.93	-24%	-2%
4	1000	5.30	6.20	0.55	0.84	0.83	-34%	+1%
5	1000	5.00	7.91	0.64	0.93	0.91	-30%	+2%
6	1000	4.56	4.80	0.48	0.78	0.92	-48%	-15%
7	10000	5.41	5.91	0.55	0.84	0.83	-34%	+1%
8	10000	6.52	6.46	0.54	0.86	0.85	-36%	+1%
9	10000	5.45	5.83	0.55	0.84	0.81	-32%	+4%

Table 4.6

Operational and Stochastic Predictions for M/G/1
with Hyperexponential Service Times
$\alpha = 0.5$ $CV^2 = 2.5$

Case	Arrival	Base-	M/G/1	OA	M/G/1	Pred.	OA	M/G/1
	Count	Line	Fit	Pred.	Pred.	Case	Error	Error
		$\bar{n}$				$\bar{n}$		
1	100	2.71	4.65	0.79	0.70	0.67	+18%	+4%
2	100	1.53	2.18	0.59	0.52	0.52	+13%	≅0%
3	100	2.88	5.73	0.84	0.75	0.68	+24%	+10%
4	1000	3.58	4.08	0.66	0.67	0.68	-3%	-1%
5	1000	4.68	4.51	0.63	0.70	0.66	-5%	+6%
6	1000	3.43	3.78	0.66	0.65	0.74	-11%	-12%
7	10000	3.66	4.01	0.67	0.67	0.70	-4%	-4%
8	10000	4.03	4.35	0.70	0.68	0.68	+3%	≅0%
9	10000	3.51	3.66	0.66	0.65	0.63	+5%	+3%

Table 4.7

Operational and Stochastic Predictions for M/G/1
with Exponential Service Times

resulting value of $\bar{n}$ must exactly matcn the observed value.

Table 4.5 summarizes the simulation experiments for the constant service time cases. We observe that the "M/G/1 Fit" value for $\bar{n}$ never exactly matches the value of $\bar{n}$ observed during the baseline simulation. This is certainly one advantage of tne operational equations. However, we see that the "M/G/1 prediction" values for $\bar{n}$ are consistently more accurate than are the "OA (operational analysis) prediction" values. Returning to Figure 4.3, we see that halving the S(n) values causes the S(n) values to be overestimated; it follows that the operational estimates of $\bar{n}$ should be on the high side. This is indeed the case in Table 4.5. Similarly, examining Figure 4.4 one would expect the operational estimates of $\bar{n}$ in the hyperexponential case to be on the low side. This is confirmed by the simulations (see Table 4.6). Finally, we would expect that the operational estimates would be most accurate in the exponential service time case. By inspecting Table 4.7 (the exponential case) we see that the OA and M/G/1 predictions have roughly the same percentage error. In both of the other tables, the operational estimates of $\bar{n}$ wer^ significantly further off than the stochastic estimates.

Thus it is clear that if the system one is modeling satisfies the assumptions of an M/G/1 system, and if the service time is non-exponential, then using the P-K mean value formula can provide more accurate predictions than the generalized birth-death formula from operational analysis (with the present method of predicting the new S(n,t) values).

Whether or not the operational or stochastic predictions are more robust in the general case has not been considered here. Sevcik and Klawe [13] have done similar analysis in the case of queueing networks. Their examples show that using operational methods to predict the performance of a two-server queueing network when the service rate of one of the servers is doubled is about as accurate as using an appropriate stochastic model. It is clear that this matter needs further investigation.

In this section we have concentrated on the utility of using adjusted values of the S(n,t) to predict performance of an M/G/1 queueing system when the mean service time is changed. In exactly the same way, we could adjust the I(n,t) values in order to attempt to predict the performance of an M/G/1 queueing system when λ changes. The results of the last section show that the relationship between I(n) and λ is independent of the service distribution. However, we know that the S(n) are independent of λ only when the service distribution is exponential. Thus it appears that similar conclusions would be reached concerning the accuracy of operational methods for predicting the performance of M/G/1 queueing systems when the arrival rate is changed.

5. CONCLUDING REMARKS

In this paper we have attempted to clarify the relationship between the assumptions of exponential service times in M/G/1 queueing systems and the conditions of homogeneous service times and on-line=off-line behavior in operational analysis. We have pointed out that the operational concepts are defined with respect to a particular behavior sequence (or set of behavior sequences). However, immediate generalizations of these concepts allow one to define what is meant by an M/G/l queueing system with homogeneous service times or on-line=off-line behavior. In the sense of these definitions, we were able to show that an M/G/l queueing system will have these properties only when the service time is exponential. More precisely, we showed that in any non-exponential M/G/l queueing system, the set of all sample paths which have these properties as t-->∞ must be a set of probability zero. Note that our discussion does not deny the existence of such sample paths, we merely regard them as very unlikely to occur. We then demonstrated that an open, feed-forward network of single server queues with external Poisson arrivals can have a product form valid across a range of arrival rates only when the service times are all exponential.

We then discussed the problems of using the generalized birth-death formula of operational analysis to predict (from measured data) the performance of an M/G/l queueing system when the service times were halved. We demonstrated that the relationship of S(n) and $\bar{x}$ depends on the service distribution. Among the distributions we considered, the relationship was linear only in the exponential case. In the non-exponential case, the standard estimate of the new values of S(n) (one-half of the old values), was seen to lead to less accurate predictions than use of the Pollaczek-Khinchin mean value formula with estimated values for the arrival rate and the mean and variance of the service time. It appears that similar results would be found for the prediction case when the arrival rate is changed.

6. ACKNOWLEDGEMENTS

The support of the Madison Academic Computing Center for the preparation of this document and for the simulations of Section 4 is gratefully acknowledged. The simulations were written in SIMSCRIPT II.5, which is a product of C. A. C. I. incorporated, and which has been supplied to the University of Wisconsin at no charge. Comments by Micha Hofri, Ken Sevcik, and the referees have helped shape the presentation of this material. In particular, I would like to thank Peter Denning, whose careful reading of and numerous comments about an earlier draft of this document have resulted in an improved paper.

REFERENCES

[1] Balbo, G. and P. J. Denning, "Homogeneous approximations of general queueing networks," Proceedings of the 4th International Symposium on Computer System Modelling and Performance Evaluation, M. Arato, A. Butrimenko, and E. Gelenbe (eds.), Vienna, February 6-8, 1979, pp. 353-374.

[2] Burke, P. J., "The output of a queueing system," Operations Research, 4, 699-704 (1956).

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

[4] Buzen, J. P., "Operational analysis: an alternative to stochastic modeling," Proceedings of the International Conference on the Performance of Computer Installations, D. Ferrari (ed.), North-Holland Publishing Co., Amsterdam, The Netherlands (1978), pp. 175-194.

[5] Buzen, J. P. and Denning, P. J., "Operational treatment of queue distributions and mean value analysis," Computer Science Department Technical Report No. 309, Purdue University, (August 1979).

[6] Denning, P. J. and J. P. Buzen, "Operational analysis of queueing networks," Proceedings of the 3rd International Symposium on Modelling and Performance Evaluation of Computer Systems, Bonn, W. Germany, (October 1977).

[7] Denning, P. J. and J. P. Buzen, "The operational analysis of queueing network models," ACM Computing Surveys, 10, 3 (September 1978), pp. 226-261.

[8] Jackson, J. R., "Networks of waiting lines," Operations Research, 5, 518-521 (1957).

[9] Kleinrock, L. Queueing Systems Volume I: Theory, John Wiley and Sons, New York (1975).

[10] Kleinrock, L. Queueing Systems Volume II: Computer Applications, John Wiley and Sons, New York (1976).

[11] Mesztenyi, C. K., "FORMAL - a formula manipulation language," Computer note CN-1.1, University of Maryland Computer Science Center, College Park, Maryland (October 1971).

[12] Ross, S. M., Applied Probability Models with Optimization Applications, Holden-Day, San Francisco (1970).

[13] Sevcik, K. C. and Klawe, M. M., "Operational analysis versus stochastic modelling of computer systems," Proceedings of the Computer Science and Statistics: 12th Annual Symposium on the Interface, J. F. Gentleman (ed.), University of Waterloo, Waterloo, Ontario, Canada, 177-184.

(4) MRC-TSR-2045

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
2045

2. GOVT ACCESSION NO.
AD-A083 827

3. RECIP~~IENT'S CATALOG NUMBER~~
(9) Technical

(6) 4. TITLE (and Subtitle)
On Homogeneity and On-Line=Off-Line Behavior in M/G/1 Queueing Systems.

5. ~~TYPE OF REPORT & PERIOD COVERED~~
Summary Report, ~~no specific reporting period~~

6. PERFORMING ORG. REPORT NUMBER

(10) 7. ~~AUTHOR(s)~~
Raymond M./Bryant

8. CONTRACT OR GRANT NUMBER(s)
(15) DAAG29-75-C-0024

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Mathematics Research Center, University of Wisconsin
610 Walnut Street
Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Work Unit Number 5 - Operations Research

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, North Carolina 27709

12. ~~REPORT DATE~~
(11) Feb~~ruary~~ 1980

13. ~~NUMBER OF PAGES~~
~~37~~ (12) 42

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. ~~(of this report)~~
UNCLASSIFIED

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
operational analysis, queueing theory, computer system modeling

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Operational analysis replaces certain classical queueing theory assumptions with the conditions of "homogeneous service times" and "on-line=off-line behavior." In the general case, it has been conjectured that these conditions hold as $t \to \infty$ only if the service times are exponentially distributed. In this paper, we show that this is correct for stable M/G/1 queueing systems. We also state dual results for inter-arrival times in G/M/1. Finally, we consider the relationship between the operational quantities

next page

221200

ABSTRACT (Continued)

S(n) and the mean service time in M/G/1. This relationship is shown to depend on the form of the service time distribution. It follows that using operational analysis to predict the performance of an M/G/1 queueing system will be most successful when the service time is exponential. Simulation evidence is presented which supports this claim.